Published in final edited form as:

Genet Epidemiol. 2019 February ; 43(1): 82–101. doi:10.1002/gepi.22162.

VARIANCE COMPONENTS GENETIC ASSOCIATION TEST FOR ZERO-INFLATED COUNT OUTCOMES

Matthew O. Goodman,

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Lori Chibnik, and

Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Tianxi Cai

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Abstract

Commonly in biomedical research, studies collect data in which an outcome measure contains informative excess zeros; for example when observing the burden of neuritic plaques in brain pathology studies, those who show none contribute to our understanding of neurodegenerative disease. The outcome may be characterized by a mixture distribution with one component being the ‘structural zero’ and the other component being a Poisson distribution. We propose a novel variance components score test of genetic association between a set of genetic markers and a zero-inflated count outcome from a mixture distribution. This test shares advantageous properties with SNP-set tests which have been previously devised for standard continuous or binary outcomes, such as the Sequence Kernel Association Test (SKAT). In particular, our method has superior statistical power compared to competing methods, especially when there is correlation within the group of markers, and when the SNPs are associated with both the mixing proportion and the rate of the Poisson distribution. We apply the method to Alzheimer’s data from the Rush University Religious Orders Study and Memory and Aging Project (ROSMAP), where as proof of principle we find highly significant associations with the APOE gene, in both the ‘structural zero’ and ‘count’ parameters, when applied to a zero-inflated neuritic plaques count outcome.

Keywords

zero-inflated Poisson; SNP sets; variance components score tests; kernel machines; omnibus test of multiple parameters

Correspondence: Matthew Goodman, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA, *Phone:* 617.432.1056, matthewgoodman@fas.harvard.edu.

Web Resources. We have implemented our methods in R. The software can be found on GitHub (<http://github.com/MatthewOGoodman/ZeroInflatedVCtest>).

1. Introduction

It is common to encounter data that would be appropriate to model with a so-called zero-inflated distribution, where some proportion of subjects obtain a zero outcome while the remainder obtain a positive outcome from some standard distribution. In analyzing such data in order to discover associations with genetic markers, it is important to use statistical tests that are adapted to this setting, both in order to appropriately control type I error and maximize power. This can present a challenge for naive approaches. For example, dichotomizing the data or disregarding the zeros is obviously problematic because each of these approaches results in data loss, ignoring information that can contribute to our understanding. Another simple approach uses the available zero-inflated data, but fails to account for zero-inflation. Namely, one could perform a multivariate Wald test using the Huber-White robust variance estimator within the usual Poisson regression. With a zero-inflated data generating mechanism, this will entail a mis-specified working model, and may lead to loss of power or increased type-I error in small samples.

A more sophisticated approach is to adapt previously proposed variance component tests, using a hybrid binary-continuous approach. Here as a first stage, the outcome is dichotomized as zeros vs non-zeros and tested using a logistic model. Then in a second stage, the data is reduced to the non-zeros and tested using a linear model, i.e. the non-zero outcome is treated as continuous with identity link. We have implemented these methods using the previously proposed SKAT logistic and linear model methods (Lee et al., 2012; Wu et al., 2010), while adapting the methods to combine the p -values from the two stages, and term this the SKAT hurdle model. With count data the second stage will be mis-specified, potentially leading to type I error. Additionally, as discussed below, the parameters in the hurdle model differ from those in zero-inflated count models. This discrepancy can lead to loss of power, since certain types of effects generated under a zero-inflated mixture distribution are difficult to detect under the hurdle model.

Lastly, in this paper, we propose a new association testing procedure, modeling the outcome via zeroinflated Poisson (ZIP) regression (Lambert, 1992), a simple yet flexible model that can be used to infer about the underlying mixture of two subgroups: (i) the ‘structural zero’ group representing a ‘healthy’ or unaffected population; and (ii) the ‘susceptible’ group with varying degrees of severity captured by a Poisson distribution.

We aim to develop testing procedures that can effectively assess the overall association between a set of genetic markers and a zero-inflated count outcome. As such, we adapt kernel and variance component ideas to the new ZIP outcome setting. SNP-set analyses have been advocated as having several advantages over standard single-SNP analyses including better reproducibility, power and interpretability (Liu et al., 2007,2008; Wu et al., 2010). For example, with gene-level SNP sets, several SNPs may affect transcription levels of a given protein. Individual SNP effects may combine additively or may involve more complex interactions. Marginal testing of individual SNPs may miss important signals, first due to low power from inability to combine weak effects across multiple SNPs, and second due to poor model fit from inability to model interactions or other complex effects. Marginal testing also suffers from additional multiple-comparisons. Variance components score tests for

semi-parametric kernel regression been previously shown to outperform standard multivariate tests with p degrees of freedom in each of these respects, within the continuous, binary, and time-to-event outcome settings (Lin et al., 2011; Shen and Cai, 2016; Wu et al., 2010, 2011).

By modeling the zero-inflated count outcome via ZIP or zero-inflated negative binomial (ZINB) (Greene, 1994), one may form multivariate Wald tests to assess the overall association between a SNP set and the outcome. However, such Wald tests have limited power when the SNPs are in high linkage disequilibrium (LD) with each other. Furthermore, these simple Wald-test procedures cannot easily accommodate nonlinear effects and are sensitive to model mis-specification. To overcome these challenges, we propose in this paper a variance component test via the ZIP framework to detect signals from both the genetic effect on the mixing proportion and on the Poisson rate. Our main contribution is to extend existing variance component testing procedures for simple outcomes to the more complex zero-inflated count outcome.

The rest of the article is organized as follows. In Section 2, we introduce the ZIP kernel-machine setting and we present the score test and procedures for approximating the null distribution of the proposed test. Simulation results are presented in Section 3 and the proposed procedures are illustrated by assessing the association of a zero-inflated neuritic plaque (NP) count phenotype with gene-level SNP-sets in the ROSMAP cohort. We conclude with a discussion.

2. Methods

2.1. Description of Data.

Suppose we have genetic data for n subjects with a set of p SNPs and q covariates. Let Y_i , $\mathbf{G}_i = (G_{i1}, G_{i2}, \dots, G_{ip})$, $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iq})$, indicate respectively the i^{th} subject's outcome, genotypes, and covariates. We assume that $\mathbb{D} = \{\mathbf{D}_i = (Y_i, \mathbf{G}_i^\top, \mathbf{X}_i^\top)^\top, i = 1, \dots, n\}$ are independent and identically distributed random vectors. Covariates might include age, gender, and, to control population stratification, top principal components of the genetic covariance matrix. For our examples, genotypes $G_{ij} = 0, 1$, or 2 represent the number of minor alleles at a given locus under the assumption of an additive genetic model. However, one can recode genotypes if a dominant, recessive, or general model is desirable. The count outcome phenotype Y_i is assumed to follow a ZIP model consisting of two components: 1) structural zeros with probability $1 - \pi_i$ and 2) a $\text{Poisson}(\lambda_i)$ distribution with probability π_i . (Note that the parameterization of π_i is nonstandard, but is here chosen so the direction of effect in the π_i regression and the λ_i regression agree qualitatively in sign.)

Our goal is to devise a testing procedure to assess whether $\mathbf{G}_i$ plays a role in either λ_i or π_i . To this end, we note that the ZIP likelihood takes the form

$$\prod_{i=1}^n \mathcal{L}(\pi_i, \lambda_i | \mathbf{D}_i) = \prod_{i=1}^n \{(1 - \pi_i)I(Y_i = 0) + \pi_i(e^{-\lambda_i} \lambda_i^{Y_i} / Y_i!)\} \quad (2.1)$$

We further parameterize the ZIP model in semi-parametric fashion, where we assume the effect of $\mathbf{G}_i$ on Y_i is fully captured by the functions $h_\pi(\mathbf{G}_i)$ and $h_\lambda(\mathbf{G}_i)$:

$$\text{logit}(\pi_i) = \mathbf{X}_i^\top \boldsymbol{\beta}_\pi + h_\pi(\mathbf{G}_i) \quad (2.2)$$

$$\log(\lambda_i) = \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda + h_\lambda(\mathbf{G}_i) \quad (2.3)$$

Hence in our genetic testing paradigm we are interested in testing $H_0 : h_\lambda(\cdot) = h_\pi(\cdot) = 0$. If $h_\pi(\cdot) \neq 0$, we say that genotypes are associated with the structural zeros, if $h_\lambda(\cdot) \neq 0$, we say that genotypes are associated with the mean of the Poisson component of the mixture. For the purposes of this paper, we assume that $h_\pi(\mathbf{G}_i) = \boldsymbol{\Psi}(\mathbf{G}_i)^\top \boldsymbol{\gamma}_\pi$ and $h_\lambda(\mathbf{G}_i) = \boldsymbol{\Psi}(\mathbf{G}_i)^\top \boldsymbol{\gamma}_\lambda$, where $\boldsymbol{\gamma}_\lambda$ is a set of genetic effect coefficients, for a given set of basis functions $\boldsymbol{\Psi}(\cdot)$. In practice, $\boldsymbol{\Psi}(\cdot)$ can be set to identity, leading to a test of linear effects, or pre-specified, using non-linear basis functions, such as SNP-SNP interactions. In the discussion, we show how to connect this with kernel regression, allowing detection of complex non-linear effects, without the need to specify a parametric model. With a given $\boldsymbol{\Psi}(\cdot)$, testing the hypotheses of H_0 is then equivalent to testing $H_0 : \boldsymbol{\gamma}_\pi = (\gamma_{\pi 1}, \dots, \gamma_{\pi K})^\top = \boldsymbol{\gamma}_\lambda = (\gamma_{\lambda 1}, \dots, \gamma_{\lambda K})^\top = 0$.

2.2. Test statistic.

To overcome the potential high dimensionality in $\boldsymbol{\Psi}(\mathbf{G}_i)$ and leverage the correlation among the SNPs, we impose a working assumption that $\gamma_{i1}, \dots, \gamma_{iK}$ are independent and identically distributed random variables with mean 0 and variance τ_i^2 , for $i \in \{\pi, \lambda\}$; and derive a score test for the variance components τ_π and τ_λ . These random effects working assumptions are merely a convenience in developing the test statistic; they are not required for the validity of the test. Under this set-up, testing H_0 is now translated into testing $H_0 : \tau_\pi = \tau_\lambda = 0$. The variance component score test paradigm built on the random effects working assumptions attains statistical efficiency gain essentially by taking advantage of the correlation within $\mathbf{G}$ to reduce the degrees of freedom. By using a score test we also have the convenience of only needing to fit the model under the null where $\text{logit}(\pi_{0,i}) = \mathbf{X}_i^\top \boldsymbol{\beta}_{0,\pi}$, and $\log(\lambda_{0,i}) = \mathbf{X}_i^\top \boldsymbol{\beta}_{0,\lambda}$, so model fit can be accomplished with standard maximum-likelihood fixed-effects ZIP regression software. This means it is computationally feasible to test a large number of SNP sets, even if, as with our method, resampling is necessary to obtain the distribution of the test statistic.

