Stony Brook University

Modeling the effect of sequencing error

A Dissertation Presented

by

Ruiqi Zhang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2014

Stony Brook University

The Graduate School

Ruiqi Zhang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen J. Finch – Dissertation Advisor
Professor of Applied Mathematics and Statistics

Nancy R. Mendell - Chairperson of Defense
Professor of Applied Mathematics and Statistics

Wei Zhu
Professor of Applied Mathematics and Statistics

Derek Gordon
Associate Professor, Department of Genetics,
Rutgers The State University of New Jersey

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Modeling the effect of sequencing error

by

Ruiqi Zhang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2014

Genotype misclassification errors are known to reduce the power to detect genetic association, but the size of the effect is not known in next generation sequencing (NGS). The non-centrality parameter (NCP) and hence power of the association test allowing for errors for a specified error model at a base pair was found. This NCP was compared to the NCP for the usual chi-square test. The asymptotic power was compared to simulated power for specific settings of the true genotype and phenotype frequencies in the case and control populations, genotype misclassification rates, and total sample size. An R script was provided for calculating the NCP. Next, the effect of misclassification error using data from NGS technology for case-control genetic association studies was modeled. The Likelihood Ratio Test Allowing for Error using NGS data (LRT_{NGS}) was derived. The estimated genotype frequencies and misclassification rates from the observed base pair reads were calculated using the expectation-maximization (EM) algorithm. This statistic allows for both non-differential and differential misclassification. The

distribution of LRT_{NGS} was studied by simulations for both null and alternative settings. The effects of genotyping misclassification rates on the sample size needed to maintain the constant asymptotic Type I and Type II error rates were studied. For at risk minor allele frequencies less than 0.01, large sample sizes were required for the asymptotic distribution to be a good approximation. Increasing the sequencing coverage increased the estimated power and the adequacy of simulated power.

To my parents with all my love

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1-1 Research Background

Many common human diseases and traits are believed to be influenced by several genetic and environment factors, but the identification of genetic variants contributing to these 'complex diseases' is slow. Genome-wide association studies (GWAS) are an examination of many common genetic variants in different individuals to see if any variant is associated with a trait and to identify common genetic factors that influence health and disease [18]. GWAS represent a powerful new method for investigating the genetic architecture of complex diseases [20].

Although GWAS have found hundreds of common variants associated with disease, there is still a large fraction of heritability that needs to be explained. The limitations of GWAS that focus on the common genetic variants have motivated scientists to consider the contribution of rare variants to phenotypic expression[23, 26]. The increasing availability of high-throughput sequencing technologies has enabled studies of rare variants [2]. Next-generation sequencing (NGS) refers to DNA sequencing technologies that highly parallelized the sequencing process and enable the sequencing of thousands to millions of molecules at once[8]. NGS technology makes it possible to directly sequence case-control samples for testing disease association including rare variants and has greatly expanded the resolution possible in GWAS[19].

Misclassification is defined as the incorrect classification of a subject. Misclassification errors are present in the majority of data and can affect the validity of a study [25]. Several researchers have investigated the effects of genotype and phenotype misclassification on the power and robustness of statistical association methods. Bross[3] studied the effect of classification errors on the chi-square test applied to a 2×2 table. Assuming that the classification error mechanism is independent of case/control status, there is no change in the probability of a type I error due to classification error. The power for the test, however, is reduced. Since the point estimates of the population frequency parameters incorporate the probability of misclassification into the expected difference between the frequencies, their expected values are not the true population frequencies. Mote and Anderson [22] extended the work of Mitra[21] and Bross[3] and proved that the power of the chi-square test with no error is always greater than or equal to the power of the test when errors are present but ignored.

Gordon et al.[7, 12] applied the results in Mitra to find the noncentrality parameter (NCP) λ of the $2 \times C$ contingency table test in the presence of misclassification error and showed that misclassification errors in genotype and phenotype can significantly reduce the power of genetic association test. Kang et al [15-17] examined the impact of each individual SNP genotyping error for the chi-square test of independence. They determined which SNP genotyping misclassification error was most deleterious in term of increase in the sample size required to maintain type I and type II error rates. Later, Gordon et al.[8, 9, 10] developed a likelihood ratio test allowing for errors (LRT_{ae}) that incorporates double sampling information to increase the power of the association test in

the presence of genotype and phenotype misclassification errors. Ji et al.[14] then calculated the corresponding NCP for LRT_{ae}.

Recently, Gordon et al.[8] present a new statistic that allows for association testing among cases and controls directly using raw base pair reads instead of genotypes produce by an intermediate algorithm.

1-2 Research Objectives

In this paper, we apply the approach in Kim et al.[18] to the work of Gordon et al.[8] That is, instead of using the probability of true sequence-read counts for individual m in the likelihood function (denoted by $P\left(A_{n_1^t(m)}^{V(m)}\right)$ in Gordon et al.[8]), we use the probability of observed sequenced- read counts from Kim et al.[17]. Then the observed number of less common alleles with coverage $V(m)$ follows a binomial distribution, with number of trials $V(m)$ and probability $\Pr\left(A_{n_1(m)}^{V(m)}|X_j^{t(m)}, Y_i^{t(m)}\right) = \binom{V(m)}{n_{1(m)}}{\mu_{ij}}^{n_{1(m)}}(1-\mu_{ij})^{V(m)-n_1(m)}$, where μ_{ij} is the probability of observing the less common allele at a base pair. We estimate the genotype frequencies and misclassification rates using the EM algorithm.

The main purpose of this dissertation is to study the effect of misclassification error of the case-control association test in the genetic analysis. Chapter 2 presents the derivation of the NCP from both the chi-square test and likelihood ratio test at a base pair. And we study the consistency of the two NCPs through simulation studies. Also we conduct simulation studies to evaluate the accuracy of the asymptotic power and

investigate how power would change as the parameter settings change. Chapter 3 introduces the likelihood ratio test allowing for differential misclassification applying to NGS data, and describes the EM algorithms for obtaining maximum likelihood estimates of genotype frequency and differential misclassification errors under the alternatives. Chapter 3 also studies the test statistic power by a simulation study. In Chapter 4, we draw conclusions and discuss possible future work.

Chapter 2 Association Test For a Base Pair

2-1 The Chi-square Test Statistic

The classical 2×2 chi-square test is often used as a test of association. At each gene locus, an individual receives two alleles, one from each parent. Consider a gene with allele types labeled as at risk (1) and not at risk (2). Let N_A be the number of cases (affected), and N_U be the number of controls (unaffected). Let n_1 be the number of participants having the at risk allele, and n_2 be the number of participants not having the at risk allele. The 2×2 table is given by:

Table 2.1 Contingency table for trait and genotype

Allele type	Cases	Controls	Total
At risk	$n_{1,A}$	$n_{1,U}$	n_1
Not at risk	$n_{2,A}$	$n_{2,U}$	n_2
Total	N_A	N_U	N

Here, $n_{1,A}$ is the number of at risk alleles in the cases, and $n_{1,U}$ is the number of at risk alleles in the controls; $n_{2,A}$ is the number of not at risk alleles in the cases, and $n_{2,U}$ is the number of not at risk alleles in the controls. The null hypothesis is that there is no association between alleles in cases and controls. Under the null hypothesis, the chi-square test has a central chi-square distribution with 1 degree of freedom asymptotically.

2-2 Asymptotic Non-Centrality Parameter

Mitra[20] derived the asymptotic power function of the chi-square test of equality of distribution for the $R \times C$ contingency table. In general, under the alternative hypothesis

that the frequencies are not the same in cases and controls, the asymptotic distribution of the chi-square test for the $2 \times C$ contingency table follows a non-central chi-square distribution with $C - 1$ degrees of freedom and non-centrality parameter (NCP) λ, where C is the number of genotypes in the model.

Let p_A^t denote the true frequency of at risk allele in the case group (affected), and let p_U^t denote the true frequency of at risk allele in the control group (unaffected). Then the true frequency of not at risk allele in the case group is $1 - p_A^t$, and the true frequency of not at risk allele in the control group is $1 - p_U^t$.

The frequency of the at risk allele in the combined case and control group is

$$p_1^t = \frac{N_A p_A^t + N_U p_U^t}{N_A + N_U};$$

and the frequency of the not at risk allele in the combined case and control group is

$$p_2^t = \frac{N_A + N_U - N_A p_A^t - N_U p_U^t}{N_A + N_U}.$$

The NCP of the 2×2 chi-square test is given by:

$$\lambda = \frac{2N_A N_U (p_A^t - p_U^t)^2 (N_A + N_U)}{(N_A p_A^t + N_U p_U^t)(N_A + N_U - N_A p_A^t - N_U p_U^t)} \qquad (2-1)$$

The asymptotic power of the association test of case-control association is then found by integrating the non-central chi-square density function:

$$Power = 1 - F(\chi_{1-\alpha}^2(1, \lambda))$$

Here, $F(\chi^2_{1-\alpha}(1,\lambda))$ is the cumulative distribution function (CDF) of the non-central chi-square distribution with 1 degree of freedom and NCP λ evaluated at the $1-\alpha$ percentile of the central χ^2 distribution.

At each gene locus, with the at risk allele labeled as 1 and not at risk allele labeled as 2, the genotype for a SNP can be $1/1\,, 1/2\ or\ 2/2$. Let k be the number of allele 2 in the genotype. Then the genotype can be identified by the value of k. The error model used here is given in Table 2.2 and is called GLHO error model[7, 11]. It is a 3×3 matrix, where each cell contains the probability that a specific genotype k is classified as k' where $k, k' = 0,1,2$. The GLHO model assumes that errors are introduced into alleles randomly and independently. Here, I assume the probability that at risk allele is misclassified as not at risk allele is equal to the probability that the not at risk allele is misclassified as the at risk allele, and is denoted by ε. I also assume that the misclassification is non-differential; that is, the misclassification error probabilities are the same for cases and controls. Let $\xi_{j|j'}$ be the probability that allele j' is coded as j, $j', j = 1, 2$. Then my assumption can be written as:

$$\xi_{j|j'} = Pr(X_j|X^t_{j'}) = \begin{cases} \varepsilon & j' \neq j \\ 1-\varepsilon & j' = j \end{cases}$$

Table 2.2 Probability of observed genotype under GLHO error matrix

Observed Recoded Genotype k'	True Recoded Genotype k		
	0	1	2
0	$(1-\varepsilon)^2$	$\varepsilon(1-\varepsilon)$	ε^2
1	$2\varepsilon(1-\varepsilon)$	$\varepsilon^2+(1-\varepsilon)^2$	$2\varepsilon(1-\varepsilon)$
2	ε^2	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)^2$

Let p_{ik}^t denote the true genotype frequency of genotype k in phenotype i, and p_{ik} denote the frequency of genotype k in phenotype i under the GLHO error model. Under HWE, we have: $p_{00}^t = (1-p_U^t)^2,\ p_{01}^t = 2p_U^t(1-p_U^t),\ p_{02}^t = (p_U^t)^2, p_{10}^t = (1-p_A^t)^2,\ p_{11}^t = 2p_A^t(1-p_A^t),\ p_{12}^t = (p_A^t)^2$. Then the genotype frequencies p_{ik} in the presence of errors are given by:

$$\left(p_{i0}, p_{i1}, p_{i2}\right)^T = E \times (p_{i0}^t, p_{i1}^t, p_{i2}^t)^T \qquad (2-2)$$

That is, using Table 2.2 and equation (2-2), the observed genotype frequencies in the presence of errors are:

$$p_{i0} = (1-\varepsilon)^2 p_{i0}^t + \varepsilon(1-\varepsilon)p_{i1}^t + \varepsilon^2 p_{i2}^t$$

$$p_{i1} = 2\varepsilon(1-\varepsilon)p_{i0}^t + [\varepsilon^2 + (1-\varepsilon)^2]p_{i1}^t + 2\varepsilon(1-\varepsilon)p_{i2}^t$$

$$p_{i2} = \varepsilon^2 p_{i0}^t + \varepsilon(1-\varepsilon)p_{i1}^t + (1-\varepsilon)^2 p_{i2}^t$$

The frequencies of allele 1 in the case group, p_A, and in the control group, p_U, can be calculated.

$$p_A = p_{00} + \frac{p_{01}}{2}$$

$$p_U = p_{10} + \frac{p_{11}}{2}$$

The corresponding NCP in presence of error is then obtained by substituting the allele frequencies p_A and p_U in equation (2-1).

$$\lambda' = \frac{2N_A N_U \left(p_A - p_U\right)^2 (N_A + N_U)}{\left(N_A p_A + N_U p_U\right)\left(N_A + N_U - N_A p_A - N_U p_U\right)}$$

$$= \frac{2N_A N_U \left(p_{00} + \frac{p_{01}}{2} - p_{10} - \frac{p_{11}}{2}\right)^2 (N_A + N_U)}{\left(N_A(p_{00} + \frac{p_{01}}{2}) + N_U(p_{10} + \frac{p_{11}}{2})\right)\left(N_A + N_U - N_A(p_{00} + \frac{p_{01}}{2}) - N_U(p_{10} + \frac{p_{11}}{2})\right)}$$

$$= \frac{2N_A N_U (N_A + N_U)}{\left(N_A((1-\varepsilon)p_{00}^t + \frac{1}{2}p_{01}^t + \varepsilon p_{02}^t) + N_U((1-\varepsilon)p_{10}^t + \frac{1}{2}p_{11}^t + \varepsilon p_{12}^t)\right)}$$

$$\times \frac{\left((1-\varepsilon)p_{00}^t + \frac{1}{2}p_{01}^t + \varepsilon p_{02}^t - \left((1-\varepsilon)p_{10}^t + \frac{1}{2}p_{11}^t + \varepsilon p_{12}^t\right)\right)^2}{\left(N_A\left(\varepsilon p_{00}^t + \frac{1}{2}p_{01}^t + (1-\varepsilon)p_{02}^t\right) + N_U(\varepsilon p_{10}^t + \frac{1}{2}p_{11}^t + (1-\varepsilon)p_{12}^t)\right)}$$

2-3 Likelihood Ratio Test Allowing for Allele Misclassification

Ji et al.[14] calculated the NCP for the Likelihood Ratio Test Allowing for Error (LRT_{ae}) in the presence of random phenotype and genotype errors when using resampling (double-sampling). Here I extend Ji et al.[14] and derive a Likelihood Ratio Test that allows for allele errors at a base pair in sequencing. A portion of the Methods and Notations are taken from Gordon et al.[13] and Ji et al.[14]

For each term used in this section, i indicates the phenotype, and j indicates the allele type. The prime superscript is used to indicate the true phenotype or genotype. For example, $i' = 0$ indicates that the true phenotype classification of an individual is not affected, and $i' = 1$ that the true phenotype is affected. The notation $j' = 1$ indicates that an individual's true allele type is 1 and $j' = 2$ indicates that the true allele type is 2.

Let X_j denote the event that an individual has observed allele type j (j=1, 2), and $X_{j'}^t$ denote the event that an individual has true allele type j' (j'=1, 2). Let Y_i denote the event that an allele has observed phenotype i ($i = 0, 1$), and $Y_{i'}^t$ denote the event that an allele has true phenotype i' (i'=0, 1). Let n_{ij} denote the number of alleles with observed phenotype i ($i = 0, 1$) and observed allele type j ($j = 1, 2$). Let $p_{j'|i'}^t = Pr(X_{j'}^t|Y_{i'}^t)$ denote the population frequency of true allele j' for individuals with true phenotype i' ($i' = 0,1$, $j' = 1, 2$). The null hypothesis that allele frequencies are equal in case groups and control groups can be written as:

$$p_{1|0}^t = p_{1|1}^t \, , p_{2|0}^t = p_{2|1}^t$$

Let $\pi_{(i|i')} = Pr(Y_i \,|Y_{i'}^t)$ denote the probability that a true phenotype i' is misclassified as i. Here I assume that there is no phenotype misclassification error; that is:

$$\pi_{i|i'} = \begin{cases} 1 & i' = i \\ 0 & i' \neq i \end{cases}$$

Let $q_{i'}^t = Pr(Y_{i'}^t)$ denote the population frequency of true phenotype$i$' ($i' = 0, 1$). The likelihood function allowing for error is given by:

$$l(\theta) = \sum_i \sum_j n_{ij} \log\big(Pr(X_j, Y_i)\big) \, ,$$

where $Pr(X_j, Y_i)$ is the probability of having observed events X_j, Y_i. We assume that conditional on the true data, the observed data are independent. That is the classification process for the phenotype is independent of the classification process for the genotype. Then, $P\,r\left(X_j, Y_i, |X_{j'}^t, Y_{i'}^t\right) = P\,r\left(X_j \middle| X_{j'}^t\right)\Pr(Y_i|\, Y_{i'}^t\big)$. It follows that:

$$Pr\left(X_j, Y_i, X_{j'}^t, Y_{i'}^t\right) = Pr\left(X_j, Y_i, |X_{j'}^t, Y_{i'}^t\right) \Pr\left(X_{j'}^t, Y_{i'}^t\right)$$

$$= Pr\left(X_j \middle| X_{j'}^t\right) \Pr(Y_i | Y_{i'}^t) \Pr\left(X_{j'}^t, Y_{i'}^t\right)$$

$$= Pr\left(X_j \middle| X_{j'}^t\right) \Pr(Y_i | Y_{i'}^t) \Pr\left(X_{j'}^t | Y_{i'}^t\right) \Pr\left(Y_{i'}^t\right) = \pi_{i|i'}, \xi_{j|j'}, p_{j'|i'}^t q_{i'}^t$$

Then

$$Pr\left(X_j, Y_i\right) = \sum_{i'} \sum_{j'} Pr(X_j, Y_i, X_{j'}^t, Y_{i'}^t).$$

The log-likelihood function is given by:

$$l(\theta) = \sum_i \sum_j n_{ij} \log\left(Pr(X_j, Y_i)\right) = \sum_i \sum_j n_{ij} \log\left(\sum_{i'} \sum_{j'} Pr(X_j, Y_i, X_{j'}^t, Y_{i'}^t)\right)$$

$$= \sum_i \sum_j n_{ij} \log\left(\sum_{i'} \sum_{j'} Pr(Y_i|Y_{i'}^t) Pr(X_j|X_{j'}^t) Pr(X_{j'}^t|Y_{i'}^t) Pr(Y_{i'}^t)\right)$$

$$= \sum_i \sum_j n_{ij} \log\left(\sum_{i'} \sum_{j'} \pi_{i|i'}, \xi_{j|j'}, p_{j'|i'}^t q_{i'}^t\right)$$

2-4 NCP for the Likelihood Ratio Test

Here I derive a closed-form expression for the NCP of Likelihood Ratio Test allowing for allele errors using Fisher's information matrix. By using the NCP, the power under different scenarios can be determined for any specified significance level.

The unit of data used here is the individual allele. The list of parameters follows Ji et al.[14] To derive the test statistic, let $\tilde{\theta}$ be:

$$\tilde{\theta} = (p^t_{1|0}, p^t_{2|0}, p^t_{1|1}, p^t_{2|1}, q^t_0, q^t_1)'$$

The parameters are subject to the constraints that:

$$p^t_{1|0} + p^t_{2|0} = 1, p^t_{1|1} + p^t_{2|1} = 1, q^t_0 + q^t_1 = 1$$

Thus, the free parameters are $\left(p^t_{1|0}, p^t_{1|1}, q^t_0\right)'$.

Fisher's information matrix of $\tilde{\theta}$ is:

$$\tilde{I}(\theta) = E(-\frac{\partial^2 l(\theta)}{\partial\tilde{\theta}\partial\tilde{\theta}'})$$

Then, we have

$$l(\theta) = \sum_i \sum_j n_{ij} \log\left(\sum_{i'}\sum_{j'} \pi_{i|i'}\xi_{j|j'}p^t_{j'|i'}q^t_{i'}\right)$$
$$= \sum_i \sum_j n_{ij} \log\left(\sum_{u'}\sum_{v'} \pi_{i|u'}\xi_{j|v'}p^t_{v'|u'}q^t_{u'}\right) \qquad (2-3)$$

where index i' and j' are replaced by the index u' and v' for clarity. u' ranges from 0 to 1, and v' ranges from 1 to 2.

