Journal of Human Genetics (2014) 59, 198–205

www.nature.com/jhg

ORIGINAL ARTICLE

A stepwise likelihood ratio test procedure for rare variant selection in case–control studies

Anthony YC Kuk[1], David J Nott[1] and Yaning Yang[2]

There is much recent interest in finding rare genetic variants associated with various diseases. Owing to the scarcity of rare mutations, single-variant analyses often lack power. To enable pooling of information across variants, we use a random effect formulation within a retrospective modeling framework that respects the retrospective data collecting mechanism of case–control studies. More concretely, we model the control allele frequencies of the variants as random effects, and the systematic differences between the case and control frequencies as fixed effects, resulting in a mixed model. The use of Poisson approximation and gamma-distributed random effects results in a generalized negative binomial distribution for the joint distribution of the control and case frequencies. Variants are selected by conducting stepwise likelihood ratio tests. The superiority of the proposed method over two existing variant selection methods is demonstrated in a simulation study. The effects of non-gamma random effects and correlated variants are also found to be not too detrimental in the simulation study. When the proposed procedure is applied to identify rare variants associated with obesity, it identifies one additional variant not picked up by existing methods.
Journal of Human Genetics (2014) **59,** 198–205; doi:10.1038/jhg.2014.1; published online 23 January 2014

Keywords: case–control studies; mixed model; negative binomial distribution; Poisson approximation; rare variant selection; retrospective likelihood

INTRODUCTION

Despite the success of genome-wide association studies in identifying genetic variants associated with many diseases and traits,[1] there are still many common diseases that cannot be explained by common genetic variants. Furthermore, the common variants identified through genome-wide association studies often account for only a small fraction of the heritability of the disease.[2] This has led to discoveries that some common diseases are caused by the aggregate effect of multiple rare variants that individually have little impact. It has also been reported that rare variants tend to be functional alleles and have stronger effects on complex diseases than common variants.[3] Recent advances in technology have made it possible to re-sequence large stretches of a genome in a cost-effective way. With the advent of the next-generation sequencing data, the time is ripe for rare variant analysis, and as a result there is a huge surge of papers on this topic.

The analysis of rare variants, however, presents many new challenges. Most existing methods of data analysis are not designed with rare attributes in mind and their naive application will lead to imprecise estimates and tests of low power. To overcome this limitation, various strategies have been proposed to handle rare variant data, including collapsing,[4] weighting,[5] thresholding[6] and pooling.[7,8] Neale *et al.*[9] propose a C-alpha test based on comparing the expected variance with the actual variance of the distribution of allele frequencies. Lin and Tang[10] propose the use of score-type tests, and Wu *et al.*[11] propose the sequence kernel association test within the framework of a random variant effects model. All the above procedures are concerned with testing the overall significance of a collection of variants rather than variant selection that is the focus of this article. Two rare variant selection procedures that we are aware of are the 'RARECOVER' method proposed by Bhatia *et al.*,[12] and the increase in score statistic procedure proposed by Hoffmann *et al.*[13] In developing RARECOVER, Bhatia *et al.*[12] propose taking the union of rare genetic variants, which they define as those with minor allele frequency (MAF) between 0.0001 and 0.1. The union variant is said to have occurred if one or more of its component variants had occurred. By taking their union, variants with low individual MAF are combined to form a union variant with a higher frequency of occurrence that is more amenable to conventional statistical analyses. RARECOVER is basically a step-up greedy procedure whereby at each step the variant that maximizes the Pearson's χ^2-statistic upon taking union with the variants selected so far in the current set S (which is set to the empty set ϕ initially) is added to S if the increase in Pearson's statistic exceeds a certain threshold c. Bhatia *et al.*[12] commented that the choice of c is not crucial, and they used $c = 0.5$. We demonstrate that this choice of c is much too liberal, leading to hugely inflated type I error and very high false selection

[1]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore and [2]Department of Statistics and Finance, University of Science and Technology, Hefei, China
Correspondence: Professor AYC Kuk, Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546, Singapore.
E-mail: stakuka@nus.edu.sg
Received 30 September 2013; revised 26 December 2013; accepted 26 December 2013; published online 23 January 2014

rates. As a remedy, we propose a random permutation approach to determine c for a given nominal level of the type I error. For data collected from a prospective study, the unweighted version of the procedure of Hoffmann *et al.*[13] is based on statistics of the form

$$T_S = \frac{\left\{\sum_{i=1}^{N}\sum_{j\in S}(R_{ij}-\bar{R}_j)(D_i-\bar{D})\right\}^2}{\sum_{i=1}^{N}\left\{\sum_{j\in S}(R_{ij}-\bar{R}_j)(D_i-\bar{D})\right\}^2},$$

where N is the sample size, D_i the disease status of subject i, $i=1,\ldots N$, $\bar{D}=\sum_{i=1}^{N}D_i/N$, $R_{ij}=1$ if subject i has rare variant j, $\bar{R}_j=\sum_{i=1}^{N}R_{ij}/N$, and S is a subset of $\{1,\ldots,J\}$. Hoffmann *et al.*[13] interpreted T_S as the score test statistic for testing H_0: $\beta=0$ in the logistic model

$$\text{logit}\{\Pr(D_i=1)\}=\alpha+\beta\sum_{j\in S}R_{ij},$$

where S denote the set of variants included in the above model. Similar to RARECOVER, Hoffmann, Marini and Witte's SCORE procedure is a step-up procedure whereby at each step, the variant which maximizes the score test statistic is added to the current set S if the increase in score statistic exceeds a certain threshold c. Again, we will use the random permutation approach to determine c.

As useful information is at a premium for rare variant analysis, it is clear that some kind of pooling of information is necessary. The way we propose to do this is to treat the control frequencies of all the rare variants, which are not of direct interest, as random effects that follow a common distribution; and the effects of disease on the causal variants that are of substantive interest as fixed effects, resulting in a mixed model. Although mixed models have been used before in the genome-wide association studies literature, they are mostly based on the prospective[14] approach of modeling the probability of disease occurrence, given the genetic variants of an individual. We choose to model instead the distribution of the genetic variants of a person given his/her disease status that is more in line with the retrospective[14] nature of a case–control study. It will be shown in the Materials and methods section that the use of Poisson approximation and gamma-distributed random effects results in a generalized negative binomial distribution for the joint distribution of the control and case frequencies. Variants are selected by conducting stepwise likelihood ratio tests (LRTs) based on the generalized negative binomial likelihood. Again, a random permutation approach is used to determine the critical value to account for multiple testing. The superiority of the proposed method over RARECOVER and SCORE is demonstrated in a simulation study. When applied to identify rare variants associated with obesity, the proposed stepwise LRT procedure identifies one additional variant that is not picked up by RARECOVER and SCORE.

