Biostatistics (2016), **17**, 2, *pp.* 221–234
doi:10.1093/biostatistics/kxv035
Advance Access publication on September 28, 2015

Group association test using a hidden Markov model

YICHEN CHENG*, JAMES Y. DAI, CHARLES KOOPERBERG

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

ycheng@fredhutch.org

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

Summary

In the genomic era, group association tests are of great interest. Due to the overwhelming number of individual genomic features, the power of testing for association of a single genomic feature at a time is often very small, as are the effect sizes for most features. Many methods have been proposed to test association of a trait with a group of features within a functional unit as a whole, e.g. all SNPs in a gene, yet few of these methods account for the fact that generally a substantial proportion of the features are not associated with the trait. In this paper, we propose to model the association for each feature in the group as a mixture of features with no association and features with non-zero associations to explicitly account for the possibility that a fraction of features may not be associated with the trait while other features in the group are. The feature-level associations are first estimated by generalized linear models; the sequence of these estimated associations is then modeled by a hidden Markov chain. To test for global association, we develop a modified likelihood ratio test based on a log-likelihood function that ignores higher order dependency plus a penalty term. We derive the asymptotic distribution of the likelihood ratio test under the null hypothesis. Furthermore, we obtain the posterior probability of association for each feature, which provides evidence of feature-level association and is useful for potential follow-up studies. In simulations and data application, we show that our proposed method performs well when compared with existing group association tests especially when there are only few features associated with the outcome.

Keywords: Finite mixture model; Genome-wide association study; Modified likelihood ratio test.

1. Introduction

With the fast growth in high-throughput sequencing technologies, identifying important features that are associated with a disease trait in a genome-wide study has become possible. Notable examples include rare variant association in human genome sequencing studies (Morgenthaler and Thilly, 2007; Morris and Pan, 2010; Wu *and others*, 2009), and sieve analyses in HIV vaccine trials comparing viral sequences between vaccines and placebo recipients (Rolland *and others*, 2012). Due to the overwhelming burden of multiple testing, testing for association between disease and individual features often lack power. To aggregate individual weak associations and reduce the number of tests, numerous methods have been proposed to perform a global association test within a functional unit, e.g. a gene or a pathway. Motivated by, though certainly not limited to, rare-variant association testing, we propose in this paper a novel group association test named "hidden Markov model variable detection method" (HMVD).

*To whom correspondence should be addressed.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Existing association detection methods belong to two general classes: one class is to aggregate individual summary statistics like p-values or Z statistics for each feature. For example, Fisher (1932) construct test statistics based on sum of the p-value for the each feature. Tippett (1931) tests based on minimum of the p-value's for all features. More recent work along includes Chen *and others* (2012) and Cheung *and others* (2012) among others. For example, SigmaP (Cheung *and others*, 2012) tests for association by constructing test statistics using a weighted linear combination of the individual p-values.

The other class involves multiple regression models for the trait and all features together, with different levels of constraints to curb the number of parameters. One group of such tests, called Burden tests, simply collapses multiple features into one explanatory variable. For example, the cohort allelic sum test (Morgenthaler and Thilly, 2007) collapses variants within one region to a new indicator variable, which is one if a subject has at least one rare variant among that region, and zero otherwise. Along this line, Morris and Pan (2010) proposed to count the number of rare variants within a region and Li and Leal (2008) proposed the combined multivariate and collapsing test. These burden tests usually works well if a large portion of the variants are associated with the outcome. Wu *and others* (2009) proposed the sequence kernel association test (SKAT) by testing for the heterogeneity of single-nucleotide polymorphism (SNP) associations. They assumed that the SNP associations follow a normal distribution with zero mean, and tested for whether the distribution of associations has zero variance. SKAT is more powerful than the Burden tests when a small portion of variants are associated with the outcome or when the association of the SNPs go in both directions. Lee *and others* (2013) combined SKAT and Burden tests and introduced an "optimal" test that works well under both of the two aforementioned settings.

One limitation of those aforementioned methods is that they do not explicitly distinguish associated features from functionally neutral features. Thus, in settings with a substantial portion of neutral features, they may not be optimal in power performance. Recently, Logsdon *and others* (2014) proposed a variational Bayes discrete mixture (VBDM) method that incorporates the mixture of associated and neutral variants in the modeling. However, due to the approximation nature of the variational Bayes method, it does not work for dichotomous traits and logistic regression. Other authors have explored model selection techniques in group association tests, including a stepwise variable-selection method (Hoffmann *and others*, 2010) and Bayesian model selection (Capanu and Begg, 2011; Liang and Xiong, 2013).

In this paper, we propose a two-step procedure to conduct group association tests: first we summarize evidence of individual associations using Z statistics or, equivalently, the observed association in a generalized linear model. We next model each feature-level association by a latent indicator variable indexing whether the feature is associated with the outcome. The sequence of estimated individual associations are described probabilistically by a hidden Markov model (HMM). This HMM allows for a flexible structure on the sequence of hidden states yet it is computationally manageable through a modified Baum–Welsh algorithm.

The primary goal of the proposed method is to test the global null hypothesis that there is no association between the set of features and the outcome. Under the alternative hypothesis, each Z statistic or estimated association follows a mixture of two normal distributions: one is centered at zero (the null), the other is centered at the alternative. The mixture reduces to a single normal distribution under the null, and therefore we aim to test whether the number of mixture components is one. A common issue when testing for the number of mixture components is that the model is non-identifiable under the null (Lo *and others*, 2001). Let p be the proportion of the second components and θ be the parameter for the second component. Then under the null when $p = 0$, θ will be non-identifiable. As a result, the standard likelihood ratio test does not have the usual chi-square distribution. Chernoff and Lander (1995) and Sen and Ghosh (1985) among others find that the asymptotic distribution of the LRT involves the supremum of a Gaussian process. Recently, Chen *and others* (2001) proposed to add a $C\log(p) + C\log(1-p)$ term to the log-likelihood function to keep p away from 0 or 1, where C is a user defined positive constant. They show that the parameters are consistently estimated under the null. In our method, we apply a similar approach by adding a

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

$C \log(p)$ term to the log-likelihood function, so the support of p is limited to (0, 1]. Correlated Z statistics are modelled by a hidden Markov chain, so that the proportion of null features and posterior probabilities can be computed by a extension of the Baum–Welsh algorithm. For hypothesis testing, we propose a modified likelihood ratio test ignoring the higher order correlation between features with the aforementioned penalty term. We show that the asymptotic distribution under the null hypothesis follows a χ^2_1 distribution.