The first order derivatives of the log-likelihood function (2-3) with respect to $p^t_{j'|i'}$ and $q^t_{i'}$ are given by:

$$\frac{\partial l(\theta)}{\partial p^t_{j'|i'}} = \frac{\partial}{\partial p^t_{j'|i'}} \sum_i \sum_j n_{ij} \log\left(\sum_{u'} \sum_{v'} \pi_{i|u'} \xi_{j|v'} p^t_{v'|u'} q^t_{u'}\right)$$
$$= \sum_i \sum_j n_{ij} \frac{\pi_{i|i'} \xi_{j|j'} q^t_{i'}}{\sum_{u'} \sum_{v'} \pi_{i|u'} \xi_{j|v'} p^t_{v'|u'} q^t_{u'}} \qquad (2-4)$$

$$\frac{\partial l(\theta)}{\partial q^t_{i'}} = \frac{\partial}{\partial q^t_{i'}} \sum_i \sum_j n_{ij} \log\left(\sum_{u'} \sum_{v'} \pi_{i|u'} \xi_{j|v'} p^t_{v'|u'} q^t_{u'}\right)$$
$$= \sum_i \sum_j n_{ij} \frac{\sum_{v'} \pi_{i|i'} \xi_{j|v'} p^t_{v'|i'}}{\sum_{u'} \sum_{v'} \pi_{i|u'} \xi_{j|v'} p^t_{v'|u'} q^t_{u'}}. \qquad (2-5)$$

Here $\pi_{i|i'}$ is not zero when $i = i'$, and $\pi_{i|u'}$ is not zero when $i = u'$. Thus, equation (2-4) and equation (2-5) are not zero when $i = i' = u'$. In this case, equation (2-4) can be simplified as:

$$\frac{\partial l(\theta)}{\partial p^t_{j'|i'}} = \sum_j n_{ij} \frac{\xi_{j|j'} q^t_{i'}}{\sum_{v'} \xi_{j|v'} p^t_{v'|i'} q^t_{u'}}$$
$$= \sum_j n_{ij} \frac{\xi_{j|j'}}{\sum_{v'} \xi_{j|v'} p^t_{v'|i'}} \qquad (2-6)$$

Likewise, equation (2-5) becomes:

$$\frac{\partial l(\theta)}{\partial q^t_{i'}} = \sum_j n_{ij} \frac{\sum_{v'} \xi_{j|v'} p^t_{v'|i'}}{\sum_{v'} \xi_{j|v'} p^t_{v'|u'} q^t_{u'}} = \frac{\sum_j n_{ij}}{q^t_{i'}} \qquad (2-7)$$

Let $\sum_j n_{ij} = N_i$, equation (2-7) can be written as:

$$\frac{\partial l(\theta)}{\partial q^t_{i'}} = \frac{\sum_j n_{ij}}{q^t_{i'}} = \frac{N_i}{q^t_{i'}}$$

Then I take the second order derivatives of the log-likelihood function, I have:

(1) When $i' = 0, 1;\ u' = 0, 1;\ j' = 1, 2;\ v' = 1, 2$, the $(2i' + j') \times (2u' + v')$ element of $\frac{\partial^2 l(\theta)}{\partial\tilde{\theta}\partial\tilde{\theta}'}$ is:

$$\frac{\partial^2 l(\theta)}{\partial p^t_{v'|u'}\partial p^t_{j'|i'}} = \frac{\partial}{\partial p^t_{v'|u'}}\left(\sum_j n_{ij}\frac{\xi_{j|j'}}{\sum_{v'}\xi_{j|v'}p^t_{v'|i'}}\right) = -\sum_j n_{ij}\frac{\xi_{j|j'}\xi_{j|v'}}{(\sum_{v'}\xi_{j|v'}p^t_{v'|i'})^2}$$

Let

$$\eta_{ij} = Pr(X_j, Y_i) = \sum_{i'}\sum_{j'} Pr(X_j, Y_i, X^t_{j'}, Y^t_{i'}) = \sum_{i'}\sum_{j'}\pi_{i|i'}\xi_{j|j'}p^t_{j'|i'}q^t_{i'}$$

I have:

$$\eta_{01} = (\varepsilon - (2\varepsilon - 1)p^t_{1|0})q^t_0$$

$$\eta_{02} = (1 - \varepsilon + (2\varepsilon - 1)p^t_{1|0})q^t_0$$

$$\eta_{11} = (\varepsilon - (2\varepsilon - 1)p^t_{1|1})q^t_1$$

$$\eta_{12} = (1 - \varepsilon + (2\varepsilon - 1)p^t_{1|1})q^t_1$$

Thus,

$$E(n_{ij}) = NPr(X_j, Y_i) = N\eta_{ij}$$

Thus, the $(2i' + j') \times (2u' + v')$ element of $\tilde{I}(\theta)$ is:

$$E\left[-\frac{\partial^2 l(\theta)}{\partial p^t_{v'|u'}\partial p^t_{j'|i'}}\right] = E\left[\sum_j n_{ij}\frac{\xi_{j|j'}\xi_{j|v'}}{\left(\sum_{v'}\xi_{j|v'}p^t_{v'|i'}\right)^2}\right] = N\sum_j \eta_{ij}\frac{\xi_{j|j'}\xi_{j|v'}}{(\sum_{v'}\xi_{j|v'}p^t_{v'|i'})^2}$$

(2) For $i' = 0, 1;\ u' = 0, 1;\ j' = 1, 2;\ v' = 1, 2$, the $(2i' + j') \times (5 + u')$ element of $\frac{\partial^2 l(\theta)}{\partial\tilde{\theta}\partial\tilde{\theta}'}$ is

$$\frac{\partial^2 l(\theta)}{\partial q^t_{u'}\partial p^t_{j'|i'}} = 0$$

By symmetry, the $(5+u')\times(2i'+j')$ element of $\frac{\partial^2 l(\theta)}{\partial\tilde{\theta}\partial\tilde{\theta}'}$ is also 0.

Thus, the $(2i'+j')\times(5+u')$ and $(5+u')\times(2i'+j')$ element of $\tilde{I}(\theta)$ is

$$E[-\frac{\partial^2 l(\theta)}{\partial q_{u'}^t \partial p_{j'|i'}^t}] = 0$$

(3) For $i'=0,1;\ u'=0,1$; the $(5+u',5+i')$ of $\frac{\partial^2 l(\theta)}{\partial\tilde{\theta}\partial\tilde{\theta}'}$ is

$$\frac{\partial^2 l(\theta)}{\partial {q_{i'}^t}^2} = \frac{\partial}{\partial q_{i'}^t}\left(\frac{N_i}{q_{i'}^t}\right) = -\frac{N_i}{{q_{i'}^t}^2}$$

I have:

$$E(N_i) = Nq_{i'}^t$$

Thus, the $(5+u',5+i')$ element of $\tilde{I}(\theta)$ is

$$E[-\frac{\partial^2 l(\theta)}{\partial {q_{i'}^t}^2}] = \frac{N}{q_{i'}^t}$$

Finally, the information matrix $\tilde{I}(\theta) = \left(s_{ij}\right)_{6\times 6}$ can be written as:

$$\begin{bmatrix} s_{11} & s_{12} & 0 & 0 & 0 & 0 \\ s_{21} & s_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & s_{33} & s_{34} & 0 & 0 \\ 0 & 0 & s_{43} & s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{66} \end{bmatrix}$$

where

$$s_{11} = Nq_0^{t\,2}\left(\frac{\varepsilon^2}{\eta_{01}} + \frac{(1-\varepsilon)^2}{\eta_{02}}\right)$$

$$s_{12} = Nq_0^{t\,3}\ \frac{\varepsilon(1-\varepsilon)}{\eta_{01}\eta_{02}}$$

$$s_{21} = N{q_0^t}^3 \frac{\varepsilon(1-\varepsilon)}{\eta_{01}\eta_{02}}$$

$$s_{22} = N{q_0^t}^2 \left(\frac{(1-\varepsilon)^2}{\eta_{01}} + \frac{\varepsilon^2}{\eta_{02}}\right)$$

$$s_{33} = N{q_1^t}^2 \left(\frac{\varepsilon^2}{\eta_{11}} + \frac{(1-\varepsilon)^2}{\eta_{12}}\right)$$

$$s_{34} = N{q_1^t}^3 \frac{\varepsilon(1-\varepsilon)}{\eta_{11}\eta_{12}}$$

$$s_{43} = N{q_1^t}^3 \frac{\varepsilon(1-\varepsilon)}{\eta_{11}\eta_{12}}$$

$$s_{44} = N{q_1^t}^2 \left(\frac{(1-\varepsilon)^2}{\eta_{11}} + \frac{\varepsilon^2}{\eta_{12}}\right)$$

$$s_{55} = \frac{N}{q_0^t}$$

$$s_{66} = \frac{N}{q_1^t}$$

The 6×3 matrix $A = \frac{\partial \tilde{\theta}}{\partial \theta'}$ is given by:

$$A = \frac{\partial\left(\left(p_{1|0}^t, p_{2|0}^t, p_{1|1}^t, p_{2|1}^t, q_0^t, q_1^t\right)'\right)}{\partial\left(\left(\psi_1, p_{1|1}^t, q_0^t\right)'\right)} = \begin{pmatrix} \frac{\partial p_{1|0}^t}{\partial \psi_1} & \frac{\partial p_{1|0}^t}{\partial p_{1|1}^t} & \frac{\partial p_{1|0}^t}{\partial q_0^t} \\ \frac{\partial p_{2|0}^t}{\partial \psi_1} & \frac{\partial p_{2|0}^t}{\partial p_{1|1}^t} & \frac{\partial p_{2|0}^t}{\partial q_0^t} \\ \frac{\partial p_{1|1}^t}{\partial \psi_1} & \frac{\partial p_{1|1}^t}{\partial p_{1|1}^t} & \frac{\partial p_{1|1}^t}{\partial q_0^t} \\ \frac{\partial p_{2|1}^t}{\partial \psi_1} & \frac{\partial p_{2|1}^t}{\partial p_{1|1}^t} & \frac{\partial p_{2|1}^t}{\partial q_0^t} \\ \frac{\partial q_0^t}{\partial \psi_1} & \frac{\partial q_0^t}{\partial p_{1|1}^t} & \frac{\partial q_0^t}{\partial q_0^t} \\ \frac{\partial q_1^t}{\partial \psi_1} & \frac{\partial q_1^t}{\partial p_{1|1}^t} & \frac{\partial q_1^t}{\partial q_0^t} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

Let $\psi_1 = p_{1|0}^t - p_{1|1}^t$, and define θ as: $\theta = \left(\psi_1, p_{1|1}^t, q_0^t\right)' = (\psi, \lambda)'$, where $\psi = (\psi_1)$, $\lambda = (p_{1|1}^t, q_0^t)$. Then Fisher's information matrix of $l(\theta)$ is:

$$I(\theta) = E(-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'})$$

After applying Lemma 1 from Ji et al., I have:

$$I(\theta) = A'\tilde{I}(\theta)A = \begin{pmatrix} i_{11} & i_{12} & i_{13} \\ i_{21} & i_{22} & i_{23} \\ i_{31} & i_{32} & i_{33} \end{pmatrix}$$

where

$$i_{11} = N(2\varepsilon - 1)^2[{q_0^t}^2(\frac{1}{p_{01}} + \frac{1}{p_{02}}) + {q_1^t}^2(\frac{1}{p_{11}} + \frac{1}{p_{12}})]$$

$$i_{21} = N[{q_0^t}^2(2\varepsilon - 1)(\frac{1}{p_{01}} - \frac{1}{p_{02}}) - {q_1^t}^2(2\varepsilon - 1)^2(\frac{1}{p_{11}} + \frac{1}{p_{12}})$$

$$i_{13} = 0$$

$$i_{21} = N(2\varepsilon - 1)[{q_0^t}^2(\frac{1}{p_{01}} - \frac{1}{p_{02}}) - {q_1^t}^2(\frac{1}{p_{11}} - \frac{1}{p_{12}})]$$

$$i_{22} = N[{q_0^t}^2\left(\frac{1}{p_{01}} + \frac{1}{p_{02}}\right) + {q_1^t}^2(2\varepsilon - 1)(\frac{1}{p_{11}} - \frac{1}{p_{12}})]$$

$$i_{23} = 0$$

$$i_{31} = i_{32} = 0$$

$$i_{33} = N(\frac{1}{q_0^t} + \frac{1}{q_1^t})$$

The non-centrality parameter $\delta^2$is given by:

$$\delta^2 = (p^t_{1|0} - p^t_{1|1})' J(\theta)(p^t_{1|0} - p^t_{1|1})$$

where

$$J(\theta) = i_{11} - (i_{12} \quad i_{13}) \begin{pmatrix} i_{22} & i_{23} \\ i_{32} & i_{33} \end{pmatrix}^{-1} \begin{pmatrix} i_{21} \\ i_{31} \end{pmatrix}$$

This test statistic is asymptotically distributed as a non-central chi-square distribution with 1 degree of freedom and NCP δ^2 under the alternative hypothesis.

2-5 Results

2-5-1 NCP Comparison between Chi-square Test and Likelihood Ratio Test

In this section, the NCP (λ) of the chi-square test and the NCP (δ^2) of the likelihood ratio test are compared. The two NCPs are calculated under different parameter settings, and the corresponding powers are calculated for significance level 0.05. The absolute difference in NCPs and the corresponding powers, defined as $|\lambda - \delta^2|$ and $|power_{\lambda} - power_{\delta^2}|$, are also calculated. Here, the sample size is 200 and the phenotype frequency is 0.5. Three sets of case and control allele frequencies are studied, i.e. (1) 0.005, 0.01 (2) 0.01, 0.05 (3) 0.005, 0.05. The error rate ranges from 0 to 0.025, in increments of 0.005. The results are shown in Table 2.3.

From Table 2.3, the two NCP calculations are consistent with each other.

Table 2.3 NCP of Chi-square Test and Likelihood Ratio Test

			NCP			Power		
$p_{1\|0}^t$	$p_{1\|1}^t$	e	λ	δ^2	$\|\lambda - \delta^2\|$	$power_\lambda$	power_{δ^2}	$\|power_\lambda - \text{power}_{\delta^2}\|$
0.005	0.01	0	0.336	0.337	0.001	0.09	0.09	0.00
		0.005	0.2	0.2	0	0.07	0.07	0.00
		0.01	0.141	0.141	0	0.07	0.07	0.00
		0.015	0.108	0.108	0	0.06	0.06	0.00
		0.02	0.087	0.087	0	0.06	0.06	0.00
		0.025	0.073	0.073	0	0.06	0.06	0.00
0.01	0.05	0	5.498	5.605	0.107	0.65	0.66	0.01
		0.005	4.682	4.759	0.077	0.58	0.59	0.01
		0.01	4.06	4.118	0.058	0.52	0.53	0.01
		0.015	3.571	3.616	0.045	0.47	0.48	0.00
		0.02	3.177	3.212	0.035	0.43	0.43	0.00
		0.025	2.852	2.88	0.028	0.39	0.40	0.00
0.005	0.05	0	7.572	7.735	0.163	0.79	0.79	0.01
		0.005	6.364	6.479	0.115	0.71	0.72	0.01
		0.01	5.465	5.55	0.085	0.65	0.65	0.01
		0.015	4.771	4.835	0.064	0.59	0.59	0.01
		0.02	4.218	4.268	0.05	0.54	0.54	0.00
		0.025	3.767	3.808	0.041	0.49	0.50	0.00

Notations: λ = NCP from Chi-square test, δ^2 = NCP from likelihood ratio test.

Notes: Sample size = 200, phenotype frequency = 0.5, significance level = 0.005.

2-5-2 Simulation

In Section 2-2, I derived the NCP from the log-likelihood ratio test in the presence of allele error. In this section, I conducted simulation studies to evaluate the accuracy of the asymptotic power of the likelihood ratio test allowing for allele errors, and investigated how the power would change as the parameter settings changed.

The total sample size was set to 10,000. The true phenotype frequency of controls was 0.5. Let $r = p_{1|1}^t / p_{1|0}^t$ denote the ratio of case allele frequency to control allele frequency. Here the values of r are 2, 3 and 4 respectively. Case and control allele frequencies were set as shown in Table 2.4.

Table 2.4 Allele frequency settings in simulation study

	$p_{1\|0}^t$	$p_{1\|1}^t$	r
Setting 1	0.005	0.0075	1.5
Setting 2	0.005	0.01	2
Setting 3	0.005	0.015	3
Setting 4	0.005	0.02	4
Setting 5	0.01	0.015	1.5
Setting 6	0.01	0.02	2
Setting 7	0.01	0.03	3
Setting 8	0.02	0.03	1.5

Notations: $p_{1|0}^t$= control at risk allele frequency. $p_{1|1}^t$= case at risk allele frequency. $r = p_{1|1}^t / p_{1|0}^t$= ratio of case allele frequency to control allele frequency.

The error rate ranges over the interval [0.005, 0.025] in increments 0.005. Thus, there were 30 different configurations of parameters settings considered, with 1000 replicates for each parameter setting.

To run each simulation, I first specified the allele frequencies for cases and controls. I calculated the true genotype frequencies assuming Hardy-Weinberg

equilibrium. Then I generated case and control data based on the true phenotype and genotype frequencies. I then introduced genotype errors into cases and controls using the GLHO error matrix as presented in Table 2.2. Then I counted all alleles based on the observed genotypes, and performed the 2×2 chi-square allelic test. The simulation power was was Power_{sim} and was defined to be the proportion of replicates for a given set of parameter specifications whose test statistic exceeded the critical value using the asymptotic null distribution. That is

$$\text{Power}_{sim} = \frac{\text{number of significant replicates}}{\text{number of total replicates}}$$

Table 2.5 presents the results of the simulation power and asymptotic power at significance level 0.05 under various parameter settings. The asymptotic power was Power_{δ^2} and was calculated from the likelihood ratio test given in Section 2-2-2. I also reported the standard error and 95% confidence intervals of the simulation power.

Table 2.5 Comparison of asymptotic and Simulation Power

Setting	$p^t_{1\|0}$	$p^t_{1\|1}$	r	e	Power		RE	SE	95% CI
					$power_{\delta^2}$	$power_{sim}$			
1	0.005	0.0075	1.5	0.005	0.38	0.35	0.00	0.02	0.34, 0.39
				0.01	0.28	0.25	0.00	0.01	0.24, 0.28
				0.015	0.22	0.23	0.00	0.00	0.21, 0.24
				0.02	0.19	0.19	0.00	0.00	0.18, 0.20
				0.025	0.16	0.16	0.00	0.00	0.13, 0.19
2	0.005	0.01	2	0.005	0.89	0.89	0.00	0.01	0.87, 0.91
				0.01	0.76	0.76	0.00	0.01	0.73, 0.79
				0.015	0.64	0.62	0.03	0.02	0.59, 0.65
				0.02	0.55	0.51	0.08	0.02	0.48, 0.54
				0.025	0.48	0.47	0.01	0.02	0.44,0.50
3	0.005	0.015	3	0.005	1.00	1.00	0.00	0.00	1.00,1.00
				0.01	1.00	1.00	0.00	0.00	1.00, 1.00
				0.015	0.99	0.99	0.00	0.00	0.99, 1.00
				0.02	0.98	0.98	0.00	0.00	0.97, 0.99
				0.025	0.96	0.95	0.00	0.01	0.94, 0.97
4	0.005	0.02	4	0.005	1.00	1.00	0.00	0.00	1.00,1.00
				0.01	1.00	1.00	0.00	0.00	1.00, 1.00
				0.015	1.00	1.00	0.00	0.00	1.00, 1.00
				0.02	1.00	1.00	0.00	0.00	1.00, 1.00
				0.025	1.00	1.00	0.00	0.00	1.00, 1.00

Notation: $p^t_{1|0}$ = true at risk allele frequency in controls. $p^t_{1|1}$ = true at risk allele frequency in cases. r = at risk allele frequency ratio of cases to controls. e = error rate. $RE = \frac{\left|\text{Power}_{sim} - \text{Power}_{\delta^2}\right|}{\text{Power}_{\delta^2}}$ = relative error between the asymptotic and simulation power.

SE: standard error. 95%CI: 95% confidence interval of the simulated power.