MATERIALS AND METHODS

One hindrance in using the retrospective modeling approach (which focuses on the distribution of the genetic variants given disease status) when there are multiple variants is the dearth of distributions for multivariate binary/discrete data. Fortunately, it is common in the rare variant literature to assume independence between rare variants (RVs); see Li and Leal,[4] Neale *et al.*,[9] Bhatia *et al.*,[12] and the references therein. With this independence assumption, the retrospective approach is greatly simplified because we can model the allele frequency given disease status one variant at a time.

Generalized negative binomial likelihood for the collapsed frequencies

Suppose there are n_0 controls, n_1 cases and J rare genetic variants under consideration. To focus on issues that are particular to rare variants, we consider only RVs in our analysis that is also what Bhatia *et al.*[12] did. As rare mutations are infrequently observed, for each individual and at each marker, we will combine the scenario of having 'two mutant alleles' with that of 'one mutant allele only' into a merged category of 'at least one mutant allele' in the hope that the merged category will have slightly larger frequency than the original MAF. Another advantage of merging genotypes 1 and 2 is that it frees us from making the assumption of Hardy–Weinberg equilibrium that is highly unlikely to be true for rare alleles. Thus, for $j=1,\ldots,J$ we define Y_{j0} as the number of individuals among the n_0 controls who have 'at least one occurrence' of the jth RV. Similarly, Y_{j1} is the number of individuals among the n_1 cases who have 'at least one occurrence' of the jth RV. To respect the data generating mechanism of case–control studies, we adopt a retrospective approach to model the data as

$$Y_{j0} \sim \text{Binomial}(n_0, p_{j0}) \tag{1}$$

independently of

$$Y_{j1} \sim \text{Binomial}(n_1, p_{j1}), \tag{2}$$

given the probabilities p_{j0} and p_{j1} of at least one occurrence of the jth RV for the controls and cases, respectively. We refer to Y_{j0} and Y_{j1} as the collapsed frequencies in this paper. As explained earlier, we assume independence between RVs that means that the (Y_{j0},Y_{j1}) for different j are independent. Now rather than treating p_{j0} and p_{j1} as fixed parameters, and there are lots of them if J is large, we reduce the number of parameters by treating $p_{10},p_{20},\ldots,p_{J0}$ as random effects generated from a common distribution that enables pooling of information across variants. The fact that the alleles are rare means that the p_{J0} and p_{J1} are small, and so if the sample sizes n_0 and n_1 are reasonably large, the two binomial distributions given by (1) and (2) can be approximated by Poisson distributions to yield

$$Y_{j0} \sim \text{Poisson}(r_{j0}), \tag{3}$$

where $r_{j0}=n_0p_{j0}$, and

$$Y_{j1} \sim \text{Poisson}(r_{j1}), \tag{4}$$

with $r_{j1}=n_1p_{j1}=fn_0p_{j1}$, and the factor $f=n_1/n_0$ reduces to 1 when $n_0=n_1$. For the sake of mathematical convenience, we will assume that

$$r_{j0} \sim \text{gamma}(\alpha,\lambda) \tag{5}$$

independently, which implies that marginally, the collapsed control frequencies Y_{j0}, $j=1,\ldots,J$, are independently distributed according to the negative binomial distribution,[15] with probability function

$$\begin{aligned}P(Y_{j0}=y) &= \frac{\Gamma(y+v^{-1})}{y!\,\Gamma(v^{-1})}\left(\frac{1}{1+v\mu}\right)^{v^{-1}}\left(\frac{v\mu}{1+v\mu}\right)^{y}\\ &= \frac{v^{y}\Gamma(y+v^{-1})}{y!\,\Gamma(v^{-1})}\left(\frac{1}{1+v\mu}\right)^{v^{-1}}\left(\frac{\mu}{1+v\mu}\right)^{y}\end{aligned} \tag{6}$$

where $\mu=\alpha/\lambda=E(Y_{j0})$ is the marginal mean of Y_{j0}, and $v=\alpha^{-1}$ is the dispersion parameter of the negative binomial distribution. Now let

$$\delta_j=\log(p_{j1})-\log(p_{j0})=\log\left(\frac{n_0p_{j1}}{r_{j0}}\right)$$

be the difference between the case and control probabilities for RV j on the log scale, so that $r_{j1}=n_1p_{j1}=fn_0p_{j1}=f\exp(\delta_j)r_{j0}$. Note that $\delta_j>0$ corresponds to a deleterious effect of the variant, and $\delta_j<0$ a protective effect, and $f=n_1/n_0$ is a factor to account for unequal sample sizes. As the δ_j are of substantive interest, we will treat them as fixed rather than random effects. According to (4), Y_{j1} is conditionally Poisson, and under the log link,

$$\log(r_{j1})=\log(fn_0p_{j1})=\log f+\delta_j+\log(r_{j0})$$

is a linear function of both the fixed effects δ_j and logarithm of the gamma-distributed random effects r_{j0}. This results in what is called a generalized linear mixed model. As the random effects r_{j0}, $j=1,\ldots,J$, are independent, it follows that marginally, the vectors (Y_{j0},Y_{j1}) are also independent. Generalizing (6),

the joint distribution of (Y_{j0},Y_{j1}) is given by

$$P(Y_{j0}=y_0, Y_{j1}=y_1) = \frac{v^{y_{j0}+y_{j1}}\Gamma(y_0+y_1+v^{-1})}{y_0!\,y_1!\,\Gamma(v^{-1})}\left\{\frac{1}{1+v\mu(1+fe^{\delta_j})}\right\}^{v^{-1}} \left\{\frac{\mu}{1+v\mu(1+fe^{\delta_j})}\right\}^{y_0}\left\{\frac{\mu fe^{\delta_j}}{1+v\mu(1+fe^{\delta_j})}\right\}^{y_1}. \quad (7)$$

