Published in final edited form as:
Ann Hum Genet. 2013 March ; 77(2): 174–182. doi:10.1111/ahg.12005.

A likelihood ratio test for genomewide association under genetic heterogeneity*

Meng Qian[1] and **Yongzhao Shao**[1]
New York University School of Medicine, 650 First Avenue, Rm 538, New York, NY 10016, USA

Summary

Most existing association tests for genome-wide association studies (GWAS) fail to account for genetic heterogeneity. Zhou and Pan proposed a binomial mixture model based association test to account for the possible genetic heterogeneity in case-control studies. The idea is elegant, however, the proposed test requires an EM-type iterative algorithm to identify the penalized maximum likelihood estimates and a permutation method to assess p-values. The intensive computational burden induced by the EM-algorithm and the permutation becomes prohibitive for direct applications to genome-wide association studies. This paper develops a likelihood ratio test (LRT) for genome-wide association studies under genetic heterogeneity based on a more general alternative mixture model. In particular, a closed-form formula for the likelihood ratio test statistic is derived to avoid the EM-type iterative numerical evaluation. Moreover, an explicit asymptotic null distribution is also obtained which avoids using the permutation to obtain p-values. Thus, the proposed LRT is easy to implement for genome-wide association studies (GWAS). Furthermore, numerical studies demonstrate that the LRT has power advantages over the commonly used Armitage trend test and other existing association tests under genetic heterogeneity. A breast cancer GWAS data set is used to illustrate the newly proposed LRT.

Keywords

association test; binomial mixture model; complex disease; genetic heterogeneity; genomewide association study

Introduction

Common and complex diseases (or traits) are often genetically heterogeneous in etiologies (Lander & Schork, 1994; Zhou & Pan, 2009). Some well-known complex diseases with genetic heterogeneity include asthma, breast cancer (Hall et al., 1990; Wooster et al., 1994; Turnbull et al., 2010), and diabetes (Hattersley, 1998; Sladek et al. 2010). As in Zhou & Pan (2009), this paper considers the situation when a complex disease (or trait) is caused by mutations in multiple unlinked loci, commonly referred to as locus heterogeneity (Ott, 1999; Abreu et al., 2002; Fu et al., 2006). As a consequence of genetic heterogeneity, the population of individuals with disease may be decomposed into various latent sub-populations, each with disease caused by mutations at different loci (or their combinations). Most of the existing association tests for population-based case-control studies, e.g. GWAS, are based on comparing the mean genotype scores (e.g. the Armitage trend test) between the

*There is not any conflict of interest for both authors

Corresponding author: Yongzhao Shao, Ph.D., Departments of Population Health and Environmental Medicine, New York University School of Medicine, 650 First Avenue, Room 538, New York, NY 10016, USA, Phone: 1-212-263-0324, Fax: 1-212-263-8570, shaoy01@nyu.edu.

case and control groups, which are not efficient in the presence of genetic heterogeneity. Zhou & Pan (2009) showed that it can be beneficial to use methods that account for genetic heterogeneity for testing association in a case-control study.

Similar to admixture mapping in linkage analysis (Smith, 1963; Abreu et al., 2002; Fu et al., 2006), Zhou & Pan (2009) proposed a binomial mixture model to account for genetic heterogeneity and developed a modified likelihood ratio test (MLRT) for a single locus (Fu et al., 2006). They also consider two methods to combine single-locus-based MLRTs across multiple loci in linkage disequilibrium to boost power when causal SNPs are not genotyped (Zhou & Pan, 2009). They illustrated, with a wide spectrum of numerical examples, that the proposed MLRT tests are more powerful than some commonly used association tests under genetic heterogeneity. Following Zhou and Pan, we define the genetic score X as the number of the minor alleles at a single locus for a subject. Zhou and Pan (2009) assumed the genetic score X_H in a healthy control subject follows a binomial distribution, that is

$$P(X_H=g)=B_2(g,\theta_b), g=0,1,2, 0<\theta_b<1, \quad (1)$$

where $B_2(g,\theta_b)=\binom{2}{g}\theta_b^g(1-\theta_b)^{2-g}$ and where θ_b represents the minor allele frequency (MAF) on that specific locus of the control subject. On the other hand, under genetic heterogeneity, the genetic score for a diseased subject, X_D, follows a simple two-component mixture binomial distribution,

$$P(X_D=g)=\alpha_1 B_2(g,\theta_b)+\alpha_2 B_2(g,\theta), g=0,1,2; 0\le\alpha_1,\alpha_2\le 1, \alpha_1+\alpha_2=1, \quad (2)$$

where θ represents the probability of having the minor allele on one chromosome for a subgroup of cases with disease associated with the minor allele. They adopt a two-step procedure for parameter estimation. First, a maximum likelihood estimate (MLE) of θ_b is obtained based only on the control sample. Then, fixing the estimated θ_b at its MLE derived from the control-group data, maximum penalized likelihood estimates of other parameters in the mixture model are obtained using an EM-type algorithm (Li et al., 2009). Subsequently, the penalized MLEs from the EM-step are plugged into a likelihood ratio to form a test statistic to detect the association between the marker genotypes and the disease status. Finally, they proposed a permutation procedure to obtain the p-value of the association test.