Another contribution of our method is its "model selection" feature. On top of its ability to test for overall association, HMVD also provides a posterior probability for each feature, summarizing how likely it is marginally associated with the outcome. This property is desirable since it allows one to identify important features amongst a pool of candidate features, thereby providing guidance for follow-up studies.

2. Methods

In this paper, we focus on testing whether a set of m predictors (or features) ($\boldsymbol{G}$) is related to an outcome variable $\boldsymbol{Y}$ while being adjusted for extra confounding variables (covariates) $\boldsymbol{X}$. For n independent subjects, let $\boldsymbol{Y}$ be an $n \times 1$ vector, $\boldsymbol{X}$ be an $n \times q$ matrix, and let $\boldsymbol{G}$ be an $n \times m$ matrix. We assume that our samples are obtained either from a case–control study or a cohort study; $\boldsymbol{Y}$ can be either a dichotomized disease variable or a continuous response. The traditional way to test association of a group of features is to fit a multiple regression model, $g\{E(\boldsymbol{Y})\} = \boldsymbol{X\beta} + \boldsymbol{G\theta}$, where $\boldsymbol{\beta}$ is the coefficient for $\boldsymbol{X}$, $\boldsymbol{\theta}$ is the coefficient for $\boldsymbol{G}$. The usual F/χ^2 statistic to test $\boldsymbol{\theta} = 0$ will have limited power when m is large.

Rather than fitting one model for all m features, which can be cumbersome when m is large, we fit a separate model for each feature:

$$g\{E(\boldsymbol{Y})\} = \beta_{0k} + \boldsymbol{X\beta}_k + \boldsymbol{G}_k\theta_k, \quad k = 1, \ldots, m, \tag{2.1}$$

where $\boldsymbol{G}_k$ is the kth column of $\boldsymbol{G}$. For the kth feature, the estimated association $\hat{\theta}_k$ and its estimated variance $\hat{\sigma}^2_k$ can be obtained through model (2.1). We assume that within a set of features, only some of these features are associated with the outcome. Let $W_k, k = 1, \ldots, m$, be an indicator whether feature k is associated with the outcome, and assume all associated features share the same effect size θ, we have

$$\hat{\theta}_k = W_k\theta + \epsilon_k, \quad k = 1, \ldots, m, \tag{2.2}$$

where $\epsilon_k \sim \mathcal{N}(0, \sigma^2_k)$. Note that without any rescaling of the features, $\sigma^2_k \to 0$ as n increases. The assumption that all associated features have the same effect size may seem restrictive. However, we note that in many applications where all features are jointly tested these features are "similar", e.g. they are all sequence variants within the same gene that may disrupt the function of that gene in a similar manner. As such, we believe that this assumption can be a reasonable approximation. We note that it is also possible to *a priori* rescale some of the features $\boldsymbol{G}_k$, for example to give them the same variance, or to incorporate some prior believes about effect sizes. In our simulation study, we will see that our method maintains correct control of the type I error and has good power, even if the assumption of the same effect size is not satisfied.

Our testing procedure has two parts: the "generic" regression model that relates $\boldsymbol{Y}$ to $\boldsymbol{G}$ and $\boldsymbol{X}$ (2.1), and the HMM that relates the sequence of estimated association θ_k (2.2). The goal here is to develop a test for association of a group of features using the second model. For this test to be valid, the actual form of (2.1) can be flexible: we require only an estimate $\hat{\theta}_k$ for k features and their estimated covariance matrix. We can also apply our procedure to a set of test statistics $Z_k, k = 1, \ldots, m$, specifying $se(\hat{\theta}'_k) = 1$, for $k = 1, \ldots, m$. Such "weighted version", can be obtained when each feature G_k is divided by the (unweighted) $se(\hat{\theta}_k)$. For the weighted version, model (2.2) gets replaced by $Z_k = W_k Z + \epsilon'_k,\ k = 1, \ldots, m$, where the covariance matrix of the ϵ'_k equals the correlation matrix for ϵ_k. In the remainder of this section, we show results for the

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

unweighted version of our procedure. The derivations for the weighted version are completely analogous and therefore omitted.

Set $\boldsymbol{\psi}_k = (\beta_{0k}, \boldsymbol{\beta}_k, \theta_k)$ and let the corresponding estimating equation be $\boldsymbol{S}_k$, which is a function of $\boldsymbol{\psi}_k$. Then the joint distribution of $(\hat{\theta}_1|W_1, \ldots, \hat{\theta}_m|W_m)$ is

$$\begin{pmatrix} \hat{\theta}_1|w_1 \\ \cdots \\ \hat{\theta}_m|w_m \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} w_1\theta \\ \cdots \\ w_m\theta \end{pmatrix}, \begin{pmatrix} \sigma_1^2, \ldots, \sigma_{1m} \\ \cdots \\ \sigma_{m1}, \ldots, \sigma_m^2 \end{pmatrix} \right),$$

where $\sigma_{kk'}$ is the last diagonal element of $E[\{\partial \hat{\boldsymbol{S}}_k/\partial \boldsymbol{\psi}_k\}^{-1} \hat{\boldsymbol{S}}_k \hat{\boldsymbol{S}}_{k'}^T \{\partial \hat{\boldsymbol{S}}_{k'}/\partial \boldsymbol{\psi}_{k'}\}^{-1}]$. Denote the variance by Σ; its estimate $\hat{\Sigma}$ can be estimated from the data.

2.1 *Hidden Markov model*

To describe the likelihood of observed data, we need to impose some structure on the sequence of latent variables W_k's to use (2.2) for group testing. The simplest structure is to assume that the W_k's are independent, and that each follows a Bernoulli distribution with $P(W_k = 1) = p$: that is, each feature has an equal "chance" of being associated with Y, independent of the other features. In this paper, we assume that the W_k follow a stationary Markov chain with transition matrix $\boldsymbol{A} = (a_{st})$, $s, t = 0, 1$, where $a_{s0} + a_{s1} = 1$ and $a_{st} = P(W_{k+1} = t|W_k = s)$. We assume that the features can be ordered in a meaningful way, and that whether one feature is associated with the response may be informative on whether the next feature is associated with the response. For example, for genetic data, there is a natural order of variants along the chromosome. The Markovian structure implies that for neighboring SNPs, if one is associated with the outcome the other one is more likely to be associated with the outcome. We note that when $a_{00} = a_{10}$, the Markov chain model is equivalent to the simple Bernoulli model with $p = a_{01}$.