Stepwise LRTs

Within the earlier described framework, the variant selection problem that we are interested in can be formulated as finding those δ_j that are not equal to 0. Our approach to variant selection is to conduct stepwise LRT in the following way. We begin with testing the complete null hypothesis

$$H_\phi : \delta_1 = \cdots = \delta_J = 0$$

against the alternative

$$H_{\{k\}} : \delta_k \neq 0;\ \delta_j = 0 \text{ for } j \neq k$$

one k at a time using the LRT, and we select the rare variant that maximizes the likelihood ratio statistic provided the value of this maximized statistic is greater than some critical value or cutoff c. We will postpone discussion of the choice of c to the next section and treat c as given for our present discussion. After we have included a variant, we will try to add one more variant by maximizing the likelihood ratio statistic of the current subset versus the current subset plus one more RV. The procedure stops when the maximized likelihood ratio statistic is $<c$. To make this operational, we need to maximize the marginal likelihood function under the generic null hypothesis

$$H_0 : \delta_j \neq 0 \text{ for } j \in S;\ \delta_j = 0 \text{ for } j \notin S,$$

where S denotes the subset of RVs with non-zero δ_j in the current model. The likelihood function under H_0 based on the observed frequencies y_{j0} and y_{j1} is given by the following product of terms like (7)

$$\begin{aligned}\text{lik}(H_0) &= \prod_{j=1}^{J} P(Y_{j0}=y_{j0}, Y_{j1}=y_{j1}) \\ &= \prod_{j\in S}\frac{v^{y_{j0}+y_{j1}}\Gamma(y_{j0}+y_{j1}+v^{-1})}{y_{j0}!\,y_{j1}!\,\Gamma(v^{-1})}\left\{\frac{1}{1+v\mu(1+fe^{\delta_j})}\right\}^{v^{-1}} \left\{\frac{\mu}{1+v\mu(1+fe^{\delta_j})}\right\}^{y_{j0}}\left\{\frac{\mu fe^{\delta_j}}{1+v\mu(1+fe^{\delta_j})}\right\}^{y_{j1}} \\ &\times \prod_{j\notin S}\frac{v^{y_{j0}+y_{j1}}\Gamma(y_{j0}+y_{j1}+v^{-1})}{y_{j0}!\,y_{j1}!\,\Gamma(v^{-1})}\left\{\frac{1}{1+v\mu(1+f)}\right\}^{v^{-1}} \left\{\frac{\mu}{1+v\mu(1+f)}\right\}^{y_{j0}}\left\{\frac{\mu f}{1+v\mu(1+f)}\right\}^{y_{j1}}.\end{aligned}$$

To obtain the maximum likelihood estimates $\hat{\mu}_0, \hat{\nu}_0$ and $\hat{\delta}_j (j\in S)$ of the parameters under H_0, we differentiate the log-likelihood with respect to μ, ν, and δ_j, $j\in S$, and set them to 0. The Newton–Raphson algorithm is used to solve these score equations.

The generic alternative hypothesis under our stepwise setup is

$$H_1 : \delta_j \neq 0 \text{ for } j \in S';\ \delta_j = 0 \text{ for } j \notin S',$$

where $S' = S\cup\{k\}$ for some $k\notin S$. The likelihood function $\text{lik}(H_1)$ under this alternative hypothesis has the same form as $\text{lik}(H_0)$ given above, but with S replaced by $S' = S\cup\{k\}$. The maximum likelihood estimates of the parameters μ, ν, and δ_j, $j\in S' = S\cup\{k\}$, can again be obtained using the Newton–Raphson algorithm. The LRT statistic of H_0 against H_1 is given by

$$W = 2\big(\log\{\hat{\text{lik}}(H_1)\} - \log\{\hat{\text{lik}}(H_0)\}\big),$$

where $\hat{\text{lik}}(H_1)$ and $\hat{\text{lik}}(H_0)$ are the maximum values of $\text{lik}(H_1)$ and $\text{lik}(H_0)$ respectively.

Choice of critical value

As the hypotheses being tested are nested, the LRT statistic $W_k = W_{\phi\text{vs}\{k\}}$ for testing the initial complete null hypothesis H_ϕ: $\delta_1 = \cdots = \delta_J = 0$ versus $H_{\{k\}}$: $\delta_k \neq 0$; $\delta_j = 0$ for $j\neq k$ is distributed asymptotically like χ^2_1 under H_ϕ, and so setting $c = 3.84$ will control the asymptotic type I error at level 0.05 one test at a time, but the familywise error rate is not controlled at 0.05 because a variant will be selected at stage 1 if the maximal statistic $M = \max_{1\leqslant k\leqslant J} W_k$ is $>c$. This implies the selection of at least one variant because more variants could have been selected at the subsequent stages of the stepwise procedure. Thus, the type I error is $P_{H_\phi}(M>c)$. The null distribution of the maximal statistic $M = \max_{1\leqslant k\leqslant J} W_k$ is, however, quite complicated because the statistics $W_1, W_2, \ldots, W_J$ are not independent even when the variants are. The use of Bonferroni inequality leads to an overly large critical value. We propose instead the following permutation approach to find the critical value adaptively. Represent the observed data by an J by (n_0+n_1) matrix $Y = \{y_{ji}\}$, where $y_{ji} = 1$ if the i^{th} individual has at least one copy of the jth RV, and $y_{ij} = 0$ otherwise. We can assume without loss of generality that $i = 1,\ldots,n_0$ correspond to the controls, and $i = n_0+1,\ldots,n_0+n_1$ correspond to the cases, so that $Y_{j0} = \sum_{i=1}^{n_0} y_{ji}$ and $Y_{j1} = \sum_{i=n_0+1}^{n_0+n_1} y_{ji}$ are the control and case frequencies, respectively. The permutation approach operates as follows.

Step 1: Generate a random permutation $r(1),\ldots,r(n_0+n_1)$ of $1,\ldots,n_0+n_1$.

Step 2: Compute $Y^*_{j0} = \sum_{i=1}^{n_0} y_{j,r(i)}$ and $Y^*_{j1} = \sum_{i=n_0+1}^{n_0+n_1} y_{j,r(i)}$ for $j = 1,\ldots,J$.

Step 3: For $k = 1,\ldots,J$ re-compute the likelihood ratio statistic for testing $S = \phi$ versus $S = \{k\}$ using the permuted frequencies $Y^*_{j0}, Y^*_{j1}, j = 1,\ldots,J$. Denote that by W^*_k, and let $M^* = \max_{1\leqslant k\leqslant J} W^*_k$.