To form the test statistic, we write $\gamma_{ik} = \tau_i \varepsilon_{ik}$ so that $E(\varepsilon_{ik}) = 0$, and $\text{Var}(\varepsilon_{ik}) = 1$, and all covariances $\text{Cov}(\varepsilon_{ik}, \varepsilon_{i'k'}) = 0$ when $i \neq i'$ or $k \neq k'$. Then for $i \in \{\pi, \lambda\}$, we define the score test statistic associated with τ_i as

$$\hat{Q}_t = E_{\varepsilon} \left[\left\| \frac{\partial \log \prod_{i=1}^n \mathcal{L}(\pi_i, \lambda_i | \mathbf{D}_i)}{\partial \tau_t} \right|_{\tau_{\pi} = \tau_{\lambda} = 0} \right]^2 = n \hat{\mathbf{S}}_t^{\top} \hat{\mathbf{S}}_t = \|\sqrt{n} \hat{\mathbf{S}}_t\|_2^2,$$

where $\hat{\mathbf{S}}_t = n^{-1} \sum_{i=1}^n \hat{r}_{t,i} \Psi(\mathbf{G}_i)$.

Here, $\hat{r}_{\pi,i}$ and $\hat{r}_{\lambda,i}$ can be thought of as a covariate-adjusted scalar residual that is typical of a score statistic.

Explicitly,

$$\begin{aligned} \hat{r}_{\pi,i} &\equiv r_{\pi}(Y_i, \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}}_{0,\pi}, \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}}_{0,\lambda}) \equiv I(Y_i = 0) \frac{\hat{\pi}_{0,i}(1 - \hat{\pi}_{0,i})(1 - \exp(-\hat{\lambda}_{0,i}))}{1 - \hat{\pi}_{0,i}(1 - \exp(-\hat{\lambda}_{0,i}))} \\ &\quad - I(Y_i > 0)(1 - \hat{\pi}_{0,i}) \end{aligned} \quad (2.4)$$

and

$$\begin{aligned} \hat{r}_{\lambda,i} &\equiv r_{\lambda}(Y_i, \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}}_{0,\pi}, \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}}_{0,\lambda}) \equiv I(Y_i = 0) \frac{\hat{\pi}_{0,i} \hat{\lambda}_{0,i} \exp(-\hat{\lambda}_{0,i})}{1 - \hat{\pi}_{0,i}(1 - \exp(-\hat{\lambda}_{0,i}))} - I(Y_i > 0)(Y_i - \hat{\lambda}_{0,i}), \\ &\quad (2.5) \end{aligned}$$

where $\text{logit} \hat{\pi}_{0,i} = \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}}_{0,\pi}$, $\text{log} \hat{\lambda}_{0,i} = \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}}_{0,\lambda}$, and $\hat{\boldsymbol{\beta}}_0 = (\hat{\boldsymbol{\beta}}_{0,\pi}^{\top}, \hat{\boldsymbol{\beta}}_{0,\lambda}^{\top})^{\top}$ is the maximum likelihood estimator of the covariate effects $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,\pi}^{\top}, \boldsymbol{\beta}_{0,\lambda}^{\top})^{\top}$ under H_0 . For $\iota \in \{\pi, \lambda\}$, the statistic $\hat{Q}_t$ can be interpreted as the L_2 norm of an empirical covariance between the residuals $\{\hat{r}_{t,i}\}$ estimated under the null of no genetic effect and the (possibly) transformed genotypes $\{\Psi(\mathbf{G}_i)\}$. Heuristically, if Y_i is related to the markers in $\mathbf{G}_i$ conditional on $\mathbf{X}_i$, the covariate-adjusted residuals from the null model will be correlated with the components of $\{\Psi(\mathbf{G}_i)\}$, leading to larger test statistics as the covariance between $\hat{r}_{t,i}$ and $\{\Psi(\mathbf{G}_i)\}$ increases (in L_2 norm).

To understand the form of $\hat{r}_{\pi,i}$ and $\hat{r}_{\lambda,i}$ note that under the null $\hat{E}(Y_i) = \hat{\pi}_{0,i}$, and $\hat{P}(Y_i > 0) = \hat{\pi}_{0,i}(1 - \exp(-\hat{\lambda}_{0,i}))$. Also, let W_i be the latent (unobserved) random variable that controls the ZIP mixture distribution, such that given $W_i = 1$, $Y_i \sim \text{Poisson}(\lambda_i)$ and given $W_i = 0$, $Y_i = 0$. Then we have $\hat{\pi}_{i,0} = \hat{P}(W_i = 1)$, $\hat{\lambda}_{0,i} = \hat{E}(Y_i | W_i = 1)$ and $\exp(-\hat{\lambda}_{0,i}) = \hat{P}(Y_i = 0 | W_i = 1)$. One can think of W_i as an indicator of whether the subject is

‘susceptible’ to the disease process responsible for the Poisson count (Lambert, 1992) Then

$$\hat{r}_{\pi,i} = I(Y_i = 0) \frac{\hat{P}(Y_i > 0)}{\hat{P}(Y_i = 0)} - I(Y_i > 0) \hat{P}(W_i = 0) \text{ and}$$

$$\hat{r}_{\lambda,i} = I(Y_i = 0) \frac{\hat{P}(Y_i = 0 | W_i = 1)}{\hat{P}(Y_i = 0)} - I(Y_i > 0) (Y_i - \hat{E}(Y_i | W_i = 1)).$$

The complex form of these residuals arises in the context of the mixture distribution which depends on the latent W_i . Nevertheless, one can easily check (see Appendix 1) that residuals $r_{\pi,i}$ and $r_{\lambda,i}$ are mean zero under the null. For a familiar example of a mean-zero residual, consider $\hat{r}_{\mu,i} = Y_i - \hat{\mu}_i$, which arises in the score equation for a linear or logistic regression targeting $\mu = E(Y_i)$, and appears in the test statistic of previous kernel regression variance component score tests such as the SKAT (Wu et al., 2010). Typically, one focuses on one residual of this form in distribution families where the mean depends on a single parameter. Here we have two mean-zero residuals corresponding to the score for τ_π and τ_λ respectively.

In Appendix 1, we show that under H_0 , the statistics $\hat{\mathbf{S}}_\pi$ and $\hat{\mathbf{S}}_\lambda$ converge to zero in probability. Furthermore, $\sqrt{n}\hat{\mathbf{S}} \equiv \sqrt{n}(\hat{\mathbf{S}}_\pi^\top, \hat{\mathbf{S}}_\lambda^\top)^\top$ converges in distribution to a multivariate

normal with covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{\pi,\pi} & \Sigma_{\pi,\lambda} \\ \Sigma_{\pi,\lambda}^\top & \Sigma_{\lambda,\lambda} \end{bmatrix}$, where

$\Sigma_{i,i'} = \text{cov}(\Phi_{i,i}, \Phi_{i',i'})$, $\Phi_{i,i} = r_{i,i} \Psi(\mathbf{G}_i) + \rho_{i,\pi} \mathbb{I}_\pi^{-1} \mathbf{U}_{\pi,i} + \rho_{i,\lambda} \mathbb{I}_\lambda^{-1} \mathbf{U}_{\lambda,i}$, $\mathbf{U}_{i,i}$ and $\mathbb{I}_i$ are the respective score and information matrix for β_i , $\rho_{i,\pi} = E[\dot{r}_{2i}(Y_i, \mathbf{X}_i^\top \beta_{0,\pi}, \mathbf{X}_i^\top \beta_{0,\lambda}) \Psi(\mathbf{G}_i) \mathbf{X}_i^\top]$, $\rho_{i,\lambda} = E[\dot{r}_{3i}(Y_i, \mathbf{X}_i^\top \beta_{0,\pi}, \mathbf{X}_i^\top \beta_{0,\lambda}) \Psi(\mathbf{G}_i) \mathbf{X}_i^\top]$, and $\dot{r}_{ki} = \partial r_i(x_1, x_2, x_3) / \partial x_k$. It follows (see Appendix 1) that $\hat{Q}_i$ converges in distribution to a mixture of χ_1^2 distributions with mixing parameters being the eigenvalues of $\Sigma_{i,i}$, for $i \in \pi, \lambda$.

To form an omnibus test for H_0 combining information from the two score statistics $\hat{Q}_\pi$ and $\hat{Q}_\lambda$, several general approaches for combining p -values can be employed, for example a min- p approach (Huang et al., 2014; Won et al., 2009), or a Fisher’s method approach (Fisher, 1925). In this case we calculate the separate p -values, $\hat{p}_\pi$ and $\hat{p}_\lambda$ for π and for λ , and obtain an estimate of their joint distribution via perturbation resampling so that a size α test can be computed for the statistics $\hat{p}_m = \min(\hat{p}_\pi, \hat{p}_\lambda)$, and $\hat{p}_F = \log(\hat{p}_\pi) + \log(\hat{p}_\lambda)$, as described below.

These approaches have the advantage of putting the two statistics $\hat{Q}_\pi$ and $\hat{Q}_\lambda$ on the same scale via p -value for combination. Alternatively, we may rescale $\hat{\mathbf{S}}_\pi$ and $\hat{\mathbf{S}}_\lambda$ as $\hat{\sigma}_\pi^{-1} \hat{\mathbf{S}}_\pi$ and

$\hat{\sigma}_\lambda^{-1} \hat{\mathbf{S}}_\lambda$ to form $\hat{\mathbf{S}}_{\pi,\lambda} = (\hat{\sigma}_\pi^{-2} \hat{\mathbf{S}}_\pi^\top, \hat{\sigma}_\lambda^{-2} \hat{\mathbf{S}}_\lambda^\top)^\top$ and then construct

$\hat{Q}_{\pi,\lambda} = n \hat{\mathbf{S}}_{\pi,\lambda}^\top \hat{\mathbf{S}}_{\pi,\lambda} = \sum_{i \in \{\pi, \lambda\}} \hat{\sigma}_i^{-2} \hat{Q}_i$, where $\hat{\sigma}_i^2 = \text{trace}(\hat{\Sigma}_{i,i})$ and $\hat{\Sigma}_{i,i}$ is the estimated $\Sigma_{i,i}$.