Note: Sample size is 10,000. Number of replicates is 1000. Phenotype frequency is 0.5. Significance level is 0.05.

Table 2.5 Comparison of asymptotic and Simulation Power (Continued)

Setting	$p_{1\|0}^t$	$p_{1\|1}^t$	r	e	Power		RE	SE	95% CI
					$power_{\delta^2}$	$power_{sim}$			
5	0.01	0.015	1.5	0.005	0.76	0.76	0.00	0.01	0.73, 0.79
				0.01	0.65	0.65	0.00	0.02	0.62, 0.68
				0.015	0.56	0.54	0.03	0.02	0.51, 0.57
				0.02	0.49	0.45	0.08	0.02	0.42, 0.48
				0.025	0.43	0.42	0.03	0.02	0.39, 0.45
6	0.01	0.02	2	0.005	1.00	1.00	0.00	0.00	1.00, 1.00
				0.01	1.00	1.00	0.00	0.00	1.00, 1.00
				0.015	1.00	1.00	0.00	0.00	1.00, 1.00
				0.02	1.00	1.00	0.00	0.00	1.00, 1.00
				0.025	1.00	1.00	0.00	0.00	1.00, 1.00
7	0.01	0.03	3	0.005	1.00	1.00	0.00	0.00	1.00, 1.00
				0.01	1.00	1.00	0.00	0.00	1.00, 1.00
				0.015	1.00	1.00	0.00	0.00	1.00, 1.00
				0.02	1.00	1.00	0.00	0.00	1.00, 1.00
				0.025	1.00	1.00	0.00	0.00	1.00, 1.00
8	0.02	0.03	1.5	0.005	0.99	0.98	0.01	0.00	0.98, 0.99
				0.01	0.97	0.96	0.01	0.00	0.96, 0.98
				0.015	0.94	0.93	0.01	0.00	0.93, 0.94
				0.02	0.91	0.92	0.00	0.00	0.90, 0.92
				0.025	0.88	0.87	0.01	0.01	0.87, 0.88

Notation: $p_{1|0}^t$ = true at risk allele frequency in controls. $p_{1|1}^t$ = true at risk allele frequency in cases. r = at risk allele frequency ratio of cases to controls. e = error rate. $RE = \frac{\left|\text{Power}_{sim} - \text{Power}_{\delta^2}\right|}{\text{Power}_{\delta^2}}$ = relative error between the asymptotic and simulation power. SE= standard error. 95%CI= 95% confidence interval of the simulated power.

Note: Sample size is 10,000. Number of replicates is 1000. Phenotype frequency is 0.5. Significance level is 0.05.

The comparison of the simulation and asymptotic power under different parameter settings is given in Table 2.5. The maximum relative error for the 5% significance level was 0.08 in two situations: (1) when the at risk allele frequency in the control group is 0.005, the at risk allele frequency in the case group is 0.01, and the error rate is 0.02. (2) when the at risk allele frequency in the control group is 0.01, the at risk allele frequency in case group is 0.015, and the error rate is 0.02. Whenever the asymptotic power exceeds 0.67, the simulation power is close to the asymptotic power. When the asymptotic power is less than 0.67, minor differences may occur. In general, the simulation power was slightly lower than the asymptotic power, and the aymptotic powers are in agreement with simulation powers under each configuration of the parameter settings.

As expected, this result is consistent with the finding (Mote and Anderson[22]; Gordon et al.[12,13]) that the power from the likelihood ratio test decreases as the error rates increases when the level of significance remains the same. For a fixed value of the control allele frequency, the asymptotic power increases as the ratio of case to control at risk allele frequency increases. For a fixed ratio of case to control at risk allele frequency, the asymptotic power increases as the control allele frequency increases. When the control allele frequency is 0.005 and the ratio between case and control allele frequency was 4, the power is always 1, so that genotype error rate and the allele frequency ratio were not significant factors in determining the power. When control allele frequency was 0.01, and the case/control allele frequency ratio was 2, power was always 1. At these settings, the error rate and case/control allele frequency ratio had no significant effect on the power.

Table 2.6 shows the asymptotic power for various sample sizes under the six parameter settings at significance level 0.05. Sample sizes are: 1000, 2000, 5000, 10000 and 20000. Power rate is defined as the ratio between power with error and power without error. The Phenotype frequency is 0.5. The error rate ranges from 0 to 0.025, in increment of 0.005.

The asymptotic power has similar pattern under different parameter settings. For all parameter settings, the power when there is no misclassification error is the upper bound for the actual power. When the sample size is small (i.e.1000), power is highly sensitive to the error rate. Even a small error rate will bring down the power significantly. When the sample size is 5000 or more, the error rate is not so important.

Table 2.6 Asymptotic power as a function of sample size

$p_{1\|0}^{t}$	r	e	Asymptotic power					Power ratio				
			Sample size					Sample size				
			1000	2500	5000	10000	20000	1000	2500	5000	10000	20000
0.005	1.5	0	0.109	0.202	0.355	0.612	0.888	1.00	1.00	1.00	1.00	1.00
		0.005	0.082	0.132	0.218	0.384	0.653	0.75	0.65	0.61	0.63	0.74
		0.01	0.072	0.106	0.163	0.280	0.494	0.66	0.52	0.46	0.46	0.56
		0.015	0.066	0.092	0.135	0.223	0.393	0.61	0.45	0.38	0.36	0.44
		0.02	0.063	0.083	0.117	0.187	0.326	0.58	0.41	0.33	0.31	0.37
		0.025	0.061	0.077	0.105	0.163	0.279	0.56	0.38	0.30	0.27	0.31
0.005	2	0	0.254	0.536	0.827	0.984	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.170	0.353	0.609	0.885	0.994	0.67	0.66	0.74	0.90	0.99
		0.01	0.134	0.264	0.467	0.756	0.964	0.53	0.49	0.56	0.77	0.96
		0.015	0.114	0.213	0.376	0.642	0.908	0.45	0.40	0.45	0.65	0.91
		0.02	0.101	0.181	0.314	0.551	0.839	0.40	0.34	0.38	0.56	0.84
		0.025	0.093	0.159	0.271	0.478	0.769	0.37	0.30	0.33	0.49	0.77
0.005	3	0	0.615	0.945	0.999	1.000	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.448	0.825	0.984	1.000	1.000	0.73	0.87	0.98	1.00	1.00
		0.01	0.350	0.702	0.941	0.999	1.000	0.57	0.74	0.94	1.00	1.00
		0.015	0.288	0.599	0.879	0.993	1.000	0.47	0.63	0.88	0.99	1.00
		0.02	0.245	0.518	0.809	0.980	1.000	0.40	0.55	0.81	0.98	1.00
		0.025	0.214	0.453	0.740	0.958	0.999	0.35	0.48	0.74	0.96	1.00
0.005	4	0	0.858	0.998	1.000	1.000	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.722	0.981	1.000	1.000	1.000	0.84	0.98	1.00	1.00	1.00
		0.01	0.608	0.942	0.999	1.000	1.000	0.71	0.94	1.00	1.00	1.00
		0.015	0.518	0.887	0.994	1.000	1.000	0.60	0.89	0.99	1.00	1.00
		0.02	0.449	0.826	0.984	1.000	1.000	0.52	0.83	0.98	1.00	1.00
		0.025	0.395	0.763	0.966	1.000	1.000	0.46	0.76	0.97	1.00	1.00

Notation: $p_{1|0}^{t}$ = control at risk allele frequency. r = case at risk allele frequency/ control allele frequency.
Note: Power ratio = Power with error / Power without error.

Table 2.6 Asymptotic power as a function of sample size (Continued)

$p_{1\|0}^t$	r	e	Asymptotic power					Power ratio				
			Sample size					Sample size				
			1000	2500	5000	10000	20000	1000	2500	5000	10000	20000
0.01	1.5	0	0.172	0.357	0.615	0.890	0.995	1.00	1.00	1.00	1.00	1.00
		0.005	0.136	0.268	0.474	0.765	0.967	0.79	0.75	0.77	0.86	0.97
		0.01	0.115	0.217	0.383	0.652	0.914	0.67	0.61	0.62	0.73	0.92
		0.015	0.102	0.184	0.321	0.561	0.848	0.59	0.52	0.52	0.63	0.85
		0.02	0.094	0.162	0.276	0.488	0.779	0.55	0.45	0.45	0.55	0.78
		0.025	0.087	0.145	0.243	0.430	0.713	0.51	0.41	0.40	0.48	0.72
0.01	2	0	0.454	0.830	0.985	1.000	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.356	0.710	0.945	0.999	1.000	0.78	0.86	0.96	1.00	1.00
		0.01	0.293	0.609	0.885	0.994	1.000	0.65	0.73	0.90	0.99	1.00
		0.015	0.250	0.527	0.818	0.982	1.000	0.55	0.63	0.83	0.98	1.00
		0.02	0.218	0.462	0.751	0.962	1.000	0.48	0.56	0.76	0.96	1.00
		0.025	0.195	0.410	0.687	0.933	0.998	0.43	0.49	0.70	0.93	1.00
0.01	3	0	0.894	0.999	1.000	1.000	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.815	0.995	1.000	1.000	1.000	0.91	1.00	1.00	1.00	1.00
		0.01	0.737	0.984	1.000	1.000	1.000	0.82	0.98	1.00	1.00	1.00
		0.015	0.665	0.965	1.000	1.000	1.000	0.74	0.97	1.00	1.00	1.00
		0.02	0.602	0.939	0.999	1.000	1.000	0.67	0.94	1.00	1.00	1.00
		0.025	0.546	0.907	0.996	1.000	1.000	0.61	0.91	1.00	1.00	1.00
0.01	4	0	0.991	1.000	1.000	1.000	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.976	1.000	1.000	1.000	1.000	0.98	1.00	1.00	1.00	1.00
		0.01	0.952	1.000	1.000	1.000	1.000	0.96	1.00	1.00	1.00	1.00
		0.015	0.920	1.000	1.000	1.000	1.000	0.93	1.00	1.00	1.00	1.00
		0.02	0.883	0.999	1.000	1.000	1.000	0.89	1.00	1.00	1.00	1.00
		0.025	0.844	0.997	1.000	1.000	1.000	0.85	1.00	1.00	1.00	1.00
0.02	1.5	0	0.300	0.622	0.894	0.995	1.000	1.00	1.00	1.00	1.00	1.00
		0.005	0.257	0.542	0.831	0.985	1.000	0.86	0.87	0.93	0.99	1.00
		0.01	0.225	0.477	0.767	0.967	1.000	0.75	0.77	0.86	0.97	1.00
		0.015	0.201	0.424	0.706	0.943	0.999	0.67	0.68	0.79	0.95	1.00
		0.02	0.182	0.381	0.649	0.912	0.997	0.61	0.61	0.73	0.92	1.00
		0.025	0.167	0.345	0.598	0.878	0.993	0.56	0.56	0.67	0.88	0.99

Notation: $p_{1|0}^t$ = control at risk allele frequency. r = case at risk allele frequency/ control at risk allele frequency.

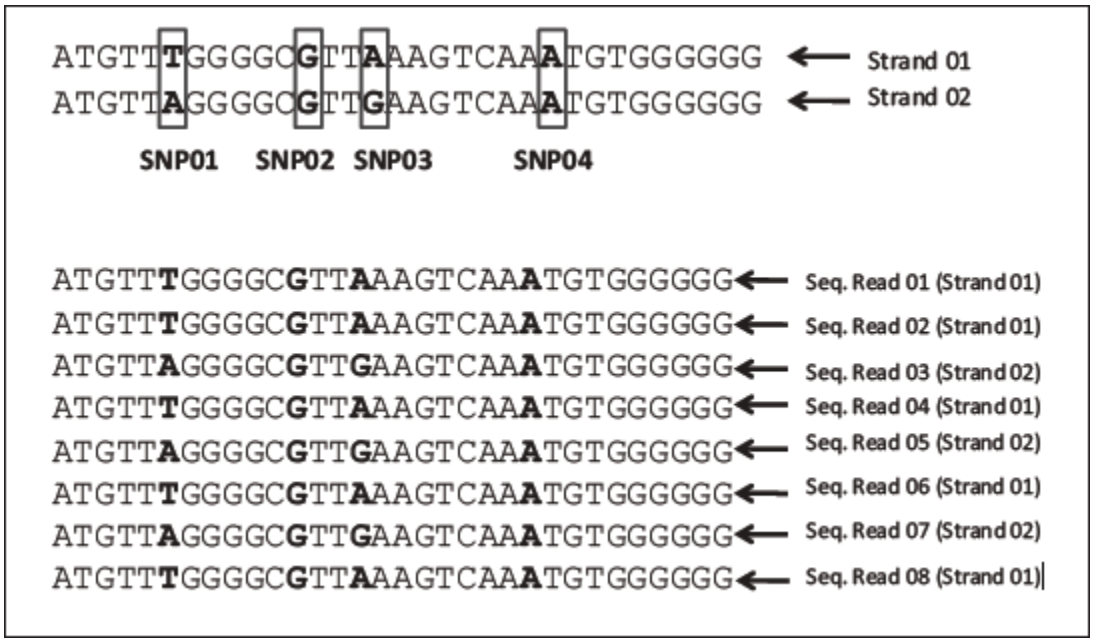
Note: Power ratio = Power with error / Power without error.

Chapter 3 Association Test using NGS data

NGS studies can identify multiple causal variants of a disease that may not be apparent from data generated by SNP-chip technology. One of the challenges of NGS is misclassification error. I consider a diploid gene in this section; that is, a gene that contains two alleles, one from the mother and the other from the father.

Figure 3.1 from Gordon et al.[7] is an example of the number of observed sequence-read counts for a single individual m at four SNP loci. The top panel shows a stretch of the two strands of the DNA sequence for this individual. In this example, they are labeled as 'strand 01' and 'strand 02'. Four SNP loci are illustrated. This individual is heterozygous at SNP01 with genotype T/A, homozygous at SNP02 with genotype G/G, heterozygous at SNP03 with genotype A/G, and homozygous at SNP04 with genotype A/A. The bottom panel shows the sequence reads consisting of random selections of one of the two strands. From the figure, the total count of reads is 8. 'Strand 01' is selected 5 times, and 'strand 02' is selected 3 times. We assume that allele A is the at-risk allele for these four SNPs. The observed numbers of allele A at the four loci is 3, 0, 5, and 8 respectively.

Figure 3.1 Example representation of sequence reads for an individual



Gordon et al.[8] provided a new test statistic for association testing in NGS studies that incorporates non-differential misclassification. These authors found that at very low error rates, misclassifying a common homozygote as a heterozygote causes loss of power to detect association, and the power loss increases as the minor allele frequency decreases. This is consistent with findings by Kang et al.[15-17]. Kim et al.[18] extended the use of NGS data to the linear trend test (LTT) developed by Cochran[4] and Armitage[1], and focused on slightly more common causal variants. In this section, I apply the approach in Kim et al.[18] to the work of Gordon et al.[8] That is, instead of using the probability of true sequence-read counts for an individual in the likelihood function, I use the probability of the observed sequence read counts. This statistic allows for differential misclassification errors in base pair reads and tests whether the true genotype frequencies differ between cases and controls. I estimate the probability of the genotype frequencies and misclassification error rates based on the observed base pair reads

using the expectation-maximization (EM) algorithm and report the power of this test for the bi-allelic case.

3-1 Likelihood Ratio Test allowing for Differential Misclassification using NGS Data

The following notation is from Gordon et al.[8]:

The index i denotes the phenotype for an individual: $i = 0$ for control group and $i = 1$ for case group. I code the two alleles at a base pair position as 1 and 2, where allele "1" is the at-risk allele, and "2" is the not at-risk allele. The index j ($j = 0,1,2$) denotes the genotype for an individual, where j is the number of 1 alleles. The superscript t denotes that the status of an individual is true. For example, $X_j^{t(m)}$ denotes the event that the true genotype of individual m is j, and $Y_i^{t(m)}$ denotes the event that the true affection status of individual m is i. Let p_{ij}^t denote the true genotype frequency of genotype j in phenotype i; that is, $p_{ij}^t = \Pr(X_j^t | Y_i^t)$.

The null hypothesis assumes that the true genotype frequencies are equal in cases and controls. Let p_j^t denote the true population frequency of genotype j under H_0. We have: $p_j^t = \Pr(X_j^t)$. Thus, the null hypothesis can be expressed as:

$$H_0: p_{0j}^t = p_{1j}^t = p_j^t,\ j = 0,1,2.$$

The alternative hypothesis (H_1) that the true genotype frequencies in cases and controls are unequal can be written as:

$$\mathrm{H}_1: p_{0j}^t \neq p_{1j}^t, \text{ for some } j, j = 0, 1, 2.$$

Let N denote the total number of individuals that are sequenced. I specify that the first N_0 subjects are in the control group and that the remaining N_1 subjects are in the case group so that: $N = N_0 + N_1$.

Let $V^{(m)}$ denote the total count of reads for individual m at the base pair position. I assume the number of reads is the same for each individual and call it "coverage" in this study. Let $n_1(m)$ and $n_2(m)$ denote the observed number of 1 and 2 alleles recorded for individual m respectively. We have: $n_1(m) + n_2(m) = V(m)$. Here $n_1(m)$and $n_2(m)$ are random variables that incorporate the misclassification errors of individual sequence reads. In the example from Figure 3, $n_1(m) = 3$ at SNP01, $n_1(m) = 0$ at SNP 02, $n_1(m) = 5$ at SNP03 and $n_1(m) = 8$ at SNP04.

Let $\varepsilon_{(\mathrm{v_1 v_2})\mathrm{i}}$ denote the probability that allele v_1 is misclassified as allele v_2, $\mathrm{v}_1, \mathrm{v}_2$ = 1, 2. The phenotype class is i, where $\mathrm{i} = 0$ refers to control group and $\mathrm{i} = 1$ refers to case group. The statistic proposed here allows for differential but symmetric misclassification. I define symmetric misclassification that the misclassification rates are the same in both directions. Let ε_0 denote the probability of misclassification in controls, and let ε_1 denote the probability of misclassification in cases. Symmetric misclassification is: $\varepsilon_{(12)0} = \varepsilon_{(21)0} = \varepsilon_0, \varepsilon_{(12)1} = \varepsilon_{(21)1} = \varepsilon_1$. I define differential misclassification as the misclassification probabilities are unequal in cases and controls, that is $\varepsilon_0 \neq \varepsilon_1$.