Repeat the above procedure independently B times to obtain $M^*_1,\ldots,M^*_B$. For a nominal type I error of ε, we will choose c to be the upper 100ε empirical percentile of $M^*_1,\ldots,M^*_B$. To save computing time, we use $B = 100$ that seems to do a reasonable job in our simulation study. With the aim of filtering the list to a manageable but sufficiently rich set of RVs for further investigation and confirmatory study, we recommend a relatively liberal nominal type I error such as 0.1 or 0.2 (because of the inherent problem of insufficient sample size for rare mutations) so as to select more variants at the expense of a possible increase in false selection rate, which hovers $\sim$15–30% in our simulation studies to be reported later. An alternative to the permutation approach is the bootstrap.[16] The way the bootstrap differs from the permutation approach is that $r(1),\ldots,r(n_0+n_1)$ are now obtained by sampling with replacement n_0+n_1 times from $1,\ldots,n_0+n_1$. We do not expect sampling with or without replacement to make a big difference, and hence we expect the permutation and bootstrap approach to produce similar critical values. This turns out to be the case in our real data examples.

To have a fair comparison between the proposed stepwise LRT procedure with RARECOVER and SCORE, we ought to control the type I error of all three procedures at the same level. Bhatia *et al.*[12] recommended the use of $c = 0.5$. Based on the evidence of our simulation study (not shown here to save space), this choice of c is much too liberal and grossly over-selects the number of variants with false selection rates easily reaching 70% or more. Hoffmann, Marini and Witte[13] seem to suggest the use of $c = 0$ that also over-selects. The solution that we propose to overcome this problem is again to use the random permutation approach to determine the cutoffs for all three procedures.

RESULTS

Selection of rare variants associated with obesity

Against the background that blockade of the endocannabinoid receptor reduces obesity and improves metabolic abnormalities, the Comprehensive Rimonabant Evaluation Study of Cardiovascular ENDpoints and Outcomes (CRESCENDO) clinical trial (trial number NCT00263042 in ClinicalTrials.gov) was conducted to assess whether rimonabant, a cannabinoid-1 receptor blocker, would improve major vascular event-free survival. The subjects in this study are patients with abdominal obesity and with previously manifested or increased risk of vascular disease. More details about the study design and protocol can be found in http://clinicaltrials.gov/ct/show/NCT00263042 and Topol *et al.*[17] Our concerns here are not on cardiovascular outcomes, but on finding rare genetic variants that are associated with obesity. Out of 2958 Caucasian individuals aged 55 years or older in the CRESCENDO cohort, Harismendy *et al.*[18] selected individuals at the two extreme ends of the body mass

index for DNA sequencing. In the end, 143 individuals (73 men and 70 men) with body mass index $>40\,kg\,m^{-2}$ were selected as the cases (obese persons); and 146 individuals (74 men and 72 women) with body mass index $<30\,kg\,m^{-2}$ were selected as the controls. So the selected samples are balanced in gender. The DNA samples of the cases and controls were re-sequenced around two genes, *FAAH* and *MGLL*, known to be involved in endocannabinoid metabolism. Bhatia *et al.*[12] also made use of this data set to illustrate their RARECOVER procedure, and we extracted the data from their online supplementary materials. Apparently, some new cases and controls have been added because the data online consist of 148 cases and 150 controls that our analysis will be based on. For each rare variant in the *FAAH* and *MGLL* regions, Table 1 lists the collapsed frequencies (that is, the number of individuals with at least 1 occurrence of the rare variant) among the 148 cases and the 150 controls. The variants in bold type, 16 of them in the *FAAH* region and 12 in the *MGLL* region, were those selected by Bhatia *et al.*[12] using RARECOVER with critical value $c=0.5$. Among the 16 RVs selected by RARECOVER in the *FAAH* region, the case versus control frequencies are 1:0 for 12 of them.

Table 1 Case- versus control-collapsed frequency comparisons for 32 rare variants near the *FAAH* gene on chromosome 1 and 25 rare variants near the *MGLL* gene on chromosome 3

FAAH			*MGLL*		
Variant	*Position*	*No. of case: no. of control*	*Variant*	*Position*	*No. of case: no. of control*
1	**chr1:46626821**	**1:0**	1	chr3:129030872	18:13
2	**chr1:46626861**	**1:0**	**2**	**chr3:129031044**	**1:0**
3	**chr1:46627175**	**1:0**	**3**	**chr3:129031107**	**15:6**
4	**chr1:46627232**	**1:0**	4	chr3:129031199	0:1
5	**chr1:46627269**	**1:0**	**5**	**chr3:129031511**	**1:0**
6	chr1:46627603	0:2	**6**	**chr3:129031590**	**10:2**
7	chr1:46627621	11:6	7	chr3:129031591	9:3
8	chr1:46627662	1:0	8	chr3:129031787	0:2
9	**chr1:46628062**	**1:0**	**9**	**chr3:129031864**	**3:0**
10	chr1:46628247	3:3	10	chr3:129032558	0:1
11	chr1:46628507	1:4	**11**	**chr3:129032662**	**1:0**
12	**chr1:46628583**	**1:0**	12	chr3:129032671	1:0
13	chr1:46628584	1:0	13	chr3:129032842	1:0
14	**chr1:46628662**	**1:0**	14	chr3:129033307	1:1
15	chr1:46628901	0:1	**15**	**chr3:129033308**	**1:0**
16	chr1:46629068	0:1	16	chr3:129033356	0:1
17	**chr1:46629129**	**1:0**	17	chr3:129033939	0:1
18	chr1:46629215	0:1	18	chr3:129034047	0:1
19	chr1:46629280	0:1	**19**	**chr3:129034092**	**2:0**
20	**chr1:46629431**	**1:0**	**20**	**chr3:129034093**	**2:0**
21	chr1:46629606	9:6	**21**	**chr3:129034259**	**1:0**
22	**chr1:46629717**	**4:0**	**22**	**chr3:129034402**	**4:1**
23	**chr1:46630187**	**1:0**	**23**	**chr3:129034757**	**1:0**
24	chr1:46630534	14:13	**24**	**chr3:129034814**	**4:3**
25	chr1:46630612	3:5	25	chr3:129035532	1:0
26	**chr1:46630633**	**2:0**			
27	chr1:46630674	1:0			
28	chr1:46630716	18:10			
29	chr1:46630719	1:0			
30	**chr1:46631328**	**4:0**			
31	**chr1:46631519**	**1:0**			
32	**chr1:46631810**	**2:0**			

The variants in bold type are those selected by RARECOVER with threshold 0.5.

Likewise, 6 of the 12 RVs selected by RARECOVER in the *MGLL* region have frequency comparison of 1:0. It is hard to justify in our view this mass selection of RVs, each of which occurs only once in the entire sample of 148 cases. As commented earlier, it is better to try to control the familywise type I error by using the permutation approach to determine the critical value. We will do this for all three selection procedures: stepwise LRT, RARECOVER and SCORE.