2.3. Resampling Methods.

To approximate the null distribution of the test statistics in practice, we employ a perturbation resampling method (Parzen et al., 1994) to approximate Σ as well as the joint

distribution of $\hat{p}_\pi$ and $\hat{p}_\lambda$ under H_0 . Specifically we generate n independent samples $\mathbf{V} = \{V_i\}_{i=1}^n$ from some fixed distribution with $E[V_i] = 1$, $Var[V_i] = 1$, such as $V_i \sim Exponential(1)$. For a given set of weights $\mathbf{V}^{(b)} = \{V_i^{(b)}\}_{i=1}^n$, we obtain the perturbed score vectors $\hat{\mathbf{S}}^{(b)} = (\hat{\mathbf{S}}_\pi^{(b)\top}, \hat{\mathbf{S}}_\lambda^{(b)\top})^\top$, where

$$\hat{\mathbf{S}}_i^{(b)} = \left(\sum_{i=1}^n V_i^{(b)} \right)^{-1} \sum_{i=1}^n \hat{r}_{i,i}^{(b)} \Psi(\mathbf{G}_i) V_i^{(b)}, \quad \hat{r}_{i,i}^{(b)} \equiv r_i(Y_i, \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{0,\pi}^{(b)}, \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{0,\lambda}^{(b)}), \quad (2.6)$$

and $\hat{\boldsymbol{\beta}}_0^{(b)} = (\hat{\boldsymbol{\beta}}_{0,\pi}^{(b)}, \hat{\boldsymbol{\beta}}_{0,\lambda}^{(b)})$ is obtained by fitting a weighted ZIP regression under the null using weights $\mathbf{V}^{(b)}$. Heuristically this can be seen as re-estimating all parameters under perturbed estimating equations with a new set of independent subject-specific continuous weights, much as the non-parametric bootstrap (Efron and Tibshirani, 1994) can be considered to re-weight individuals using multinomial weights. It can be shown that, under suitable regularity conditions, $\sqrt{n}(\hat{\mathbf{S}}^{(b)} - \hat{\mathbf{S}}) | \mathbf{D} \xrightarrow{\mathcal{D}} MVN(0, \boldsymbol{\Sigma})$ (Parzen et al., 1994). Hence, for some large number B of repetitions of the resampling procedure, we use the empirical distribution $\{\hat{\mathcal{Q}}_i^{(b)} \equiv n(\hat{\mathbf{S}}_i^{(b)} - \hat{\mathbf{S}}_i)^\top (\hat{\mathbf{S}}_i^{(b)} - \hat{\mathbf{S}}_i)\}_{b=1}^B$ to approximate the asymptotic distribution of $\hat{\mathcal{Q}}$. In order to calculate p -values with greater precision than $1/B$ we use Davies' method (Davies, 1980) to obtain the distribution of the quadratic form $\hat{\mathcal{Q}}$, under the established multivariate normality on the half-statistic $\hat{\mathbf{S}}$, via the eigendecomposition of the empirical covariance of the resampling distribution $\{(\hat{\mathbf{S}}_i^{(b)} - \hat{\mathbf{S}}_i)\}_{b=1}^B$. This procedure is facilitated by the *R* package 'CompQuadForm' (Duchesne and de Micheaux, 2010).

To calculate a p -value for the omnibus tests based on $\hat{p}_{min}$ and $\hat{p}_{Fisher}$, appropriately controlling type I error in the presence of correlation between $\hat{p}_\pi$ and $\hat{p}_\lambda$, we take advantage of the same resampling procedure described above. For each $\hat{\mathbf{S}}_\pi^{(b)}$ and $\hat{\mathbf{S}}_\lambda^{(b)}$, $b = 1, \dots, B$, we use Davies' method to calculate a corresponding $\hat{p}_\pi^{(b)}$ and $\hat{p}_\lambda^{(b)}$. Define $\hat{p}_m = \min(\hat{p}_\pi, \hat{p}_\lambda)$. Then the null distribution of $\hat{p}_m$ can be approximated using the resampling distribution $\{\hat{p}_m^{(b)}\}_{b=1}^B$ where $\hat{p}_m^{(b)} = \min(\hat{p}_\pi^{(b)}, \hat{p}_\lambda^{(b)})$. We can obtain a min- p p -value as $\hat{p}_{min} = \frac{1}{B} \sum_b I(\hat{p}_m^{(b)} \leq \hat{p}_m)$. The null distribution of $\hat{p}_F$ is similarly obtained, leading to an alternative omnibus p -value: $\hat{p}_{Fisher} = \frac{1}{B} \sum_b I(\hat{p}_F^{(b)} \leq \hat{p}_F)$. In our simulations we show that both $\hat{p}_{min}$ and $\hat{p}_{Fisher}$ have appropriate size under the null. Heuristically, under the alternative we expect $\hat{p}_{Fisher}$ to perform worse than $\hat{p}_{min}$ with respect to power in cases when the signal is either through π or λ but not both, and to outperform $\hat{p}_{min}$ when an association signal appears in both π and λ . A limitation with the statistics $\hat{p}_{Fisher}$ and $\hat{p}_{min}$ is that we here rely on resampling to

directly calculate their distribution, and hence are limited to precision $1/B$, unlike for $\hat{p}_\pi$, $\hat{p}_\lambda$, and $\hat{p}_{\pi,\lambda}$, corresponding to the statistics $\hat{\mathbf{S}}_\pi$, $\hat{\mathbf{S}}_\lambda$ and $\hat{\mathbf{S}}_{\pi,\lambda}$, which can be obtained with high precision using Davies' method as described above.

3. Results

3.1. Simulations.

3.1.1. Simulation Setting.—We generated data from a zero-inflated Poisson model of the form 2.1 with $\text{logit}(\pi_i) = \alpha_\pi + \mathbf{X}_i^\top \boldsymbol{\beta}_\pi + \mathbf{G}_{Yi}^\top \boldsymbol{\gamma}_\pi$, and $\log(\lambda_i) = \alpha_\lambda + \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda + \mathbf{G}_{Yi}^\top \boldsymbol{\gamma}_\lambda$. We varied both the genetic data distribution and parameter settings across five overall simulation scenarios, the first four of which examine size of the test under the null. In setting 1.1), we fix the sample size at $n = 1000$ but considered simple cases where $\mathbf{X}$ is either absent or is independent of $\mathbf{G}$. In setting 1.2) we let the sample size $n = 1000, 2000$ and 4000 , for $\mathbf{X}$ independent of $\mathbf{G}$. In setting 2), we let $\mathbf{X}$ to be moderately correlated with $\mathbf{G}$ and let the sample size $n = 1000, 2000$ and 4000 . In setting 3), we varied the correlation strength between $\mathbf{X}$ and $\mathbf{G}$, up to a maximum pairwise correlation of approximately 0.9 to investigate the impact of correlation on the results. In setting 4), we simulated data with overdispersion by including an unobserved covariate associated with the outcome. This was to determine how robust our method is to this important deviation from the assumptions of the specified model. In setting 5), we generated data under the alternative, where the outcome was associated with (a) both π_i and λ_i using consonant effects (described further below), (b) both π_i and λ_i using dissonant effects (c) only with π_i , and (d) only with λ_i , in order to investigate how power depends on the source of the signal. Here, the covariates $\mathbf{X}$ were generated independent of $\mathbf{G}$.

Across all scenarios to obtain genetic marker data $\mathbf{G}_j$ we used HAPGEN2 (Su et al., 2011) to generate haplotypes from 1000 genomes phase one data, CEU panel. We selected candidate SNP sets based on previously reported GWAS associations in the genes APOE, CD33, and MAPT, from analyses of the Alzheimer's casecontrol phenotype. The selection criteria are described in detail below in the data analysis section. For APOE we used the 8 SNPs in the set {rs10119, **rs429358**, **rs7412**, 19–50106239, rs445925, rs483082, rs59325138, **rs438811**}, and for CD33, we used the 22 SNPs in the set {rs273637, rs273638, rs273639, **rs273640**, rs1399837, rs3826656, rs1710398, rs1697553, 19–56419774, rs2459141, 19–56420453, rs7245846, **rs34813869**, rs1354106, rs35112940, **rs10409348**, rs273652, rs1697531, rs169275, rs273649, rs273648, rs273646}. For MAPT we used the 54 SNPs in the set {rs7210728, rs3744456, rs4792891, rs9303523, rs28646281, rs1560312, rs3785879, rs9915721, rs9899833, rs3785880, rs66660193, rs8080903, rs1560313, rs9904290, rs71375325, rs1001945, rs8078967, rs2435205, rs242557, rs242562, rs878918, rs242554, rs2664006, rs9896485, **rs1800547**, **rs3785883**, rs754593, rs2435206, rs2435207, rs7209707, rs63750072, rs6503453, rs2435211, rs713522, rs2471737, rs2435200, rs11568305, rs2435201, rs2435202, **rs8070723**, rs73317025, rs73317026, rs56087321, rs41543317, rs11079728, rs41543512, rs60969130, rs66499584, rs67676322, rs916896, rs16940797, rs7521, rs7220988, rs1078997}. In each case under the alternative, 3 SNPs $\mathbf{G}_{Yi}$ (in bold) were associated with the outcome under the true data generation model. We used

these varying SNP sets to reveal the effect of differing linkage disequilibrium (LD) structures (correlation) and genotype dimension on the performance of our method when applied to various genes. The APOE gene has a low-moderate LD across all selected SNPs while the CD33 gene has two independent haploblocks with strong LD. MAPT has several small LD blocks with less than 10 SNPs each, with overall low-moderate LD across the remaining SNPs. LD plots for each gene can be viewed in the supplement, where we also present results for a set of 601 SNPs taken from MAPT, to investigate type I error with larger SNP sets with high LD.

In each scenario we compared our method to the multivariate Wald test from three models: the zeroinflated Poisson, the zero-inflated negative binomial, and a simple Poisson model using the robust HuberWhite (HW) variance. For each competing model except the single-parameter Poisson model, we performed multivariate Wald tests on $\hat{\gamma}_{\pi}$ and $\hat{\gamma}_{\lambda}$ separately, as well as the combined test of the vector concatenation $(\hat{\gamma}_{\pi}^{\top}, \hat{\gamma}_{\lambda}^{\top})^{\top}$. The Poisson, ZIP, and ZINB models fit under the alternative for the Wald test often failed to achieve convergence due to collinearity, especially in the larger SNP sets, a problem that does not affect the variance component-based methods, as they are score tests fit under the null. In order to allow models to converge, for the Wald tests only, we performed an LD pruning on the matrix $\mathbf{G}$, with SNPs chosen so that no pairwise correlation was above 0.99 for APOE and CD33 and 0.90 for MAPT. For APOE, this LD pruning resulted in no removal of SNPs, however for CD33 this reduced the number of SNPs in the model from 22 to approximately 12, and for MAPT from 54 to approximately 33, but this dimension reduction varied depending on the simulated genetic data.

