Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 468-478
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Genome-Wide Likelihood Ratio Tests under Heterogeneity

Xiaoxia Han[1], Yongzhao Shao[2*]

[1]Public Health Sciences Department, Henry Ford Health System, Detroit, MI, USA
[2]Department of Population Health, New York University School of Medicine, New York, NY, USA
Email: *shaoy01@nyu.edu

How to cite this paper: Han, X.X. and Shao, Y.Z. (2018) Genome-Wide Likelihood Ratio Tests under Heterogeneity. *Open Journal of Statistics*, **8**, 468-478.
https://doi.org/10.4236/ojs.2018.83030

Received: April 25, 2018
Accepted: June 8, 2018
Published: June 11, 2018

Abstract

The commonly used statistical methods in medical research generally assume patients arise from one homogeneous population. However, the existence and importance of significant heterogeneity have been widely documented. It is well known that common and complex human diseases usually have heterogeneous disease etiology, which often involves interplay of multiple genetic and environmental factors, leading to latent population substructure. Genome-wide association studies (GWAS) is a useful tool to uncover genetic association with disease of interest, while linkage analysis is a commonly used method to identify statistical association between the inheritance of a human disease and inheritance of marker loci that are in linkage with disease causing loci. We propose a likelihood ratio test for genome-wide linkage analysis under genetic heterogeneity using family data. We derive a closed-form formula for the LRT test statistic and provide explicit asymptotic null distribution. The closed form asymptotic distribution allows easy determination of the asymptotic p-values. Our extensive simulation studies indicate that the proposed test has proper type I error and good power under genetic heterogeneity. In order to simplify application of the proposed method for non-statisticians, we develop an R package gLRTH to implement the proposed LRT for genome-wide linkage analysis as well as Qian and Shao's LRT for GWAS under heterogeneity. The newly developed open source R package gLRTH is available at CRAN.

Keywords

Genetic Heterogeneity, Transmission Heterogeneity, Complex Disease, Genome-Wide Association Study, Genetic Linkage Analysis, R Software Package

1. Introduction

The commonly used statistical methods in medical research generally assume that patients in the study arise from one homogeneous population. However, the existence and importance of significant heterogeneity are well known and have been documented in literature for many diseases, including Alzheimer's disease [1] [2], asthma [3] [4], diabetes [5] [6], and multiple cancer types [7] [8] [9] [10]. These common and complex human diseases usually have non-unique disease etiology, which also frequently involve interplay of multiple genetic and environmental factors, leading to latent population substructure [11] [12] [13]. Therefore, it is common that the patient population of a complex disease consists of various latent subpopulations, each with disease caused by mutations at different loci. Yet each of the unobservable subgroups is relatively homogeneous in etiology or diagnosis.

The genome-wide association study (GWAS) and linkage analysis are two classical approaches for studying human genetic disorders. GWAS is an experimental design (typically case-control) used to detect associations between genetic variants and diseases/traits from a study population [14]. The ultimate goal of the population-based GWAS is to assist researchers to have a better understanding of the biology of the disease and develop better prevention or treatment strategies for common and complex diseases. However, the standard GWAS analysis methods ignore the widely existing genetic heterogeneity. To account for latent genetic heterogeneity in GWAS, Qian and Shao [15] recently developed a novel likelihood ratio test under genetic heterogeneity (LRT-H). This methods has been shown to have superior power advantage over the commonly used Cochran-Armitage trend test (CATT) in GWAS for complex diseases where genetic heterogeneity commonly exists [15] [16].

Linkage analysis is a commonly used method to identify statistical association between the inheritance of a human disease and inheritance of marker loci before the era of GWAS. In the last two decades, linkage based gene mapping has been marginalized by the population-based genome-wide association study. Association analysis uses common variants and allows for finer mapping than linkage analysis in general. However, one major problem for association study is population stratification, which can lead to increased number of false negative as well as false positive findings if latent heterogeneity is not properly controlled for [17]. Yet this is not a concern for family-based linkage analysis, as children's genotypes only depend on their parents but not on the population genotype frequencies [18] [19]. Recent advancement in next generation sequencing (NGS) has made it technologically feasible and financially affordable to determine mutation profiles for families. Linkage analysis again becomes important to identify causal variants using family-based deep sequencing data. Ott *et al.* [20] and Shao [21] presented reviews of genetic linkage analysis in the age of NGS.