Rare variants in the *MGLL* and *FAAH* regions

Even though it is commonly believed that RVs are at linkage equilibrium (that is, occur independently), we should check whether this assumption holds for the data at hand. One way to test whether two RVs are at linkage disequilibrium (LD) is to compute their correlation r from a sample of n individuals. As the data are bivariate binary rather than bivariate normal, we do not use the usual test statistic $(n-2)\times r^2/(1-r^2)$ to test the significance of r. Rather, we use the Pearson test of independence in the corresponding 2×2 table of frequencies, which for the present case of binary variables, is numerically equal to nr^2. As we are dealing with RVs, the expected frequencies will be low in some cells, and the χ^2 approximation may not be accurate. As a remedy, we use the option provided in the R function 'chisq.test' to compute P-values by $B=10\,000$ Monte Carlo simulations. We consider the *MGLL* region first, with 25 RVs in this region, there are ${}_{25}C_2=300$ pairs of RVs to be tested for LD. To correct for multiple testing, we use Bonferroni's correction and declare LD for a pair of RVs only when the P-value is $<0.05/300=0.000167$. Of all the 300 pairs of correlations, only the correlation between RV6 and RV7 is declared significant by Bonferroni's method ($n=298$, $r=0.913$, P-value $=0.0001$). It may not be appropriate to combine the cases with controls to test for LD between two RVs because there may be a systematic difference between the case and control frequencies, but the same conclusion is reached if we test LD using the control data ($n=148$) or the case data only ($n=150$). Looking more closely at RVs 6 and 7, out of the total of 298 individuals, the two variants occur simultaneously for 11 individuals, and separately for only two individuals (one for each variant), with neither variant occurring for the remaining 285 individuals. Furthermore, the position of the two variants differ only by one (chr3:129031590 versus chr3:129031591), and so they are in tight LD. Thus, we drop RV7 (with frequencies 9:3) and keep RV6 (with frequencies 10:2) in our analysis. RVs 19 and 20 also differ by one position only, and they both have collapsed frequency ratio of 2:0. However, their sample correlation is 0.497 only, with P-value 0.014 that is not significant when the familywise type I error is set at level 0.05. Thus, we keep both RVs 19 and 20 in our analysis. Ignoring significance for the time being, the order in which RVs are selected by stepwise LRT (Materials and Methods) can be found in the top panel of Table 2. The first three are RVs 6, 3 and 1 with collapsed frequency ratio of 10:2, 15:6 and 18:13, respectively. The associated stepwise (maximal) likelihood ratio statistics are 7.22, 7.45 and 5.07. The permutation-based critical values turn out to be 6.28 for nominal level 0.1 and 6.90 for level 0.05. Thus, whether we aim to control the type I error at 0.05 or 0.1, the proposed stepwise LRT procedure will select RVs 6 and 3, but not RV1. As commented before, the bootstrap critical values (6.01 and 6.90) are similar to the permutation critical values and the conclusion remains unchanged for this example. We will use the permutation approach to determine critical values in the remainder of this paper.

The middle and lower panels of Table 2 show the results for RARECOVER and SCORE using permutation-based critical values. It can be seen that whether at level 0.05 or 0.1, both procedures select

Table 2 Selection of RVs associated with obesity in the *MGLL* region using stepwise LRT, RARECOVER and SCORE procedures with critical values obtained using random permutations and case- versus control-collapsed frequencies given in parentheses

Stepwise LRT

Step	*RV added*	$\hat{\nu}$	$\hat{\mu}$	$\hat{\delta}_6$	$\hat{\delta}_3$	$\hat{\delta}_1$	$\hat{\delta}_8$	$2\times$ *increase in log-likelihood*
0	None	1.22	2.12					$c_{0.1}=6.28$, $c_{0.05}=6.90$
1	6 (10:2)	1.15	1.95	1.63				7.22
2	**3 (15:6)**	**0.98**	**1.72**	**1.68**	**1.24**			**7.45**
3	1 (18:13)	0.72	1.49	1.75	1.38	0.89		5.07
4	8 (0:2)	0.73	1.53	1.74	1.37	0.88	−11.7	3.27

RARECOVER

Step	*RV added*	*Increase in Pearson statistic*
		$c_{0.1}=5.02$, $c_{0.05}=5.67$
1	**6 (10:2)**	**5.67**
2	3 (15:6)	4.99
3	9 (3:0)	2.92

SCORE

Step	*RV added*	*Increase in score statistic*
		$c_{0.1}=5.08$, $c_{0.05}=5.62$
1	**6 (10:2)**	**5.62**
2	3 (15:6)	3.23
3	9 (3:0)	2.58

Abbreviations: LRT, likelihood ratio test; RV, rare variants.
The models in bold type are those selected at nominal level 0.05.

RV 6 only, but not RV3. It appears that by pooling information across variants, the proposed stepwise LRT procedure was able to identify one more potential causal variant.

In the *FAAH* region, only the correlation between RV4 and RV7 is found significant at level 0.05 after applying Bonferroni's correction, but we will not merge these two variants and keep them separate as they are quite far apart. It can be seen from Table 3 that at level 0.1, no variant is selected by all procedures. If one is willing to increase the level to $\sim$0.2 to induce more discoveries, then stepwise LRT will flag out RV28 (with frequency ratio 18:10) as a variant worthy of further investigation, but RARECOVER and SCORE will select no variant even at level 0.2.

Our findings so far are purely empirical. Bhatia *et al.*[12] and Harismendy *et al.*[18] have reported findings that corroborate with ours, and they also offer some scientific conjectures to explain how the selected RVs in the *MGLL* and *FAAH* regions could cause obesity.