We define the emission probabilities, i.e. the conditional distribution of θ_k given $\theta_1, \ldots, \theta_{k-1}$ and the hidden states up to k $(W_1, \ldots, W_k)$ as $p(\hat{\theta}_k|\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}, W_1, \ldots, W_k) \triangleq b_k(\hat{\theta}_k)$. Let $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_m)$ and $\hat{\boldsymbol{\Theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$. When n is sufficiently large, $\hat{\boldsymbol{\Theta}}$ is a multivariate Gaussian distribution given all the hidden states. We obtain the likelihood of $\hat{\boldsymbol{\Theta}}$ by summing over all hidden states: $P(\hat{\boldsymbol{\Theta}}|p, \theta) = \sum_{\mathbf{w}} P(\hat{\boldsymbol{\Theta}}, \boldsymbol{W}|p, \theta) = \sum_{\boldsymbol{w}} \pi_{w_1} b_1(\hat{\theta}_1) \prod_{k=2}^m a_{w_{k-1}, w_k} b_k(\hat{\theta}_k)$, where $\pi_{w_1} = P(W_1 = w_1)$. Contrary to the usual first-order HMMs, the emission probability of a feature depends on the hidden states of the feature itself and all prior features. Therefore, the full likelihood of $\hat{\theta}_1, \ldots, \hat{\theta}_m$ involves summation over all possible combinations of the prior hidden states for each W_k, which is computationally infeasible. Some simplification is needed. We consider two strategies for accounting for the dependency between features: for the hypothesis testing purpose, we develop a modified likelihood ratio test that ignores the higher order dependency and we prove that the resulting test is still valid. For estimating the posterior probability of an individual feature, we use a modified Baum–Welsh algorithm that reduced the dependency in emission probabilities to the first order only, thereby substantially reducing the burden of computation.

We formulate a model similar to the standard hidden Markov model (HMM) with one key difference. In the standard HMM, all observations $\hat{\theta}_k$ are assumed to be conditionally independent, that is, it is assumed that $p(\hat{\theta}_k|\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}, W_1, \ldots, W_k) = p(\hat{\theta}_k|W_k)$. These assumptions are usually not satisfied in our setting, since the effects from neighboring features (e.g. SNPs) are likely to be correlated. We generalize the HMM by allowing first-order conditional dependence between neighboring features and ignoring any higher order correlations.

In particular, we assume that $p(\hat{\theta}_k|\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}, W_1, \ldots, W_k) = p(\hat{\theta}_k|\hat{\theta}_{k-1}, W_{k-1}, W_k) = \phi(\hat{\theta}_k - B_k \hat{\theta}_{k-1}, (W_k - B_k W_{k-1})\theta, \sigma^2_{k|k-1})$, where $B_k = \sigma_{k,k-1}/\sigma^2_{k-1}$, $\sigma^2_{k|k-1} = \sigma^2_k - B_k \sigma_{k-1,k}$ and $\phi(x, \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2/2\sigma^2\}$. For simplicity, we define $b_{w_{k-1}, w_k}(\hat{\theta}_k) = p(\hat{\theta}_k|\hat{\theta}_{k-1}, W_{k-1} = w_{k-1},$

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

$W_k = w_k)$, for $k = 2, \ldots, m$, $b_{w_1}(\hat{\theta}_1) = \phi(\hat{\theta}_1, w_1, \sigma_1^2)$ and $\pi_{w_1} = P(W_1 = w_1)$. Then the likelihood function, by ignoring higher order conditional correlations, can be expressed as

$$P(\hat{\boldsymbol{\Theta}}|p, \theta) = \sum_{\boldsymbol{W}} P(\hat{\boldsymbol{\Theta}}, \boldsymbol{W}|p, \theta) = \sum_{\boldsymbol{w}} \pi_{w_1} b_{w_1}(\hat{\theta}_1) \prod_{k=2}^{m} a_{w_{k-1}, w_k} b_{w_{k-1}, w_k}(\hat{\theta}_k). \tag{2.3}$$

2.2 *Hypothesis testing*

The primary goal of this paper is group testing for association. Assume that the Markov chain W_k is stationary, then $p \triangleq P(W = 1) = a_{01}/(a_{10} + a_{01})$. Testing for association between outcome and features is equivalent to testing for the composite null hypothesis

$$H_0 : \theta = 0 \quad \text{or} \quad p = 0.$$

This set-up has two complications. First, the parameters are not identifiable under H_0; when $p = 0$, θ vanishes in the likelihood, and when $\theta = 0$, p vanishes. Second, there are no existing results about the distribution for the likelihood ratio test statistics under the HMM setting described here. To eliminate the identifiability problem, one approach is to add a small penalty term, that pushes one component of the null parameters away from its boundary value (e.g. Chen *and others*, 2001; Fu *and others*, 2009). In the HMM setting, tests based on the marginal distribution of $\hat{\boldsymbol{\Theta}}$, ignoring dependency, have been proposed (Dannemann and Holzmann, 2008). The rational for this approximation is that the number of states in the HMM is equivalent to the number of components for each $\hat{\theta}_k$. Under H_0, when the HMM only has one state, marginally $\hat{\theta}_k$ follows a normal distribution with mean 0. Under the alternative, when the HMM has two states, marginally $\hat{\theta}_k|\hat{\theta}_{k-1}$ follows a mixture of four normal distributions, with the mixture probabilities being $p_{00} = (1 - p)a_{00}$, $p_{01} = (1 - p)a_{01}$, $p_{10} = pa_{10}$, and $p_{11} = pa_{11}$.