To specify covariates, for those scenarios where $\mathbf{X}_i$ is independent of $\mathbf{G}_i$, we set $\mathbf{X}_{Ind,i} = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i})^{\top}$, where each X_{ki} is a complex linear combination of independent binomial and normal random variables, with pairwise correlations that ranged from approximately 0.1 to 0.9. The complete specification is shown in the Appendix. When $\mathbf{X}_i$ is dependent on $\mathbf{G}_i$, we set $\mathbf{X}_{i,Dep} = \rho_{XG} \mathbf{A} \mathbf{G}_i + \mathbf{X}_{Ind,i}$ where the $\mathbf{X} : \mathbf{G}$ association is through a matrix $\mathbf{A}$ shown in the Appendix, and ρ_{XG} is a scalar correlation strength parameter chosen to be 0.25, 0.5, or 1, representing a mild, moderate, or strong correlation, respectively. When $\rho_{XG} = 0.25, 0.5$ and 1, the respective maximal pairwise correlations in the $\mathbf{X} : \mathbf{G}$ correlation matrix were approximately 0.23, 0.44 and 0.64 for APOE, 0.58, 0.84, and 0.92 for CD33, and 0.28, 0.49, and 0.74 for MAPT.

For all settings, $\alpha_{\pi} = 1.5$, $\alpha_{\lambda} = 1.3$, $\beta_{\pi} = (0.75, 0.5, 0.25, 1.0, 1.0)$, $\beta_{\lambda} = (0.25, 0.5, 0.75, 1.0, 1.0)$. For settings where $\mathbf{G}_{Y_i}$ is associated with the outcome, for APOE, we set $\gamma_{\pi} = (0.06, 0.3, 0.6)$, and $\gamma_{\lambda} = (0.15, 0.75, 0.015)$, for CD33, we set $\gamma_{\pi} = (0.024, 0.12, 0.24)$, and $\gamma_{\lambda} = (0.06, 0.03, 0.006)$, and for MAPT we set $\gamma_{\pi} = (0.04, 0.20, 0.4)$, and $\gamma_{\lambda} = (0.10, 0.05, 0.01)$. Different parameter values are chosen such that the power is not too close to 0 or 1 across these settings. Depending on the gene, the dimension of $\mathbf{G}_i$ was either $p = 8$, $p = 22$, or $p = 54$, but in each case there were 3 SNPs $\mathbf{G}_{Y_i}$ associated with Y_i under the alternative, while the rest were not included in the true association model.

For setting 4) in which the model is mis-specified due to overdispersion, the true data generating mechanism is parameterized by $\text{logit}(\pi_i) = \alpha_\pi + \mathbf{X}_i^\top \boldsymbol{\beta}_\pi + \mathbf{G}_{Y_i}^\top \boldsymbol{\gamma}_\pi$, (as above), but $\text{log}(\lambda_i) = \alpha_\lambda + \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda + \mathbf{G}_{Y_i}^\top \boldsymbol{\gamma}_\lambda + \varepsilon_{\lambda,i}$ where $\varepsilon_{\lambda,i} \sim \text{Normal}(0, 0.3^2)$.

All testing was performed using the linear kernel test statistic with $\boldsymbol{\Psi}(\mathbf{G}) = \mathbf{G}$. In each scenario, the nominal size of the type I error is $\alpha = 0.05$. Unless otherwise stated, the simulated data had sample size $n = 1000$.

3.1.2. Competing Methods.—We first compare our proposed variance component (VC) tests of the individual parameters $\boldsymbol{\pi}$ and for λ described above (here denoted VC_π and VC_λ) with the corresponding multivariate Wald tests from the zero-inflated Poisson model and the zero-inflated negative binomial model (denoted $\text{Wald}_{\pi, \text{ZIP}}$, $\text{Wald}_{\pi, \text{ZINB}}$, $\text{Wald}_{\lambda, \text{ZIP}}$ and $\text{Wald}_{\lambda, \text{ZINB}}$). Additionally we compared six tests of the global null: three methods of combining our tests on $\boldsymbol{\pi}$ and λ , namely the min- p approach ($\text{VC}_{\min}$), the Fisher's method approach ($\text{VC}_{\text{Fisher}}$), and the variance-standardized approach (VC_{std}) based on the statistic $\hat{Q}_{\pi, \lambda}$ proposed in Section 2, along with the Wald test using the Huber-White robust variance from the Poisson model (HW_{Pois}) and the Wald tests of combined $(\boldsymbol{\gamma}_\pi^\top, \boldsymbol{\gamma}_\lambda^\top)^\top$ vector in the ZIP and ZINB settings (Wald_{ZIP} , $\text{Wald}_{\text{ZINB}}$).

We additionally compare with the two-stage hybrid binary-continuous SKAT hurdle model approach. In the first stage, this amounts to a variance component test of the parameter $\boldsymbol{\gamma}_{\pi_2}$ in the binary model $\text{logit}(\pi_{2,i}) = \mathbf{X}_i^\top \boldsymbol{\beta}_{\pi_2} + \boldsymbol{\Psi}(\mathbf{G}_i)^\top \boldsymbol{\gamma}_{\pi_2}$, where $\pi_{2,i} = P(Y_i > 0)$. We denote this test SKAT_{π_2} . Note that the parameter $\boldsymbol{\pi}_2$ is not equivalent to the parameter $\boldsymbol{\pi}$ in our method, since $\pi_{2,i}$ is the probability of observing a non-zero outcome, whereas π_i is the probability of observing a Poisson distributed outcome, which may include zeros. Then in the second stage, we compute the variance component test in the linear model $\mu_i = \mathbf{X}_i^\top \boldsymbol{\beta}_\mu + \boldsymbol{\Psi}(\mathbf{G}_i)^\top \boldsymbol{\gamma}_\mu$, where $\mu_i = E(Y_i^* | Y_i^* > 0)$, where we use the transformed $Y^* = \sqrt{Y}$ in order to reduce skewness in Y as a linear regression outcome. We denote this test as SKAT_μ . Note that the second-stage model is fit on the data such that $Y_i > 0$, hence there are differing sample sizes in the two stages. In our simulations, we use the small sample adjustment for the logistic model implemented in the SKAT *R* package (Lee et al., 2012), and the standard output for the linear model test (adopting the SKAT, rather than the SKAT-O method). To unify the resulting p -values, despite differing sample sizes, we use perturbation resampling to account for potential covariance in the test statistics for $\boldsymbol{\pi}_2$ and for μ . We obtain the p -value for the global test using Brown's method, an extension of the traditional Fisher's χ^2 statistic for independent p -values to correlated tests, as implemented in the *R* package Empirical Browns Method (Poole et al., 2016). For comparison we also compute a global p -value via Brown's method for our approach (denoted VC_{Brown}).

3.1.3. Size of the tests under correct model specification.—Settings 1) - 3) allow us to determine whether the size of the test is maintained in finite sample under correct specification of the model but with varying configurations for sample sizes and correlation structures. Results summarized in Figures 3.2, 3.3 and 3.5 suggest that when there is no

correlation between $\mathbf{X}$ and $\mathbf{G}$, type I errors of all methods except for the Poisson HW procedure are well-controlled, with empirical sizes close to the nominal α of 0.05 across all scenarios, under the APOE, CD33, and MAPT LD structures. The HW procedure appears to yield inflated type I errors for a majority of the settings and is thus removed from subsequent power comparisons. With correlation between $\mathbf{X}$ and $\mathbf{G}$, the analytic p -values for the second stage SKAT model for μ are no longer valid for our ZIP context, with type I errors above 0.2 for moderate and strong correlation in the APOE and CD33 LD structures. For the MAPT LD structure (several small LD blocks), SKAT μ performed noticeably better, but was still inferior to our method under high $\mathbf{X} : \mathbf{G}$ correlation. It is worth noting that the correlation between the covariates $\mathbf{X}_i$ and the tested genetic markers $\mathbf{G}_i$ plays relatively little role in affecting type I error of our method unless levels of correlation are quite high, e.g. where the maximal element of the $\mathbf{X} : \mathbf{G}$ correlation matrix is above 0.8.

3.1.4. Effect of mis-specification (overdispersion) on size of tests.—In this scenario we investigate the robustness of different testing procedures in preserving type I errors in the presence of overdispersion and hence model mis-specification. Specifically, as described above, we included a Normally distributed covariate in the ZIP data generating model for λ that was not controlled for in the models we used to derive the tests. We see in Figure 3.6 that the type I error of our VC test was relatively robust to overdispersion, but performance was compromised especially in APOE with increased $\mathbf{X} : \mathbf{G}$ correlation. As a comparison, the ZIP Wald test was badly compromised in terms of type I error with the amount of overdispersion in this setting, resulting in sizes ranging from approximately 0.4 to above 0.9. With up to moderate $\mathbf{X} : \mathbf{G}$ correlation, both the Wald test based on ZINB and our VC tests have at most slight inflation of type I error in the presence of mis-specification due to the observed overdispersion.

3.1.5. Power.—Figure 3.7 shows empirical power of different testing procedures when $\mathbf{G}$ is associated with (a) both π and λ via consonant effects (each component of both γ_π and γ_λ is positive); (b) both π and λ via dissonant effects (each component of γ_π is negative, whereas each component of γ_λ is positive); (c) π only; and (d) λ only. We find that our method outperforms the competing methods on power in many scenarios, and does not have substantially lower power in any scenario we investigated. Results for APOE suggest that in some LD scenarios the power of our method can outperform the Wald tests when the true signal is from π , but underperform when the signal is from λ . Note that this does not carry through to the higher dimensional CD33 and MAPT cases, where performance is superior. Of note is the fact that the SKAT hurdle model is incorrectly specified in our data and is not equipped to account properly for potentially conflicting effects in the two sources of zeros (structural vs Poisson zeros from the ZIP mixture distribution) resulting in loss of power in this setting. This discrepancy between the ZIP and the hurdle model has been previously discussed (Lachenbruch, 2001; Xu et al., 2015). In addition, the SKAT test for μ has previously been shown to have inflated type I error in cases with $\mathbf{X} : \mathbf{G}$ correlation. The global VC test of π and λ has comparable performance to the global ZINB and ZIP Wald tests in the lower-dimensional APOE case, whereas in the higher dimensional CD33 and MAPT scenarios we find that the power of our method is consistently superior. As expected, we confirm $\hat{p}_{Fisher}$ modestly outperforms $\hat{p}_{min}$ when the association signal is through both

parameters, while the reverse is true when the signal is through a single parameter. Additionally, we find that the global test via VC_{std} has comparable performance to the leader in both of these scenarios.