Simulation results when the assumed model is correct

Mimicking the *MGLL* example, we generate data for $n_0=150$ controls, $n_1=148$ cases, and 24 RVs in the following way. First, we simulate $r_{j0}=n_0p_{j0}=150p_{j0}$ $(j=1,\ldots,24)$ according to the gamma $(\alpha,\ \lambda)$ distribution with $\nu=\alpha^{-1}=0.98$ and $\mu=\alpha/\lambda=1.72$ as in the model selected by stepwise LRT based on the original *MGLL* data (given in bold type in Table 2). We then divide r_{j0} by $n_0=150$ to get the p_{j0} and then generate the control data in the form of binary vectors $(Y_{ij0},\ j=1,\ldots,24)$, $i=1,\ldots,150$, with $Y_{ij0}\sim$Bernoulli(p_{j0}) independently. Summing over individuals, we obtain $Y_{j0}=\sum_i Y_{ij0}$. The case data $(Y_{ij1},\ j=1,\ldots,24)$, $i=1,\ldots,148$, are generated with $Y_{ij1}\sim$Bernoulli(p_{j1}) independently, where $p_{j1}=\exp(\delta_j)p_{j0}$. We consider two parameter settings for the δ_j. In setting (a), we have $\delta_1=\ldots=\delta_{24}=0$, which corresponds to the situation of no causal variant. For this setting, we focus on type I error. In setting (b), we set $\delta_3=1.24$, $\delta_6=1.68$, and $\delta_j=0$ for $j\neq3,6$ as in the model selected by stepwise LRT based on the original *MGLL* data. For this setting, our interest will focus on the procedure's ability to select RVs 3 and 6, and the false selection rate. To compare selection procedures on equal footing, we will use the random permutation approach to determine the critical values, with nominal type I error set at a liberal level of 0.2 (as our aim is to select a sufficiently rich set of potential causal RVs for further investigation).

The results based on 100 sets of simulations are present in the top panel of Table 4. It can be seen that the type I errors of the three procedures range from 0.12 to 0.15 when there is no casual variant and so are conservative. In setting (b), where RVs 3 and 6 are the casual variants with $\delta_3=1.24$ and $\delta_6=1.68$, RV6 is selected more often by stepwise LRT than by RARECOVER and SCORE (47 times as compared with 40 and 39). The number of times that RVs 3 and 6 are selected together is also highest for stepwise LRT (15 versus 11 and 10). Over the 100 samples simulated, RV3 is selected by stepwise LRT 26 times, RV6 is selected 47 times, whereas the other non-casual variants are selected a total of 15 times; thus, the false selection rate for stepwise LRT is $15/(26+47+15)=0.17$. The false selection rates of both RARECOVER and SCORE are 0.167. Although the set of figures given above seems to suggest only modest power for stepwise LRT to select the right variants, we expect the power to improve with either larger sample sizes and/or larger sizes of the variant effects (that is, larger values for δ_3 and δ_6). To illustrate this point, we conduct a simulation study with the same model parameters but double the

Table 3 Selection of RVs associated with obesity in the *FAAH* region using stepwise LRT, RARECOVER and SCORE procedures with critical values obtained using random permutations and case- versus control-collapsed frequencies given in parentheses

Stepwise LRT

Step	*RV added*	$\hat{\nu}$	$\hat{\mu}$	$\hat{\delta}_{28}$	$\hat{\delta}_{22}$	$\hat{\delta}_{30}$	$\hat{\delta}_{7}$	$2\times$ *increase in log-likelihood*
0	None	1.24	2.20					$c_{0.2}=5.24$, $c_{0.1}=7.22$
1	28 (18:10)	1.12	2.04	0.88				5.17
2	22 (4:0)	1.16	1.98	0.88	1.91			3.51
3	30 (4:0)	1.20	1.93	0.88	1.94	1.94		3.57
4	7 (11:6)	1.12	1.82	0.91	1.91	1.91	0.88	3.27

RARECOVER

Step	*RV added*	*Increase in Pearson statistic*
		$c_{0.2}=5.18$, $c_{0.1}=6.21$
1	22 (4:0)	4.11
2	30 (4:0)	3.16
3	32 (2:0)	2.14

SCORE

Step	*RV added*	*Increase in score statistic*
		$c_{0.2}=5.53$, $c_{0.1}=6.12$
1	22 (4:0)	4.06
2	30 (4:0)	2.49
3	26 (2:0)	2.03

Abbreviations: LRT, likelihood ratio test; RV, rare variants.

sample sizes (that is, 300 controls and 296 cases), and the results of stepwise LRT improve to selecting RV3 35 times, RV6 55 times, and together 21 times, with a false selection rate of 0.167 that is almost unchanged.

Simulation results when the assumed model is incorrect

To investigate how the performance of the three procedures is affected by violations of the model assumptions, we conduct two more sets of simulations. For the first set of extra simulations, we focus on non-gamma r_{j0}. A general class of mixtures of Poisson distributions to model over-dispersed count data has been proposed by Hougaard *et al.*[19] A prominent member of that class is the inverse Gaussian–Poisson distribution. The second panel of Table 4 summarizes the simulation results when the r_{j0} are simulated from an inverse Gaussian distribution with the same mean and variance as the gamma distribution used to generate the results of the top panel of Table 4. We can see that the powers of all three procedures are slightly reduced, but stepwise LRT is still marginally more powerful and has lower false selection rate (0.186 versus 0.215 and 0.261).

Our last set of simulations is designed to look into the effect of LD (that is, correlated variants) on our procedure and its competitors. Instead of simulating independent binary observations given the p_{j0}, we simulate correlated binary observations given the p_{j0}. As there is a dearth of distributions for correlated discrete data, we resort to the familiar technique of dichotomizing multivariate normal latent variables as is commonly done in multivariate probit models. To be specific, we first generate the r_{j0} from the same gamma distribution as before. Given the r_{j0}, and for each $i=1,\ldots,n_0$, rather than generating $Y_{ij0}\sim\text{Bernoulli}(p_{j0}=r_{j0}/n_0)$ independently for $j=1,\ldots,24$, we simulate a multivariate normal vector $(Z_{ij0},\ j=1,\ldots,24)\sim N_{24}(\eta,\Sigma)$, where $\Sigma=(\sigma_{ij})$ is a correlation matrix with $\sigma_{ii}=1$, and $\sigma_{ij}=\rho^{|c_i-c_j|}$, where c_i and c_j are the positions of RVs i and j listed in Table 1. This seems to be a sensible correlation structure where the correlation decays exponentially with inter-loci distance. We set ρ to 0.99, so that $\rho^{200}=0.134$, where 200 is roughly the average distance between successive RVs in the *MGLL* region; Table 1. The mean vector η of the multivariate normal distribution is chosen to make the marginal distribution of the dichotomized variable $Y^*_{ij0}=I\{Z_{ij0}>0\}$ the same as that of Y_{ij0}, namely, Bernoulli(p_{j0}). But unlike the Y_{ij0}, that are mutually independent, the $Y^*_{ij0}=I\{Z_{ij0}>0\}$ are correlated because the Z_{ij0} are. We will treat Y^*_{ij0} as the control data in this set of simulations. Similarly, to simulate the case data, we generate $(Z_{ij1},\ j=1,\ldots,24)\sim N_{24}(\eta,\Sigma)$, where the correlation matrix Σ is as defined above, and the mean vector η is chosen to make $Y^*_{ij1}=I\{Z_{ij0}>0\}\sim\text{Bernoulli}(p_{j1})$, where $p_{j1}=\exp(\delta_j)p_{j0}$. As before, we consider two settings: (a) all $\delta_j=0$, and (b) $\delta_3=1.24$, $\delta_6=1.68$, and $\delta_j=0$ for $j\neq3,6$. The results based on 100 simulations are given in the bottom panel of Table 4 under the headings 3(a) and 3(b). It can be seen that for setting 3(a), the type I error is inflated to 0.23 for RARECOVER, 0.25 for stepwise LRT, and 0.28 for SCORE. In setting 3(b), just as in settings 1(b) and 2(b), stepwise LRT selects the correct RVs more often than RARECOVER and SCORE, but the false selection rates increase to 0.283, 0.25 and 0.241, respectively.