In our method, $\eta_k \triangleq \hat{\theta}_k - B_k\hat{\theta}_{k-1}$ has a distribution that is a mixture of four normal distributions:

$$\begin{aligned} f_{\text{mix}}(\hat{\theta}_k|\hat{\theta}_{k-1}) = {} & p_{00}\phi(\eta_k|0, \sigma^2_{k|k-1}) + p_{01}\phi(\eta_k|\theta, \sigma^2_{k|k-1}) + p_{10}\phi(\eta_k| - B_k\theta, \sigma^2_{k|k-1}) \\ & + p_{11}\phi(\eta_k|(1 - B_k)\theta, \sigma^2_{k|k-1}). \end{aligned} \tag{2.4}$$

Here $\phi(x|\mu, \sigma^2)$ denotes the normal probability density function evaluated with mean μ and variance σ^2. Then the likelihood function for $\hat{\boldsymbol{\Theta}}$, ignoring higher order dependency, is

$$l_m^I(p, \theta) = \sum_{k=1}^{m} \log f_{\text{mix}}(\hat{\theta}_k|\hat{\theta}_{k-1}). \tag{2.5}$$

We propose the following modified log-likelihood function:

$$l_m^{IP}(\theta, A) = \sum_{k=1}^{m} \log f_{\text{mix}}(\hat{\theta}_k|\hat{\theta}_{k-1}) + C\log(a_{01}) + C\log(a_{11}), \tag{2.6}$$

and we construct a likelihood ratio test based on this expression.

As discussed in Chen *and others* (2001), the best choice of the tuning parameter C depends on the model. In our simulation study we find that the value of C affects both the type I error and the power: setting C too large will reduce the power while setting C too small will lead to some inflation of the type I error. We found that $C = 1$ yields a good balance between type I error and power.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

Interestingly, maximizing (2.6) will still give us a consistent estimator of θ. Related results on the consistency for MLEs while assuming independence have been discussed in Chandler and Bate (2007) and Zhou *and others* (2001).

Note both $\hat{\theta}_k$ and $\sigma_{k|k-1}$ are of order $n^{-1/2}$, it is natural to make the following transformation and rewrite the log-likelihood function. Define $\tilde{\eta}_k = \sqrt{n}\eta_k$, $\tilde{\theta} = \sqrt{n}\theta$, and $\tilde{\sigma}_k^2 = n\sigma_{k|k-1}^2$. Then equation (2.4) can be written as $f_{\text{mix}}(\tilde{\eta}_k) = p_{00}\phi(\tilde{\eta}_k|0, \tilde{\sigma}_k^2) + p_{01}\phi(\tilde{\eta}_k|\tilde{\theta}, \tilde{\sigma}_k^2) + p_{10}\phi(\tilde{\eta}_k| - B_k\tilde{\theta}, \tilde{\sigma}_k^2) + p_{11}\phi(\tilde{\eta}_k|(1 - B_k)\tilde{\theta}, \tilde{\sigma}_k^2)$ and the modified log-likelihood function can be rewritten as

$$l_m^{IP}(\tilde{\theta}, A) = \sum_{k=1}^{m} \log f_{\text{mix}}(\tilde{\eta}_k) + C\log(a_{01}) + C\log(a_{11}), \tag{2.7}$$

Now define the modified likelihood ratio test (MLRT) statistics as

$$R_m = 2\{l_m^I(\hat{\tilde{\theta}}, \tilde{A}) - l_m^I(\theta = 0, A)\}\frac{\sum \text{diag}(\boldsymbol{\Omega}_\theta)}{\sum \boldsymbol{\Omega}_\theta}, \tag{2.8}$$

where $\boldsymbol{\Omega}_\theta$ is the covariance matrix for $(1 - B_k)\tilde{\eta}_k/\tilde{\sigma}_k^2, k = 1, \ldots, m$. $\sum \text{diag}(\boldsymbol{\Omega}_\theta)$ is the sum of all diagonal elements of $\boldsymbol{\Omega}_\theta$ and $\sum \boldsymbol{\Omega}_\theta$ is the sum of all elements in $\boldsymbol{\Omega}_\theta$. Then the following two theorems show the consistency of the estimator of $\tilde{\theta}$ and the asymptotic distribution of R_m under the null hypothesis.

THEOREM 2.1 If Conditions 1–5 (given in Supplementary material available at *Biostatistics* online) hold, the maximum modified likelihood estimator is consistent under the null.

THEOREM 2.2 If Conditions 1–5 hold, then under the null hypothesis the modified likelihood ratio test statistics defined in equation (2.8) follows a χ_1^2 distribution as $n \to \infty$ and $m \to \infty$.

The proofs of the theorems are given in Supplementary material available at *Biostatistics* online. For data with limited sample size and small number of features, we observed small inflation of type I errors. We refer to the p-values that are obtained using Theorem 2.2 as approximate p-values. In cases where accurate p-values are needed, we propose the following adaptive permutation strategy. For each test, we obtain $\hat{\boldsymbol{\Theta}}$ and $\hat{\Sigma}$ as described in Section 2 and the approximate p-value p^0. Under the null hypothesis, $\hat{\boldsymbol{\Theta}}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\hat{\Sigma}$, denoted by MVN($\mathbf{0}$, $\hat{\Sigma}$). Given a significance level α, the adaptive permutation tests proceeds as follows. The first step is to determine whether a permutation test is needed. If p^0 is large enough compared with α, then there is no need to conduct a permutation test. In both the power comparison and type I error calculation in Section 3, we do not initiate the permutation test if $p^0 > 5\alpha$. Otherwise, we set the number of permutations nperm $= \min([N/p^0], [N/\alpha])$, where $[a]$ is the greatest integer that is less than a and $N > 1$ is a user defined parameter. Throughout the paper, we set N to be 10. For permutation i, we generate $\boldsymbol{\Theta}^i$ from MVN($\mathbf{0}$, $\hat{\Sigma}$) and calculate the corresponding approximate p-value p^i, $i = 1, \ldots,$ nperm. Then a test is claimed to be significant at level α if $(\sum_i (p^0 < p^i) + 1)/(\text{nperm} + 1) < \alpha$. The proposed adaptive permutation procedure is similar to the method introduced in Besag and Clifford (1991).