Zhou and Pan's idea is applicable to an association study for a limited number of candidate markers, however, there are several challenges in applying their proposed method to genome-wide association studies (GWAS). First, the computation of their proposed MLRT for a vast number of SNPs in a typical GWAS would be very intensive. Since the penalized MLEs are obtained by an EM algorithm for maximization of the penalized mixture likelihood, there are known complexities and caveats associated with the EM or other iterative methods for identifying MLEs and penalized MLEs in mixture models including the challenges in selecting multiple starting points for parameter estimation. Moreover, the p-value of the MLRT is proposed to be attained by permutation, which is also difficult to apply directly to detect the SNP-disease association in GWAS with a vast number of SNPs, where the significance level is usually set to be less than 10^{-6}. In addition, it is widely believed that complex diseases and traits are caused by interplays of a large number of genetic loci and environmental risk factors. The simple binomial mixture model with two-components in equation (2) may be too simple to capture the complex heterogeneity for many complex diseases. A more general form of binomial mixture model can be written as follows

$$P_\eta(X_D=g)=\sum_{j=1}^{J}\alpha_j B_2(g,\theta_j), J \geq 2, \sum_{j=1}^{J}\alpha_j=1, \alpha_j \geq 0, \quad (3)$$

where $\eta = (\eta_j)_{j\leq J}$, $\eta_j = (\theta_j, \alpha_j)^T$, $j = 1, \ldots, J$, and $\theta_i = \theta_j$ if and only if $i = j$. In particular, for many of the complex diseases with genetic heterogeneity, it is likely that J is quite large. Since it is hard to know the number of the sub-populations J under genetic heterogeneity, it is desirable to have a new test that is applicable without the need to know the exact value of J while allowing $J \geq 2$.

In this paper, we developed a likelihood ratio test (LRT) for genome-wide association studies (GWAS) based on the more flexible binomial mixture models in (3). It is widely believed that complex diseases and traits are caused by interplays of a large number of genetic loci and environmental risk factors. Thus, we assume that the genetic score in the case group, X_D, follows a general binomial mixture distribution in (3) which allows the possibility of a large and unknown J. The proposed LRT overcomes the above mentioned challenges of using Zhou and Pan's method for testing association of a vast number of SNPs in a typical GWAS. In particular, we derived the closed-form formula for the likelihood ratio test statistic even though the maximum likelihood estimates (MLEs) of parameters in the binomial mixture model are non-regular with loss of identifiability (Liu & Shao, 2003). We further derived the simple closed-form asymptotic null distribution of the LRT which avoids the intensive numerical calculations, such as the EM based iterations for identification of MLEs and the permutations for evaluation of p-values. Additionally, the LRT can be implemented without the requirement of knowing the number of components J in the mixture model (3). We conducted extensive simulation studies to show that the LRT has power advantages over the Armitage trend test (ATT) and some other association tests under genetic heterogeneity. We applied our test to a real dataset from a breast cancer GWAS to illustrate that it can achieve a much smaller *p*-value than some commonly used tests when there is evidence of genetic heterogeneity. Thus, the proposed LRT might be used to scan SNPs in GWAS to make novel discoveries by taking account of genetic heterogeneity.

Method

Notation and set-up

We focus on detecting marker-disease association at a single locus with two alleles A and a, such as a SNP in a case-control genome-wide association study (GWAS). Suppose m_+ controls and n_+ cases are sampled from the population. For each SNP, the genotype frequencies in a case-control study can be summarized as in the following 2 × 3 table.

Let the genetic score X_H and X_D denote the number of minor alleles, say a, at a single locus for a healthy control and a diseased case, respectively. It is clear that $\Sigma X_H = 2m_2 + m_1$, $\Sigma X_D = 2n_2 + n_1$. Similar to Zhou and Pan's set-up, we assume that under the null hypothesis, both X_H and X_D have the same binomial distribution $B_2(g, \theta_b)$ as described in equation (1). As in Zhou & Pan (2009), X_H is assumed to have a binomial distribution under H_1. Under the alternative hypothesis of genetic heterogeneity, we assume X_D has a mixture distribution as described in equation (3). This last assumption is worthy of further comments. On one hand, it is possible to have $J > 2$ in equation (3) under H_1 both in practice and in theory, thus it is conceptually desirable to allow $J > 2$ in equation (3). On the other hand, for likelihood inference, it is not necessary to have $J > 2$ in the model in order to achieve the maximum of the likelihood because the model is actually saturated with $J = 2$. In other words, for a given dataset, posing a model (3) with $J = 2$ or with $J \geq 2$, the testing results from the LRT are not

going to be different. In fact, as will be seen in next section, our proposed likelihood ratio test actually has the "non-parametric" nature because it has a closed-form formula, with a simple null distribution shown to be valid, thus it will be valid for testing any alternative models including the common models and those under heterogeneity. In this paper we will establish that the test is actually a likelihood ratio test under the specified set-up motivated by the elegant work of Zhou & Pan (2009) and by the fact that the likelihood ratio test has well-known optimalities in terms of statistical power and efficiency.

Mixture binomial and maximum likelihood

Assuming the set-up in the previous subsection, under H_0, using the notation in Table 1 and denoting the true value of θ_b as P_0, the maximum likelihood estimate (MLE) of P_0 for the overall combined case-control data in Table 1 is

$$\hat{p}_0=[\Sigma X_D+\Sigma X_H]/(2n_++2m_+)=[n_2+m_2+(n_1+m_1)/2]/(n_++m_+]). \quad (4)$$