For marker alleles that are associated with inheritance of complex disease, it is not uncommon that the transmission probabilities of a marker allele of interest vary across heterozygous parents, due to locus heterogeneity, etiologic heterogeneity, and many other complexities and/or combinations of them [11].

For example, breast cancer as a complex disease, is well known to be heterogeneous. Some cases of breast cancer are due to the inherited mutations in BRCA1/2 in some families [7], while in other families due to mutations in other genes (e.g. PTEN) [22]. These genetic heterogeneities are often not directly observable from linkage data or GWAS data. The current available genetic linkage methods that account for latent genetic heterogeneity are based on mixture models and generally are computational expensive for genome-wide or NGS data [13] [23] [24] [25], yet ignoring heterogeneity can cause loss of efficiency in statistical test with increased numbers of false negative findings or missed opportunities.

In the era of whole genome sequencing, it is important to have statistical tests that are 1) computationally efficient even for genome-wide data, 2) robust under genetic heterogeneity and 3) statistically powerful. Motivated by the Qian and Shao's [15] LRT-H for GWAS, in this paper we propose a powerful and computational efficient likelihood ratio test under genetic heterogeneity for linkage analysis based on a binomial mixture model, using family data with parental marker genotypes and genotypes of two affected siblings. We have developed an R package gLRTH to implement the newly proposed LRT for genome-wide linkage analysis under genetic heterogeneity as well as Qian and Shao's [15] LRT-H for GWAS. The package is freely available on CRAN. The purpose of this R package is to simplify the application of these two methods for non-specialists. The rest of paper is organized as follows. In Section 2, we introduce the LRT for linkage analysis under genetic heterogeneity. We derive the closed-form test statistic and provide explicit asymptotic null distribution that simplify the computations for p-values. In Section 3, we present numerical simulation studies for type I error and power analysis. In Section 4, we describe the R functions and their arguments. The paper is concluded in Section 5.

2. Methods

Genetic markers can have multiple alleles. In next generation sequencing (NGS), GWAS and other genome-wide studies, markers with two alleles are most common. Thus, without much loss of generality, we focus on markers with two alleles. Here we consider a binary trait and focus on detecting linkage under genetic heterogeneity at a single marker locus with two alleles *A* and *a*. We consider independent families each with one marke-homozygous (*AA*) parent, one marker-heterozygous parent (*Aa*) and two diseased children. Let *X* denote the total number of allele *a* inherited by the two affected children from their heterozygous parent (*Aa*). Then *X* has a binomial distribution $B_2(g,\theta_b)$,

$$P(X=g)=B_2(g,\theta_b),\ g=0,1,2,\ \theta_b\in[0,1], \tag{1}$$

where

$$B_2(g,\theta_b)=\binom{2}{g}\theta_b^g(1-\theta_b)^{(2-g)}$$

and θ_b is the transmission probability for the marker-heterogeneous patient to pass allele *a* to a child.

Under the null hypothesis H_0 of no linkage between the marker and any disease-causing loci, $\theta_b = 0.5$ for all families, *i.e.* $X \sim B_2(g, 0.5)$. One can test linkage by detecting distribution departure from the null $B_2(g, 0.5)$:

$$H_0 : X \sim B_2(g, 0.5) \text{ against } H_a : X \nsim B_2(g, 0.5). \tag{2}$$

However, transmission heterogeneity, *i.e.*, variations among θ_b generally exists in complex diseases. For example, any combination of the complexities listed in Lander and Schork [12] can result in transmission heterogeneity. Thus, under transmission heterogeneity, we assume *X*, the number of allele *a*, follows a binomial mixture distribution in the population, that is

$$\begin{aligned} &P_\eta(X = g) = \sum_{j=1}^{J} \alpha_j B_2(g, \theta_j), \ g = 0,1,2, \\ &J \ge 2, 1 > \theta_j > 0, \sum_{j=1}^{J} \alpha_j = 1, \alpha_j \ge 0, \end{aligned} \tag{3}$$

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

Suppose *n* independent families each with one marker homozygous (*AA*) parent, one marker heterozygous parent (*Aa*) and two diseased children are sampled from the population. For each locus, the observed genotype frequencies inherited from the heterozygous *Aa* parent in the two diseased children are summarized in the first row in Table 1. Under H_0, the expected genotype frequencies are summarized in the second row in Table 1.