3.2. Data Application.

3.2.1. Data Set.—The ROSMAP study is a longitudinal, epidemiologic clinical-pathologic cohort study following subjects who are free of cognitive impairment at baseline. It focuses on cognition, tested annually, while also investigating correlated pathology in postmortem brain tissue analysis. Further details of the study can be found in Bennett et al (Bennett et al., 2012b,a). Genotype information was generated using the Affymetrix GeneChip 6.0 genotype platform, as described in Chibnik et al (Chibnik et al., 2011). Genotyping in the original study was limited to self-identified non-Hispanic Caucasians to reduce population heterogeneity. Remaining population heterogeneity was controlled by using the first three eigenvectors of an eigendecomposition of the genotypes. Genotypes were imputed using BEAGLE software, version 3.3.2 and 1000 Genomes Project Consortium interim phase I haplotypes, 2010–2011 data freeze. Association analysis was performed using the imputed genotype dosages.

Among 983 subjects who had died and had postmortem NP pathology measurements as well as imputed genotypes as of 2015, we restricted to the 970 subjects for whom the covariate ‘packyears at baseline’, a smoking status indicator, was recorded. We also include sex and age at death as covariates. The covariates ‘packyears at baseline’ and ‘age at death’ are included in the models as linear terms.

To illustrate the utility of our method we applied the test to several candidate genes, including ABCA7, APOE, CD33, MAPT and PTPRD, which have previously been shown to be associated with Alzheimer’s disease (AD) risk, from studies of the case-control phenotype. APOE on cytoband 19q13.32 (Zhong, 2017) is a well known gene previously reported to be associated with AD risk (Lambert, 2013). The previously reported GWAS SNPs for APOE were rs429358, and rs7412, which we used to obtain a set of additional SNPs in LD. The additional genes analyzed, along with their cytoband and prior case-control GWAS hits, were ABCA7 (19p13.3; rs3752246, rs3764650), CD33 (19q13.41; rs3865444, rs3826656), MAPT (17q21.31; rs1800547, rs3785883, rs8070723) and PTPRD (9p23; rs560380, rs3764650). These genes and SNPs have also been previously found to be associated with AD risk (Lambert, 2013; Giri et al., 2016). To select a SNP set for analysis we started with all BEAGLE imputed SNPs lying within the boundaries of the gene along with those $\pm 10,000$ bp upstream and downstream. We restricted to those SNPs which matched the candidate gene on the flag ‘ensembl gene id’ in the Ensembl.org GRCh38.11 data. We further restricted to SNPs with a variance equivalent to an MAF of 0.01 or above. We also restricted to SNPs that were in LD, with correlation ranging from 0.4 to 0.95, with SNPs previously reported to be significantly associated with the AD case-control phenotype in GWAS. These restrictions yields a final set of 46 SNPs for ABCA7, 8 SNPs for APOE, 22 for CD33, 54 for MAPT, and 62 for PTPRD. The complete list of SNPs analyzed in each SNP set is listed in the supplement.

The phenotype outcome we analyze is a total count measure of neuritic plaques (NP), taken from five sites: hippocampus CA1, entorhinal cortex, inferior parietal cortex, midfrontal cortex, and midtemporal cortex.

The site-specific counts are recorded by a technician who observes tissue pathology slides from each site, and are then summed to obtain a total count. We observe that 203 of the 970 selected subjects (21%) have zero neuritic plaques. The total count distribution shows evidence of both zero-inflation and overdispersion (see figure 3.9).

3.2.2. Data Analysis Results.—In Table 1, we present p -values for these genes obtained via our proposed VC test, as well as the ZINB Wald test and the hurdle model SKAT test. Other competing tests were excluded from this analysis due to their susceptibility to inflated type I error. We take $\alpha = 0.05$ as the nominal size.

Looking at the results from our proposed method, significant p -values for APOE are consistent with past case-control GWAS studies which have shown it to be strongly related to Alzheimer's risk. In addition, recent studies of pathology outcomes in Alzheimer's (Shulman et al., 2013; Beecham et al., 2014) have found APOE to be associated with plaques, although neither study used zero-inflated count outcomes. Shulman et al analyze the non-zero square-root transformed count, whereas Beecham et al analyze binary presence/absence and ordinal NP outcomes. In our study, we find significant associations in both the π and λ parameters, suggesting that there is strong evidence APOE is associated with both the mean count of neuritic plaques among the 'susceptible' population and the total absence of such plaques. Although the p -values for π and λ are different orders of magnitude, it is probably advisable not to try to assign this difference to biological mechanisms. Notably the Wald test shows the opposite pattern, and is consistent with the relative power observed in the tests for π and λ in our simulations. Although a strong signal in one parameter may suggest a biological mechanism affecting 'immunity' (structural zeros) or 'severity' (Poisson rate) statistical considerations in addition to effect size determine the magnitude of the p -value for each parameter. For example it is well known that in tests of binary regression, power is a function of π , with lower power as π approaches 0 or 1, and the same pattern is expected to hold here.

To illustrate the individual genetic effects within a SNP set, we performed ZINB regression for the NP outcome to estimate genetic effects through π and λ for each SNP in APOE, using a single-SNP marginal model approach (but still controlling for the above-mentioned covariates). We also computed the effect estimate for the regression against the APOE $\epsilon 4$ allele count, to compare results for individual SNPs with those for this well-known haplotype. In Figure 3.10 we report the maximum likelihood estimates of genetic effects γ_π and γ_λ , along with their standard errors, and the corresponding VC $_\pi$ or VC $_\lambda$ p -value. Results for previously reported case-control GWAS SNPs rs429358, and rs7412 show a strong signal in our analysis in both π and λ . In addition, the main finding from Shulman et al, who also analyze NP data from the ROSMAP cohort, is that the SNP rs4420638, downstream from ApoC1 and 14kb from APOE, has a p -value of 1.5×10^{-17} , whereas for the same SNP we find a marginal omnibus p -value of 7.8×10^{-16} .

Farebrother's method (Farebrother, 1984), implemented in the R statistical software package 'CompQuadForm', was used to calculate the distribution of $\hat{Q}_i$ in both the single-SNP and the gene-level data application in order to obtain additional precision for the p -values. Nevertheless, the p -value was calculated as 0 for the test of π for the APOE $\epsilon 4$ allele.

4. Discussion

We have derived novel VC tests that can efficiently detect the association between a set of genetic markers and a zero-inflated count outcome. We have shown the test to adequately control type I error in many simulation scenarios and have shown our method to obtain superior power to competing tests especially with large or sparse sets of markers with certain LD structures. Additionally we have seen that the proposed method has feasible application to genomic data from an Alzheimer's data study. The proposed VC testing procedures appear to adequately control type I error in all of the correctly specified simulation settings we investigated. The test is also not sensitive to the presence of model mis-specification due to overdispersion, which is of practical importance. This is in part due to the ability of perturbation resampling to ascertain the correct asymptotic distribution for the test statistic in the mis-specified model, provided that $\mathbf{X}$ is independent of $\mathbf{G}$, and under this assumption our proposed procedure can be shown to preserve the type I error.

As previous variance component approaches to SNP-set testing have shown, properly accounting for the LD structure in the SNP set improves power due to appropriately reducing the degrees of freedom of the test statistic. We see this in our simulation results comparing our VC method to the ZIP and ZINB Wald tests, particularly as the size of the SNP set increased. The popular linear and logistic SKAT tests can be adapted to our scenario, however, the performance is compromised in two respects. First the test of μ is mis-specified with our zero-truncated Poisson count data and was shown to have inflated type I error in simulations in the context of $\mathbf{X} : \mathbf{G}$ correlation. This may disqualify the SKAT hurdle model in many situations. Second, power in the hurdle model can be compromised by potentially dissonant effects, i.e. where a SNP set increases the likelihood of a structural zero, while at the same time increasing the mean of the Poisson count. For illustration, in our application setting one could imagine that a gene could be protective for developing neuritic plaques, but could change behavior under active disease, contributing to an increase in the mean number of plaques, perhaps by interacting with a molecular pathway activated in the disease process. In simulations with consonant effects, we find the power of our test to be comparable to the hurdle model, whereas it appears superior under dissonant effects. Hence, in a GWAS, where the nature of effects are not known in advance, a test based on ZIP parameters may have advantage with respect to power over one based on the hurdle model parameters, if it is believed that the underlying data generating mechanism is a mixture distribution with both structural and count-based zeros.

The standard GWAS remains one that considers testing individual SNPs separately. It is possible to take this marginal approach using our method. However, we believe that the potential to combine SNP effects at the gene level is likely to increase power in many circumstances. The linear model proposed in this paper to test the association of a SNP set with a ZIP phenotype is probably adequate for a proportion of such associations, in

particular we detect a signal in the APOE gene, despite not explicitly coding the APOE $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ haplotypes but instead using additive SNP effects. However, if it is important to capture interactions or other complex SNP effects, our method easily admits a generalization which makes this possible. The proposed test is expected to maintain the correct size when the dimension of $\Psi(\mathbf{G}_i)$ is not too high compared to the sample size since the validity of our test relies on the central limit theorem. In practice, when the correlation within $\Psi(\mathbf{G}_i)$ is high, the effective degrees of freedom may remain low and so in this scenario we also expect our test to work well. In the supplement, we have provided results for a high dimensional SNP set (601 SNPs) based on MAPT, which demonstrates that our tests maintains the desired empirical size. However, further research is warranted to investigate the conditions under which such variance component tests are valid when the dimension of $\Psi(\mathbf{G}_i)$ is not small relative to the sample size.