For individual m, let $A_{n_1(m)}^{V(m)}$ denote the event that the number of at-risk alleles recorded at a single base pair is n_1 with coverage $V(m)$. Let $Y_i^{t(m)}$ denote that the true affection status is i. Let $X_j^{t(m)}$ denote the true genotype is j. Kim et al.[18] found that the probability of event $A_{n_1(m)}^{V(m)}$ given the event $(X_j^{t(m)}, Y_i^{t(m)})$ is:

$$\Pr\left(A_{n_1(m)}^{V(m)} | X_j^{t(m)}, Y_i^{t(m)}\right) = \binom{V(m)}{n_{1(m)}} {\mu_{ij}}^{n_{1(m)}} \left(1-\mu_{ij}\right)^{V(m)-n_1(m)} \quad (7)$$

Here μ_{ij} denotes the probability of recording the at-risk allele ("1") given that the affection status is i and the genotype is j. For individual m, when the true genotype is "22"($j = 0$), every observed allele "1" is a "2" that has been misclassified. Thus, $\mu_{i0} = \varepsilon_{(21)i}$. When the true genotype is "11"($j = 2$), every allele recorded as "1" is correctly read. Thus $\mu_{i2} = 1 - \varepsilon_{(12)i}$. When the true genotype is "12"($j = 1$), then:

$$\begin{aligned}
&\mu_{i1} \\
&= \Pr(\text{allele 2 is recorded as allele 1} | \text{correct allele is 2}) \Pr(\text{correct allele is 2}) \\
&+ \Pr(\text{allele 1 is correctly recorded as allele 1} | \text{correct allele is 1}) \Pr(\text{correct allele is 1}) \\
&= \varepsilon_{(21)i} \times \frac{1}{2} + \left(1 - \varepsilon_{(12)i}\right) \times \frac{1}{2}
\end{aligned}$$

In general, μ_{ij} can be written as:

$$\mu_{ij} = \frac{2-j}{2} \varepsilon_{(21)i} + \frac{j}{2} \left(1 - \varepsilon_{(12)i}\right)$$

Under the assumption of symmetric but differential misclassification, μ_{ij} is:

$$\mu_{ij} = \frac{2-j}{2}\varepsilon_i + \frac{j}{2}(1-\varepsilon_i) = \begin{matrix} \varepsilon_i & j=0 \\ \frac{1}{2} & j=1 \\ 1-\varepsilon_i & j=2 \end{matrix} \qquad (3-2)$$

Let $q_i^t = Pr(Y_i^t)$ denote the true sampling frequency of phenotype i. Here, I assume that the phenotype is measured without error. Then,

$$q_0^t = \frac{N_0}{N}, q_1^t = \frac{N_1}{N}$$

Let $\left(\mathrm{A}_{\mathrm{n_1(m)}}^{\mathrm{V(m)}}, Y_i^{t(m)}\right)$ be the observed data. That is, the observed number of alleles of reads, the true genotype, and the true affection status. In general, the log-likelihood of the observed data under hypothesis H_u ($u = 0$ for null hypothesis, $u = 1$ for alternative hypothesis) is:

$$\begin{aligned} l_u(\theta) = \ln(\mathrm{L_u}) &= \sum_{m=1}^{N}\sum_{i=0}^{1} \mathrm{I}\left(\mathrm{Y_i}^{t(m)}\right) \ln\left(\mathrm{P_u}\left(\mathrm{A}_{\mathrm{n_1(m)}}^{\mathrm{V(m)}}, Y_i^{t(m)}\right)\right) \\ &= \sum_{m=1}^{N}\sum_{i=0}^{1} \mathrm{I}\left(\mathrm{Y_i}^{t(m)}\right) \ln\left(\sum_{j=0}^{2} \mathrm{P_u}\left(\mathrm{A}_{\mathrm{n_1(m)}}^{\mathrm{V(m)}}, X_j^{t(m)}, Y_i^{t(m)}\right)\right) \\ &= \sum_{\mathrm{m=1}}^{\mathrm{N_0}} \ln\left(\sum_{\mathrm{j=0}}^{2} \mathrm{P_u}\left(\mathrm{A}_{\mathrm{n_1(m)}}^{\mathrm{V(m)}}, \mathrm{X_j^{t(m)}}, \mathrm{Y_i^{t(m)}}\right)\right) \\ &+ \sum_{\mathrm{m=N_0+1}}^{\mathrm{N}} \ln\left(\sum_{\mathrm{j=0}}^{2} \mathrm{P_u}\left(\mathrm{A}_{\mathrm{n_1(m)}}^{\mathrm{V(m)}}, \mathrm{X_j^{t(m)}}, \mathrm{Y_i^{t(m)}}\right)\right) \qquad (3-3) \end{aligned}$$

Here, $I()$ denotes the indicator function so that $I(Y_i^{t(m)})$ is 1 when the true affection status is i for individual m and is 0 otherwise.

From the chain rule for conditional probabilities, I have:

$$P_u\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right) = P_u\left(A_{n_1(m)}^{V(m)} | X_j^{t(m)}, Y_i^{t(m)}\right) P_u\left(X_j^{t(m)}, Y_i^{t(m)}\right)$$

$$= P_{u,c}\left(A_{n_1(m)}^{V(m)} | X_j^{t(m)}, Y_i^{t(m)}\right) P_u\left(X_j^{t(m)} | Y_i^{t(m)}\right) P_u\left(Y_i^{t(m)}\right)$$

$$= \binom{V(m)}{n_{1(m)}} {\mu_{ij}}^{n_{1(m)}} \left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{ij}^t q_i^t \qquad (3-4)$$

Thus, from equation (3-3), the log-likelihood of the observed data under the alternative hypothesis is:

$$l_1(\theta) = \ln(\mathrm{L}_1) = \sum_{m=1}^{N_0} \ln\left(\sum_{j=0}^{2}\left(\binom{V(m)}{n_1(m)} {\mu_{0j}}^{n_1(m)} \left(1-\mu_{0j}\right)^{\left(V(m)-n_1(m)\right)} p_{0j}^t q_0^t\right)\right)$$

$$+ \sum_{m=N_0+1}^{N} \ln\left(\sum_{j=0}^{2}\left(\binom{V(m)}{n_1(m)} {\mu_{1j}}^{n_1(m)} \left(1-\mu_{1j}\right)^{\left(V(m)-n_1(m)\right)} p_{1j}^t q_1^t\right)\right) (3-5\mathrm{a})$$

And the log-likelihood of the observed data under the null hypothesis is:

$$l_0(\theta) = \ln(\mathrm{L}_0) = \sum_{m=1}^{N_0} \ln\left(\sum_{j=0}^{2}\left(\binom{V(m)}{n_1(m)} {\mu_{0j}}^{n_1(m)} \left(1-\mu_{0j}\right)^{\left(V(m)-n_1(m)\right)} p_j^t q_0^t\right)\right)$$

$$+ \sum_{m=N_0+1}^{N} \ln\left(\sum_{j=0}^{2}\left(\binom{V(m)}{n_1(m)} {\mu_{1j}}^{n_1(m)} \left(1-\mu_{1j}\right)^{\left(V(m)-n_1(m)\right)} p_j^t q_1^t\right)\right) (3-5b)$$

3-2 Test Statistic using NGS data from Likelihood Ratio Test Allowing for Misclassification

The test statistic using NGS data is:

$$LRT_{NGS} = 2[\ln(\widehat{L_1}) - \ln(\widehat{L_0})],$$

where $\ln(\widehat{L_1}) \geq \ln(\widehat{L_0})$. $\ln(\widehat{L_1})$ and $\ln(\widehat{L_0})$ are the maximum log-likelihood values under the alternative and null hypothesis respectively. There are three parameters p_{i0}, p_{i1}, p_{i2} in the test to estimate with the restriction $\sum_{j=0}^{2} p_{ij} = 1$. Thus, degrees of freedom is $3 - 1 = 2$. Under the null hypothesis, this test statistic follows a central chi-square distribution. The maximum log-likelihood for each hypothesis is determined from the Expectation Maximization (EM) algorithm as described in the following section.

3-3 The Expectation Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm[6] is used to find the maximum likelihood estimates of a statistical model that contains latent variables where the equations cannot be solved directly. The EM algorithm consists of the iteration of two steps: the Expectation (E) step and the Maximization (M) step. The E step calculates the expected value of the log-likelihood function with respect to the conditional distribution of the latent data. The M step finds the parameters that maximize the expectation of the complete log-likelihood. These two steps are applied iteratively until the difference between the successive log-likelihood values is less than a specified value.

Let $()^{(r)}$ denote the r^{th} step estimate of a parameter. For example, $\tau_{m,i,j}^{(r)}$ and $p_{ij}^{t(r)}$ denote the r^{th} step estimates of the parameters $\tau_{m,i,j}$ and p_{ij}^{t}, respectively. These values are updated in each iteration of the EM-algorithm.

3-3-1 EM Algorithm for Obtaining the Maximum Likelihood Estimates (MLEs) under Alternative Hypothesis

Expectation (E) step:

Define the complete data to be the observed data and unobserved data. That is, the observed number of alleles of reads, the true genotype, and the true affection status.

The log-likelihood of the complete data under the hypothesis H_u is:

$$\ln(\mathrm{L_{u,c}}) = \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1}\left[I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right) \times \ln\left(\mathrm{P_u}\left(\mathrm{A}_{\mathrm{n_1(m)}}^{\mathrm{V(m)}}, X_j^{t(m)}, Y_i^{t(m)}\right)\right)\right] \qquad (3-6)$$

Here, $I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right)$ is 1 when the true genotype is j is and the true affection status is i of individual m, $I\left(X_j^{t(m)}, Y_i^{t(m)}\right)$ is 0 otherwise.

Let Q_u denote the expected value of the log-likelihood of the complete data, conditional on the observed data under the hypothesis H_u, where $u = 0$ for null hypothesis, and $u = 1$ for alternative hypothesis. I have:

$$Q_u = E\left[\ln(\mathrm{L_{u,c}})\middle|\text{Observed Data}\right]$$

$$= \mathrm{E}\left[\left(\sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1}\left[I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right)\ln\left(P_u\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right)\right)\right]\right) \middle| \left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right]$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} E\left[I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right) \middle| \left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right] \times \ln\left(P_u\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right)\right)$$

From equation (9), Q_u is:

$$Q_u = \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} E\left[I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right)\middle|\left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right] \times \ln\left(\binom{V(m)}{n_{1(m)}}{\mu_{ij}}^{n_{1(m)}}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{ij}^t q_i^t\right) \quad (3-7)$$

Let $\tau_{i,j,H_u}^{(m)} = E\left[I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right)\middle|\left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right]$. Since the expected value of an indicator function is equal to the probability of the event, $\tau_{m,i,j,u}$ can be written as:

$$\tau_{i,j,H_u}^{(m)} = E\left[I_u\left(X_j^{t(m)}, Y_i^{t(m)}\right)\middle|\left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right] = \mathrm{P_u}\left(\left(X_j^{t(m)}, Y_i^{t(m)}\right)\middle|\left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right)$$

$$= \frac{\mathrm{P_u}\left(\mathrm{A_{n_1(m)}^{V(m)}}, \mathrm{X_j^{t(m)}}, \mathrm{Y_i^{t(m)}}\right)}{\mathrm{P_u}\left(\mathrm{A_{n_1(m)}^{V(m)}}, \mathrm{Y_i^{t(m)}}\right)} = \frac{\mathrm{P_u}\left(\mathrm{A_{n_1(m)}^{V(m)}}, \mathrm{X_j^{t(m)}}, \mathrm{Y_i^{t(m)}}\right)}{\sum_{k=0}^{2}\mathrm{P_u}\left(A_{n_1(m)}^{V(m)}, X_k^{t(m)}, Y_i^{t(m)}\right)}$$

From equation (3-4), I have:

$$\tau_{i,j,H_u}^{(m)} = \frac{\binom{\mathrm{V(m)}}{\mathrm{n_{1(m)}}}\mu_{\mathrm{ij}}^{\mathrm{n_{1(m)}}}\left(1-\mu_{ij}\right)^{\mathrm{V(m)-n_1(m)}}\mathrm{p_{ij}^t q_i^t}}{\sum_{\mathrm{k=0}}^{2}\left[\binom{\mathrm{V(m)}}{\mathrm{n_{1(m)}}}\mu_{\mathrm{ik}}^{n_1(m)}(1-\mu_{\mathrm{ik}})^{\mathrm{V(m)-n_1(m)}}\mathrm{p_{ik}^t q_i^t}\right]}$$

$$= \frac{\mu_{\mathrm{ij}}^{n_1(m)}\left(1-\mu_{ij}\right)^{\mathrm{V(m)-n_1(m)}}\mathrm{p_{ij}^t}}{\sum_{\mathrm{k=0}}^{2}[\mu_{\mathrm{ik}}^{n_1(m)}(1-\mu_{\mathrm{ik}})^{\mathrm{V(m)-n_1(m)}}\mathrm{p_{ik}^t}]} \quad (3-8)$$

The quantity $\tau_{i,j,H_u}^{(m)}$ denotes the Bayesian posterior probability (BPP) that individual *m* has true phenotype *i* and true genotype *j* , given the observed data $(A_{n_1(m)}^{V(m)}, Y_i^{t(m)})$, under the hypothesis H_u. I have $\tau_{i,0,H_u}^{(m)} + \tau_{i,1,H_u}^{(m)} + \tau_{i,2,H_u}^{(m)} = 1$, where $i = 0$ or 1, $j = 0,1,$ or 2 and $u = 0$ or 1.

Under the alternative hypothesis H_1, I have:

$$\tau_{i,j,H_1}^{(m)} = \frac{\mu_{ij}{}^{n_1(m)}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{ij}^t}{\sum_{k=0}^{2}\left[\mu_{ik}{}^{n_1(m)}(1-\mu_{ik})^{V(m)-n_1(m)} p_{ik}^t\right]} \qquad (3-9a)$$

Under the null hypothesis H_0, I have :

$$\tau_{i,j,H_0}^{(m)} = \frac{\mu_{ij}{}^{n_1(m)}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{ij}^t}{\sum_{k=0}^{2}\left[\mu_{ik}{}^{n_1(m)}(1-\mu_{ik})^{V(m)-n_1(m)} p_{ik}^t\right]}$$

$$= \frac{\mu_{ij}{}^{n_1(m)}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{j}^t}{\sum_{k=0}^{2}\left[\mu_{ik}{}^{n_1(m)}(1-\mu_{ik})^{V(m)-n_1(m)} p_{k}^t\right]} \qquad (3-9b)$$

The difference between equation (3-9 a) and equation (3-9 b) is that genotype frequencies in case group and control group are the same under the null hypothesis. i.e. $p_{0j}^t = p_{1j}^t = p_j^t$.

Therefore, from equation (3-7), under alternative hypothesis, I have:

$$Q_1 = E\left[\ln\left(L_{1,c}\right)\middle|\text{Observed Data}\right]$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_1}^{(m)} \times \ln\left[\binom{V(m)}{n_1(m)}\mu_{ij}{}^{n_1(m)}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{ij}^t q_i^t\right] \quad (3-10a)$$

Under null hypothesis, I have:

$$Q_0 = E\left[\ln\left(L_{0,c}\right)\middle|\text{Observed Data}\right]$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_0}^{(m)} \times \ln\left[\binom{V(m)}{n_1(m)}\mu_{ij}{}^{n_1(m)}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_{j}^t q_i^t\right] \quad (3-10b)$$

I discussed the E step above, and I now consider the M step of the EM algorithm. I use the superscript (r) to indicate the r-th step iteration.

From equation (3-10a), the r[th] step of Q_1 can be written as:

$$Q_1^{(r)} = \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_1}^{(m)(r-1)} \times \ln\left[\binom{V(m)}{n_1(m)}\left(\mu_{ij}^{(r)}\right)^{n_1(m)}\left(1-\mu_{ij}^{(r)}\right)^{(V-n_1(m))} p_{ij}^{(r)} q_i^t\right]$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_1}^{(m)(r-1)} \ln\left(p_{ij}^{(r)} q_i^t\right)$$

$$+ \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_1}^{(m)(r-1)} \ln\left(\binom{V}{n_1}\mu_{ij}^{(r-1)^{n_1}}\left(1-\mu_{ij}^{(r-1)}\right)^{V-n_1}\right)$$

$$= \sum_{m=1}^{N_0}\sum_{j=0}^{2} \tau_{0,j,H_1}^{(m)(r-1)}\left[\ln p_{0j}^{(r)} + \ln q_0^t\right] + \sum_{m=N_0+1}^{N}\sum_{j=0}^{2} \tau_{1,j,H_1}^{(m)(r-1)}\left[\ln p_{1j}^{(r)} + \ln q_1^t\right]$$

$$+ \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_1}^{(m)(r-1)} \ln\left(\binom{V(m)}{n_1(m)}\left(\mu_{ij}^{(r)}\right)^{n_1(m)}\left(1-\mu_{ij}^{(r)}\right)^{V-n_1(m)}\right) \quad (3-11)$$

where

$$\tau_{i,j,H_1}^{(m)(r-1)} = \frac{\left(\mu_{ij}^{(r-1)}\right)^{n_1(m)}\left(1-\mu_{ij}^{(r-1)}\right)^{V(m)-n_1(m)} p_{ij}^{(r-1)}}{\sum_{k=0}^{2}\left[\left(\mu_{ik}^{(r-1)}\right)^{n_1(m)}\left(1-\mu_{ik}^{(r-1)}\right)^{V(m)-n_1(m)} p_{ik}^{(r-1)}\right]}$$

Next I take the partial derivatives of equation (3-11) with respect to $p_{00}^{(r)}, p_{01}^{(r)}, p_{02}^{(r)}$, while holding $\tau_{0,j,H_1}^{(m)(r-1)}$ constant, then:

$$\frac{\partial Q_1^{(r)}}{\partial p_{00}^{(r)}} = \frac{\partial\left(\sum_{m=1}^{N_0}\sum_{j=0}^{2}\tau_{0,j,H_1}^{(m)(r-1)}\left[\ln p_{0j}^{(r)} + \ln q_0^{t}\right]\right)}{\partial p_{00}^{(r)}} = \sum_{m=1}^{N_0}\frac{\partial\left(\sum_{j=0}^{2}\tau_{0,j,H_1}^{(m)(r-1)}\left[\ln p_{0j}^{(r)} + \ln q_0^{t}\right]\right)}{\partial p_{00}^{(r)}}$$

$$= \sum_{m=1}^{N_0}\frac{\partial\left(\sum_{j=0}^{2}\tau_{0,j,H_1}^{(m)(r-1)}\ln p_{0j}^{(r)}\right)}{\partial p_{00}^{(r)}}$$

$$= \sum_{m=1}^{N_0}\frac{\partial\left(\tau_{0,0,H_1}^{(m)(r-1)}\ln p_{00}^{(r)} + \tau_{0,1,H_1}^{(m)(r-1)}\ln p_{01}^{(r)} + \tau_{0,2,H_1}^{(m)(r-1)}\ln p_{02}^{(r)}\right)}{\partial p_{00}^{(r)}}$$

$$= \sum_{m=1}^{N_0}\left[\frac{\partial\left(\tau_{0,0,H_1}^{(m)(r-1)}\ln p_{00}^{(r)}\right)}{\partial p_{00}^{(r)}} + \frac{\partial\left(\tau_{0,1,H_1}^{(m)(r-1)}\ln p_{01}^{(r)}\right)}{\partial p_{00}^{(r)}} + \frac{\partial\left(\tau_{0,2,H_1}^{(m)(r-1)}\ln p_{02}^{(r)}\right)}{\partial\left(1 - p_{01}^{(r)} - p_{02}^{(r)}\right)}\right]$$

$$= \sum_{m=1}^{N_0}\left[\frac{\tau_{0,0,H_1}^{(m)(r-1)}}{p_{00}^{(r)}} + 0 - \frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}}\right] = \sum_{m=1}^{N_0}\left[\frac{\tau_{0,0,H_1}^{(m)(r-1)}}{p_{00}^{(r)}} - \frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}}\right]$$

I set the partial derivative to be 0 to give:

$$\sum_{m=1}^{N_0}\left[\frac{\tau_{0,0,H_1}^{(m)(r-1)}}{p_{00}^{(r)}} - \frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}}\right] = 0 (3-12a)$$

Similarly from

$$\frac{\partial Q_1^{(r)}}{\partial p_{01}^{(r)}} = \frac{\partial\left(\sum_{m=1}^{N_0}\sum_{j=0}^{2}\tau_{0,j,H_1}^{(m)(r-1)}\left[\ln p_{0j}^{(r-1)} + \ln q_0^{t}\right]\right)}{\partial p_{01}^{(r)}},$$

I got:

$$\sum_{m=1}^{N_0}\left[\frac{\tau_{0,1,H_1}^{(m)(r-1)}}{p_{01}^{(r)}} - \frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}}\right] = 0\ (3-12b)$$

From

$$\frac{\partial Q_1^{(r)}}{\partial p_{02}^{(r)}} = \frac{\partial \left(\sum_{m=1}^{N_0} \sum_{j=0}^{2} \tau_{0,j,H_1}^{(m)(r-1)} \left[\ln p_{0j}^{(r)} + \ln q_0^t \right] \right)}{\partial p_{02}^{(r)}},$$

I got:

$$\sum_{m=1}^{N_0} \left[\frac{\tau_{0,0,H_1}^{(m)(r-1)}}{p_{00}^{(r)}} - \frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}} \right] = 0 \ (3-12c)$$