Thus, the binomial likelihood function for the overall combined case-control data evaluated at $\hat{p_0}$, L_0, is given by

$$L_0=\prod_{g=0}^{2} B_2(g,\hat{p}_0)^{m_g+n_g}, \quad (5)$$

where $B_2(g,\hat{p}_0)=\binom{2}{g}\hat{p}_0^g(1-\hat{p}_0)^{2-g}$ and $\hat{p_0}$ is defined in (4). Following Zhou & Pan (2009), in the control group, the genetic score X_H is assumed to follow a binomial distribution under the alternative hypothesis, say

$$P(X_H=g)=B_2(g,P_H), g=0,1,2. \quad (6)$$

Using the notation in Table 1, the maximum likelihood estimate of P_H within the healthy control group only is given by

$$\hat{p}_H=\Sigma X_H/(2m_+)=(m_2+m_1/2)/m_+. \quad (7)$$

The binomial likelihood function of the healthy controls data evaluated at $\hat{p_H}$, L_H, is given by

$$L_H=\prod_{g=0}^{2} B_2(g,\hat{p}_H)^{m_g} \quad (8)$$

Similarly, in the case group, if the genetic score X_D has the distribution $B_2(g; P_D)$, the maximum likelihood estimate $\hat{p_D}$ of P_D within the diseased case group only would be

$$\hat{p}_D=\Sigma X_D/(2n_+)=(n_2+n_1/2)/n_+. \quad (9)$$

However, as in Zhou & Pan (2009), we assume that under genetic heterogeneity, the cases can be divided into multiple latent sub-populations. Thus, under the alternative hypothesis of genetic heterogeneity, we assume X_D has a mixture distribution as described in equation (3). It can be shown that (see Appendix 1), using the above notation, the maximum of the mixture likelihood for X_D has an explicit formula:

$$L_D = \sup_{\eta} \prod_{g=0}^{2} P_\eta(X_D = g)^{n_g} = \begin{cases} \prod_{g=0}^{2} (n_g/n_+)^{n_g} & \text{if} 4n_0 n_2 > n_1^2; \\ \prod_{g=0}^{2} B_2(g; \hat{p}_D)^{n_g} & \text{if} 4n_0 n_2 \le n_1^2. \end{cases} \quad (10)$$

The derivation of the above equation can be found in Appendix 1. It is also clear from the derivation in Appendix 1 that the mixture likelihood function of the parameter vector $\eta = (\theta_j, \alpha_j)_{j \le J}$ in the mixture model (3) can have many local maxima due to the lack of identifiability in parameters (Liu & Shao, 2003). Nevertheless, the supremum of the mixture likelihood L_D for X_D has a single unique value for each dataset and can be obtained from the explicit formula in equation (10). In the typical case-control study design, L_D is independent of L_H.

The likelihood ratio test

Using the maximum of the likelihood L_0, L_H and L_D in equations (5), (8) and (10), respectively, we can write down the explicit formula of the log likelihood ratio test statistic as follows

$$2\lambda_N = 2[\log(L_D L_H) - \log L_0]. \quad (11)$$

No iterative numerical maximization of the mixture likelihood function is needed for the evaluation of the LRT statistic in (11). Thus the LRT statistic is easy to compute even for GWAS. It is known that the LRT statistics for testing homogeneity in mixture models often have complicated asymptotic distributions that typically lack closed-form representations. However, the above statistic $2\lambda_N$ can be shown to have an explicit form of asymptotic distribution under the null hypothesis. More specifically, under H_0, as $n_+ \to \infty$ and $m_+ \to \infty$, we have

$$2\lambda_N \to \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2, \quad (12)$$

where χ_d^2 denotes a chi squared distribution with d degrees of freedom, d=1, 2. Although the above asymptotic null distribution can be derived from general results such as those in Chernoff & Lander (1995), Chiano & Yates (1995), or Liu & Shao (2003), an elementary and detailed direct derivation of the above asymptotic null distribution is given in Appendix 2 for readers who are interested in a direct derivation based on first principles.

It is worth pointing out that our extensive numerical simulations discussed in the next section indicate that the simple asymptotic null distribution in (12) approximates the exact finite sample null distribution very well. The asymptotic formula is only slightly conservative. Therefore, the p-values of the likelihood ratio test can be easily read off from the above simple closed-form asymptotic null distribution. For example, given any observed data in Table 1, one can first evaluate the value of $2\lambda_N$ in (11), then can obtain the p-value using the following simple command in the widely used *R*-platform:

```
{pchisq(2λN; 1; lower:tail = F) + pchisq(2λN; 2; lower:tail =
F)}/2.
```

Last but not least, it is well known that the likelihood ratio test generally has better power than other ad hoc tests. Thus it should not be a surprise to see that the LRT can be more powerful than other commonly used tests which ignore the genetic heterogeneity that exists for many common complex diseases such as breast cancer. Finally, to implement the LRT, there is no need to identify the exact number of mixture components J in (3), which is desirable because J is hard to determine in practice.

Numerical Results

Type I Errors

The LRT has an explicit asymptotic distribution under H_0. Consequently, it is convenient to evaluate the p-value and type I errors. We conducted comprehensive simulations to compare the empirical type I error of the LRT to the nominal significance level ranging from 10^{-2} to 10^{-8}. In the Monte Carlo simulations, the genotype data for both the control group and the case group were generated from the same binomial distribution $B_2(g; \theta_b)$, where θ_b takes some fixed value P_0, which represents the minor allele frequency (MAF). A number of simulation set-ups, which varied over a range of minor allele frequency and sample size were selected. The control and case sample sizes are set to be equal. The nominal significance levels were taken to be 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7} and 10^{-8}, respectively. For each set-up, 10^{11} samples were generated. We found that the empirical type I error is slightly smaller than the nominal level, but they are extremely close to each other. Thus using the asymptotic null distribution for the LRT is valid. For illustration, an example with $\theta_b = 0.4$ and sample size $n_+ = m_+ = 1000$ is shown in Table 2.