By specifying a non-linear transformation $\Psi(\cdot)$, our test can be used to detect non-linear effects. With proper choices of $\Psi(\cdot)$, our proposed procedure corresponds to a kernel machine score test, a feature shared with previous variance component SNP-set tests for other outcome distributions such as the SKAT Wu et al. (2010, 2013). Specifically, suppose $h_\ell(\mathbf{G}) \in \mathcal{H}_K$ with $\mathcal{H}_K = \text{span}\{\psi_\ell(\mathbf{G} \cdot), \ell = 1, \dots, \mathcal{J}\}$ being a reproducible kernel space generated by a given positive definite kernel function $K(\cdot, \cdot)$ (Cristianini and Shawe-Taylor, 2000). Then letting $\Psi(\cdot) = [\psi_\ell(\cdot), \dots, \psi_{\mathcal{J}}(\cdot)]^\top$ leads to a score test of the form $\hat{\mathbf{r}}_i^\top \mathbb{K} \hat{\mathbf{r}}_i$, where $\hat{\mathbf{r}}_i = (\hat{r}_{i,1}, \dots, \hat{r}_{i,n})^\top$ and $\mathbb{K} = [K(\mathbf{G}_i, \mathbf{G}_j)]_{n \times n}$ is the observed kernel matrix. Intuitively $K(\cdot, \cdot)$ defines a measure of genetic similarity between different subjects' genotypes. Moreover, this transformation projects the original genotype information via a nonlinear transformation to a larger function space which can be incorporated into the model linearly. Different choices of kernel (see Wu et al. 2010) allow different bases for the non-parametric function modeling the association between the ZIP outcome and the q markers in the SNP set, in turn allowing for power to test complex relationships and interactions. An alternative strategy to choose $\Psi(\cdot)$ is through kernel principal component analysis as suggested in Shen and Cai (2016). More specifically, for a given kernel function and the observed data on $\mathbf{G}_i$, one may approximate the basis functions associated with the corresponding reproducing kernel Hilbert space obtaining $\hat{\Psi}_k(\mathbf{G}_i)$ as the k leading eigenvectors of $\mathbb{K}$.

Another generalization direction is to consider multivariate $\mathbf{Y}$'s. For example in our data, since the total count of neuritic plaque's is obtained from separate regions, an alternative analysis would be to perform a multivariate version of our test, based on $\mathbf{Y} = (Y_1, \dots, Y_5)^\top$, the vector of counts at the five NP pathology sites. To obtain a global test one could proceed according to the above approaches for combining tests, for example, by combining p -values, or by using a combined statistic such as $\mathbf{S}_{\pi, \lambda, \mathbf{Y}} = (\mathbf{S}_{\pi, \lambda, Y_1}^\top, \dots, \mathbf{S}_{\pi, \lambda, Y_5}^\top)^\top$, where one may additionally decide to standardize each component by its standard deviation to put the statistics from each of the five sites on the same scale.

The simulation results under the settings with overdispersion suggests that our VC test is robust against mild model mis-specification. When the overdispersion is a potentially

significant issue, one may employ existing methods to detect overdispersion (Loeys et al., 2012; Min and Agresti, 2002; Ridout et al., 2001; Van den Broek, 1995). To further increase the robustness of the proposed procedure, a possible extension is to derive a comparable VC test for ZINB-distributed outcomes, which warrants future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

This research was funded in part by NIH grant T32NS048005. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors report no conflict of interests.

Appendix 1: asymptotic distribution

We next outline the key steps for deriving the asymptotic distribution for $\sqrt{n}\hat{\mathbf{S}} \equiv \sqrt{n}(\hat{\mathbf{S}}_\pi^\top, \hat{\mathbf{S}}_\lambda^\top)^\top$ under H_0 . To this end, we write $\hat{\mathbf{S}}_i = \tilde{\mathbf{S}}_i(\hat{\boldsymbol{\beta}}_0)$ and $\bar{\mathbf{S}}_i = \tilde{\mathbf{S}}_i(\boldsymbol{\beta}_0) = n^{-1} \sum r_{i,i} \boldsymbol{\Psi}(\mathbf{G}_i)$, where $\tilde{\mathbf{S}}_i(\boldsymbol{\beta}) = n^{-1} \sum r_i(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_\pi, \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda) \boldsymbol{\Psi}(\mathbf{G}_i)$, and $r_{i,i} = r_i(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_{0,\pi}, \mathbf{X}_i^\top \boldsymbol{\beta}_{0,\lambda})$. First, by a standard law of large numbers and properties of the maximum likelihood estimator, we have $\hat{\boldsymbol{\beta}}_0 \rightarrow \boldsymbol{\beta}_0$ in probability and $\sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) = n^{-1/2} \sum_{i=1}^n \{(\mathbb{I}_\pi^{-1} \mathbf{U}_{\pi,i})^\top, (\mathbb{I}_\lambda^{-1} \mathbf{U}_{\lambda,i})^\top\}^\top + o_p(1)$ which converges in distribution to a zero-mean multivariate normal. On the other hand, it follows from a uniform law of large numbers (Pollard, 1990) that $\tilde{\mathbf{S}}_i(\boldsymbol{\beta}) - \mathbf{s}_i(\boldsymbol{\beta}) \rightarrow 0$ in probability uniformly in $\boldsymbol{\beta}$, where $\mathbf{s}_i(\boldsymbol{\beta}) = E\{r_i(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_\pi, \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda) \mathbf{G}_i\}$. Therefore, $|\hat{\mathbf{S}}_i| = |\hat{\mathbf{S}}_i - \mathbf{s}_i(\boldsymbol{\beta}_0)| \leq \sup_{\boldsymbol{\beta}} |\tilde{\mathbf{S}}_i(\boldsymbol{\beta}) - \mathbf{s}_i(\boldsymbol{\beta})| + |\mathbf{s}_i(\hat{\boldsymbol{\beta}}_0) - \mathbf{s}_i(\boldsymbol{\beta}_0)| \rightarrow 0$ in probability. The first equality is due to the fact that $\mathbf{s}_i(\boldsymbol{\beta}_0) = 0$, which can be verified as follows:

$$\begin{aligned} E[r_{\pi,i}] &\equiv E\left[I_{Y_i=0} \frac{P(Y_i > 0)}{P(Y_i = 0)} (1 - \pi_{0,i}) - I_{Y_i > 0} (1 - \pi_{0,i})\right] \\ &= P(Y_i > 0)(1 - \pi_{0,i}) - P(Y_i > 0)(1 - \pi_{0,i}) = 0 \\ E[r_{\lambda,i}] &\equiv E\left[I_{Y_i=0} \frac{\lambda_{0,i} \pi_{0,i} \exp(-\lambda_{0,i})}{P(Y_i = 0)} - I_{Y_i > 0} (Y_i - \lambda_{0,i})\right] \\ &= \left(\frac{\lambda_{0,i} \pi_{0,i} \exp(-\lambda_{0,i})}{P(Y_i = 0)}\right) P(Y_i = 0) - \left(E[Y_i I_{Y_i > 0}] - \lambda_{0,i} E[I_{Y_i > 0}]\right) \\ &= \lambda_{0,i} \pi_{0,i} e^{-\lambda_{0,i}} - \lambda_{0,i} \pi_{0,i} + \lambda_{0,i} \pi_{0,i} (1 - e^{-\lambda_{0,i}}) = 0. \end{aligned}$$

To establish the asymptotic distribution for $\sqrt{n}\hat{\mathbf{S}}$, we first note that by the functional central limit theorem (Pollard, 1990), $\sqrt{n}\{\tilde{\mathbf{S}}(\boldsymbol{\beta}) - \mathbf{s}(\boldsymbol{\beta})\}$ converges weakly to a zero-mean Gaussian

process and thus is equicontinuous in β . This, together with a Taylor series expansion, implies that

$$\begin{aligned}\sqrt{n}\widehat{\mathbf{S}}_t &\equiv \sqrt{n}\{\widetilde{\mathbf{S}}_t(\widehat{\beta}_0) - \widetilde{\mathbf{S}}_t(\beta_0)\} + \sqrt{n}\widetilde{\mathbf{S}}_t \\ &= \sqrt{n}(\rho_{t,\pi}, \rho_{t,\lambda})(\widehat{\beta}_0 - \beta_0) + \sqrt{n}\widetilde{\mathbf{S}}_t \\ &= n^{-1/2} \sum_i \left\{ \rho_{t,\pi} \mathbb{I}_{\pi}^{-1} \mathbf{U}_{\pi,i} + \rho_{t,\lambda} \mathbb{I}_{\lambda}^{-1} \mathbf{U}_{\lambda,i} + r_{t,i} \Psi(\mathbf{G}_i) \right\} + o_p(1).\end{aligned}$$

It then follows from a central limit theorem that $\sqrt{n}\widehat{\mathbf{S}}$ converges in distribution to a multivariate normal distribution $MVN(\mathbf{0}, \Sigma)$.

One can show that if $\mathbf{S} \sim MVN(\mathbf{0}, \Sigma)$, then $\mathbf{S}^T \mathbf{S}$ has an asymptotic distribution as a linear combination of independent random variables. For completeness, we sketch the argument

here. First note $\mathbf{S}^T \mathbf{S} = \left(\mathbf{S}^T \Sigma^{-\frac{1}{2}} \right) \Sigma \left(\Sigma^{-\frac{1}{2}} \mathbf{S} \right)$ so we can write $\mathbf{S}^T \mathbf{S} = \mathbf{U}^T \Sigma \mathbf{U}$ where

$\mathbf{U} = \Sigma^{-\frac{1}{2}} \mathbf{S} \sim MVN(\mathbf{0}, I_{p \times p})$. Then find the spectral decomposition $\Sigma = \mathbf{A}^T \mathbf{\Lambda} \mathbf{A}$, where $\mathbf{\Lambda}$ is diagonal and $\mathbf{A}^T \mathbf{A} = \mathbf{A} \mathbf{A}^T = I_{p \times p}$. Hence for $\mathbf{V} = \mathbf{A} \mathbf{U}$, $Var[\mathbf{V}] = \mathbf{A} Var[\mathbf{U}] \mathbf{A}^T = \mathbf{A} \mathbf{A}^T = I$. So $\mathbf{V}$ is multivariate normal with identity covariance, and the k^{th} element of $\mathbf{V}$, V_k is standard normal. So we have $\mathbf{S}^T \mathbf{S} = \mathbf{V}^T \mathbf{\Lambda} \mathbf{V} = \sum_k \lambda_k V_k \cdot V_k = \sum_k \lambda_k \chi_1^2$.

Appendix 2: simulation settings

Specification of $\mathbf{X}_j = (X_{1j}, X_{2j}, X_{3j}, X_{4j}, X_{5j})^T$:

$X_{1i} \sim \text{binomial}(k=2, p=0.5)$

$W_{2i} \sim \text{binomial}(2, 0.5)$, $Z_{2i} \sim \text{normal}(\mu=0, \sigma^2=.25)$

$X_{2i} = 0.5(W_{2i}) + Z_{2i}$

$Z_{3i} \sim \text{normal}(\mu=0, \sigma^2=.25)$

$X_{3i} \sim 0.1X_{1i}X_{2i} + Z_{3i}$

$W_{4i} \sim \text{binomial}(2, 0.4)$, $Z_{4i} \sim \text{normal}(\mu=0, \sigma^2=.25)$

$X_{4i} = 0.1(W_{4i}) - 0.2(X_{2i}) + 0.2(X_{3i}) + Z_{4i}$

$W_{5i} \sim 0.5(\text{binomial}(n, 2, 0.4))$, $Z_{5i} \sim \text{normal}(\mu=0, \sigma^2=.25)$

$X_{5i} = 0.5(W_{5i}) + 0.15(X_{2i}) + 0.2(X_{3i}) + Z_{5i}$

Genotype-Covariate Dependence Matrix:

$$A = \begin{bmatrix} 0.08 & 0.5 & 0.0 & 0.5 & 0.8 \\ 0.09 & 0.4 & 0.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.6 & 0.9 \end{bmatrix}$$

Appendix 3: table of results

Table 1:

Empirical size (α) or power ($1 - \beta$) for APOE under various simulation scenarios. The following abbreviations are used: overdispersion (OD), under the alternative (Alt.), consonant effects (cons.) dissonant effects (diss.). See text for detailed descriptions of numbered simulation scenarios.