I rewrite equation (3-12a) :

$$\sum_{m=1}^{N_0} \left(\frac{\tau_{0,0,H_1}^{(m)(r-1)}}{p_{00}^{(r)}} - \frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}} \right) = \sum_{m=1}^{N_0} \left(\frac{\tau_{0,0,H_1}^{(m)(r-1)} \left(1 - p_{00}^{(r)} - p_{01}^{(r)} \right) - \tau_{0,2,H_1}^{(m)(r-1)} p_{00}^{(r)}}{p_{00}^{(r)} p_{02}^{(r)}} \right)$$

$$= \frac{1}{p_{00}^{(r)} p_{02}^{(r)}} \sum_{m=1}^{N_0} \left(\tau_{0,0,H_1}^{(m)(r-1)} \left(1 - p_{00}^{(r)} - p_{01}^{(r)} \right) - \tau_{0,2,H_1}^{(m)(r-1)} p_{00}^{(r)} \right)$$

$$= \frac{1}{p_{00}^{(r)} p_{02}^{(r)}} \sum_{m=1}^{N_0} \left(\tau_{0,0,H_1}^{(m)(r-1)} \left(1 - p_{01}^{(r)} \right) - \left(\tau_{0,0,H_1}^{(m)(r-1)} + \tau_{0,2,H_1}^{(m)(r-1)} \right) p_{00}^{(r)} \right)$$

$$= \frac{1}{p_{00}^{(r)} p_{02}^{(r)}} \sum_{m=1}^{N_0} \left(\tau_{0,0,H_1}^{(m)(r-1)} \left(1 - p_{01}^{(r)} \right) - \left(1 - \tau_{0,1,H_1}^{(m)(r-1)} \right) p_{00}^{(r)} \right) = 0.$$

That is:

$$\sum_{m=1}^{N_0} \left(\tau_{0,0,H_1}^{(m)(r-1)} \left(1 - p_{01}^{(r)} \right) - \left(1 - \tau_{0,1,H_1}^{(m)(r-1)} \right) p_{00}^{(r)} \right) = 0 \ (3-13a).$$

Similarly, equation (3-12b) can be rewritten as:

$$\sum_{m=1}^{N_0}\left(\frac{\tau_{0,1,H_1}^{(m)(r-1)}}{p_{01}^{(r)}}-\frac{\tau_{0,2,H_1}^{(m)(r-1)}}{p_{02}^{(r)}}\right)=\sum_{m=1}^{N_0}\left(\frac{\tau_{0,1,H_1}^{(m)(r-1)}p_{02}^{(r)}-\tau_{0,2,H_1}^{(m)(r-1)}p_{01}^{(r)}}{p_{01}^{(r)}p_{02}^{(r)}}\right)$$

$$=\frac{1}{p_{01}^{(r)}p_{02}^{(r)}}\sum_{m=1}^{N_0}\left(\tau_{0,2,H_1}^{(m)(r-1)}\left(1-p_{00}^{(r)}\right)-\left(\tau_{0,1,H_1}^{(m)(r-1)}+\tau_{0,2,H_1}^{(m)(r-1)}\right)p_{01}^{(r)}\right)=0.$$

That is:

$$\sum_{m=1}^{N_0}\left(\tau_{0,1,H_1}^{(m)(r-1)}\left(1-p_{00}^{(r)}\right)-\left(1-\tau_{0,0,H_1}^{(m)(r-1)}\right)p_{01}^{(r)}\right)=0\ (3-13b).$$

From equation (3-13a) and (3-13b):

$$\sum_{m=1}^{N_0}\left(\tau_{0,0,H_1}^{(m)(r-1)}\left(1-p_{01}^{(r)}\right)-\left(1-\tau_{0,1,H_1}^{(m)(r-1)}\right)p_{00}^{(r)}\right)+\sum_{m=1}^{N_0}\left(\tau_{0,1,H_1}^{(m)(r-1)}\left(1-p_{00}^{(r)}\right)-\left(1-\tau_{0,0,H_1}^{(m)(r-1)}\right)p_{01}^{(r)}\right)$$

$$=\sum_{m=1}^{N_0}\left(\tau_{0,0,H_1}^{(m)(r-1)}+\tau_{0,1,H_1}^{(m)(r-1)}-p_{00}^{(r)}-p_{01}^{(r)}+\tau_{0,1,H_1}^{(m)(r-1)}p_{00}^{(r)}-\tau_{0,1,H_1}^{(m)(r-1)}p_{00}^{(r)}\right.$$

$$\left.+\tau_{0,0,H_1}^{(m)(r-1)}p_{01}^{(r)}-\tau_{0,0,H_1}^{(m)(r-1)}p_{01}^{(r)}\right)$$

$$=\sum_{m=1}^{N_0}\left(\tau_{0,0,H_1}^{(m)(r-1)}+\tau_{0,1,H_1}^{(m)(r-1)}-p_{00}^{(r)}-p_{01}^{(r)}\right)$$

$$=\sum_{m=1}^{N_0}\left(1-\tau_{0,2,H_1}^{(m)(r-1)}-\left(1-p_{02}^{(r)}\right)\right)=\sum_{m=1}^{N_0}\left(p_{02}^{(r)}-\tau_{0,2,H_1}^{(m)(r-1)}\right)=0$$

Thus,

$$p_{02}^{(r)}=\frac{\sum_{m=1}^{N_0}\tau_{0,2,H_1}^{(m)(r-1)}}{N_0}$$

Similarly, I have:

$$p_{01}^{(r)} = \frac{\sum_{m=1}^{N_0} \tau_{0,1,H_1}^{(m)(r-1)}}{N_0} p_{00}^{(r)} = \frac{\sum_{m=1}^{N_0} \tau_{0,0,H_1}^{(m)(r-1)}}{N_0}$$

Analogous results will be achieved after taking the partial derivatives to equation (3-11) with respect to $p_{10}^{(r)}, p_{11}^{(r)}$ and $p_{12}^{(r)}$. In general, the r-th iteration updates of the estimated true genotype frequencies in the control group and in the case group are:

$$p_{0j}^{(r)} = \frac{\sum_{m=1}^{N_0} \tau_{0,j,H_1}^{(m)(r-1)}}{N_0}, j = 0,1,2 \ (3-14 \text{ a}) p_{1j}^{(r)} = \frac{\sum_{m=N_0+1}^{N} \tau_{1,j,H_1}^{(m)(r-1)}}{N_1}, j = 0,1,2 \ (3-14 \text{ b})$$

Under the assumption of symmetric but differential misclassification rate, from equation (3-11), the r-th step log-likelihood $Q_1^{(r)}$ can also be written as:

$$Q_1^{(r)} = \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{\mathrm{i,j},H_1}^{(m)(\mathrm{r}-1)} \times \ln\left[\binom{V(m)}{n_1(m)}\left(\mu_{ij}^{(r)}\right)^{n_1(m)}\left(1-\mu_{ij}^{(r)}\right)^{V(m)-n_1(m)} p_{ij}^{(r)} q_i^t\right]$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{\mathrm{i,j},H_1}^{(m)(\mathrm{r}-1)}$$

$$\times\left[\ln\binom{V(m)}{n_1(m)} + n_1(m)\ln\left(\mu_{ij}^{(r)}\right) + (V(m)-n_1(m))\ln\left(1-\mu_{ij}^{(r)}\right) + \ln p_{ij}^{(r)}\right.$$

$$\left. + \ln q_i^t\right]$$

$$= \sum_{m=1}^{N_0}\sum_{j=0}^{2} \tau_{0,\mathrm{j},H_1}^{(m)(\mathrm{r}-1)} \times\left[n_1 \ln \mu_{0j}^{(r)} + (V-n_1)\ln\left(1-\mu_{0j}^{(r)}\right)\right] + \sum_{m=N_0+1}^{N}\sum_{j=0}^{2} \tau_{1,\mathrm{j},H_1}^{(m)(\mathrm{r}-1)}$$

$$\times\left[n_1(m)\ln {\mu_{1j}}^{(r)} + (V(m)-n_1(m))\ln\left(1-\mu_{1j}^{(r)}\right)\right]$$

$$+ \text{terms that do not contain } \varepsilon_i$$

$$= \sum_{m=1}^{N_0}\left[\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)}\left(n_1(m)\ln\varepsilon_0^{(r)} + (V(m)-n_1(m))\ln\left(1-\varepsilon_0^{(r)}\right)\right)\right.$$

$$+ \tau_{0,1,H_1}^{(m)(\mathrm{r}-1)}\left(n_1\ln\frac{1}{2} + (V(m)-n_1(m))\ln\frac{1}{2}\right)$$

$$\left. + \tau_{0,2,H_1}^{(m)(\mathrm{r}-1)}\left(n_1\ln(1-\varepsilon_0^{(r)}) + (V(m)-n_1(m))\ln\varepsilon_0^{(r)}\right)\right]$$

$$+ \sum_{m=N_0+1}^{N}\left[\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)}\left(n_1(m)\ln\varepsilon_1^{(r)} + (V(m)-n_1(m))\ln\left(1-\varepsilon_1^{(r)}\right)\right)\right.$$

$$+ \tau_{1,1,H_1}^{(m)(\mathrm{r}-1)}\left(n_1(m)\ln\frac{1}{2} + (V(m)-n_1(m))ln\frac{1}{2}\right)$$

$$\left. + \tau_{1,2,H_1}^{(m)(\mathrm{r}-1)}\left(n_1(\mathrm{m})\ln(1-\varepsilon_1^{(r)}) + (V(m)-n_1(m))\ln\varepsilon_1^{(r)}\right)\right]$$

$$+ \text{terms that do not contain } \varepsilon_i$$

Differentiating Q_1 with respect to ε_0 and ε_1, and setting it to be zero, I have:

$$\frac{\partial Q_1^{(r)}}{\partial \varepsilon_0^{(r)}} = \sum_{m=1}^{N_0} \left[\frac{\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)} n_1(m)}{\varepsilon_0^{(r)}} - \frac{\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)} (V(m) - n_1(m))}{1 - \varepsilon_0^{(r)}} - \frac{\tau_{0,2,H_1}^{(m)(\mathrm{r}-1)} n_1(m)}{1 - \varepsilon_0^{(r)}} + \frac{\tau_{0,2,H_1}^{(m)(\mathrm{r}-1)} (V(m) - n_1(m))}{\varepsilon_0^{(r)}} \right] = 0 \ (3-15a)$$

$$\frac{\partial Q_1}{\partial \varepsilon_1^{(r)}} = \sum_{m=N_0+1}^{N} \left[\frac{\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)} n_1(m)}{\varepsilon_1^{(r)}} - \frac{\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)} (V(m) - n_1(m))}{1 - \varepsilon_1^{(r)}} - \frac{\tau_{1,2,H_1}^{(m)(\mathrm{r}-1)} n_1(m)}{1 - \varepsilon_1^{(r)}} + \frac{\tau_{1,2,H_1}^{(m)(\mathrm{r}-1)} (V(m) - n_1(m))}{\varepsilon_1^{(r)}} \right] = 0 \ (3-15b)$$

Since $n_1(m) + n_2(m) = V(m)$, equation (3-15a), (3-15b) can be simplified as:

$$\sum_{m=1}^{N_0} \frac{\left(n_1(m)\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)} + n_2(m)\tau_{0,2,H_1}^{(m)(\mathrm{r}-1)} \right)\left(1 - \varepsilon_0^{(r)}\right) - \varepsilon_0^{(r)} (n_1(m)\tau_{0,2,H_1}^{(m)(\mathrm{r}-1)} + n_2(m)\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)})}{\varepsilon_0^{(r)}(1 - \varepsilon_0^{(r)})} = 0 \quad (3-16a)$$

$$\sum_{m=N_0+1}^{N} \frac{\left(n_1(m)\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)} + n_2(m)\tau_{1,2,H_1}^{(m)(\mathrm{r}-1)} \right)\left(1 - \varepsilon_1^{(r)}\right) - \varepsilon_1^{(r)} \left(n_1(m)\tau_{1,2,H_1}^{(m)(\mathrm{r}-1)} + n_2(m)\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)} \right)}{\varepsilon_1^{(r)} \left(1 - \varepsilon_1^{(r)}\right)} = 0 (3-16b)$$

Solving equation (3-16a), I have:

$$\varepsilon_0^{(r)} = \sum_{m=1}^{N_0} \frac{n_1(m)\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)} + n_2(m)\tau_{0,2,H_1}^{(m)(\mathrm{r}-1)}}{V(m)(\tau_{0,0,H_1}^{(m)(\mathrm{r}-1)} + \tau_{0,2,H_1}^{(m)(\mathrm{r}-1)})} \ (3-17a)$$

Similarly, after solving equation (16 b), I have:

$$\varepsilon_1^{(r)} = \sum_{m=N_0+1}^{N} \frac{n_1(m)\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)} + n_2(m)\tau_{1,2,H_1}^{(m)(\mathrm{r}-1)}}{V(m)\left(\tau_{1,0,H_1}^{(m)(\mathrm{r}-1)} + \tau_{1,2,H_1}^{(m)(\mathrm{r}-1)}\right)} (3-17b)$$

3-3-2 EM Algorithm for Obtaining the MLEs under the Null hypothesis

The log-likelihood of the complete data under H_0 is:

$$l_0 = \ln(L_0) = \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1}\left[I\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right) \times \ln\left(P_0\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right)\right)\right] \quad (3-18)$$

As described in section 3-3-1, let Q_0 denote the expected value of the log-likelihood of the complete data, conditional on the observed data under the null hypothesis. I have:

$$Q_0 = E[\ln(L_0)|\text{Observed Data}]$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} P_0\left(\left(X_j^{t(m)}, Y_i^{t(m)}\right)\middle|\left(A_{n_1(m)}^{V(m)}, Y_i^{t(m)}\right)\right) \times \ln\left(P_0\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right)\right)$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_0}^{(m)} \times \ln\left(P\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right)\right)$$

$$= \sum_{m=1}^{N}\sum_{j=0}^{2}\sum_{i=0}^{1} \tau_{i,j,H_0}^{(m)} \times \ln\left[\binom{V(m)}{n_{1(m)}} \mu_{ij}^{\,n_{1(m)}}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_j^t q_i^t\right] \quad (3-19)$$

Here,

$$\tau_{i,j,H_0}^{(m)} = \frac{P_0\left(A_{n_1(m)}^{V(m)}, X_j^{t(m)}, Y_i^{t(m)}\right)}{\sum_{k=0}^{2} \Pr\left(A_{n_1(m)}^{V(m)}, X_k^{t(m)}, Y_i^{t(m)}\right)} = \frac{\mu_{ij}^{\,n_{1(m)}}\left(1-\mu_{ij}\right)^{V(m)-n_1(m)} p_j^t}{\sum_{k=0}^{2}\left[\mu_{ik}^{\,n_{1(m)}}\left(1-\mu_{ik}\right)^{V(m)-n_1(m)} p_k^t\right]}$$

The r-th step Q_0can be written as :

$$Q_0^{(r)} = \sum_{\mathrm{m=1}}^{\mathrm{N}} \sum_{\mathrm{i=0}}^{1} \tau_{\mathrm{i,0},H_0}^{(m)(r)} (\ln \mathrm{p}_0^{(\mathrm{r})} + \ln \mathrm{q}_\mathrm{i}^{\mathrm{t}}) + \sum_{\mathrm{m=1}}^{\mathrm{N}} \sum_{\mathrm{i=0}}^{1} \tau_{\mathrm{i,1},H_0}^{(m)(r)} (\ln \mathrm{p}_1^{(\mathrm{r})} + \ln \mathrm{q}_\mathrm{i}^{\mathrm{t}})$$

$$+ \sum_{\mathrm{m=1}}^{\mathrm{N}} \sum_{\mathrm{i=0}}^{1} \tau_{\mathrm{i,2},H_0}^{(m)(r)} \left(\ln \mathrm{p}_2^{(\mathrm{r})} + \ln \mathrm{q}_\mathrm{i}^{\mathrm{t}}\right).$$

Here

$$\tau_{\mathrm{i,j},H_0}^{(m)(r)} = \frac{\left(\mu_{\mathrm{ij}}^{(\mathrm{r})}\right)^{n_1} \left(1 - \mu_{\mathrm{ij}}^{(\mathrm{r})}\right)^{\mathrm{V-n_1}} \mathrm{p}_{\mathrm{ij}}^{(\mathrm{r})}}{\sum_{\mathrm{k=0}}^{2} \left[\left(\mu_{\mathrm{ik}}^{(\mathrm{r})}\right)^{n_1} \left(1 - \mu_{\mathrm{ik}}^{(\mathrm{r})}\right)^{\mathrm{V-n_1}} \mathrm{p}_{\mathrm{ik}}^{(\mathrm{r})}\right]}$$

$$= \frac{\left(\mu_{\mathrm{ij}}^{(\mathrm{r})}\right)^{n_1} \left(1 - \mu_{\mathrm{ij}}^{(\mathrm{r})}\right)^{\mathrm{V-n_1}} \mathrm{p}_{\mathrm{j}}^{(\mathrm{r})}}{\sum_{\mathrm{k=0}}^{2} \left[\left(\mu_{\mathrm{ik}}^{(\mathrm{r})}\right)^{n_1(m)} \left(1 - \mu_{\mathrm{ik}}^{(\mathrm{r})}\right)^{\mathrm{V-n_1}} \mathrm{p}_{\mathrm{k}}^{(\mathrm{r})}\right]}. \qquad (3-20)$$

Using the same algebra as in section 3-3-1, the r-th step genotype frequency is:

$$\mathrm{p}_{\mathrm{j}}^{(\mathrm{r})} = \frac{\sum_{\mathrm{m=1}}^{\mathrm{N}} \sum_{\mathrm{i=0}}^{1} \tau_{\mathrm{i,j},H_0}^{(m)(r-1)}}{\mathrm{N}}, \text{for j} = 0,1,2.$$

The r-th step differential but symmetric misclassification error rate is:

$$\varepsilon_i^{(r)} = \frac{\sum_{m=1}^{N_i} n_1(m) \tau_{i,0,H_0}^{(m)(r-1)} + n_2(m) \tau_{i,2,H_0}^{(m)(r-1)}}{\sum_{m=1}^{N_i} V(m) (\tau_{i,0,H_0}^{(m)(r-1)} + \tau_{i,2,H_0}^{(m)(r-1)})}, \text{for } i = 0,1 \qquad (3-21)$$

Thus, from equation (3-5a) and (3-5b), the r-th step log-likelihood of the observed data under the alternative hypothesis is:

$$l_1^{(r)}(\theta) = \ln\left(\mathrm{L}_1^{(\mathrm{r})}\right) = \sum_{m=1}^{N_0} \ln\left(\sum_{j=0}^{2} \left(\binom{V(m)}{n_1(m)} \left(\mu_{0j}^{(r)}\right)^{n_1(m)} \left(1 - \mu_{0j}^{(r)}\right)^{(V(m) - n_1(m))} p_{0j}^{(r)} q_0^t\right)\right)$$

$$+\sum_{m=N_0+1}^{N}\ln\left(\sum_{j=0}^{2}\left(\binom{V(m)}{n_1(m)}\left(\mu_{1j}^{(r)}\right)^{n_1(m)}\left(1-\mu_{1j}^{(r)}\right)^{(V(m)-n_1(m))}p_{1j}^{(r)}q_1^t\right)\right)\ (3-22\text{a})$$

The log-likelihood of the observed data under the null hypothesis is:

$$l_0^{(r)}(\theta)=\ln\left(\mathrm{L}_0^{(\mathrm{r})}\right)=\sum_{m=1}^{N_0}\ln\left(\sum_{j=0}^{2}\left(\binom{V}{n_1}\left(\mu_{0j}^{(r)}\right)^{n_1(m)}\left(1-\mu_{0j}^{(r)}\right)^{(V(m)-n_1(m))}p_j^{(r)}q_0^t\right)\right)$$

$$+\sum_{m=N_0+1}^{N}\ln\left(\sum_{j=0}^{2}\left(\binom{V(m)}{n_1(m)}\left(\mu_{1j}^{(r)}\right)^{n_1(m)}\left(1-\mu_{1j}^{(r)}\right)^{(V(m)-n_1(m))}p_j^{(r)}q_1^t\right)\right).(3-22b)$$

Under hypothesis H_u, when the difference of log-likelihoods of observed data between the r_u^{th} step and the $(r_u-1)^{st}$ step is less than the tolerance, i.e., when $|\ln(L_u^{(r_u)})-\ln(L_u^{(r_u-1)})|<\text{tolerance}$. I report $\ln(L_u^{r_u})$ as the maximum likelihood under H_i, that is: $\ln\left(\widehat{L_u}\right)=\ln(L_u^{r_u})$. Here, I use a tolerance of 10^{-5} and , $u=0$ for null hypothesis and $u=1$ for alternative hypothesis.