Power Comparison

The significance level of the association test is usually set very small for genome-wide association studies (GWAS). For example, the genome-wide significance level of 5×10^{-8} is being increasingly used for arrays that contain one million SNPs. The most commonly used association tests for GWAS include Armitage's trend test (ATT) and the χ^2_2 test, both applicable for testing association in a 2 × 3 table between the case-control status and the three genotypes, as illustrated in Table 1. Accordingly, we designed simulation studies to evaluate and compare the powers of the LRT, the ATT and the χ^2_2 test when the significance level is set to be 5×10^{-8}. Note that, the MLRT of Zhou & Pan (2009) was not included for comparison due to its severe computational challenge when the significance level is very small. In the first set of Monte Carlo simulations, the control sample was generated from a binomial distribution $B_2(g, \theta_b)$; the case sample was generated from a two-component mixture binomial distribution as described in (3) with $J = 2$:

$$P_\eta(X_D=k)=\sum_{j=1}^{2}\alpha_j B_2(k,\theta_j), 0 \le \alpha_1, \alpha_2 \le 1, \alpha_1+\alpha_2=1.$$

20000 replicate data sets of $n_+ = m_+ = N$ controls and cases were simulated for each of the eight simulations set-up and the empirical power for each test are shown in Table 3. The simulation results indicate that the LRT has power advantage over the Armitage trend test (ATT) and the χ^2_2 test under genetic heterogeneity.

Similar power advantages of the LRT over other tests are also observed when the alternative mixture model has three components ($J = 3$) as demonstrated in Table 4 where θ_3 for the cases is set as equal to θ_b for the control group.

Note that the Armitage trend test (ATT), also called Cochran-Armitage trend test (CATT) by many researchers, has good power only when the disease risk of the genotypes *AA*, *Aa*, *aa* is monotone increasing or decreasing under the alternative hypothesis (Armitage, 1955; Freidlin et al., 2002). Thus, ATT can have very low power when there is a violation of a linear trend in the disease risk across the ordered genotypes *AA*, *Aa* and *aa*, as in the case of both set-ups #3 and #5 in Table 3.

It is clear from the power simulations across the multiple simulation set-ups that the LRT can be much more powerful than the commonly used Armitage trend test (ATT) and the χ^2_2 test in GWAS in the presence of genetic heterogeneity.

A Breast Cancer GWAS

Breast cancer is the most common cancer among women. Many genes on different chromosomes that underlie breast cancer have been identified including many well-known studies conducted two decades ago (Hall et al., 1990; Wooster et al., 1994). Many more genetic variants underlying breast cancer are still being discovered nowadays, thus there is little doubt about the existence of genetic heterogeneity in the case of breast cancer. For illustration, we applied the newly proposed likelihood ratio test to a breast cancer GWAS dataset. In particular, Turnbull et al. (2010) conducted a genome-wide association study to identify breast cancer susceptibility alleles. They studied 582886 SNPs in 3659 breast cancer cases and 4897 controls in the first stage, and evaluated promising SNPs that were identified in Stage 1 in a second stage with 12576 cases and 12223 controls. In the paper they reported five new susceptibility SNPs with summary genotype data of the five SNPs made publicly available. A literature search indicates that four of the five SNPs (rs1011970, rs10995190, rs704010 and rs614367) have been independently confirmed by other studies since the publication of their GWAS results in 2010 (Lambrechts et al., 2012; Peng et al., 2011). We evaluated the p-values of the LRT, Armitage trend test (ATT) and χ^2_2 test for these four SNPs for comparison. The results are summarized in Table 5.

Note that for the SNP rs10995190 and SNP rs614367, the p-values are smaller than the genome-wide significance level 5×10^{-8} for the newly proposed LRT and the ATT, and for each of the two stages. The performance of the LRT is as good as or better than the other two tests. In particular, the LRT has an extremely small p-value 6×10^{-15} for Stage 2 data of SNP rs614367 showing statistical significance at even lower levels. It is thus not surprising that these SNPs are independently replicated by other GWAS. For the SNP rs704010 and SNP 1011970, a simple combined p-value (for combining the two stages), e.g. Fisher's meta p-value, indicates both SNPs are significant even using the genome-wide significance level 5×10^{-8} for all three tests. The newly proposed LRT is also very competitive for these two SNPs. For example, for the Stage 1 data of SNP rs704010, only the p-value of the LRT is smaller than the genome-wide significance level 5×10^{-8}. As an indication of overall strength of the test, Fisher's meta p-value of the LRT from the combined Stages 1 and 2 is smaller than those of the other two tests, and the LRT is clearly the most competitive test among the three competitors. This example indicates the potential value of the proposed LRT for GWAS data to detect association of complex diseases where the presence of genetic heterogeneity is always a possibility.

Discussion

In the analysis of GWAS data, potential latent genetic heterogeneity has been largely ignored by researchers. Zhou & Pan (2009) first proposed mixture models to account for genetic heterogeneity. However, for the analysis of a vast number of SNPs in GWAS, the MLRT of Zhou and Pan has major computational challenges. In this paper, using a more general binomial mixture model, we have derived a likelihood ratio test for case-control association studies that improves the MLRT by Zhou and Pan on computational efficiency and multiple other aspects. In particular, the likelihood ratio test statistic has a simple closed-form formula, which could avoid intensive computation, such as the EM algorithm for penalized maximum likelihood estimates. Additionally, we have derived an explicit asymptotic null distribution for the proposed LRT, which is convenient to obtain p-values even at a small significance level. Moreover, to perform the LRT, there is no need to decide

the exact number of mixture components, which is convenient in practice. Therefore, the new LRT has computational advantages over the MLRT proposed by Zhou and Pan and is suitable for scanning SNPs in GWAS data.