Simulation Setting	π : VC	π : SKAT	π : ZINB Wald	π : ZIP Wald	λ : VC	μ : SKAT	λ : ZINB Wald	λ : ZIP Wald	π : VC std.	π : VC Fisher's	π : VC min-p	π : VC Brown's	π : SKAT	π : ZINB	π : ZIP	λ : Wald HW
1.1a: without X	0.055	0.053	0.038	0.038	0.053	0.048	0.055	0.050	0.055	0.051	0.049	0.055	0.047	0.047	0.045	0.069
1.1b/1.2a: N = 1000	0.051	0.047	0.033	0.033	0.057	0.050	0.049	0.052	0.055	0.050	0.048	0.057	0.048	0.041	0.043	0.078
1.2b: N = 2000	0.053	0.051	0.043	0.043	0.058	0.049	0.049	0.052	0.054	0.051	0.050	0.055	0.050	0.045	0.047	0.066
1.2c: N = 4000	0.051	0.050	0.050	0.050	0.053	0.049	0.045	0.048	0.049	0.050	0.051	0.051	0.045	0.046	0.048	0.057
2a: corr. N=1000	0.052	0.051	0.035	0.035	0.064	0.220	0.050	0.053	0.060	0.053	0.051	0.061	0.165	0.038	0.041	0.089
2b: corr. N=2000	0.050	0.050	0.045	0.045	0.054	0.413	0.045	0.048	0.056	0.053	0.048	0.057	0.306	0.043	0.045	0.079
2c: corr. N=4000	0.049	0.050	0.047	0.047	0.056	0.758	0.046	0.049	0.054	0.053	0.049	0.054	0.621	0.048	0.049	0.081
3a: low corr.	0.054	0.049	0.032	0.036	0.062	0.079	0.061	0.385	0.059	0.053	0.048	0.060	0.067	0.045	0.277	0.090
3b: mod. corr.	0.056	0.051	0.032	0.037	0.066	0.183	0.062	0.481	0.064	0.056	0.051	0.064	0.138	0.047	0.365	0.109
3c: high corr.	0.054	0.052	0.034	0.039	0.111	0.467	0.069	0.852	0.098	0.084	0.074	0.093	0.347	0.052	0.770	0.231
4a: OD+low corr.	0.053	0.055	0.038	0.038	0.061	0.075	0.048	0.050	0.060	0.055	0.053	0.061	0.068	0.039	0.042	0.078
4b: OD+mod. corr.	0.052	0.051	0.035	0.035	0.064	0.220	0.050	0.053	0.060	0.053	0.051	0.061	0.165	0.038	0.041	0.089
4c: OD+high corr.	0.050	0.048	0.034	0.033	0.069	0.569	0.044	0.046	0.064	0.057	0.051	0.065	0.427	0.035	0.036	0.165
5a: Alt. π & λ cons.	0.711	0.664	0.357	0.359	0.649	0.466	0.647	0.659	0.879	0.868	0.787	0.879	0.784	0.791	0.796	0.843
5b: Alt. π & λ diss.	0.750	0.750	0.739	0.739	0.473	0.233	0.625	0.638	0.861	0.849	0.766	0.861	0.770	0.889	0.893	0.440
5c: Alt. π	0.647	0.589	0.346	0.347	0.065	0.052	0.048	0.050	0.528	0.523	0.538	0.543	0.468	0.224	0.231	0.299
5d: Alt. λ	0.055	0.050	0.031	0.031	0.569	0.385	0.629	0.640	0.444	0.430	0.445	0.450	0.299	0.481	0.494	0.472

Table 2: Empirical size or power for CD33. See text for detailed descriptions of simulation scenarios

Simulation Setting	π : VC	π : SKAT	π : ZINB Wald	π : ZIP Wald	λ : ZIP Wald	π : ZINB Wald	λ : ZINB Wald	μ : SKAT	λ : VC	π : VC	π : Fisher's VC	π : VC min-p	π : VC Brown's	π : μ : SKAT	π : ZINB	π : λ : ZIP	λ : Poisson: Wald HW
1.1a: without X	0.057	0.056	0.032	0.034	0.051	0.046	0.052	0.053	0.058	0.058	0.058	0.055	0.059	0.051	0.038	0.044	0.093
1.1b/1.2a: N = 1000	0.054	0.051	0.029	0.030	0.057	0.049	0.049	0.053	0.059	0.059	0.059	0.053	0.060	0.054	0.035	0.036	0.113
1.2b: N = 2000	0.058	0.054	0.041	0.041	0.052	0.047	0.047	0.050	0.052	0.054	0.054	0.054	0.055	0.052	0.037	0.040	0.079
1.2c: N = 4000	0.050	0.049	0.042	0.042	0.055	0.048	0.050	0.053	0.052	0.053	0.053	0.050	0.052	0.050	0.044	0.046	0.065
2a: corr. N=1000	0.054	0.047	0.028	0.028	0.060	0.343	0.046	0.050	0.057	0.053	0.053	0.054	0.058	0.243	0.033	0.035	0.139
2b: corr. N=2000	0.056	0.052	0.036	0.037	0.053	0.698	0.046	0.048	0.051	0.051	0.051	0.052	0.052	0.516	0.040	0.041	0.128
2c: corr. N=4000	0.053	0.052	0.040	0.040	0.050	0.970	0.046	0.048	0.054	0.055	0.055	0.050	0.057	0.892	0.043	0.044	0.156
3a: low corr.	0.054	0.050	0.028	0.031	0.057	0.095	0.065	0.416	0.059	0.058	0.058	0.056	0.059	0.074	0.040	0.290	0.138
3b: mod. corr.	0.053	0.050	0.026	0.028	0.059	0.237	0.061	0.576	0.060	0.057	0.057	0.054	0.059	0.176	0.042	0.436	0.171
3c: high corr.	0.051	0.045	0.022	0.024	0.068	0.192	0.065	0.918	0.061	0.057	0.057	0.053	0.061	0.139	0.040	0.846	0.317
4a: OD+low corr.	0.056	0.051	0.031	0.030	0.058	0.113	0.049	0.052	0.057	0.054	0.054	0.054	0.058	0.087	0.034	0.037	0.115
4b: OD+mod. corr.	0.054	0.047	0.028	0.028	0.060	0.343	0.046	0.050	0.057	0.053	0.053	0.054	0.058	0.243	0.033	0.035	0.139
4c: OD+high corr.	0.054	0.050	0.027	0.027	0.063	0.314	0.049	0.052	0.059	0.056	0.056	0.056	0.060	0.210	0.031	0.032	0.347
5a: Alt. π & λ cons.	0.768	0.768	0.416	0.418	0.825	0.728	0.552	0.564	0.952	0.954	0.954	0.897	0.955	0.927	0.779	0.786	0.884
5b: Alt. π & λ diss.	0.492	0.445	0.321	0.321	0.658	0.108	0.563	0.578	0.809	0.810	0.810	0.704	0.817	0.411	0.607	0.619	0.152
5c: Alt. π	0.668	0.654	0.409	0.411	0.064	0.046	0.052	0.054	0.551	0.544	0.544	0.556	0.553	0.516	0.288	0.293	0.259
5d: Alt. λ	0.061	0.058	0.029	0.029	0.754	0.550	0.567	0.580	0.634	0.622	0.622	0.647	0.628	0.452	0.394	0.407	0.426

Table 3: Empirical size or power for 54 SNPs from MAPT. See text for detailed descriptions of scenarios.

Simulation Setting	π : VC	π : SKAT	π : ZINB Wald	π : ZIP Wald	λ : VC	μ : SKAT	λ : ZINB Wald	λ : ZIP Wald	π : VC std.	π : VC Fisher's	π : VC min-p	π : VC Brown's	π : SKAT	π : ZINB	π : ZIP	λ : Poisson Wald HW
1.1a: without X	0.056	0.053	0.009	0.010	0.054	0.046	0.053	0.059	0.056	0.053	0.052	0.058	0.053	0.016	0.017	0.161
1.1b/1.2a: N = 1000	0.051	0.049	0.006	0.006	0.056	0.049	0.051	0.053	0.055	0.052	0.048	0.058	0.050	0.016	0.016	0.194
1.2b: N = 2000	0.045	0.043	0.019	0.019	0.053	0.050	0.048	0.050	0.049	0.048	0.042	0.050	0.045	0.024	0.025	0.116
1.2c: N = 4000	0.054	0.054	0.034	0.034	0.047	0.050	0.052	0.055	0.050	0.050	0.050	0.052	0.053	0.040	0.042	0.095
2a: corr. N=1000	0.053	0.051	0.005	0.005	0.060	0.056	0.056	0.057	0.057	0.054	0.051	0.059	0.051	0.015	0.015	0.211
2b: corr. N=2000	0.052	0.046	0.016	0.017	0.054	0.053	0.055	0.056	0.051	0.050	0.049	0.053	0.053	0.031	0.031	0.128
2c: corr. N=4000	0.050	0.050	0.033	0.033	0.051	0.053	0.050	0.052	0.050	0.048	0.051	0.050	0.050	0.037	0.039	0.092
3a: low corr.	0.055	0.052	0.004	0.006	0.052	0.053	0.107	0.733	0.056	0.050	0.044	0.058	0.051	0.031	0.468	0.268
3b: mod. corr.	0.053	0.048	0.002	0.006	0.067	0.057	0.113	0.827	0.063	0.055	0.048	0.065	0.050	0.032	0.600	0.311
3c: high corr.	0.051	0.047	0.003	0.007	0.070	0.075	0.110	0.983	0.061	0.053	0.049	0.061	0.065	0.033	0.922	0.456
4a: OD+low corr.	0.048	0.048	0.005	0.005	0.058	0.052	0.057	0.060	0.060	0.053	0.049	0.061	0.051	0.019	0.020	0.196
4b: OD+mod. corr.	0.053	0.051	0.005	0.005	0.060	0.056	0.056	0.057	0.057	0.054	0.051	0.059	0.051	0.015	0.015	0.211
4c: OD+high corr.	0.054	0.048	0.005	0.005	0.062	0.080	0.055	0.056	0.060	0.052	0.050	0.061	0.070	0.016	0.018	0.269
5a: Alt. π & λ cons.	0.610	0.601	0.050	0.051	0.796	0.612	0.330	0.336	0.894	0.887	0.826	0.894	0.812	0.281	0.288	0.731
5b: Alt. π & λ diss.	0.468	0.444	0.049	0.048	0.587	0.374	0.306	0.312	0.767	0.756	0.668	0.767	0.601	0.194	0.199	0.216
5c: Alt. π	0.491	0.460	0.044	0.044	0.060	0.050	0.059	0.058	0.389	0.386	0.388	0.402	0.358	0.054	0.054	0.272
5d: Alt. λ	0.065	0.065	0.006	0.006	0.713	0.534	0.323	0.327	0.601	0.588	0.611	0.606	0.424	0.126	0.128	0.393

Table 4:

Empirical size or power for 601 SNPs from MAPT. See text for detailed descriptions of scenarios.