3-4 Simulation

3-4-1 Generating the samples

Genotype frequencies were determined from the allele frequencies by assuming HWE. For each individual m, let $\alpha_i^{t(m)}$ be the true at-risk allele frequency for affection status i, and let $p_{ij}^{t(m)}$ be the true genotype frequency for affection status i and genotype j. At a single locus, $p_{i0}^{t(m)}=\left(1-\alpha_i^{t(m)}\right)^2, p_{i1}^{t(m)}=2\alpha_i^{t(m)}(1-\alpha_i^{t(m)}), p_{i2}^{t(m)}=\left(\alpha_i^{t(m)}\right)^2$. First, I set the true

genotype for individual m by generating the random number $r^{(m)}$ that followed a uniform distribution from 0 to 1. If $r^{(m)} \leq p_{i0}^{t(m)}$, then I set the true genotype to "0". If $p_{i0}^{t(m)} < r^{(m)} \leq p_{i0}^{t(m)} + p_{i1}^{t(m)}$, then I set the true genotype to "1". If $p_{i0}^{t(m)} + p_{i1}^{t(m)} < r^{(m)} \leq 1$, then I set the true genotype to "2". Next, I generated the read data. I generated a random variable X, which represented the number of reads of the at-risk allele in V reads for individual m. Then $n_1(m) = X$. From equation (8), X followed a binomial distribution with number of trials V and probability μ_{ij}. (1) When the true genotype was "0", $j = 0$, and $\mu_{ij} = \frac{2-0}{2}\varepsilon_i + \frac{0}{2}(1-\varepsilon_i) = \varepsilon_i$; (2) When the true genotype was "2", $j = 2$, then $\mu_{ij} = \frac{2-2}{2}\varepsilon_i + \frac{2}{2}(1-\varepsilon_i) = 1-\varepsilon_i$; (3) When the true genotype was "1", $j = 1$ then $\mu_{ij} = \left(\frac{1}{2}\right)\varepsilon_i + \left(\frac{1}{2}\right)(1-\varepsilon_i) = \frac{1}{2}$. The distribution of $n_1(m)$ is summarized in Table 3.1.

Table 3.1 Distribution of X, the number of observed reads of the at-risk allele in V reads

	True genotype	Distribution
$n_1(m) = X$	0	$X \sim Bin(V, \varepsilon_i)$
	1	$X \sim Bin(V, 1-\varepsilon_i)$
	2	$X \sim Bin(V, {}^1/_2)$

Notation: $n_1(m)$= number of reads of at-risk allele in V reads for individual m.

For each setting of the parameters, under the null hypothesis, let $\alpha^{(0)}$ be the starting point of the at-risk allele frequency. Let $\varepsilon_i^{(0)}$ be the starting point of the misclassification rate for

affection status i. I first calculated the starting values of the genotype frequencies $p_j^{(0)}: p_0^{(0)} = \left(1-\alpha^{(0)}\right)^2, p_1^{(0)} = 2\alpha^{(0)}(1-\alpha^{(0)}), p_0^{(0)} = \left(\alpha^{(0)}\right)^2$. At step 0, with the values of $n_1(m)$, $p_j^{(0)}$ and $\varepsilon_i^{(0)}$, I calculated the 0th step log-likelihood $lnL_0^{(0)}$ by applying equation (3-22b). Then I calculated the 0-th step Bayesian Posterior Probability (BPP) $\tau_{m,i,j,H_0}^{(0)}$ from equation (3-20). At step 1, I calculated the genotype frequencies $p_i^{(1)}$and misclassification rates $\varepsilon_i^{(1)}$ from equation (3-14a), (3-14b), (3-17a) and (3-17b). Then by using $n_1(m)$, $p_i^{(1)}$ and $\varepsilon_i^{(1)}$, I calculated the the 1st step BPP $\tau_{m,i,j,H_0}^{(1)}$from equation (3-20). Next, the 1st step log-likelihood $lnL_0^{(0)}$was calculated from equation (3-22b). I stopped the iteration sequence when the difference between log-likelihood $\ln {L_0}^{(r_0+1)}$ and $\ln {L_0}^{r_0}$ step is smaller than the tolerance 10^{-5}. The value of $\ln {L_0}^{(r_0+1)}$ was then used as the maximized log-likelihood for the random staring value $(\alpha^{(0)}, \varepsilon_i^{(0)})$. The maximum of all random staring values' maximized log-likelihood was then used as the global maximum log-likelihood and was denoted as $\ln \widehat{L_0}$.

A similar method was applied for the alternative hypothesis. Under H_1, let $\alpha_i^{(0)}$ be the starting point of the at-risk allele frequency with affection status i, and let $\varepsilon_i^{(0)}$ be the starting point of the misclassification rate with affection status i. I calculated the starting values of the genotype frequencies $p_{ij}^{(0)}$using HWE: $p_{i0}^{(0)} = \left(1-\alpha_i^{(0)}\right)^2, p_{i1}^{(0)} = 2\alpha_i^{(0)}(1-\alpha_i^{(0)}), p_{i2}^{(0)} = \left(\alpha_i^{(0)}\right)^2$. As with the null hypothesis, at step 0, by applying equation (3-22 a), I calculated the 0th step log-likelihood $lnL_1^{(0)}$ from the values of $n_1(m)$, $p_{ij}^{(0)}$ and $\varepsilon_i^{(0)}$. Then I calculated the 0th step Bayesian Posterior Probability (BPP) $\tau_{m,i,j,H_1}^{(0)}$ from equation (3-20). At step 1, I calculated the 1st

step genotype frequencies $p_{ij}^{(1)}$and misclassification rates $\varepsilon_i^{(1)}$ from equation (3-14a), (3-14b), (3-17a) and (3-17b). By using $n_1(m)$, $p_{ij}^{(1)}$ and $\varepsilon_i^{(1)}$, I calculated the 1st step BPP $\tau_{m,i,j,H_1}^{(1)}$ from equation (3-20). Then the 1st step log-likelihood $\ln L_1^{(1)}$was calculated from equation (3-22a). Continued the iterations until the difference between (r_1+1) step log-likelihood $\ln {L_1}^{(r_1+1)}$ and r_1 step $\ln {L_1}^{r_1}$ step is small than the tolerance 10^{-5}. The value of $\ln {L_1}^{(r_1+1)}$ was then used as the maximized log-likelihood for this random starting values $(\alpha_{ij}^{(0)}, \varepsilon_i^{(0)})$. The maximum of all maximized log-likelihood values was then used as the global maximum log-likelihood, and was denoted as $\ln \widehat{L_1}$.

3-4-2 Choosing initial values for the EM algorithm

Before performing the simulation study for various parameter settings, I performed a simulation study to assess the adequacy of the distribution and number of random starting values for finding the global maximum likelihood and the convergence rate of my EM algorithm.

The true at-risk allele frequencies were set to 0.005 in controls and 0.005 in cases; that is, the null hypothesis was true. The true misclassification error rates were 0.001 in controls and 0.001 in cases. The total number of cases was 1000 and the total number of controls was 1000.

For both the null and alternative likelihood functions, I chose 500 independent random starting points for the at-risk allele frequencies in case and control group respectively. Each starting point followed a uniform distribution from 0 to 1. I chose 500 independent random starting points for the misclassification rate in the case group and control group. Each starting point followed a uniform distribution $U(0,b)$. I compared three distributions here: (1) $U(0,1)$; (2) $U(0,0.5)$; (3) $U(0,0.1)$. The maximum EM step per starting value was 100. The total number of samples was 50.

Table 3.2 is a summary of the number of iterations to convergence under the three distributions of random staring values for the null and alternative maximum likelihood functions.

Table 3.2 Summary of number of iterations until tolerance limit achieved

Under null hypothesis			
	misclassification rate distribution		
Number of iterations	U(0,1)	U(0,0.5)	U(0,0.1)
Mean	6.20	5.96	5.51
Standard Deviation	0.79	0.83	0.72
Minimum	4	4	4
Maximum	13	12	11
Under the alternative hypothesis			
	misclassification rate distribution		
Mean	5.93	5.93	5.75
Standard Deviation	0.63	0.63	0.67
Minimum	4	4	4
Maximum	10	12	10

From Table 3.2, the maximum number of iterations was 13 and the minimum was 4. All three distributions had a small number of iterations before reaching tolerance.

Table 3.3 shows the descriptions of the maximized likelihood functions for selected replicates. The first column "Rep" shows replicate number (1 to 10 and 50). The second column shows the likelihood function using the true parameters as calculated from equation (22). The next set of columns summarizes the maximum log-likelihood values using the starting values of misclassification rate from three distributions under the null hypothesis. There are two columns under each distribution for each replicate. The first column is the number of maximized log-likelihoods observed, which is 2 for the $U(0,1)$ and $U(0,0.5)$ distributions of starting values. It is 1 for the $U(0,0.1)$ distribution. The second column is the value of the local maximized log-likelihood, and the entry underneath is the corresponding frequency of the local maximum. The next set of columns is the summary of maximum log-likelihood values for the three distributions of starting values under the alternative hypothesis. The first column is the number of observed maximized log-likelihoods. All three distributions have only 1 value under each parameter values. The second column is the value of the local maximized log-likelihood, and as for null hypothesis, the underneath value is the corresponding frequency of each local maximum log-likelihood.

For example, for replicate 1, the likelihood value from the true parameter values was -1637.27. Under the null hypothesis, when the starting value of misclassification rate followed $U(0,1)$, two values of the maximized likelihood were obtained: -3005.97 and -1636.99. Among 500 random starting values, 243 of 500 converged to -3005.97, 257 of 500 converged to -1636.99. When the starting value followed $U(0,0.5)$, the same two maximized log-likelihoods

were obtained, 60 of 500 were -3005.97, and 440 of 500 were -1636.99. When the starting value followed $U(0,0.1)$, all 500 converged to -1636.99. Under the alternative hypothesis, 500 random starting value settings all converged to the log-likelihood -1636.58 for each distribution.

Table 3.3 Maximized log-likelihood values under three distributions of misclassification rate initial values

Rep	$\ln L$	$\ln \widehat{L_0}$								$\ln \widehat{L_1}$	
		U(0,1)			U(0,0.5)			U(0,0.1)		all dist^n	
		N			N			N		N	
1	-1637.27	2	-3005.97	-1636.99	2	-3005.97	-1636.99		-1636.99		-1636.58
	500		243	257		60	440	1	500	1	500
2	-1618.77	2	-2992.22	-1618.34	2	-2992.22	-1618.34		-1618.34		-1618.08
	500		243	257		73	427	1	500	1	500
3	-1576.76	2	-2939.37	-1574.01	2	-2939.37	-1574.01		-1574.01		-1573.29
	500		243	257		58	442	1	500	1	500
4	-1599.75	2	-2970.61	-1599.19	2	-2970.61	-1599.19		-1599.19		-1598.92
	500		223	277		52	448	1	500	1	500
5	-1639.43	2	-3010.62	-1638.18	2	-3010.62	-1638.18		-1638.18		-1637.49
	500		220	280		55	445	1	500	1	500
6	-1649.69	2	-3020.33	-1648.69	2	-3020.33	-1648.69		-1648.69		-1647.03
	500		214	286		59	441	1	500	1	500
7	-1598.67	2	-2972.31	-1598.39	2	-2972.31	-1598.39		-1598.39		-1598.37
	500		206	294		53	447	1	500	1	500
8	-1617.04	2	-2982.48	-1616.18	2	-2982.48	-1616.18		-1616.18		-1614.17
	500		246	254		58	442	1	500	1	500
9	-1614.28	2	-2982.49	-1613.34	2	-2982.49	-1613.34		-1613.34		-1610.46
	500		236	264		60	440	1	500	1	500
10	-1637.37	2	-2997.02	-1635.64	2	-2997.02	-1635.64		-1635.64		-1634.97
	500		241	259		63	437	1	500	1	500
…	…		…	…		…	…		…		…
50	-1570.22	2	-2935.33	-1567.07	2	-2935.33	-1567.07		-1567.07		-1566.07
	500		227	273		54	446	1	500	1	500

Notation: $\ln L$ = log-likelihood from parameters; $\ln \widehat{L_0}$=maximum log-likelihood under null hypothesis; $\ln \widehat{L_1}$=maximum log-likelihood under alternative hypothesis; N = number of maximized log-likelihood value

Figure 3.2 Comparison of the global max rate of three distributions under the null hypothesis

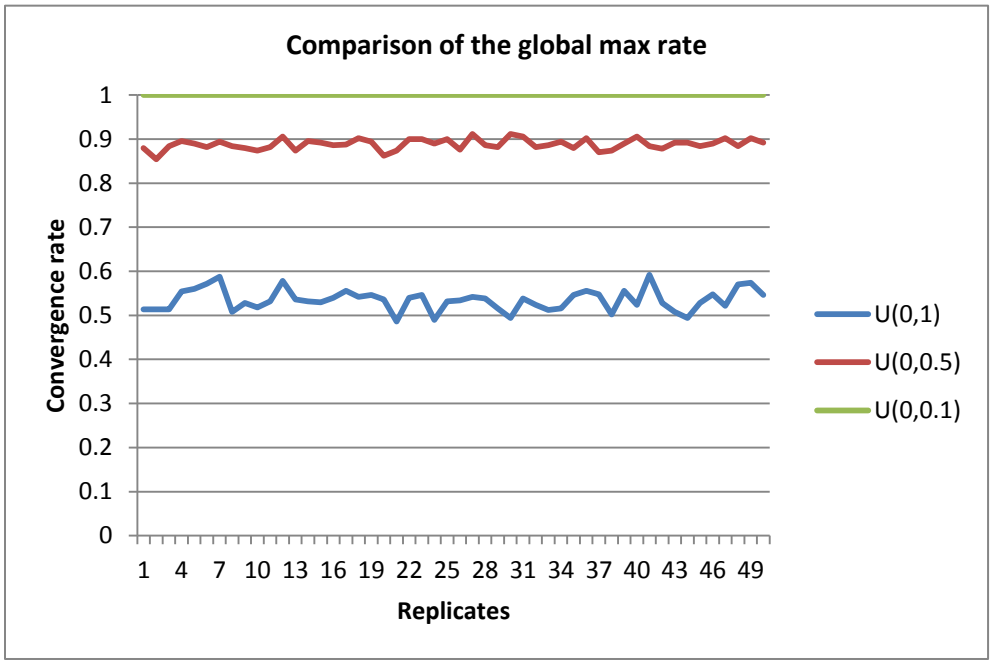


Figure 3.2 is the plot of the global max rate for each distribution of random starting values for the null hypothesis. Here global max rate = number of global maximized log-likelihood/ total number of maximized log-likelihoods = number of global maximum log-likelihood/500. For example, under the null hypothesis, for replicate 1 of distribution $(0, 1)$, 257 out of 500 has global maximized log-likelihood of -1636.99. Then the global max rate = 257/500=0.514; similarly, for distribution $U(0, 0.5)$, the global max rate = 440/500=0.88; for $U(0, 0.1)$ the global max rate = 500/500=1. Under the alternative hypothesis, the global max rate = 500/500 = 1 for all distributions of random starting values.

In conclusion, no more than two maximized log-likelihood values were obtained under the various settings of the distribution of random starting values. Random starting values that follow the $U(0,0.1)$ distribution always had maximized value that is equal to the global maximum. Under H_0, both $U(0,1)$ and $U(0,0.5)$ obtained two 'local maximum' log-likelihood

values. Distribution U(0, 1) had around 50% global max rate, and distribution U(0, 0.5) had around 88% global max rate. Starting values of the misclassification rate that follow distribution $U(0, 0.1)$ showed the best behavior with respect to the estimation of each parameter and the maximized likelihood. It did not require more iterations till convergence, compared to the other two distributions. Hence, in the following simulations I chose $U(0,0.1)$ as the distribution of misclassification rate starting values.

3-4-3 Results

In this section, I report several simulations to evaluate the performance of the likelihood ratio test using NGS data in the presence of both non-differential and differential misclassification at a single locus. The parameter values used in this section are given in Table 3.4.

Table 3.4 Parameter settings of the simulation studies

Parameter	Notation	Value
α_0^t	True at-risk allele frequency in control group	0.005, 0.02, 0.05, 0.1
α_1^t	True at-risk allele frequency in case group	$\alpha_0^t + d$
d	The difference of at-risk allele frequency between control and case group	0, 0.01, 0.025
V	Coverage	8, 40
e_0^t	True misclassification rate in control group	0.001, 0.04
e_1^t	True misclassification rate in case group	0.001, 0.04
N_0	Number of controls	1000, 2500, 5000, 10000
N_1	Number of cases	1000, 2500, 5000, 10000
R	Number of replicates at each setting	100
S	Number of starting points	100
M	Number of maximum EM steps per starting point	100

For each simulation, the starting values of the at-risk allele frequency followed the uniform distribution $\mathrm{U}(0, 1)$, and the starting value of misclassification rate followed the uniform distribution $\mathrm{U}(0, 0.1)$. I considered both non-differential and differential misclassification rates, that is, the misclassification rate in control and case groups were (0.001, 0.001), (0.04, 0.04), (0.001, 0.04) and (0.04, 0.001).

3-4-3-1 The estimated genotype frequencies and misclassification rate from EM

I first performed a simulation to study the estimation of misclassification rates and genotype frequencies from my EM algorithm. I considered 24 parameter settings with 1000 observations from the control group and 1000 from the case group. Table 3.5 presents the results of estimated misclassification ratios. The misclassification ratio is defined as the ratio of estimated value of the misclassification rate to the true value of misclassification rate. In Table 3.5, the first four columns are the correct values of the parameter settings. The following four columns contain the average and standard deviation of the estimated misclassification ratios for both null and alternative hypotheses. Table 3.6a presents the results of the estimated genotype frequency ratios under the null hypothesis. Table 3.6b contains the values of the estimated genotype frequency ratios under the alternative hypothesis. Here, the genotype frequency ratio is defined to be the ratio of the estimated value of genotype frequency to the true value of genotype frequency. In Table 3.6a and Table 3.6b, the first four columns contain the true parameter values. The next six columns contain the average and standard deviation of the corresponding estimated genotype frequency ratios under the specific hypothesis. Figure 3.3 and Figure 3.4 are the graphical representations of Table 3.6a and Table 3.6b.

From the ratio of misclassification rate results, the observed average was between 0.98 and 1.05, and the observed standard deviation was between 0.02 and 0.46. The observed value from EM algorithm was a good estimate of the true misclassification value under both the null and alternative hypotheses.

The average of the estimated genotype ratios ranged from 0.8 to 1.03 under the null hypothesis, and ranged from 0.40 to 2.80 under the alternative hypothesis. The standard deviation of the estimated genotype ratios was between 0.42 to 5.22 under the null hypothesis, and was between 0.00 and 10.44 under the alternative hypothesis. From Figure 3.4a and 3.4b, the EM algorithm provided good estimates of the common homozygote and heterozygote genotype frequencies under different parameter settings. The estimation of the rare homozygote genotype depended on the true at-risk allele frequencies. As the true at-risk allele frequency increased, the accuracy of the estimation of the rare homozygote genotype frequencies increased. There was no significant pattern of parameter estimations for differential and non-differential misclassification rates.