2.1. Mixture Binomial and Maximum Likelihood

Assuming the setup in the previous subsection and using notations in Table 1, the maximum likelihood estimator (MLE) of θ under the binomial likelihood in Equation (1) is

$$\hat{\theta} = (n_2 + n_1/2)/n. \tag{4}$$

Thus, the binomial likelihood in Equation (1) evaluated at $\hat{\theta}$ is

$$L_M = \prod_{g=0}^{2} B_2(g, \hat{\theta})^{n_g}, \tag{5}$$

Table 1. Genotype frequencies inherited from the heterozygous *Aa* parents for *n* affected sibling pairs.

	AA	Aa	aa	total
Observed	n_0	n_1	n_2	n
Expected	$n/4$	$n/2$	$n/4$	n

where $B_2\left(g,\hat{\theta}\right)=\binom{2}{g}\hat{\theta}^g\left(1-\hat{\theta}\right)^{2-g}$.

Under H_0, $\theta=1/2$, the binomial likelihood value is

$$L_0=\prod_{g=0}^{2}B_2\left(g,\frac{1}{2}\right)^{n_g}. \tag{6}$$

The maximum of the mixture likelihood for X in Equation (3) has an explicit formula [15], that is

$$L_D=\sup_{\eta}\prod_{g=0}^{2}P_\eta\left(X=g\right)^{n_g}=\begin{cases}\prod_{g=0}^{2}\left(n_g/n\right)^{n_g} & \text{if } 4n_0n_2>n_1^2\\ \prod_{g=0}^{2}B_2\left(g;\hat{\theta}\right)^{n_g} & \text{if } 4n_0n_2\le n_1^2,\end{cases} \tag{7}$$

where $\hat{\theta}$ is defined in Equation (4).

2.2. The Likelihood Ratio Test

Using the maximum of the likelihood L_0, L_M and L_D, respectively, we can write down the explicit formula of the log-LRT statistic $2\lambda_N$ as follows,

$$2\lambda_N=2\left(\log L_D-\log L_0\right) \tag{8}$$

Equation (8) can be written as following

$$2\lambda_N=2\log\frac{L_D}{L_0}=2\log\frac{L_D}{L_M}+2\log\frac{L_M}{L_0}. \tag{9}$$

First, we may consider a classic problem for testing $H_0:B_2\left(g;1/2\right)$ against $H_a:B_2\left(g;\theta_b\right),\theta_b\in\left(0,1\right)$. The LRT statistic is well known to have a χ_1^2 distribution.

$$2\log\frac{L_M}{L_0}=2\sum_{g=0}^{2}n_g\log\frac{B_2\left(g,1/2\right)}{B_2\left(g,\hat{\theta}\right)}=2n\frac{\left(\hat{\theta}-1/2\right)^2}{1/2\left(1-1/2\right)}+o_p\left(1\right), \tag{10}$$

where $\hat{\theta}=\left(n_2+n_1/2\right)/n$ is the MLE of θ_b defined in Equation (4).

When $4n_0n_2\le n_1^2$, we have $L_D=L_M$. Thus,

$$2\log\frac{L_D}{L_M}=0. \tag{11}$$

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

When $4n_0n_2>n_1^2$, we can consider testing of goodness-of-fit of $H_0:B_2\left(g;1/2\right)$. The LRT statistic has a χ_2^2 asymptotic distribution and can be written as

$$\begin{aligned}&2\sum_{g=0}^{2}n_g\log\frac{n_g/n}{B_2\left(g;1/2\right)}\\&=2\sum_{g=0}^{2}n_g\log\frac{n_g/n}{B_2\left(g;\hat{\theta}\right)}+2\sum_{g=0}^{2}n_g\log\frac{B_2\left(g;\hat{\theta}\right)}{\left(1/2\right)\left(1-1/2\right)}\\&=2n\frac{\sum_{g=0}^{2}\left(n_g/n-\hat{\theta}\right)^2}{\hat{\theta}\left(1-\hat{\theta}\right)}+2n\frac{\left(\hat{\theta}-1/2\right)^2}{\left(1/2\right)\left(1-1/2\right)}+o_p\left(1\right).\end{aligned} \tag{12}$$