2.3 *Estimation*

Recall that the likelihood function based on the HMM structure, while ignoring higher order dependency, is $P(\hat{\boldsymbol{\Theta}}|p, \theta) = \sum_{\boldsymbol{w}} P(\hat{\boldsymbol{\Theta}}, \boldsymbol{W}|p, \theta) = \sum_{\boldsymbol{w}} \pi_{w_1} b_{w_1}(\hat{\theta}_1) \prod_{k=2}^{m} a_{w_{k-1}, w_k} b_{w_{k-1}, w_k}(\hat{\theta}_k)$. To deal with the composite null issue discussed in the previous section, we add a penalty term to the above likelihood function and

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

obtain the following penalized version:

$$P^P(\hat{\boldsymbol{\Theta}}|p,\theta) = \sum_{\boldsymbol{w}} \pi_{w_1} b_{w_1}(\hat{\theta}_1) \prod_{k=2}^{m} a_{w_{k-1},w_k} b_{w_{k-1},w_k}(\hat{\theta}_k) a_{01}^C a_{11}^C. \tag{2.9}$$

We can estimate θ by maximizing expression (2.9). Maximization can be done using a modified Baum–Welsh algorithm. A detailed description of this algorithm is given in Supplementary material available at *Biostatistics* online.

3. Simulation study

In this section, we conduct simulation study to evaluate the performance of the proposed HMVD method in settings with continuous and binary features and in settings with continuous and binary outcomes.

3.1 *Type I error control for finite sample*

We perform HMVD tests under the null hypothesis and calculate type I error rates in various settings. For each setting, the empirical type I error rate is obtained based on independent simulated datasets. Sample sizes are $n = 2000$ and $n = 4000$. The outcome Y is generated either from a standard normal distribution or from a Bernoulli distribution with $p = 0.5$.

For continuous features, we generated 20 and 40 features with two different covariance matrix structures: independent features or features with an "AR1" correlation structure $\text{cor}(G_k, G_{k+d}) = 0.25^{|d|}$. The marginal variance of each feature was one for all simulations.

For discrete features, we generate genetic data mimicking real SNP data. We use HAPGEN2 (Spencer *and others*, 2009) and the CEU sample from the Hapmap project (data release #24) to generate genotypes of the Von Willebrand Factor (*VWF*) gene on chromosome 12. To generate rare variants, we only select those variants with minor allele frequency (MAF) < 0.05. We select the first 20 or 40 variants to be consistent with the setting for continuous features.

For each simulation, we calculate the proportion of times the group of features is declared as significantly associated with the outcome at three thresholds ($\alpha = 0.01$, 0.001, and 0.0001). For $\alpha = 0.01$, we generated 10^5 simulated datasets and for $\alpha = 0.001$ and $\alpha = 0.0001$, we generated 10^6 simulated datasets to test for the type I error rates. Results for independent features and discrete features are given in Tables 1–2. Overall, the type I error rates are well controlled. The p-values are conservative for small sample size when outcome is binary and features are generated based on VWF gene. This is due to the conservative standard error obtained using logistic regression for small sample size when MAFs are small. Results for features generated from "AR1" correlation structure are given in Supplementary material available at *Biostatistics* online. The type I error rates are also well controlled.

3.2 *Power comparison*

To study the power of our proposed approach, we compare HMVD with other popular group testing methods: SKAT (Wu *and others*, 2009), SKAT-O (Lee *and others*, 2013), VBDM (Logsdon *and others*, 2014), SigmaP (Cheung *and others*, 2012), and the Burden test (Morris and Pan, 2010). We examine the power for the different methods across a variety of settings. The features that are associated with the response are selected independently with probability $p = 0.1$ (sparse signal) or 0.2 (denser signal). The number of features is again set to be either 20 or 40. Note that for $m = 20$ and $p = 0.1$ in $\sim$12% of the simulations there is no signal, so for this setting the maximum power is around 88%. The sample size is set to be 4000

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

Table 1. *Empirical type I error rates for continuous features with identity covariance matrix*

Outcome	m	n	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$
Continuous	20	2000	1.00×10^{-2}	1.04×10^{-3}	1.06×10^{-4}
Continuous	20	4000	1.00×10^{-2}	1.00×10^{-3}	0.87×10^{-4}
Continuous	40	2000	0.95×10^{-2}	0.98×10^{-3}	1.01×10^{-4}
Continuous	40	4000	1.02×10^{-2}	0.99×10^{-3}	1.11×10^{-4}
Binary	20	2000	1.04×10^{-2}	1.13×10^{-3}	0.94×10^{-4}
Binary	20	4000	1.11×10^{-2}	1.11×10^{-3}	0.96×10^{-4}
Binary	40	2000	0.98×10^{-2}	1.01×10^{-3}	1.08×10^{-4}
Binary	40	4000	1.06×10^{-2}	1.06×10^{-3}	0.91×10^{-4}

Table 2. *Empirical type I error rates for discrete features generated from the VWF gene*

Outcome	m	n	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$
Continuous	20	2000	0.95×10^{-2}	0.96×10^{-3}	0.93×10^{-4}
Continuous	20	4000	1.00×10^{-2}	0.99×10^{-3}	1.13×10^{-4}
Continuous	40	2000	0.97×10^{-2}	0.98×10^{-3}	0.98×10^{-4}
Continuous	40	4000	0.97×10^{-2}	0.96×10^{-3}	1.08×10^{-4}
Binary	20	2000	0.84×10^{-2}	0.66×10^{-3}	0.60×10^{-4}
Binary	20	4000	0.90×10^{-2}	0.81×10^{-3}	0.76×10^{-4}
Binary	40	2000	0.82×10^{-2}	0.71×10^{-3}	0.66×10^{-4}
Binary	40	4000	0.89×10^{-2}	0.81×10^{-3}	0.82×10^{-4}

for all simulations. Comparisons are made for both continuous and ordinal features, and for continuous and binary outcomes Y.

For continuous outcomes, the data are generated from the model $Y = \sum \theta_k G_k + \epsilon_k$, where ϵ_k are taken iid normal with mean 0 and variance 4. For binary outcomes, the data are generated from the model $\text{logit}\{P(Y = 1)\} = \sum \theta_k G_k$.

The effect sizes θ_k are generated under two scenarios: In the first scenario, we assume that the features associated with the outcome share the same effect size, i.e., $\theta_k = 0$ or $\theta_k = \theta$ (when our model is correct); In the second scenario we assume that the associated features have different effect sizes, such that $\theta_k = 0$ or $\theta_k \sim \theta \times \mathcal{N}(1, 2)$. This scenario is used to study the robustness of HMVD against model misspecification. We note that in the second scenario on average 31% of the truly associated θ_k will have effects that are in the opposite direction from the other θ_k. Under each scenario, the associated features are chosen randomly and we report the percentages of tests that are significant at level $\alpha = 0.005$ over 1000 runs.