Table 3.5 Estimated misclassification ratio distribution under simulation settings

True at-risk allele frequency		True misclassification rate		Estimated misclassification ratio			
				Control		Case	
Control	Case	Control	Case	AVE	SD	AVE	SD
0.005	0.005	0.001	0.001	1.02	0.20	1.01	0.24
		0.04	0.04	1.00	0.04	0.99	0.03
		0.001	0.04	1.00	0.03	1.01	0.24
		0.04	0.001	1.03	0.24	1.01	0.04
0.02	0.02	0.001	0.001	0.99	0.17	0.98	0.19
		0.04	0.04	1.00	0.03	1.00	0.03
		0.001	0.04	1.00	0.03	0.99	0.16
		0.04	0.001	1.00	0.18	1.00	0.03
0.005	0.015	0.001	0.001	1.01	0.15	0.99	0.17
		0.04	0.04	1.00	0.02	1.00	0.03
		0.001	0.04	1.00	0.02	0.99	0.16
		0.04	0.001	0.99	0.17	1.00	0.02
0.005	0.03	0.001	0.001	0.98	0.17	0.98	0.17
		0.04	0.04	1.00	0.02	1.00	0.02
		0.001	0.04	1.00	0.02	0.99	0.13
		0.04	0.001	0.99	0.16	1.00	0.02
0.02	0.03	0.001	0.001	0.99	0.16	1.02	0.14
		0.04	0.04	1.00	0.03	1.00	0.02
		0.001	0.04	0.99	0.03	0.99	0.16
		0.04	0.001	1.01	0.17	1.00	0.03
0.02	0.045	0.001	0.001	1.00	0.03	1.00	0.02
		0.04	0.04	1.00	0.03	1.00	0.02
		0.001	0.04	1.00	0.03	0.99	0.14
		0.04	0.001	1.02	0.19	1.00	0.02

Note: Number of cases = 1000. Number of controls = 1000. Coverage = 8. Number of replicates = 100. Number of starting points = 100.
Notation: Estimated misclassification ratio = Estimated misclassification rate/ true misclassification rate. SD = Standard

Table 3.6a Estimated genotype frequency ratios under the null hypothesis

True at-risk allele frequency		True misclassification rate		Under the null hypothesis					
				Estimated genotype frequency ratio					
				p_0		p_1		p_2	
Control	Case	Control	Case	AVE	SD	AVE	SD	AVE	SD
0.005	0.005	0.001	0.001	1.00	0.00	0.99	0.24	0.80	3.94
		0.04	0.04	1.00	0.00	0.99	0.23	1.00	4.38
		0.001	0.04	1.00	0.00	0.98	0.20	1.00	5.22
		0.04	0.001	1.00	0.00	0.99	0.18	1.00	4.38
0.02	0.02	0.001	0.001	1.00	0.00	1.00	0.10	1.02	1.29
		0.04	0.04	1.00	0.01	1.01	0.12	0.92	1.09
		0.001	0.04	1.00	0.00	1.01	0.12	0.83	1.04
		0.04	0.001	1.00	0.00	0.98	0.12	0.89	1.04
0.05	0.05	0.001	0.001	1.00	0.01	0.99	0.07	0.99	0.47
		0.04	0.04	1.00	0.01	1.00	0.06	1.00	0.42
		0.001	0.04	1.00	0.01	1.00	0.08	0.99	0.46
		0.04	0.001	1.00	0.01	1.01	0.07	1.03	0.45

Note: Number of cases = 1000. Number of controls = 1000. Coverage = 8. Number of replicates = 100. Number of starting points = 100.

Notation: p_j= frequency of genotype j ($j = 0,1,2$). Estimated genotype frequency ratio = Average estimated genotype value/ true genotype value. AVE = Average of estimated genotype frequency ratio. SD = Standard deviation of estimated genotype frequency ratio.

Figure 3.3a Average of estimated genotype frequency ratio under the null hypothesis

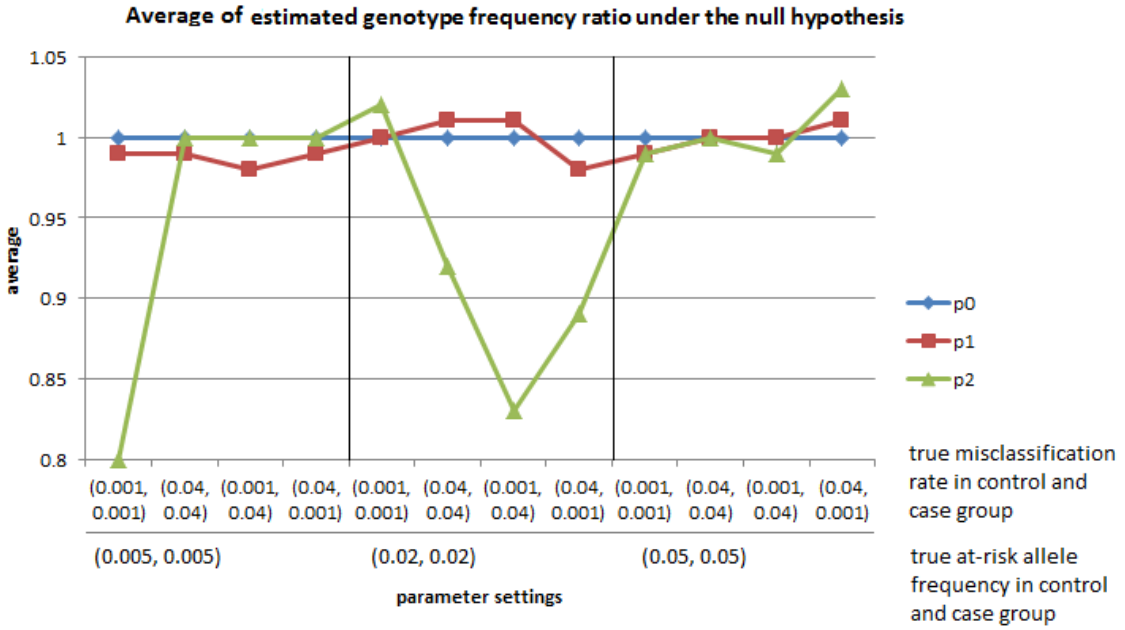


Figure 3.3b Standard deviation of estimated genotype frequency ratio under the null hypothesis

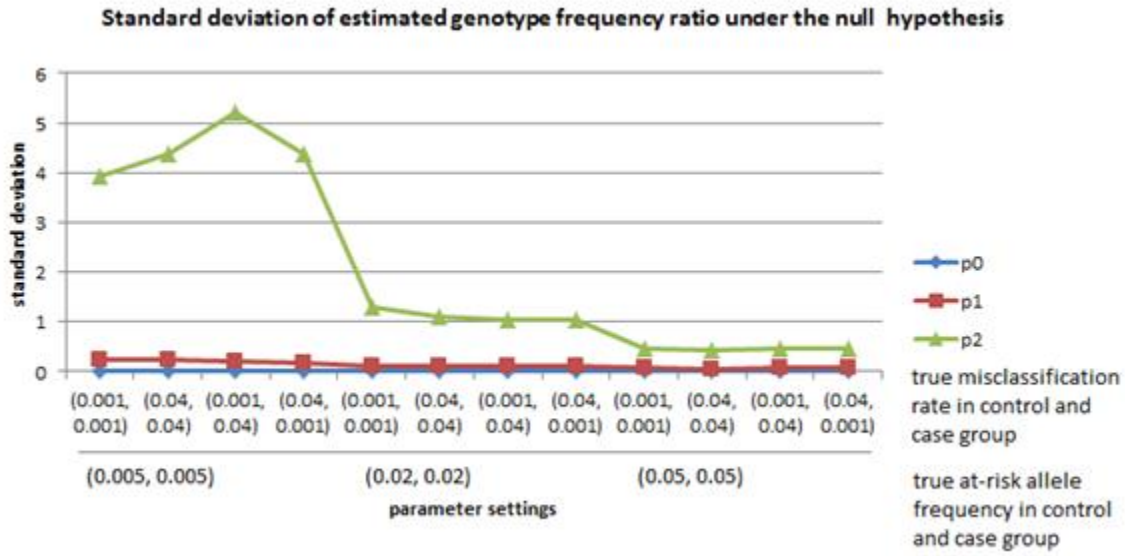


Table 3.6b Estimated genotype frequency ratios under the alternative hypothesis

True at-risk allele frequency		True misclassification rate		Under the alternative hypothesis											
				Estimated genotype frequency ratio											
				p_{00}		p_{01}		p_{02}		p_{10}		p_{11}		p_{12}	
Control	Case	Control	Case	AVE	SD	AVE	SD	AVE	SD	AVE	SD	AVE	SD	AVE	SD
0.005	0.015	0.001	0.001	1.00	0.00	0.98	0.31	0.80	5.63	1.00	0.00	1.00	0.17	1.02	1.98
		0.04	0.04	1.00	0.00	1.03	0.32	0.80	5.63	1.00	0.01	1.01	0.19	0.80	1.93
		0.001	0.04	1.00	0.00	0.99	0.37	1.20	6.86	1.00	0.01	1.00	0.17	1.11	2.13
		0.04	0.001	1.00	0.00	1.03	0.30	0.80	5.63	1.00	0.01	1.00	0.18	0.98	1.85
0.005	0.03	0.001	0.001	1.00	0.00	1.01	0.36	2.00	10.44	1.00	0.01	0.99	0.14	1.02	0.97
		0.04	0.04	1.00	0.00	1.07	0.30	1.60	7.88	1.00	0.01	1.01	0.13	1.08	1.13
		0.001	0.04	1.00	0.00	0.99	0.32	0.40	4.00	1.00	0.01	0.98	0.11	0.96	0.99
		0.04	0.001	1.00	0.00	1.00	0.29	2.80	10.26	1.00	0.01	1.00	0.13	1.04	1.02
0.02	0.03	0.001	0.001	1.00	0.01	1.00	0.14	1.00	1.38	1.00	0.01	1.00	0.12	1.12	1.13
		0.04	0.04	1.00	0.01	0.99	0.17	1.12	1.93	1.00	0.01	0.98	0.11	1.01	1.23
		0.001	0.04	1.00	0.01	1.01	0.16	0.87	1.39	1.00	0.01	1.02	0.14	1.03	1.03
		0.04	0.001	1.00	0.01	1.02	0.16	1.02	1.43	1.00	0.01	1.01	0.14	1.09	1.00
0.02	0.045	0.001	0.001	1.00	0.01	1.01	0.16	1.12	1.82	1.00	0.01	1.00	0.10	1.02	0.63
		0.04	0.04	1.00	0.01	1.02	0.17	0.95	1.66	1.00	0.01	1.01	0.10	1.04	0.70
		0.001	0.04	1.00	0.01	1.00	0.16	0.85	1.47	1.00	0.01	0.98	0.10	0.94	0.66
		0.04	0.001	1.00	0.01	1.00	0.14	1.08	1.64	1.00	0.01	1.01	0.11	0.96	0.61
0.05	0.1	0.001	0.001	1.00	0.01	1.01	0.10	0.98	0.64	1.00	0.01	1.00	0.07	0.98	0.30
		0.04	0.04	1.00	0.01	0.99	0.08	1.06	0.65	1.00	0.01	1.00	0.06	0.95	0.33
		0.001	0.04	1.00	0.01	1.01	0.09	0.96	0.58	1.00	0.01	1.00	0.06	0.99	0.34
		0.04	0.001	1.00	0.01	1.00	0.10	0.91	0.63	1.00	0.01	1.00	0.06	0.99	0.32

Note: Number of cases = 1000. Number of controls = 1000. Coverage = 8. Number of replicates = 100. Number of starting points = 100.Notation: Estimated genotype frequency ratio = average estimated genotype value/ true genotype value. AVE = average of estimated genotype frequency ratio. SD = standard deviation of estimated genotype frequency ratio. p_{ij}= frequency of genotype j with affect status i. ($i = 0,1. j = 0,1,2$).

Figure 3.4a Average of estimated genotype frequency ratio under the alternative hypothesis

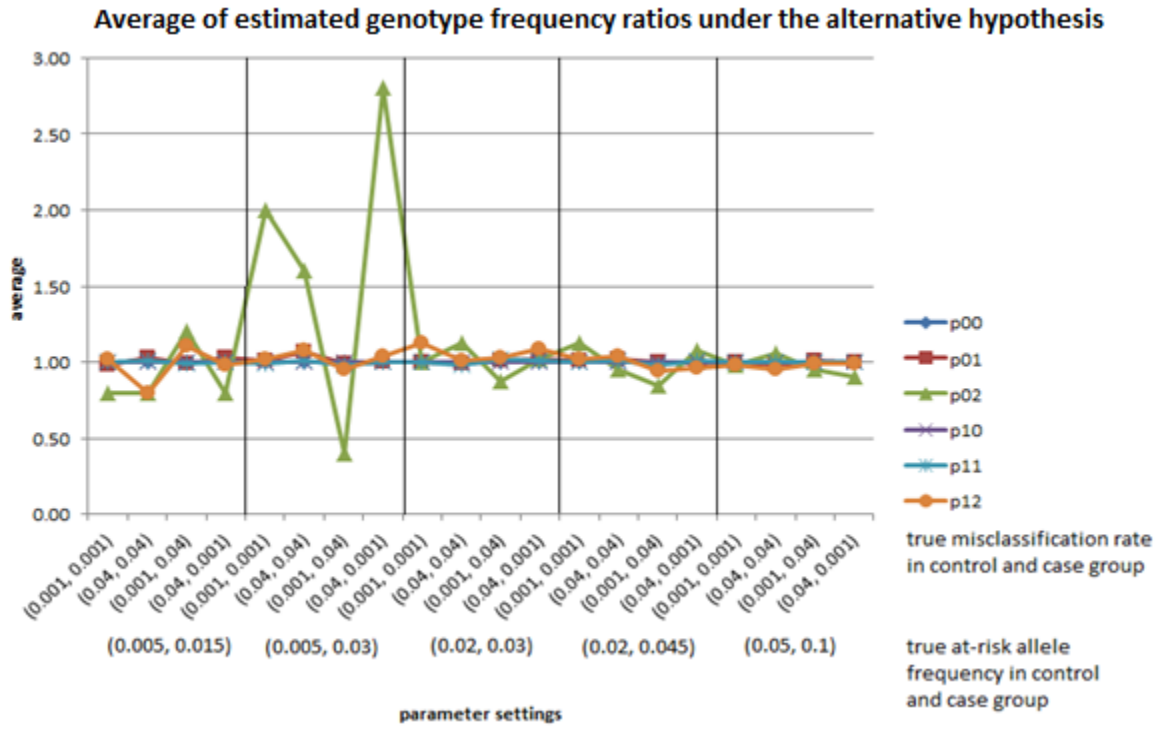
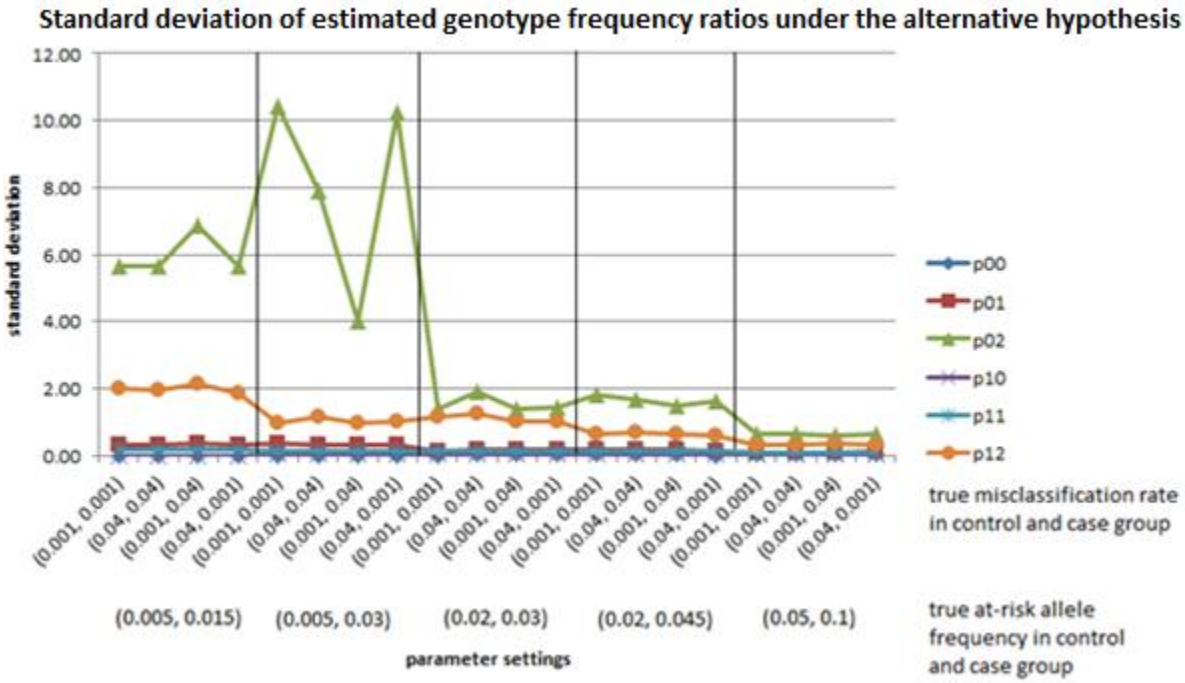


Figure 3.4b Standard deviation of estimated genotype frequency ratio under the alternative hypothesis



3-4-3-2 The properties of LRT_{NGS}

Under the null hypothesis, the test statistic LRT_{NGS} appeared to follow an asymptotic central chi-square distribution with two degrees of freedom. Under the null hypothesis, the expected value of LRT_{NGS} should be equal to its degrees of freedom, i.e. 2. I studied the properties of the test statistic LRT_{NGS} from the simulations as described in section 3-4-1. I defined the average of LRT_{NGS} as the sum of simulated LRT_{NGS} values divided by the total number of replicates.

Table 3.7 contains the results of the average LRT_{NGS} with different parameter settings under the null hypothesis. The first two columns are the values of the true misclassification rates, the next two columns are the true at-risk allele frequencies, and the last two columns are the average and standard deviation of the LRT_{NGS} from simulations. The comparisons of the average LRT_{NGS} with different at-risk allele frequencies in Table 3.7 are displayed in Figure 3.5.

Table 3.7 Results of LRT_{NGS} under the null hypothesis

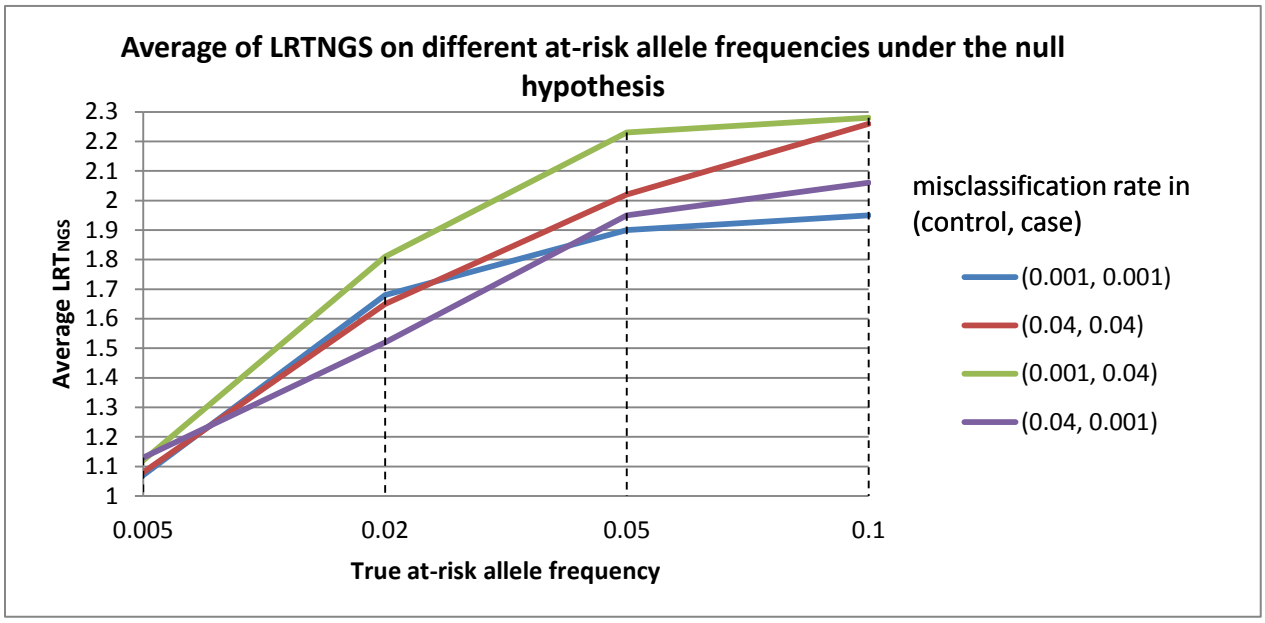
True misclassification rate		True at-risk allele frequency	LRT_{NGS}	
Control	Case		Average	SD
0.001	0.001	0.005	1.07	1.34
		0.02	1.68	1.48
		0.05	1.90	1.97
		0.1	1.95	1.87
0.04	0.04	0.005	1.08	1.11
		0.02	1.65	1.61
		0.05	2.02	2.01
		0.1	2.26	2.56

Table 3.7 (Continued) Results of LRT_{NGS} under the null hypothesis

Misclassification rate		At-risk allele frequency	LRT_{NGS}	
Control	Case		mean	SD
0.001	0.04	0.005	1.12	1.29
		0.02	1.81	1.47
		0.05	2.28	1.75
		0.1	2.23	2.61
0.04	0.001	0.005	1.13	1.28
		0.02	1.52	1.49
		0.05	1.95	2.36
		0.1	2.06	1.85

Note: Number of controls is 1000, number of cases is 1000. Number of replicates is 100. Coverage = 8.