The first term at the right-hand side of the last equality is equivalent to the Pearson's classic χ^2 statistic (via comparing observed to expected cell frequencies) for testing Hardy-Weinberg equilibrium which is know to have the χ_1^2 distribution. Note that the two terms in the right hand side of equation (12) are well known to be asymptotically independent. Therefore, when $4n_0n_2 > n_1^2$, we have

$$2\lambda_N = 2\log\frac{L_D}{L_0} \sim \chi_2^2. \tag{13}$$

It is easy to show that $P\left(4n_0n_2 > n_1^2 \mid H_0\right) \to 1/2$ as $n \to \infty$ as in Qian and Shao [15]. Thus, we obtain the explicit form of asymptotic distribution under the null hypothesis. That is, under H_0,

$$2\lambda_N \to \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2. \tag{14}$$

Importantly, to implement the LRT, there is no need to identify the exact number of mixture components J in equation (3).

3. Simulations

3.1. Type I Errors

As the LRT $2\lambda_N$ has an explicit asymptotic distribution under H_0, it is convenient to evaluate p-value and type I error. We conducted simulations to compare the empirical type I error of the LRT to the nominal significant level ranging from 10^{-2} to 10^{-8}. The genotype data were generated from binomial distribution $B_2\left(g;1/2\right)$. The simulation was replicated 10^{11} times. As shown in Table 2, the empirical type I error is slightly smaller than the nominal level, but they are very close to each other. Therefore, using the asymptotic null distribution for the LRT is valid. The closed form asymptotic distribution allows easy determination of the asymptotic p-values.

3.2. Power Comparison

In the simulation studies for power comparison, the sample was generated from a two-component mixture binomial distribution as described in equation (3) with $J = 2$, *i.e.*,

$$P_\eta\left(X = g\right) = \sum_{j=1}^{2}\alpha_j B_2\left(g,\theta_j\right),\ g = 0,1,2,\ 1 > \theta_j > 0,\ \alpha_1 + \alpha_2 = 1,\ \alpha_j \geq 0.$$

One hundred-thousand replicate dataset of n disease cases $\left(n = 800, 1000 \text{ or } 1200\right)$ were simulated for each of the seven simulation setup

Table 2. Empirical type I error and nominal significant level at $\theta = 1/2$ and $n = 1000$ with 10^{11} replications.

Nominal level	0.01	0.001	1×10^{-4}	1×10^{-5}	1×10^{-6}	1×10^{-7}	1×10^{-8}
Empirical level	0.0097	9.9×10^{-4}	9.5×10^{-5}	9.8×10^{-6}	9.4×10^{-7}	9.6×10^{-8}	9.1×10^{-9}

and the empirical power for LRT and χ_2^2 are shown in Table 3. The simulation results indicate that the LRT has power advantage over the χ_2^2 test under genetic heterogeneity.

4. The R Package Description and Examples

The gLRTH R package is available on CRAN and the installation is standard. The purpose of this package is to implement the previously discussed two methods, *i.e.*, LRT for genome-wide linkage analysis under genetic heterogeneity and Qian and Shao's LRT-H for GWAS [15]. The gLRTH R package is composed of two main functions: gLRTH_L for linkage analysis under heterogeneity and gLRTH_A for association studies.

The gLRTH_L function calculates the test statistic and asymptotic p-value for the likelihood ratio test for testing linkage. The gLRTH_L function in the package can be called with the following syntax:

$$\text{gLRTH\_L}(n_0, n_1, n_2)$$

The required arguments are:

1) n0: Number of affected sibling pairs that both inherited A from their heterozygous parent Aa

2) n1: Number of affected sibling pairs that one inherited A and the other inherited a from their heterozygous parent Aa

3) n2: Number of affected sibling pairs that both inherited a from their heterozygous parent Aa

To illustrate the gLRTH_L function, suppose we have hypothetical genetic marker $M1/M2$ information from a sample of $n = 1000$ independent families, with $M2$ be the marker of interest. Each family has one marker homozygous ($M1/M1$) parent, one marker heterozygous parent ($M1/M2$) and two diseased children. Suppose we have $n_0 = 100$ families with both sibling inherited $M1$ from their heterozygous parent ($M1/M2$), $n_1 = 650$ families have one sibling inherited $M2$ and one sibling inherited $M1$ from their heterozygous parent ($M1/M2$), and $n_2 = 250$ families have both siblings inherited $M2$ from their heterozygous parent ($M1/M2$).