We generate continuous features G_k using the same covariance matrix structures as we used for the simulations for evaluating type I error. The effect size θ take the value 0.04, 0.08, 0.12, or 0.16. These effect sizes are chosen such that the resulting powers cover a reasonable range. Figure 1 shows the comparison for the independent covariance structure. The top 2 panels are for continuous outcomes. All associated features share the common effect in the first panel and have different effect sizes in the second panel. The bottom 2 panels are for binary outcomes with common and different effect sizes. Results for features generated from "AR1" correlation structure are given in Supplementary material available at *Biostatistics* online.

For ordinal features, we compare power based on data generated using the *VWF* gene, as described in previous subsection. The MAF spectrum for selected SNPs in *VWF* gene is given in Appendix of

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

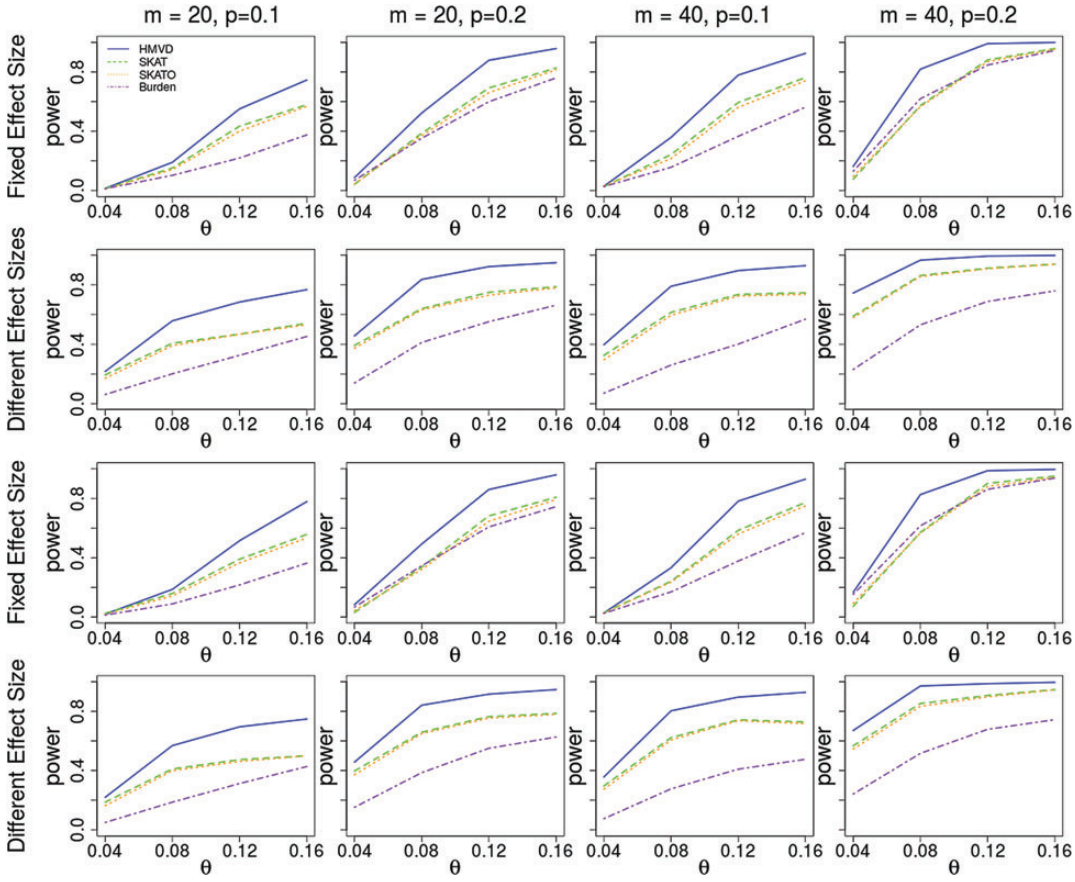

Fig. 1. Power comparison for continuous features with independent covariance structure. Top 2 panels are for continuous outcomes and bottom 2 panels are for binary outcomes.

Supplementary material available at *Biostatistics* online. The effect size θ takes value 0.3, 0.6, 0.9, and 1.2. Results for ordinal features are given in Figure 2 with the same layout as Figure 1.

The proposed method and VBDM yield the greatest power for the continuous outcome binary predictor case (Figure 2). For all other set ups, HMVD has the best power, even for settings with model misspecification. The setting with "$m = 20,\ p = 0.2$" and the setting with "$m = 40,\ p = 0.1$" provide equivalent numbers of truly associated predictors but with different numbers of noise features (predictors that are not associated with the outcome). The effect of more noise features can be observed by comparing column 2 and column 3 in each figure. Noise features have the strongest effect on the Burden test: the power is greatly reduced when the number of noise features increases. Noise features have some effect on all other methods, though the effect on HMVD is minimal in many cases, suggesting that HMVD has the ability to better discern signals from noise. As mentioned earlier, model misspecification does not affect the performance of the proposed method. Another interesting observation is that our method show substantial advantages when the features are generated independently. This is plausible since we first conduct a variable by variable regression for each feature. The correlation structure is not explicitly modeled in the likelihood ratio test. We observe a slightly smaller advantage when all features are correlated. This is