As demonstrated by our numerical studies, in the presence of genetic heterogeneity, the LRT can be much more powerful than either Armitage's trend test or the χ^2_2 test, both of which are among the most widely used tests in GWAS. Given that most complex diseases are widely believed to be polygenic and have environmental components, genetic heterogeneity is a hallmark of complex diseases. As illustrated using the GWAS data for breast cancer, newly proposed LRT can be easily used for any GWAS data, thus researchers can use the simple algorithm to scan their SNPs as a cost-effective way to potentially make novel and important discoveries using existing data already collected in the large number of GWAS. Given that there are already about 1000 published GWAS, and many more genome-wide studies are being planned and conducted, the new LRT has the potential to become one of the useful tests to scan the SNPs in these GWAS, maybe as a secondary analysis to account for genetic heterogeneity. Thus the new user-friendly LRT can potentially be used to increase the impact of existing and future genome-wide association studies.

Acknowledgments

This research is partially supported by the NIH Cancer Center Supporting Grant to NYU (2P30 CA16087), and the NIEHS Center Grant to NYU (5P30 ES00260), as well as a Stony Wold-Herbert Foundation grant to YS. The authors would like to thank the reviewers for insightful suggestions that lead to improvement of the paper.

References

Armitage P. Tests for linear trends in proportions and frequencies. Biometrics. 1955; 11:375386.

Abreu PC, Hodge SE, Greenberg DA. Quantification of type I error probabilities for heterogeneity lod scores. Genet Epidemiol. 2002; 22:156–169. [PubMed: 11788961]

Bickel, PJ.; Doksum, KA. Mathematical Statistics: Basic Ideas and Selected Topics. Vol. Vol I. New Jersey: Prentice Hall; 2000.

Chernoff H, Lander E. Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. J Statist Plann Inference. 1995; 43:19–40.

Chiano MN, Yates JRW. Linkage detection under heterogeneity and the mixture problem. Ann Hum Genet. 1995; 59:83–95. [PubMed: 7762986]

Emigh TH. A comparison of tests for Hardy-Weinberg equilibrium. Biometrics. 1980; 36:627–642.

Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002; 53:146–152. [PubMed: 12145550]

Fu Y, Chen J, Kalbfleisch JD. Testing for Homogeneity in Genetic Linkage Analysis. Stat Sinica. 2006; 16:805–823.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990; 250:1684–1689. [PubMed: 2270482]

Hattersley AT. Maturity-onset diabetes of the young: clinical heterogeneity explained by genetic heterogeneity. Diabet Med. 1998; 15(1):15–24. [PubMed: 9472859]

Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994; 265:2037–2048. [PubMed: 8091226]

Lambrechts D, Truong T, Justenhoven C, Humphreys MK, Wang J, Hopper JL, Dite GS, Apicella C, Southey MC, Schmidt MK, et al. 11q13 is a susceptibility locus for hormone receptor positive breast cancer. Hum Mutat. 2012; 33(7):1123–1132. [PubMed: 22461340]

Li P, Chen JH, Marriott P. Non-finite Fisher information and homogeneity: an EM approach. Biometrika. 2009; 96:411–426.

Liu X, Shao Y. Asymptotics for likelihood ratio tests under loss of indentifiability. Ann Stat. 2003; 31:807–832.

Ott, J. Analysis of Human Genetic Linkage. Third Edtition. Baltimore: The John Hopkins University Press; 1999.

Peng S, L B, Ruan W, Zhu Y, Sheng H, Lai M. Genetic polymorphisms and breast cancer risk: evidence from meta-analyses, pooled analyses, and genome-wide association studies. Breast Cancer Res Treat. 2011; 127(2):309–324. [PubMed: 21445572]

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2010; 445:881–885. [PubMed: 17293876]

Smith CA. Testing for heterogeneity of recombination fraction values in Human Genetics. Ann Hum Genet. 1963; 27:175–182. [PubMed: 14081488]

Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, Hughes D, Warren-Perry M, Tapper W, Eccles D, Evans DG, Hooning M, Schutte M, van den Ouweland A, Houlston R, Ross G, Langford C, Pharoah PD, Stratton MR, Dunning AM, Rahman N, Easton DF. Breast Cancer Susceptibility Collaboration (UK). Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet. 2010; 42:504–507. [PubMed: 20453838]

Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, Fields P, Marshall G, Narod S, Lenoir GM, Lynch H, Feunteun J, Devilee P, Cornelisse CJ, Menko FH, Daly PA, Ormiston W, McManus R, Pye C, Lewis CM, Cannon-Albright LA, Peto J, Ponder BAJ, Skolnick MH, Easton DF, Goldgar DE, Stratton MR. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science. 1994; 265:2088–2090. [PubMed: 8091231]

Zhou H, Pan W. Binomial Mixture Model-based Association Tests under Genetic Heterogeneity. Ann Hum Genet. 2009; 73:614–630. [PubMed: 19725835]

Appendix 1

Derivation of the test statistic of the LRT

To prove our proposed association test is indeed a LRT under the given set-up, we just need to establish equation (10), that is, when X_D follows the mixture distribution in (3),

$$L_D=\sup_{\eta}\prod_{g=0}^{2}P_\eta(X_D=g)^{n_g}=\begin{cases}\prod_{g=0}^{2}(n_g/n_+)^{n_g} & \text{if} 4n_0n_2>n_1^2;\\ \prod_{g=0}^{2}B_2(g;\hat{\mathrm{p}}_D)^{n_g} & \text{if} 4n_0n_2\le n_1^2;\end{cases}$$

where $\hat{p_D}$ is defined as in equation (9). First we we want to show that when $4n_0n_2>n_1^2$, the maximum likelihood estimates $\hat{\eta}$ of the η in (3) satisfy

$$L_D=\prod_{g=0}^{2}P_{\hat{\eta}}(X_D=g)^{n_g}=\prod_{g=0}^{2}(n_g/n_+)^{n_g}. \quad (13)$$