Table 3. Empirical power (significant level is set at 5×10^{-8}) when X has a mixture distribution with $J = 2$.

Setup	1	2	3	4	5	6	7	8	9	10
α_1	0.45	0.4	0.35	0.25	0.45	0.4	0.45	0.35	0.5	0.25
α_2	0.55	0.6	0.65	0.75	0.55	0.6	0.55	0.65	0.5	0.75
θ_1	0.24	0.2	0.2	0.12	0.24	0.22	0.24	0.2	0.3	0.14
θ_2	0.72	0.69	0.66	0.63	0.69	0.69	0.69	0.66	0.72	0.63
N	800	800	1000	1000	1000	1000	1200	1200	1200	1200
Power										
LRT	0.80	0.83	0.68	0.72	0.80	0.87	0.93	0.86	0.72	0.75
χ_2^2	0.77	0.80	0.64	0.68	0.77	0.84	0.91	0.83	0.68	0.71

The LRT for linkage under transmission heterogeneity for this genetic marker can be done as following:

$$\text{gLRTH\_L}(\text{n}_0 = 100,\ \text{n}_1 = 650,\ \text{n}_2 = 250)$$

The output is:

```
$test.stat
[1] 45.17029
$pval
[1] 8.672153e-11
```

In this example, for linkage analysis under transmission heterogeneity of marker $M1/M2$, the test statistic is 24.04 and the p-value is 8.7×10^{-11}.

The gLRTH_A function calculates the test statistic and asymptotic p-value for the likelihood ratio test for GWAS. The gLRTH_A function in the package can be called with the following syntax:

$$\text{gLRTH\_A}(\text{n}_0,\ \text{n}_1,\ \text{n}_2,\ \text{m}_0,\ \text{m}_1,\ \text{m}_2)$$

The required arguments are:

1) n0: *AA* genotype frequency in case
2) n1: *Aa* genotype frequency in case
3) n2: *aa* genotype frequency in case
4) m0: *AA* genotype frequency in control
5) m1: *Aa* genotype frequency in control
6) m2: *aa* genotype frequency in control

To illustrate the gLRTH_A function, we consider a SNP called SNP rs429358 in gene Apolipoprotein E (ApoE), which is a well-known common variants that is associated with late-onset Alzheimer's diseases (AD). We use APOE $\epsilon4$ variants frequency in Han *et al.* [26] to determine SNP rs429358 in AD converters and AD non-converters. The LRT-H for SNP rs429358 can be done as following:

$$\text{gLRTH\_A}(\text{n}_0 = 89,\ \text{n}_1 = 139,\ \text{n}_2 = 47,\ \text{m}_0 = 266,\ \text{m}_1 = 153,\ \text{m}_2 = 39)$$

The output is:

```
$test.stat
[1] 46.02864
$pval
[1] 5.640675e-11
```

In this example, for SNP rs429358, the test statistic is 46.02 with a p-value 5.6×10^{-11}. We conclude that SNP rs429358 reaches genome-wide significance 5×10^{-8}.

5. Conclusions

The commonly used statistical methods in medical research often assume patients arise from one homogeneous population. However, the impact of heterogeneity is well known and has been document in much of the existing literature for common and complex diseases. Inadequate attention to the heterogeneity inherent in the complexity of complex human disease could lead to increased number of false negatives and missed opportunities in research. To solve this problem, using finite mixture models to account for latent genetic heterogeneity is an intuitive strategy. However, there are well known difficulties associated with likelihood-based inference in the context of finite mixture due to issues regarding parameter identifiability and degenerate Fisher information [27]. The mixture likelihood often has many local maximum values making the numerical maximization complicated. Moreover, the likelihood irregularities lead to great challenges in deriving the limit distribution of the LRT statistic under loss of identifiability. The strength of the proposed method is that we are able to derive closed form formula for the LRT statistic and its simple closed form asymptotic distribution despite the loss of identifiability in parameters in the context of mixture likelihood. This leads to efficient computation of the test statistic and its asymptotic p-values; and thus, it is suitable for high throughput data and genome-wide studies. The proposed method also works for a single marker or a few markers. There are a few existing methods for linkage analysis that account for latent heterogeneity [13] [23] [24] [25], but the existing methods are computationally expensive for NGS and genome-wide studies.