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

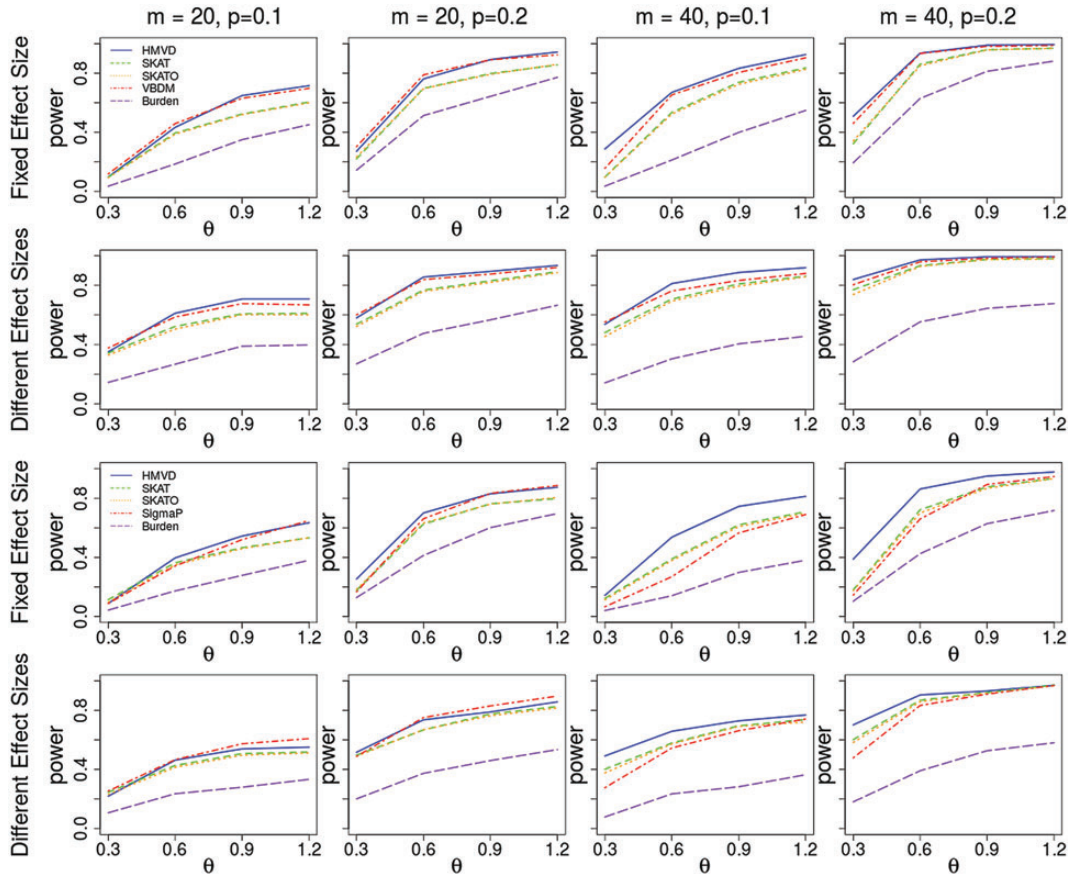

Fig. 2. Power comparison for simulations based on simulated rare variants (MAF < 0.05) of *VWF* gene. Top 2 panels are for continuous outcomes and bottom 2 panels are for binary outcomes.

Table 3. *Aggregate association test results for association between imputed missense variants within the VWF gene and log-transformed VWF levels in African Americans in the CARe consortium*

	Burden	SKAT	SKAT-O	VBDM	HMVD
Unweighted	0.56	2.8×10^{-4}	6.9×10^{-4}	0.60	$< 1 \times 10^{-7}$
Weighted	0.58	1.6×10^{-7}	1.1×10^{-6}	$< 1 \times 10^{-7}$	$< 1 \times 10^{-6}$

expected since we are conducting tests based on estimated effects: when all features are positively correlated, the effective sample size is reduced, which affects the power.

4. Data analysis

We apply our method to the same dataset from the Exome Sequencing Project (ESP) African American (AA) participants as studied in Logsdon *and others* (2014) and Johnsen *and others* (2013), which study the relationship between variants in the Von Willebrand Factor *(VWF)* gene and Von Willebrand Factor level. The Von Willebrand Factor (VWF) is a blood glycoprotein. It plays important roles in platelet adhesion.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

Table 4. *VWF variants statistics and posterior probabilities. Variants with posterior probabilities greater than 0.5 are highlighted in bold face*

Variant	Position	Annotation	MAF	PP
rs143743709	chr12:6058245	Val2793Ala	0.0069	4.0e−7
rs7962217	chr12:6061559	Gly2705Arg	0.016	6.4e−8
rs151129435	chr12:6061636	Asn2679Ser	0.002	1.5e−7
rs35335161	chr12:6078424	Phe2561Tyr	0.025	5.9e−5
rs145697622	chr12:6091089	Arg2384Trp	0.0023	1.7e−2
rs112319661	chr12:6092338	Glu2353Asp	0.0066	4.7e−6
rs61750625	**chr12:6094771**	**Arg2287Trp**	**0.0076**	**1.00**
rs34230288	chr12:6103094	Ala2178Ser	0.0038	1.4e−3
rs61750615	chr12:6103650	Pro2063Ser	0.0022	1.3e−6
rs146729537	chr12:6125326	Ala1795Val	0.0017	5.7e−4
rs78302129	chr12:6125820	Pro1725Ser	0.023	1.6e−6
rs149424724	**chr12:6128127**	**Ser1486Leu**	**0.0080**	**1.00**
rs150077670	**chr12:6128269**	**Val1439Met**	**0.0046**	**1.00**
rs141211612	**chr12:6128454**	**Ala1377Val**	**0.0023**	**1.00**
rs138900040	chr12:6128716	Glu1290Lys	0.0011	1.6e−3
—	chr12:6128898	Val1229Gly	0.0018	4.1e−5
rs145125264	chr12:6135091	Gln1030Arg	0.0077	3.05e−5
rs141087261	chr12:6138575	Gly967Asp	0.027	6.7e−6
rs143762054	chr12:6143913	Leu876Phe	0.00118	4.3e−6
rs143904314	chr12:6153559	Asn780Lys	0.0035	6.5e−9
—	chr12:6166076	Ala631Val	0.0013	7.4e−6
—	chr12:6166151	Arg606Gln	0.0023	7.7e−6
rs144817575	chr12:6172190	Ala488Gly	0.0032	5.5e−6
rs111971143	chr12:6181569	Thr346Ile	0.013	3.3e−7
rs71582882	chr12:6219663	Val137Leu	0.0031	1.5e−6
rs76505074	chr12:6219681	Gly131Ser	0.020	3.7e−2
rs61753991	chr12:6219687	Leu129Met	0.0089	1.5e−6
rs147514785	chr12:6220051	Val102Met	0.0056	3.6e−7

Low levels of VWF are associated with higher risk of Von Willebrand disease (VWD), a type of bleeding disorder. Johnsen *and others* (2013) and Logsdon *and others* (2014) showed that VWF missense variants are associated with VWF levels within the AA population.