A simple application of Jensen's inequality yields that, for any η,

$$\log\prod_{g=0}^{2}P_\eta(X_D=g)^{n_g}\le\log\prod_{g=0}^{2}(n_g/n_+)^{n_g}. \quad (14)$$

The right-hand side of the above inequality is an upper bound which may not be achievable in general. However, when $4n_0n_2>n_1^2$, we can show that the equality in (13) is achievable. In fact, when $4n_0n_2>n_1^2$, there are infinitely many values of the MLE $\hat{\eta}$ can make (13) an

equality. It is straightforward and elementary to verify that one set of solutions for MLE is given as follows:

$$\begin{aligned}
&\hat{\theta}_1 \in (\frac{2n_2}{2n_2+n_1}, 1),\\
&\hat{\theta}_2=\frac{(n_1+2n_2)\hat{\theta}_1 - 2n_2}{2n_+\hat{\theta}_1 - n_1 - 2n_2},\\
&\hat{\alpha}_1=\frac{4n_0n_2 - n_1^2}{2n_+(2n_+\hat{\theta}_1^2 - 2(n_1+2n_2)\hat{\theta}_1+2n_2)},\\
&\hat{\alpha}_2=1 - \hat{\alpha}_1,\\
&\hat{\alpha}_i=0, \hat{\theta}_i=1/2, \text{for} i \neq 1, 2.
\end{aligned}$$

As indicated above, $\hat{\theta_1}$ can take any values in an interval, thus there are infinitely many sets of solutions for the MLE. Thus equation (13) is proved. Next we show that when $4n_0n_2 \leq n_1^2$,

$$L_D=\sup_{\eta}\prod_{g=0}^{2}P_\eta(X_D=g)^{n_g}=\prod_{g=0}^{2}B_2(g;\hat{\mathrm{p}}_D)^{n_g}. \quad (15)$$

First, we show that, for any fixed η,

$$\prod_{g=0}^{2}P_\eta(X_D=g)^{n_g} \leq \prod_{g=0}^{2}B_2(g;\hat{\mathrm{p}}_D)^{n_g}.$$

Using the inequality $\log x \leq x - 1$, we get

$$\sum_{g=0}^{2}n_g\log\frac{P_\eta(X_D=g)}{B_2(g,\hat{\mathrm{p}}_D)} \leq \sum_{g=0}^{2}n_g\left(\frac{P_\eta(X_D=g)}{B_2(g,\hat{\mathrm{p}}_D)} - 1\right)$$

It is straightforward to verify that

$$\sum_{g=0}^{2}n_g\left(\frac{P_\eta(X_D=g)}{B_2(g,\hat{\mathrm{p}}_D)} - 1\right)=\frac{n_+(4n_0n_2 - n_1^2)}{(2n_0+n_1)^2(2n_2+n_1)^2}\sum_{j=0}^{J}\hat{\alpha}_j(2n_2+n_1 - 2n_+\hat{\theta}_j)^2$$

Therefore, when $4n_0n_2 \leq n_1^2$, and for any η

$$\log=\prod_{g=0}^{2}[P_\eta(X_D=g)]^{n_g} \leq \log\prod_{g=0}^{2}B_2(g;\hat{\mathrm{p}}_D)^{n_g}.$$

Finally, it is obvious that

$$L_D=\sup_{\eta}\prod_{g=0}^{2}P_\eta(X_D=g)^{n_g} \geq \prod_{g=0}^{2}B_2(g;\hat{\mathrm{p}}_D)^{n_g}.$$

This finishes the proof (15), thus also (10).

Appendix 2

The asymptotic null distribution of the LRT

Under H_0, both X_D and X_H have the same binomial distribution $B_2(g; \theta_b)$. We denote the true null value for θ_b as P_0. Without loss of generality, we assume $0 < P_0 < 1$ to avoid $P_0(1 - P_0) = 0$ appearing in any denominator. First, we may consider testing $H_0 : B_2(g; P_0)$ against $H_0 : B_2(g; \theta_b)$, $\theta_b \in (0, 1)$, using only the healthy controls. This is a classic problem, the likelihood ratio test statistic is well known to have a χ_1^2 distribution. It is well known that, under $H_0 : B_2(g; P_0)$, the LRT statistic can be written as

$$2\sum_{g=0}^{2} m_g \log\frac{B_2(g;\hat{p}_H)}{B_2(g;P_0)} = 2m_+\frac{(\hat{p}_H - P_0)^2}{P_0(1-P_0)} + o_p(1), \quad (16)$$

where $\hat{p_H} = (m_2 + m_1/2)/m_+$ is the MLE of θ_b using only the healthy controls. Similarly, we may consider testing $H_0 : B_2(g; P_0)$ against $H_0 : B_2(g; \theta_b)$, $\theta_b \in (0, 1)$, using only the diseased cases. Then, under $H_0 : B_2(g; P_0)$, the LRT statistic has χ_1^2 distribution and can be written as

$$2\sum_{g=0}^{2} n_g \log\frac{B_2(g;\hat{p}_D)}{B_2(g;P_0)} = 2n_+\frac{(\hat{p}_D - P_0)^2}{P_0(1-P_0)} + o_p(1), \quad (17)$$

where $\hat{p_D} = (n_2 + n_1/2)/n_+$ is the MLE of θ_b using only the diseased cases. Similarly, we may consider testing $H_0 : B_2(g; P_0)$ against $H_0 : B_2(g; \theta_b)$, $\theta_b \in (0, 1)$, using the overall sample combining both the diseased cases and health controls. Then the MLE for $\theta_b = P_0$ from the combined sample is $\hat{p_0}$ as defined in (4). The LRT statistic can be written as

$$2\sum_{g=0}^{2}\left[m_g \log\frac{B_2(g;\hat{p}_0)}{B_2(g;P_0)} + n_g \log\frac{B_2(g;\hat{p}_0)}{B_2(g;P_0)}\right] = 2(m_+ + n_+)\frac{(\hat{p}_0 - P_0)^2}{P_0(1-P_0)} + o_p(1). \quad (18)$$