The rapid development of next generation whole-genome sequencing (WGS) has revived family-based linkage analysis for identification and characterization of functional variants. Our proposed LRT for linkage analysis under genetic heterogeneity will likely to be a powerful tool for genetic mapping of complex traits [20]. In the era of precision medicine, using individual variations in genes and environment to develop diagnostics, prognostics, and therapies is the primary approach for disease prevention and treatment. For example, instead of using "one-size-fits-all-approach", "precision medicine" based on genetic markers can be used to optimize effectiveness of disease prevention and treatment as well as minimize side effects for persons less likely to respond to a particular therapeutic. Reliable disease associated SNPs could serve as predictive markers that inform our decisions about numerous aspects of medical care, including specific diseases, effectiveness of various drugs and adverse reactions to specific drugs. We believe that with the reduction in cost of whole-genome sequencing (WGS), genome-wide linkage analysis of family based WGS data as well as GWAS will facilitate the identification of causal variants and may contribute tremendously to the advancement of precision medicine. Our open source R package gLRTH is meant to be a valuable package to help researchers perform GWAS and genome-wide linkage analysis accounting for the ubiquitous genetic heterogeneity in common and complex human diseases without a lot of programming and computational burden.

Acknowledgements

This research was partially supported by a pilot project from NYU Alzheimers Disease Center under the NIH/NIA grant P30 AG08051.

References

[1] Drummond, E., Nayak, S., Faustin, A., Pires, G., Hickman, R.A., Askenazi, M., *et al.* (2017) Proteomic Differences in Amyloid Plaques in Rapidly Progressive and Sporadic Alzheimers Disease. *Acta Neuropathologica*, **133**, 933-954. https://doi.org/10.1007/s00401-017-1691-0

[2] Lee, H.B. and Lyketsos, C.G. (2003) Depression in Alzheimers Disease: Heterogeneity and Related Issues. *Biological Psychiatry*, **54**, 353-362. https://doi.org/10.1016/S0006-3223(03)00543-2

[3] Fitzpatrick, A.M., Teague, W.G., Meyers, D.A., Peters, S.P., Li, X., Li, H., *et al.* (2017) Heterogeneity of Severe Asthma in Childhood: Confirmation by Cluster Analysis of Children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *Journal of Allergy and Clinical Immunology*, **127**, 382-389. https://doi.org/10.1016/j.jaci.2010.11.015

[4] Drazen, J.M., Silverman, E.K. and Lee, T.H. (2002) Heterogeneity of Therapeutic Responses in Asthma. *British Medical Bulletin*, **56**, 1054-1070. https://doi.org/10.1258/0007142001903535

[5] Hattersley, A.T. (1998) Maturity-Onset Diabetes of the Young: Clinical Heterogeneity Explained by Genetic Heterogeneity. *Diabetic Medicine*, **1**, 15-24. https://doi.org/10.1002/(SICI)1096-9136(199801)15:1%3C15::AID-DIA562%3E3.0.CO;2-M

[6] Sladek, R., Rocheleau, G., Rung, J., Christian, D., Shen, L., Serre, D., *et al.* (2007) A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes. *Nature*, **445**, 881-885. https://doi.org/10.1038/nature05616

[7] Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., *et al.* (1998) Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. *The American Journal of Human Genetics*, **62**, 676-689. https://doi.org/10.1086/301749

[8] Polyak, K. (2011) Heterogeneity in Breast Cancer. *The Journal of Clinical Investigation*, **121**, 3786-3788. https://doi.org/10.1172/JCI60534