Simulation results for the case of protective variants

To investigate the power of the proposed stepwise LRT procedure in picking out protective variants, we conduct extra simulations with parameter values set to $\nu=1$, $\mu=5$, $J=24$, $\delta_6=-2,-3$, $\delta_j=0$ $(j\neq6)$

Table 4 Operating characteristics of stepwise LRT, RARECOVER and SCORE procedures with nominal level 0.2 and critical values obtained using random permutations, based on 100 sets of data (148 cases, 150 controls and 24 variants) simulated under both null and non-null models, gamma and non-gamma random effects, and independent and correlated rare variants

	Stepwise LRT	*RARECOVER*	*SCORE*
Gamma-distributed p_{j0}			
All $\delta_j=0$			
Type I error	0.15	0.12	0.15
$\delta_3=1.24$, $\delta_6=1.68$, $\delta_j=0$ for $j\neq3,6$			
RV 3 selected	26	25	26
RV 6 selected	47	40	39
RVs 3, 6 selected	15	11	10
False selection rate	15/88 = 0.170	13/78 = 0.167	13/78 = 0.167
Inverse Gaussian-distributed p_{j0}			
All $\delta_j=0$			
Type I error	0.14	0.16	0.21
$\delta_3=1.24$, $\delta_6=1.68$, $\delta_j=0$ for $j\neq3,6$			
RV 3 selected	18	15	15
RV 6 selected	39	36	36
RVs 3, 6 selected	6	4	3
False selection rate	13/70 = 0.186	14/65 = 0.215	18/69 = 0.261
Gamma-distributed p_{j0}*, correlated variants*			
All $\delta_{j=0}$			
Type I error	0.25	0.23	0.28
$\delta_3=1.24$, $\delta_6=1.68$, $\delta_j=0$ for $j\neq3,6$			
RV 3 selected	27	17	17
RV 6 selected	49	43	43
RVs 3, 6 selected	10	5	4
False selection rate	30/106 = 0.283	20/80 = 0.250	19/79 = 0.241

Abbreviations: LRT, likelihood ratio test; RV, rare variants.

Table 5 Power of the stepwise LRT procedure with nominal level 0.1 and permutation-based threshold to detect a protective variant based on 100 sets of data simulated from the gamma random effects model with $v=1$, $J=24$ and all $\delta_j=0$ except δ_6

	$\delta_{6=}-2$	$\delta_6=-3$
$n_0=150$, $n_1=148$, $\mu=5$		
RV 6 selected	24	32
Other RVs selected	7	6
False selection rate	7/31 = 0.226	6/38 = 0.158
$n_0=300$, $n_1=296$, $\mu=10$		
RV 6 selected	48	57
Other RVs selected	12	12
False selection rate	12/60 = 0.2	12/69 = 0.174

Abbreviations: LRT, likelihood ratio test; RV, rare variants.

with $n_0=150$, $n_1=148$ as in the CRESCENDO study. This amounts to an average MAF of 5/150 = 0.033 in the control population. Strictly speaking, a variant with MAF 0.033 is not very rare, but we need to leave some room for the variant to be even rarer among the cases if it is in fact protective. The nominal level is 0.1 and the critical value is determined adaptively using the random permutation approach. The number of times RV6 is picked out in 100 simulations, and the number of times other variants are wrongly picked, together with the false selection rate are shown in the top panel of Table 5 under the heading (a). The power is moderate but this is to be expected as it is very difficult to detect the difference between the rare and the rarer. Increasing the sample size will obviously help. Part (b) of Table 5 depicts the results when the sample sizes are doubled to $n_0=300$, $n_1=296$, and μ increased to 10 correspondingly. It can be seen that the power is now increased to 0.48 and 0.57 for the cases $\delta_6=-2$ and $\delta_6=-3$ respectively, whereas the false selection rate is kept at 0.2 or less.

DISCUSSION

Due to the scarcity of information in rare variant analysis, it is important to pool information across variants. We show that one way to do this is to treat all the rare variant occurrence probabilities in the control sample as random effects from a common distribution, resulting in a mixed model. Even though the retrospective likelihood approach and its advantages have been advocated by Epstein and Satten[20] and Satten and Epstein,[21] genetic variant analyses are currently still mostly based on prospective likelihoods even when the study is retrospective, with justifications provided by Prentice and Pyke.[22] One reason for this tendency is that it is easier to model disease status given genotypes by some kind of binary regression model than to model the genotypes at multiple sites given disease status. Our mixed models are based on a retrospective formulation to better reflect the sampling given disease status nature of case–control studies. The justifiable and commonly made assumption of independence between rare genetic variants offers great simplification to the retrospective likelihood that we take full advantage of. We also prefer to treat the variant effects that are of substantive interest as fixed rather than random effects, which we think is the sensible thing to do. As we are dealing with RVs, we collapse frequencies as it is highly unlikely for an individual to have two mutant alleles at the same locus. Another advantage of modeling the collapsed frequencies is that it does not assume Hardy–Weinberg equilibrium, which is unlikely to be true for rare alleles. Efficiency calculation reported by Kuk, Xu and Li[23] within the context of haplotype frequency estimation when there is no random effect demonstrated that collapsing frequencies will not lead to much loss of estimation efficiency when the alleles are rare. We take rarity explicitly into account by making use of Poisson approximation in Equations (3) and (4). Out of convenience, we assume in (5) that the $r_{j0}=n_0p_{j0}$ are gamma distributed to result in the generalized negative binomial distribution (7) for (Y_{j0},Y_{j1}). Our simulation results show that the resulting stepwise LRT procedure is more powerful than RARECOVER and SCORE in selecting the correct variants. When applied to the *MGLL* data, stepwise LRT picks up one more variant, namely, RV3, than RARECOVER and SCORE. LRT is computationally more demanding as it involves finding the maximum likelihood estimates under both the null and alternative models. But as pointed out on p.9, the log-likelihood function has the same form under both models, and both can be fitted by the Newton–Raphson algorithm without too much difficulty. We outline in the appendix how to obtain the required first and second derivatives of the log-likelihood function. The proposed LRT approach is a parametric one and may not be robust against departures from the parametric assumptions. The fact that we obtain critical values by the random permutation approach rather than from the asymptotic distribution of LRT should make the procedure a bit more robust, and this is corroborated by the findings of our simulation study to a certain degree.