From the above three equations, and the equations (5), (8), (10), we have, when $4n_0n_2 < n_1^2$,

$$2\log\frac{L_D L_H}{L_0} = 2m_+\frac{(\hat{p}_H - P_0)^2}{P_0(1-P_0)} + 2n_+\frac{(\hat{p}_D - P_0)^2}{P_0(1-P_0)} - 2(m_+ + n_+)\frac{(\hat{p}_0 - P_0)^2}{P_0(1-P_0)} + o_p(1). \quad (19)$$

Denote $\rho = n_+/(m_+ + n_+)$. Then it is straightforward to verify that

$$\hat{p}_0 - P_0 = \rho(\hat{p}_D - P_0) + (1-\rho)(\hat{p}_H - P_0)$$

and

$$2\log\frac{L_D L_H}{L_0} = (1-\rho)\frac{2n_+(\hat{p}_D - P_0)^2}{P_0(1-P_0)} + \rho\frac{2m_+(\hat{p}_H - P_0)^2}{P_0(1-P_0)} - 2\sqrt{\rho(1-\rho)}\frac{\sqrt{2n_+}(\hat{p}_D - P_0)}{\sqrt{P_0(1-P_0)}}\frac{\sqrt{2m_+}(\hat{p}_H - P_0)}{\sqrt{P_0(1-P_0)}} + o_p(1).$$

Denote

$$Z_H = \frac{\sqrt{2m_+}(\hat{p}_H - P_0)}{\sqrt{P_0(1-P_0)}}$$

and

$$Z_D = \frac{\sqrt{2n_+}(\hat{p}_D - P_0)}{\sqrt{P_0(1-P_0)}}.$$

Then

$$2\log\frac{L_D L_H}{L_0} = (\sqrt{\rho}Z_H + \sqrt{1-\rho}Z_D)^2 + o_\rho(1). \quad (20)$$

Note that $Z_H \sim N(0, 1)$ and $Z_D \sim N(0, 1)$ and Z_H and Z_D are independent. Thus

$$\sqrt{\rho}Z_H + \sqrt{1-\rho}Z_D \sim N(0,1).$$

Therefore, when $4n_0n_2 \le n_1^2$, we have $2\lambda_N = 2(\log L_D + \log L_H - \log L_0) \sim \chi_1^2$.

On the other hand, under H_0, when $4n_0n_2 > n_1^2$, we can first consider testing goodness-of-fit of $H_0 : B_2(g; P_0)$ using only the diseased cases. The likelihood ratio test statistic has a χ_2^2 asymptotic distribution and can be written as

$$2\sum_{g=0}^{2} n_g \log\frac{n_g/n_+}{B_2(g;P_0)} = 2\sum_{g=0}^{2} n_g \log\frac{n_g/n_+}{B_2(g;\hat{p}_D)} + 2\sum_{g=0}^{2} n_g \log\frac{B_2(g;\hat{p}_D)}{B_2(g;P_0)} = 2n_+\frac{(n_g/n_+ - \hat{p}_D)^2}{\hat{p}_D(1-\hat{p}_D)} + 2n_+\frac{(\hat{p}_D - P_0)^2}{P_0(1-P_0)} + o_p(1). \quad (21)$$

The first term at the right-hand side of the last equality is equivalent to the Pearson's classic chi-square statistic (via comparing observed to expected cell frequencies) for testing Hardy-Weinberg equilibrium which is well-known to have the χ_1^2 distribution (Emigh 1980). Using the above equations, when $4n_0n_2 > n_1^2$, we have

$$2\log\frac{L_D L_H}{L_0} = 2m_+\frac{(\hat{p}_H - P_0)^2}{P_0(1-P_0)} + 2n_+\frac{(\hat{p}_D - P_0)^2}{P_0(1-P_0)} - 2(m_+ + n_+)\frac{(\hat{p}_0 - P_0)^2}{P_0(1-P_0)} + 2n_+\frac{(n_g/n_+ - \hat{p}_D)^2}{\hat{p}_D(1-\hat{p}_D)} + o_p(1). \quad (22)$$

By equations (19) and (20), from the above equation, we have

$$2\log\frac{L_D L_H}{L_0} = (\sqrt{\rho}Z_H + \sqrt{1-\rho}Z_D)^2 + 2n_+\frac{(n_g/n_+ - \hat{p}_D)^2}{\hat{p}_D(1-\hat{p}_D)} + o_p(1). \quad (23)$$

Note that the two terms in the right-hand side of (21) are well known to be asymptotically independent which, in turn, implies asymptotic independence of the two terms at the right-hand side of (23). Therefore, when $4n_0n_2 > n_1^2$, we have

$$2\lambda_N = 2\log\frac{L_D L_H}{L_0} = 2(\log L_D + \log L_H - \log L_0) \sim \chi_2^2.$$