[9] Lachiewicz, A.M., Berwick, M., Wiggins, C.L., Thomas, N.E., Goldgar, D. and Devilee, P. (1998) Epidemiologic Support for Melanoma Heterogeneity Using the Surveillance, Epidemiology, and End Results Program. *Journal of Investigative Dermatology*, **128**, 1340-1342. https://doi.org/10.1038/jid.2008.18

[10] Yancovitz, M., Litterman, A., Yoon, J., Ng, E., Shapiro, R.L., Berman, R.S., *et al.* (2012) Intra- and Inter-Tumor Heterogeneity of $BRAF^{V600E}$ Mutations in Primary and Metastatic Melanoma. *PLoS ONE*, **7**, 676-689. https://doi.org/10.1371/journal.pone.0029336

[11] Shao, Y. (2005) Adjustment for Transmission Heterogeneity in Mapping Complex Genetic Diseases Using Mixture Models and Score Tests. *Proceeding of the American Statistical Association*, 383-393.

[12] Lander, E.S. and Schork, N.J. (1994) Genetic Dissection of Complex Traits. *Science*, **265**, 2037-2037. https://doi.org/10.1126/science.8091226

[13] Ott, J. (1999) Analysis of Human Genetic Linkage. JHU Press, Baltimore.

[14] Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, **101**, 5-22. https://doi.org/10.1016/j.ajhg.2017.06.005

[15] Qian, M. and Shao, Y. (2013) A Likelihood Ratio Test for Genome-Wide Associa-

tion under Genetic Heterogeneity. *Annals of Human Genetics*, **77**, 174-182. https://doi.org/10.1111/ahg.12005

[16] Xu, Z. and Pan, W. (2016) Binomial Mixture Model Based Association Testing to Account for Genetic Heterogeneity for GWAS. *Genetic Epidemiology*, **40**, 202-209. https://doi.org/10.1002/gepi.21954

[17] McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P. and Hirschhorn, J.N. (2008) Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges. *Nature Reviews Genetics*, **9**, 356-369. https://doi.org/10.1038/nrg2344

[18] Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM). *American Journal of Human Genetics*, **52**, 506-516. http://europepmc.org/articles/pmc1682161

[19] Risch, N.J. (2000) Searching for Genetic Determinants in the New Millennium. *Nature*, **405**, 847-856.

[20] Ott, J., Wang, J. and Leal, S.M. (2015) Genetic Linkage Analysis in the Age of Whole-Genome Sequencing. *Nature Reviews Genetics*, **16**, 275-284. https://doi.org/10.1038/nrg3908

[21] Shao, Y. (2018) Linkage Analysis, Encyclopedia of Quantitative Risk Analysis and Assessment. Wiley StatsRef: Statistics Reference Online.

[22] Lynch, E.D., Ostermeyer, E.A., Lee, M.K., Arena, J.F., Ji, H., Dann, J., *et al.* (1997) Inherited Mutations in PTEN That Are Associated with Breast Cancer, Cowden Disease, and Juvenile Polyposis. *The American Journal of Human Genetics*, **61**, 1254-1260. https://doi.org/10.1086/301639

[23] Lo, S., Liu, X. and Shao, Y. (2017) A Marginal Likelihood Model for Family-Based Data. *Annals of Human Genetics*, **67**, 357-366. https://doi.org/10.1046/j.1469-1809.2003.00032.x

[24] Fu, Y., Chen, J. and Kalbfleisch, J.D. (2006) Testing for Homogeneity in Genetic Linkage Analysis. *Statistica Sinica*, **16**, 805-823. http://www.jstor.org/stable/24307575

[25] Han, J. and Shao, Y. (2012) The Transmission Disequilibrium/Heterogeneity Test with Parental-Genotype Reconstruction for Refined Genetic Mapping of Complex Diseases. *Journal of Probability and Statistics*, **2012**, Article ID: 256574. https://doi.org/10.1155/2012/256574

[26] Han, X., Zhang, Y., Shao, Y. and the Alzheimer's Disease Neuroimaging Initiative (2017) Application of Concordance Probability Estimate to Predict Conversion from Mild Cognitive Impairment to Alzheimer's Disease. *Biostatistics & Epidemiology*, **1**, 105-118.

[27] Liu, X. and Shao, Y. (2003) Asymptotics for Likelihood Ratio Tests under Loss of Identifiability. *The Annals of Statistics*, **31**, 807-832.