Non-gamma distributions could have been used for $r_{j0}=n_0p_{j0}$ in our mixed model. Our modest simulation study suggests that the proposed stepwise LRT procedure based on gamma random effects is

not too adversely affected by departure from the gamma assumption. Furthermore, the resulting negative binomial distribution passed the Pearson goodness of fit test when fitted to the control data in both the *MGLL and FAAH* regions.

Rather than integrating out the random effects r_{j0} to result in the generalized negative binomial distribution (7), an alternative is to eliminate the random effects r_{j0} as nuisance parameters by conditioning on the sum of the two independent Poisson counts Y_{j0} and Y_{j1} to obtain a binomial distribution for Y_{j1} conditionally. Although not mentioned explicitly, this is the theoretical basis for the C-alpha test of Neale *et al.*[9] But as r_{j0} has been eliminated from the conditional likelihood, there is no pooling of information across variants.

Our framework allows both deleterious ($\delta>0$) and protective ($\delta<0$) variants. It is also possible to incorporate covariate effects into our retrospective modeling framework.

ACKNOWLEDGEMENTS

We would like to thank the referees for their helpful comments and suggestions. The research of the third author was supported by the National Science Foundation of China Grant 11271346.

1 Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. *et al.* Potential etiologic and functional implications of genome-wide- association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
2 Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11,** 446–450 (2010).
3 Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **82,** 100–112 (2008).
4 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).
5 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5,** e1000384 (2009).
6 Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L. J. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86,** 832–838 (2010).
7 Kim, S. Y., Lim, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T. *et al.* Design of association studies with pooled or un-pooled next-generation sequencing data. *Nat. Biotechnol.* **34,** 479–491 (2010).
8 Liang, W. E., Thomas, D. C. & Conti, D. V. Analysis and optimal designs for association studies using next-generation sequencing with case-control pools. *Genet. Epidemiol.* **36,** 870–881 (2012).
9 Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7,** e1001322 (2011).
10 Lin, D. Y. & Tang, Z. Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89,** 354–367 (2011).
11 Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.* **89,** 82–93 (2011).
12 Bhatia, G., Bansal, V., Harismendy, O., Schork, N. J., Topol, E., Frazer, K. *et al.* A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.* **6,** e1000954 (2010).
13 Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive approach to analyzing rare genetic variants. *PLoS One* **5,** e13584 (2010).
14 Breslow, N. E. & Day, N. E. *Statistical Methods in Cancer Research. Volume I—The Analysis of Case-Control Studies* (IRAC Publications, 1980).
15 Saha, K. & Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61,** 179–185 (2005).
16 Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, New York, NY, USA, 1994).
17 Topol, E. J., Bousser, M. G., Fox, K. A., Creager, M. A., Despres, J. P., Easton, J. D. *et al.* CRESCENDO investigators. Rimonabant for prevention of cardiovascular events (CRESCENDO): a randomised, multicentre, placebo-controlled trial. *Lancet* **376,** 517–523 (2010).
18 Harismendy, O., Bansal, V., Bhatia, G., Nakano, M., Scott, M., Wang, X. *et al.* Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite. *Genome. Biol.* **11,** R118 (2010).
19 Hougaard, P., Lee, M. L. T. & Whitmore, G. A. Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. *Biometrics* **53,** 1225–1238 (1997).
20 Epstein, M. P. & Satten, G. A. Inference on haplotype effects in case-control studies using genotype data. *Am. J. Hum. Genet.* **75,** 35–43 (2003).
21 Satten, G. A. & Epstein, M. P. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet. Epidemiol.* **27,** 192–201 (2004).
22 Prentice, R. L. & Pyke, R. Logistic disease incidence model and case-control studies. *Biometrika* **66,** 403–411 (1979).
23 Kuk, A. Y. C., Li, X. & Xu, J. A fast collapsed data method for estimating haplotype frequencies from pooled genotype data with applications to the study of rare variants. *Stat. Med.* **32,** 1343–1360 (2013).

APPENDIX

Appendix First and second derivatives of the log-likelihood function

It can be assumed without loss of generality that $\delta_j=0$ for $J\geqslant j>k$, hence the model involves only the parameters ν, μ and $\delta_1,\ldots,\delta_k$. The log-likelihood function can be written as $l=l_1+l_2$, where

$$l_1=\sum_{j=1}^{k}\{y_{j0}\log\mu+y_{j1}(\delta_j+\log\mu)+(y_{j0}+y_{j1}+\nu^{-1})\times\log[1+\nu\mu(1+fe^{\delta_j})]\}+\sum_{j=k+1}^{J}\{(y_{j0}+y_{j1})\log\mu+(y_{j0}+y_{j1}+\nu^{-1})\log[1+\nu\mu(1+f)]\},$$

and

$$l_2=\sum_{j=1}^{J}\sum_{i=0}^{y_{j0}+y_{j1}-1}\log(1+i\nu).$$

The first and second derivatives of l_1 can be obtained readily by symbolic differentiation. Note in particular that $\partial^2 l_1/\partial\delta_i\partial\delta_j=0$ for $i\neq j$. The only thing left to do is to find the derivatives of l_2. As l_2 is a function of ν only, the only non-zero derivatives are

$$\frac{\partial l_2}{\partial\nu}=\sum_{j=1}^{J}\sum_{i=0}^{y_{j0}+y_{j1}-1}\frac{i}{1+i\nu}$$

and

$$\frac{\partial^2 l_2}{\partial\nu^2}=-\sum_{j=1}^{J}\sum_{i=0}^{y_{j0}+y_{j1}-1}\left(\frac{i}{1+i\nu}\right)^2.$$