Finally, it suffices to show that $P(4n_0n_2>n_1^2|H_0) \to 1/2$ as $n_+ \to \infty$. Note that, under H_0, (n_0, n_1, n_2) follow a multinomial distribution (n_+, π_0, π_1, π_2), where $\pi_g = P(X_D = g)$, for g = 0, 1, 2. Let U^T be the random vector $(\frac{n_0}{n_+}, \frac{n_1}{n_+}, \frac{n_2}{n_+})$. Then we have (Bickel & Docksum, 2000)

$$E(U)^T=\Pi=(\pi_0,\pi_1,\pi_2),\ Var(U)=\Sigma/n_+,$$

where

$$\Sigma=\begin{pmatrix} \pi_0(1-\pi_0) & -\pi_0\pi_1 & -\pi_0\pi_2 \\ -\pi_0\pi_1 & \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_0\pi_2 & -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}. \quad (24)$$

Let $G(U)$ denote $\frac{4n_0n_2-n_1^2}{{n_+}^2}$, $G(\Pi)$ denote $4\pi_0\pi_2-\pi_1^2$. Under H_0, then

$$\pi_0=(1-P_0)^2, \pi_1=2P_0(1-P_0), \pi_2=P_0^2.\ G(\Pi)=4\pi_0\pi_2-\pi_1^2=0.$$

By the central limit theorem and the multivariate delta method, $G(U)$ has an asymptotic normal distribution with mean 0. That is

$$\sqrt{N}(G(U)-G(\Pi)) \to N(0, G'(U)\,Var(U)\,G'(U)^T). \quad (25)$$

Thus, under H_0, as $n_+ \to \infty$,

$$P(4n_0n_2-n_1^2<0)=P(G(U)<0) \to 1/2.$$

This finishes the proof of the following convergence in distribution, under H_0,

$$2\lambda_N \to \frac{1}{2}\chi_1^2+\frac{1}{2}\chi_2^2. \quad (26)$$

Table 1

The genotype frequencies for case-control data of a SNP.

	AA	***aA***	***aa***	**total**
case	n_0	n_1	n_2	n_+
control	m_0	m_1	m_2	m_+

Table 2

Empirical type I error and nominal significance level at $\theta_b = 0.4$ and $n_+ = m_+ = 1000$.

Nominal level	.01	.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
Empirical level	.0098	.00099	9.7×10^{-5}	9.8×10^{-6}	9.7×10^{-7}	9.5×10^{-8}	9.8×10^{-9}

Table 3

empirical power* when X_D has a mixture distribution with $J = 2$

Set-up	1	2	3	4	5	6	7	8
θ_b	0.1	0.1	0.2	0.2	0.25	0.25	0.3	0.3
θ_1	0.12	0.08	0.18	0.23	0.20	0.30	0.32	0.28
θ_2	0.50	0.50	0.60	0.60	0.70	0.70	0.70	0.70
α_1	0.90	0.85	0.80	0.9	0.8	0.9	0.9	0.8
α_2	0.10	0.15	0.20	0.1	0.2	0.1	0.1	0.2
N	1000	1000	1500	1500	1500	1000	2000	1500
Power								
LRT	0.717	0.840	0.987	0.833	0.997	0.829	0.673	0.918
ATT	0.439	0.059	0.576	0.717	0.087	0.756	0.490	0.362
χ^2_2	0.399	0.150	0.810	0.687	0.718	0.709	0.484	0.604

*Significance level is set at 5×10^{-8}

Table 4

empirical power* when X_D has a mixture distribution with $J = 3$

Set-up	1	2	3	4	5	6	7	8
$\theta_b = \theta_3$	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.3
θ_1	0.13	0.15	0.15	0.1	0.25	0.22	0.2	0.33
θ_2	0.5	0.5	0.4	0.6	0.6	0.6	0.7	0.7
α_1	0.35	0.4	0.3	0.2	0.3	0.35	0.4	0.4
α_2	0.15	0.1	0.2	0.2	0.1	0.15	0.2	0.15
α_3	0.5	0.5	0.5	0.6	0.6	0.5	0.4	0.45
N	800	1200	800	1000	2000	1500	1500	1500
Power								
LRT	0.883	0.903	0.843	0.829	0.846	0.937	0.864	0.852
ATT	0.554	0.707	0.713	0.125	0.612	0.696	0.011	0.631
χ^2_2	0.528	0.683	0.641	0.32	0.64	0.757	0.274	0.665

*Significance level is set at 5×10^{-8}

Table 5

Comparison of P-values for the four SNPs reported in Turnbull et al. (2010).

SNP	Stage	LRT P-values	ATT P-values	χ^2 test P-values
rs10995190	Stage 1	7×10^{-9}	5×10^{-9}	3×10^{-8}
	Stage 2	2×10^{-8}	10^{-8}	2×10^{-8}
	Fisher P-value	5×10^{-15}	2×10^{-15}	2×10^{-14}
rs614367	Stage 1	6×10^{-10}	2×10^{-10}	3×10^{-10}
	Stage 2	6×10^{-15}	10^{-8}	6×10^{-8}
	Fisher P-value	2×10^{-22}	10^{-16}	7×10^{-16}
rs704010	Stage 1	3×10^{-8}	3×10^{-6}	7×10^{-7}
	Stage 2	7×10^{-4}	3×10^{-4}	4×10^{-4}
	Fisher P-value	5×10^{-10}	2×10^{-8}	6×10^{-9}
rs1011970	Stage 1	3×10^{-5}	9×10^{-6}	5×10^{-5}
	Stage 2	7×10^{-5}	2×10^{-4}	4×10^{-4}
	Fisher P-value	4×10^{-8}	4×10^{-8}	4×10^{-7}