The dataset consists of 2487 AA subjects from the CARe consortium who were imputed using the 1000-genomes reference panel as part of the ESP. Details of the imputation are in Auer *and others* (2012). There are a total of 30 imputed missense VWF variants. Two pairs of the variants are highly correlated with correlation > 0.9; we remove one from each pair to reduce the collinearity and end up with 28 imputed missense VWF variants. We apply our method to study the association between 28 imputed missense VWF variants and log-transformed VWF level. In this section, we apply both weighted and unweighted version of our method for testing association. We obtain strong evidence of association with a χ^2 p-value of 9.1×10^{-8} and 4.2×10^{-9} and permutation based p-value $< 10^{-6}$ and $< 10^{-7}$ for the weighted and unweighted procedures, respectively. SKAT and SKAT-O also yield significant results, but the significance levels are not as strong as those for HMVD. For the VBDM method of Logsdon *and others* (2014), the weighted version found a significant association; however, the unweighted version failed to establish association. In Table 3, p-values using different methods are summarized for comparison. It should also be noted that all other methods seem to be sensitive to the choice of weights.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

Our method identifies four low-frequency variants that have a strong association with the outcome (with posterior probabilities >0.99) and all the other variants seems to contribute little to the association (with posterior probabilities close to 0); see Table 4. Three of the four variants selected by HMVD are Arg2287Trp, Ser1486Leu, and Val1439Met, which concord perfectly with the results reported in Johnsen *and others* (2013). Interestingly, in Johnsen *and others*'s report, the effects of these three variants are of similar size, so it is not surprising that the unweighted version of our test establishes a significant effect (the weighted version provided similar results). The variants that show strong evidence of association are shown in bold.

5. Discussion

In this paper, we propose HMVD, a general group association test. By introducing a hidden indicator variable, our method explicitly accounts for the fact that a portion of variables might not be associated with the outcome. Compared with other group association tests HMVD yields increased power in a variety of scenarios. Our method is especially powerful when only a small fraction of all candidate features are associated with the outcome. Under both the HMM framework and the marginal independent likelihood framework, we obtain a posterior probability for each predictor being associated with the outcome. These probabilities serve as evidence of association for follow-up studies. We establish asymptotic distribution results of the HMVD test statistics for any generalized linear model.

As is evident from our example, HMVD is not very sensitive to the choice of weighting scheme, while other methods may provide different conclusions based on the choices of weights. This property is desirable since the "true" weight is never known, and ideally choice of weights should not affect the conclusions.

There are several possible extensions of our approach. First, in this paper, we assume all the associated features to have effects in one direction. Extensions can be made to allow both features with positive association and features with negative association. Second, although we focus on group-wise association test, our framework can be easily extended to other settings such as the identification of gene–environment interaction, or group-wise testing when features have entirely different scales.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors thank Benjamin Logsdon and the authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO), and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). In addition, this research was supported by National Institute of Health grants R01 HG-006124, P01 CA-53996, and R01 HL-114901. *Conflict of Interest*: None declared.

References

Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., Zhang, G., Franceschini, N., Fox, K., Lange, E. M. *and others*. (2012). Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in african americans: NHLBI go exome sequencing project. *American Journal of Human Genetics* **91**, 794–808.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

Besag, J. and Clifford, P. (1991). Sequential monte carlo p-values. *Biometrika* **78**, 301–304.

Capanu, M. and Begg, C. B. (2011). Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics* **67**, 371–380.

Chandler, R. E. and Bate, S. (2007). Inference for clustered data using the independence log-likelihood. *Biometrika* **94**, 167–183.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of Royal Statistical Society, Series B* **63**, 19–29.

Chen, L. S., Hsu, L., Gamazon, E. R., Cox, N. J. and Nicolae, D. L. (2012). An exponential combination procedure for set-based association tests in sequencing studies. *American Journal of Human Genetics* **91**, 977–986.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference* **25**, 579–586.

Cheung, Y. H., Wang, G., Leal, S. M. and Wang, S. (2012). A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genetic Epidemiology* **36**, 675–685.

Fisher, R. A. (1932) *Statistical Methods for Research Workers*, 4th edition. NewYork: Springer.

Fu, Y., Chen, J. and Kalbfleisch, J. D. (2009). Modified likelihood ratio test for homogeneity in a two-sample problem. *Statistica Sinica* **19**, 1603–1619.

Hoffmann, T. J., Marini, N. J. and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* **5**, e13584.

Johnsen, J. M., Auer, P. L., Morrison, A. C., Jiao, S., Wei, P., Haessler, J., Fox, K., McGee, S. R., Smith, J. D., Carlson, C. S. *and others*. (2013). Common and rare von willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in african americans: the NHLBI exome sequencing project. *Blood* **122**, 590–597.

Lee, S., Wu, M. C. and Lin, X. (2013). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775.

Li, B. and Leal, S. M.. (2008). Optimal tests for rare variant effects in sequencing association studies. *The American Journal of Human Genetics* **3**, 311–321.

Liang, F. and Xiong, M. (2013). Bayesian detection of disease–associated rare variants under posterior consistency. *PLoS ONE* **8**, e69633.

Lo, Y., Mendell, N. R. and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* **88**, 767–778.

Logsdon, B. A., Dai, J. Y., Auer, P. L., Johnsen, J. M., Ganesh, S. K., Smith, N. L., Wilson, J. G., Tracy, R. P., Lange, L. A., Jiao, S. *and others*. (2014). A variational Bayes discrete mixture test for rare variant association. *Genetic Epidemiology* **38**, 21–30.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research* **615**, 28–56.

Morris, A. P. and Pan, W. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* **34**, 188–193.

Rolland, M., Edlefsen, P. T., Larsen, B. B., Tovanabutra, S., Sanders-Buell, E., Hertz, T., deCamp, A. C., Carrico, C., Menis, S., Magaret, C. A. *and others*. (2012). Increased HIV-1 vaccine efficacy against viruses with genetic signatures in env v2. *Nature* **490**, 417–420.

Sen, P. K. and Ghosh, J. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer Volume II, 789–806.

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016

SPENCER, C., SU, Z., DONNELLY, P. AND MARCHINI, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* **5**, e1000477.

TIPPETT, L. H. C. (1931) *The Methods of Statistics*. London: Williams and Norgate.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. AND LIN, X. (2009). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93.

ZHOU, H., XIE, M., SIMPSON, D. G. AND WEINBERG, C. R. (2001). A generalized likelihood ratio approach for cluster-correlated data from human fertility studies. *The Indian Journal of Statistics* **63**, 56–68.

[*Received May 11, 2015; revised August 24, 2015; accepted for publication August 25, 2015*]

Downloaded from http://biostatistics.oxfordjournals.org/ at University of Washington on March 4, 2016