Simulation Setting	π : VC	π_2 : SKAT	π : ZINB Wald	π : ZIP Wald	λ : VC	μ : SKAT	λ : ZINB Wald	λ : ZIP Wald	$\pi \& \lambda$: VC std.	$\pi \& \lambda$: VC Fisher's	$\pi \& \lambda$: VC min-p	$\pi \& \lambda$: VC Brown's	$\pi_2 \& \mu$: SKAT	$\pi \& \lambda$: ZINB	$\pi \& \lambda$: ZIP	$\lambda_{Poisson}$: Wald HW
1.1b/1.2a: N = 1000	0.059	0.053	0.000	0.000	0.058	0.052	0.095	0.092	0.065	0.063	0.063	0.065	0.057	0.007	0.006	0.735
1.2b: N = 2000	0.054	0.051	0.001	0.002	0.055	0.052	0.060	0.060	0.056	0.053	0.056	0.055	0.050	0.009	0.009	0.483
1.2c: N = 4000	0.049	0.048	0.013	0.013	0.050	0.047	0.048	0.050	0.050	0.050	0.051	0.051	0.049	0.020	0.021	0.217
2a: corr. N=1000	0.058	0.056	0.000	0.000	0.055	0.064	0.099	0.097	0.059	0.055	0.055	0.058	0.058	0.005	0.004	0.789
2b: corr. N=2000	0.049	0.049	0.002	0.002	0.054	0.068	0.056	0.057	0.047	0.048	0.051	0.048	0.061	0.010	0.010	0.504
2c: corr. N=4000	0.053	0.050	0.013	0.013	0.056	0.082	0.055	0.056	0.051	0.050	0.051	0.052	0.066	0.019	0.019	0.240

References

- Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, Corneveaux JJ, Hardy J, Vonsattel J-P, Younkin SG, et al. (2014). Genome-wide association meta-analysis of neuropathologic features of alzheimer's disease and related dementias. *PLoS genetics*, 10(9):e1004606. [PubMed: 25188341]
- Bennett D, Schneider J, Arvanitakis Z, and Wilson R (2012a). Overview and findings from the religious orders study. *Current Alzheimer Research*, 9(6):628–645. [PubMed: 22471860]
- Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, and Wilson RS (2012b). Overview and findings from the rush memory and aging project. *Curr Alzheimer Res*, 9(6):646–663. [PubMed: 22471867]
- Chibnik LB, Shulman JM, Leurgans SE, Schneider JA, Wilson RS, Tran D, Aubin C, Buchman AS, Heward CB, Myers AJ, Hardy JA, Huentelman MJ, Corneveaux JJ, Reiman EM, Evans DA, Bennett DA, and De Jager PL (2011). Cr1 is associated with amyloid plaque burden and age-related cognitive decline. *Ann Neurol*, 69(3):560–569. [PubMed: 21391232]
- Cristianini N and Shawe-Taylor J (2000). An introduction to support vector machines and other kernelbased learning methods. Cambridge university press.
- Davies RB (1980). Algorithm as 155: Distribution of a linear combination of χ^2 random variables. *Applied Statistics*, 29(3):323–33.
- Duchesne P and de Micheaux PL (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54:858–862.
- Efron B and Tibshirani RJ (1994). An introduction to the bootstrap Number 57 in Monographs on statistics and applied probability. Chapman & Hall, Boca Raton, Fla [u.a.], [nachdr.] edition. Literaturverz S [413]–425.
- Farebrother R (1984). Algorithm as 204: the distribution of a positive linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):332–339.
- Fisher RA (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.
- Giri M, Zhang M, and Lü Y (2016). Genes associated with alzheimer's disease: an overview and current status. *Clinical interventions in aging*, 11:665. [PubMed: 27274215]
- Greene WH (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.
- Huang Y-T, Vanderweele TJ, and Lin X (2014). Joint analysis of snp and gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat. The Annals of Applied Statistics*, 8(1):352–376.
- Lachenbruch PA (2001). Comparisons of two-part models with competitors. *Statistics in medicine*, 20(8):1215–1234. [PubMed: 11304737]
- Lambert (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45:1452–1458. [PubMed: 24162737]
- Lambert D (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.
- Lee S, Wu MC, and Lin X (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775. [PubMed: 22699862]
- Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, and Lin X (2011). Kernel machine snp-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology Genet. Epidemiol*, 35(7):620–631. [PubMed: 21818772]
- Liu D, Ghosh D, and Lin X (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292. [PubMed: 18577223]
- Liu D, Lin X, and Ghosh D (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088. [PubMed: 18078480]

- Loeys T, Moerkerke B, De Smet O, and Buysse A (2012). The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180. [PubMed: 21950803]
- Min Y and Agresti A (2002). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, 1(1):7–33.
- Parzen MI, Wei LJ, and Ying Z (1994). A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350.
- Pollard D (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS regional conference series in probability and statistics. IMS and ASA.
- Poole W, Gibbs DL, Shmulevich I, Bernard B, and Knijnenburg TA (2016). Combining dependent p-values with an empirical adaptation of brown's method. *Bioinformatics*, 32(17):i430–i436. [PubMed: 27587659]
- Ridout M, Hinde J, and Demétrio CG (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1):219–223. [PubMed: 11252601]
- Shen Y and Cai T (2016). Identifying predictive markers for personalized treatment selection. *Biometrics*, 72(4):1017–1025. [PubMed: 26999054]
- Shulman JM, Chen K, Keenan BT, Chibnik LB, Fleisher A, Thiyyagura P, Roontiva A, McCabe C, Patsopoulos NA, Corneveaux JJ, et al. (2013). Genetic susceptibility for alzheimer disease neuritic plaque pathology. *JAMA neurology*, 70(9):1150–1157. [PubMed: 23836404]
- Su Z, Marchini J, and Donnelly P (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305. [PubMed: 21653516]
- Van den Broek J (1995). A score test for zero inflation in a poisson distribution. *Biometrics*, pages 738–743. [PubMed: 7662855]
- Won S, Morris N, Lu Q, and Elston RC (2009). Choosing an optimal method to combine p-values. *Statistics in medicine*, 28(11):1537–1553. [PubMed: 19266501]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, and Lin X (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, and Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93. [PubMed: 21737059]
- Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Molldrem JJ, and Armistead PM (2013). Kernel machine snp-set testing under multiple candidate kernels. *Genetic epidemiology*, 37(3):267–275. [PubMed: 23471868]
- Xu L, Paterson AD, Turpin W, and Xu W (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7):e0129606. [PubMed: 26148172]
- Zhong K (2017). Locate the cytoband. <https://cytob.herokuapp.com/>. Accessed: 2017–09-10.

- π : VC
- π_2 : SKAT
- π : ZINB Wald
- π : ZIP Wald
- λ : VC
- μ : SKAT
- λ : ZINB Wald
- λ : ZIP Wald
- $\pi \& \lambda$: VC std.
- $\pi \& \lambda$: VC Fisher
- $\pi \& \lambda$: VC min-p
- $\pi \& \lambda$: VC Brown
- $\pi_2 \& \mu$: SKAT Brown
- $\pi \& \lambda$: ZINB Wald
- $\pi \& \lambda$: ZIP Wald
- λ_{Poisson} : Wald HW

Figure 3.1.
Legend for the following plots of simulation results.

Figure 3.2.
Empirical sizes of different testing procedures for Setting 1.1) (a) with **X**; and (b) without **X**.

Figure 3.3.
Empirical sizes of different testing procedures for Setting 1.2) (with $\mathbf{X}$ independent of $\mathbf{G}$) when $n=1000$, 2000, and 4000..

Figure 3.4.
Empirical sizes of different testing procedures for Setting 2) (moderate $\mathbf{X} : \mathbf{G}$ correlation) when $n=1000$, 2000, and 4000..

Figure 3.5.

Empirical sizes of different testing procedures for Setting 3) when the correlation between **X** and **G** is weak, moderate or strong..

Figure 3.7.

Power of different testing procedures (Setting 5) where **G** affects (a) both mixing proportion and Poisson rate via consonant effects; (b) both mixing proportion and Poisson rate via dissonant effects..

Figure 3.9.
Histogram of the neuritic plaque (NP) outcome..

Figure 3.10.

Plot of APOE SNP and APOE- ϵ 4 haplotype (denoted APOE4N) effects from marginal models..

Table 1.Data analysis results: p -values of selected SNP sets in ROSMAP data.

	π : VC	π : ZINB Wald	π_2 : SKAT	λ : VC	λ : ZINB Wald	μ : SKAT	$\pi \& \lambda$: VC std.	$\pi \& \lambda$: ZINB	$\pi_2 \& \mu$: SKAT
ABCA7	2.1E-01	2.1E-02	2.3E-01	1.9E-01	6.6E-01	7.4E-01	1.8E-01	1.1E-01	4.7E-01
APOE	1.1E-08	2.0E-06	1.7E-06	1.6E-06	1.4E-08	2.0E-06	5.8E-12	4.4E-13	2.9E-10
CD33	3.5E-01	6.7E-01	3.6E-01	9.9E-01	3.9E-01	1.0E+00	7.4E-01	5.8E-01	7.3E-01
MAPT	8.0E-01	4.7E-01	8.1E-01	4.1E-01	9.8E-01	3.2E-01	6.7E-01	9.1E-01	6.1E-01
PTPRD	4.0E-01	7.6E-01	4.1E-01	3.0E-04	2.7E-01	1.2E-03	1.6E-03	5.2E-01	4.1E-03