Figure 3.5 Comparison of the average LRTNGS for different at-risk allele frequency



Note: Number of controls = 1000, number of cases = 1000. Number of replicates = 100. Coverage = 8.

Under the null hypothesis, for sample size 1000 in each group, with both non-differential and differential misclassification rate, the value of the average LRT_{NGS} was between 1 and 2.3. As the true at-risk allele frequency increased, the value of the average LRT_{NGS} increased. When the true at-risk allele frequency was less than 0.05, the average LRT_{NGS} was less than 2. When the true at-risk allele frequency was equal or greater than 0.05, the average LRT_{NGS} was around 2, which was the degrees of freedom of central chi-square test. When the sample size was small and the true at-risk allele frequency was small, the cell count of the rare homozygotes was extremely low. The effective degree of freedom was less than 2, so that the average of LRT_{NGS} was less than 2. A possible future study would be to evaluate the properties of LRT_{NGS} using permutation testing.

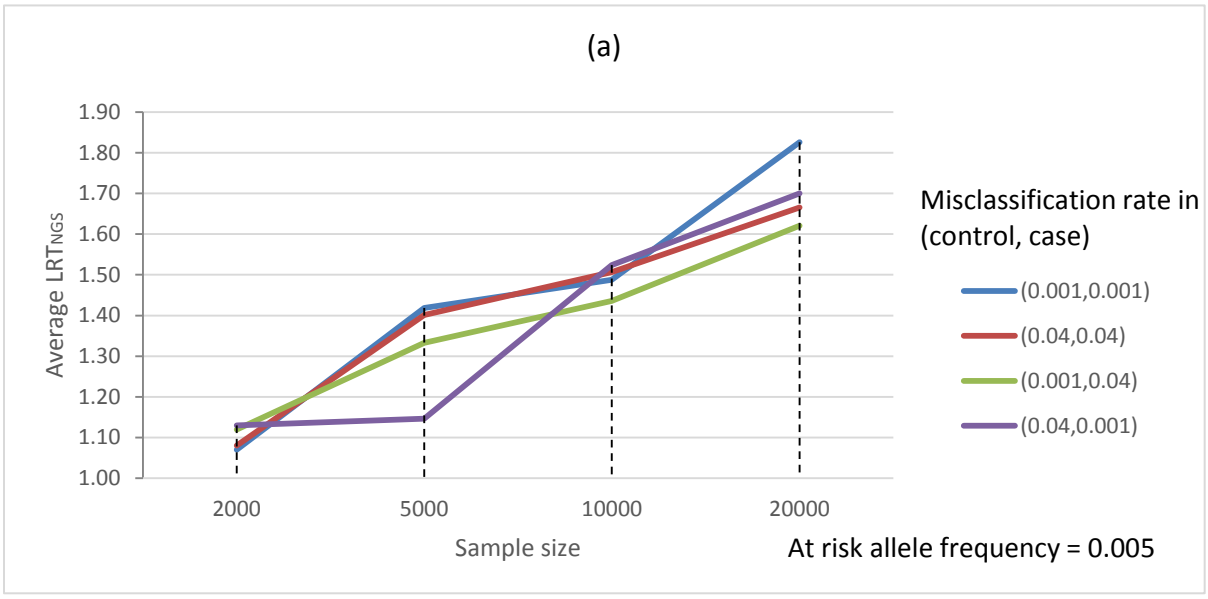
I then performed another set of simulations to study the effect of sample size on the average of the LRT_{NGS} under the null hypothesis. Three sample sizes in each group were considered here: 2500, 5000 and 10000. The simulation results are shown in Table 3.8. The first column is the true at-risk allele frequency. The second column contains the setting of sample size. The next 12 columns contain the results of LRT_{NGS} under different misclassification rates. Under each setting of misclassification rates, the first column was the average of the simulated LRT_{NGS}; the second column was the standard deviation of the simulated LRT_{NGS}, and the third column was the 95 percentile of the simulated LRT_{NGS}. The last row of Table 3.8 contains the average, standard deviation, and 95 percentile of the asymptotic LRT_{NGS} under the null hypothesis. Figure 3.6 shows the relationship between average LRT_{NGS} and sample size with different at-risk allele frequencies under the null hypothesis.

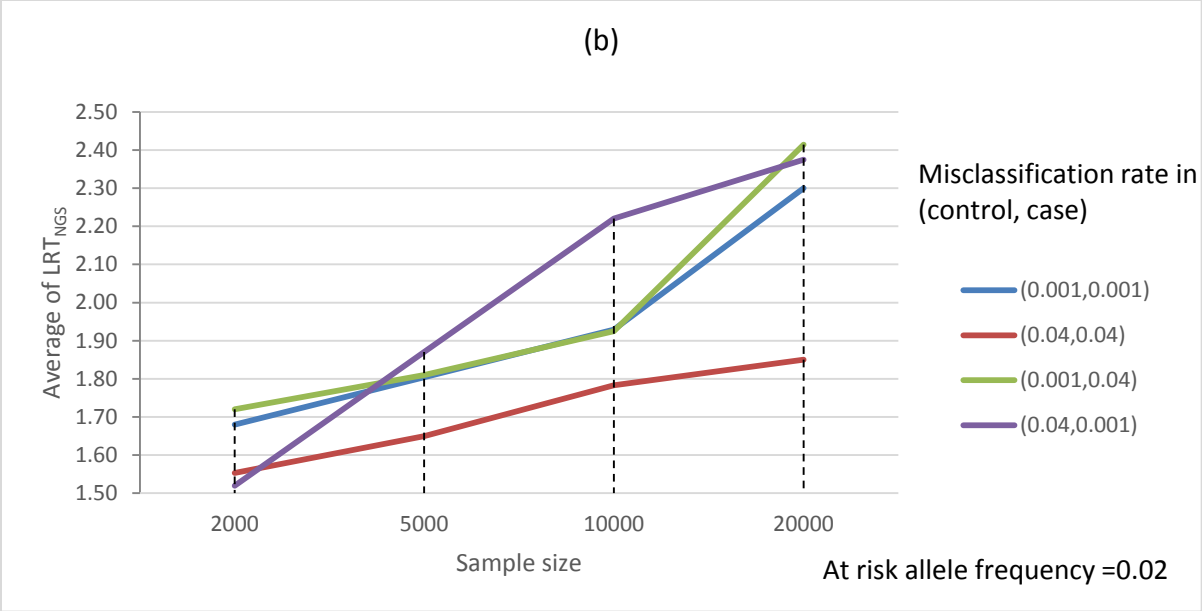
Table 3.8 Simulated LRTNGS values for different sample size under the null hypothesis

True at-risk allele frequency	Sample size	Simulated LRT_{NGS}											
		True misclassification rate in control and case group											
		(0.001,0.001)			(0.04,0.04)			(0.001,0.04)			(0.04,0.001)		
		AVE	SD	95%	AVE	SD	95%	AVE	SD	95%	AVE	SD	95%
0.005	2000	1.07	1.33	3.24	1.08	1.11	3.33	1.12	1.29	3.87	1.13	1.28	3.92
	5000	1.42	1.43	3.96	1.40	1.51	4.49	1.33	1.68	3.62	1.15	1.33	3.83
	10000	1.49	1.45	3.83	1.51	1.59	4.02	1.44	1.50	4.72	1.52	1.49	4.18
	20000	1.83	1.80	5.35	1.67	1.58	4.78	1.62	1.91	4.58	1.70	1.75	5.58
0.02	2000	1.68	1.48	4.88	1.55	1.36	4.38	1.72	2.15	4.66	1.52	1.49	4.50
	5000	1.80	1.72	4.93	1.65	1.60	4.53	1.81	1.47	4.40	1.87	2.11	4.99
	10000	1.93	1.73	5.14	1.80	1.53	5.91	1.93	1.51	5.77	2.22	1.59	5.51
	20000	2.30	2.33	6.51	1.90	1.90	5.12	2.41	2.48	7.50	2.37	2.46	7.23
Asymptotic LRT_{NGS}		**2.00**	**2.00**	**5.99**	**2.00**	**2.00**	**5.99**	**2.00**	**2.00**	**5.99**	**2.00**	**2.00**	**5.99**

Note: Coverage = 8
Notation: LRT_{NGS} = Statistic from log-likelihood ratio test using NGS data. AVE= Average of the simulated $LRT_{NGS.}$ SD = Standard deviation of the simulated $LRT_{NGS.}$ 95% = 95% percentile of the simulated LRT_{NGS}

Figure 3.6 Distribution of the average LRTNGS for different sample size with at-risk allele frequency 0.005 (figure (a)) and 0.02 (figure (b)) under the null hypothesis





Using the results from Table 3.8, I then performed a linear regression analysis on the average LRT_{NGS} and the sample size given the fixed at-risk allele frequencies and misclassification rate. The general form of each linear regression can be written as:

$$\overline{\text{LRT}_{\text{NGS}}} = \alpha + \beta \frac{1}{\sqrt{N}}$$

where $\overline{\text{LRT}_{\text{NGS}}}$ is the average of the simulated LRT_{NGS} and N is the sample size N. The terms α and β are the regression coefficients.

Table 3.9 contains the results of the linear regression analysis under each parameter setting. The first column contains the values of true at-risk allele frequency, the second and third column contain the value of misclassification rates in control and case group. Column $\hat{a}$ contain the estimated value of regression intercept, α, and column $\hat{b}$ contain the estimated value of regression coefficient β. The column labeled Pr(> |t|) is the corresponding p-value. The column labeled R^2 is the coefficient of determination.

Table 3.9 Linear regression of average LRT_{NGS} on sample size

At-risk allele frequency	Misclassification		Model: $\overline{LRT_{NGS}} = \alpha + \beta \frac{1}{\sqrt{N}}$				
	Control	Case	$\hat{a}$	Pr(>|t|)	$\hat{b}$	Pr(>|t|)	R^2
	Non-differential						
0.005	0.001	0.001	2.06	0.004	-45.06	0.039	0.92
	0.04	0.04	1.92	0.000	-37.37	0.004	0.99
0.02	0.001	0.001	2.40	0.007	-35.32	0.126	0.76
	0.04	0.04	1.93	0.000	-17.26	0.023	0.96
	Differential						
0.005	0.001	0.04	1.79	0.001	-30.88	0.019	0.96
	0.04	0.001	1.82	0.022	-33.10	0.218	0.61
0.02	0.001	0.04	2.47	0.012	-37.81	0.183	0.67
	0.04	0.001	2.75	0.001	-56.53	0.011	0.98

From Table 3.9, the estimated intercept of the regression $\hat{a}$ was between 1.79 and 2.75, and was significant at level 0.05. The estimated coefficient $\hat{b}$ was between -17.26 and -56.53. Under non-differential misclassification, the average proportion of variation explained by the linear regression was 0.91 (=(0.92+0.99+0.76+0.96)/4); Under differential misclassification, the average proportion of variation explained was 0.81 ((0.96+0.61+0.67+0.98)/4). This linear regression provided a better fit for non-differential misclassification scenario than differential misclassification. In general, a low frequency of the at-risk allele requires a greater sample size than a high frequency of the at-risk allele for the average LRT_{NGS} to be close to 2. As sample size increased, the average of LRT increased with 2 as a plausible limit.

3-4-3-3 Simulation Power

The previous studies were performed using coverage equaled 8. In this section, I performed another set of simulations using coverage equaled 40. I then combined the results from the two scenarios together to study the effects of coverage on the power of the test statistic.

Simulation power ($\text{Power}_{\text{sim}}$) was defined as the proportion of test statistic values that exceed the critical value from the simulation. The critical value from the simulation was the value on the scale of the simulated LRT_{NGS} beyond which I rejected the null hypothesis at the level of significance α of the test. i.e.:

$$\text{Power}_{\text{sim}} = \frac{\#\text{ of (simulated LRT}_{\text{NGS}} > \text{critical value from simulation)}}{\text{Total number of replicates}}$$

The Power estimated by method of moments ($\text{Power}_{\text{estimated}}$) was defined as:

$$\text{Power}_{\text{estimated}} = P(\chi^2(2, \hat{\lambda}) \geq \chi^2_{1-\alpha}(2))$$

Where $\chi^2(2, \hat{\lambda})$ was the non-central chi-squared distribution with non-centrality parameter $\hat{\lambda}$ and 2 degrees of freedom, $\chi^2_{1-\alpha}(2)$ was the critical value of a central chi-squared distribution with 2 degrees of freedom at significance level α. From the method of moments, $\hat{\lambda}$ was calculated as the difference between average of simulated LRT_{NGS} and the degrees of freedom, i.e.: $\hat{\lambda} = \text{Average}(\text{LRT}_{\text{NGS}}) - 2$.

Table 3.10 contains the power of the simulation and the power estimated by the method of moments under different parameter settings at significance level 0.05. The first four columns

showed the true parameter values under each setting. The next columns showed the critical value, the simulation power and the asymptotic power with coverage = 8 and coverage = 40 respectively.

Table 3.10 Distribution of directly simulated power and estimated power from NCP using method of moments

At-risk allele frequency		Misclassification rate		Coverage					
				8			40		
Control	Case	Control	Case	critical value	$Power_{sim}$	$Power_{Asym}$	critical value	$Power_{sim}$	$Power_{Asym}$
0.005	0.015	0.001	0.001	3.24	0.96	0.80	3.99	0.96	0.81
		0.04	0.04	3.33	0.84	0.70	3.70	0.93	0.79
		0.001	0.04	3.87	0.90	0.75	4.42	0.91	0.81
		0.04	0.001	3.92	0.87	0.70	4.23	0.89	0.78
0.005	0.03	0.001	0.001	3.24	1.00	1.00	3.99	1.00	1.00
		0.04	0.04	3.33	1.00	1.00	3.70	1.00	1.00
		0.001	0.04	3.87	1.00	1.00	4.42	1.00	1.00
		0.04	0.001	3.92	1.00	1.00	4.23	1.00	1.00
0.02	0.03	0.001	0.001	4.68	0.55	0.40	4.88	0.58	0.47
		0.04	0.04	4.38	0.32	0.42	5.31	0.52	0.43
		0.001	0.04	4.66	0.51	0.40	5.15	0.54	0.45
		0.04	0.001	4.50	0.50	0.40	4.98	0.51	0.42
0.02	0.045	0.001	0.001	4.68	0.99	0.98	4.88	1	0.98
		0.04	0.04	4.38	0.98	0.97	5.31	1	0.99
		0.001	0.04	4.66	0.97	0.97	5.15	0.99	0.98
		0.04	0.001	4.50	0.99	0.97	4.98	1	0.98

Note: Number of cases = 1000. Number of controls = 1000. Significance level = 0.05.

Table 3.10 showed that the power determined by the simulation was always larger than power determined by the NCP using the method of moments. As the difference between at-risk allele frequencies in control and case group increased, the power increased. In the situation where the difference in allele frequencies was the same, low at-risk allele frequencies always had higher power at the same significance level. The power in the presence of differential misclassification was always in between the high and low non-differential misclassification powers when all other parameters were the same.

Comparing the results of 40 coverage to 8 coverage, when all other parameters were the same, and the difference of the at-risk allele frequency in control and case group was greater than 0, the 40 coverage had the larger critical value and larger power than the 8 coverage. for the same significance level. For example, when the at-risk allele frequencies in control and case group were 0.005 and 0.015 respectively, and the misclassification rates in control and case groups were 0.04 and 0.04, the 40 coverage power was 0.93 at significance level 0.05, while the 8 coverage power was 0.84 at significance level 0.05. The critical values of the coverage 40 seemed to be closer to the asymptotic distribution than the critical values for coverage 8.

Chapter 4 Discussions and Future Work

Variants with low frequency may contribute a large fraction of risk in genetically associated diseases. Misclassification of the genotype reduces the power and increases the bias of the estimation of genetic parameters. In this dissertation, I studied the effect of misclassification error for case-control association studies with low frequency of at-risk allele.

I extended the result of Ji et al. for the NCP for the Likelihood Ratio Test Allowing for allelic errors. The NCP was a function of the sample size, genotype frequencies and genotype misclassification errors. My simulation study showed that the asymptotic power using this NCP predicted the simulation power. The NCP thus can be used to calculate the asymptotic power for a fixed sample size and/or sample size for a fixed power for the association test at any significance level. For smaller sample sizes, power decreased as the misclassification rate increased for fixed minor allele frequencies in cases and controls. A large sample size was required to maintain the power for small minor allele frequencies. An R script that considered the genotype misclassification at a base pair was given in Appendix A to calculate the power gain or loss based on given parameter values.

I then studied the association testing with NGS technology. I present a test statistic that allows for genotype misclassification using base pair reads directly from sequencing. This statistic can test for association with observed genotype and misclassification errors in the data. This statistic provided asymptotically unbiased estimations of genotype frequencies and genotype misclassification rates using the Bayesian posterior probability. Based on the results of simulation studies, low frequency of at-risk allele required a greater sample size than high

frequency of at-risk allele to maintain power. Also, genotyping misclassification error resulted in an increase in the required sample size to maintain power at a fixed level of significance.

In this dissertation, I noticed when the at-risk allele frequency was small, the average of test statistic was less than 2 in the simulation studies. Possible future work would be to evaluate the properties of the test statistic using permutation testing. Additionally, the NCP for the likelihood ratio test using NGS data could be developed. In this way, power and sample size calculations for any parameter settings could be determined at any significance level.

Reference

1. Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3), pp.375--386.

2. Bansal, V., Libiger, O., Torkamani, A. and Schork, N. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11), pp.773--785.

3. Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics*, 10, pp.478--486.

4. Cochran, WG (1954). Some methods for strengthening the common chi-squared tests. *Biometrics* 10 (4), pp.417--451.

5. Copel, Checkoway, H., McMichael, A. and Holbrook, R. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*, 105(5), pp.488--495.

6. Dempster, A., Laird, N., Rubin, D. and others, (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1), pp.1--38.

7. Gordon, D. and Ott, J. (2001). Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. 6, pp.18--29.

8. Gordon, D., Finch, S. and De La Vega, F. (2011). A new expectation-maximization statistical test for case-control association studies considering rare variants obtained by high-throughput sequencing. *Human heredity*, 71(2), pp.113--125.

9. Gordon, D., Finch, S., Nothnagel, M. and Ott, J. (2002). Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human heredity*, 54(1), pp.22--33.

10. Gordon, D., Haynes, C., Blumenfeld, J. and Finch, S. (2005). PAWE-3D: visualizing power for association with error in case--control genetic studies of complex traits. *Bioinformatics*, 21(20), pp.3935--3937.

11. Gordon, D., Heath, S., Liu, X. and Ott, J. (2001). A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *The American Journal of Human Genetics*, 69(2), pp.371--380.

12. Gordon, D., Levenstien, M., Finch, S. and Ott, J. (2003). Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. 8, pp.490--501.

13. Gordon, D., Yang, Y., Haynes, C., Finch, S., Mendell, N., Brown, A. and Haroutunian, V. (2004). Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling. *Statistical applications in genetics and*

molecular biology, 3(1), p.1085.

14. Ji, F., Yang, Y., Haynes, C., Finch, S. and Gordon, D. (2006). Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. *Statistical applications in genetics and molecular biology*, 4(1).

15. Kang, S., Finch, S., Haynes, C. and Gordon, D. (2005). Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. *Human heredity*, 58(3-4), pp.139--144.

16. Kang, S., Gordon, D. and Finch, S. (2004). What SNP genotyping errors are most costly for genetic association studies?. *Genetic epidemiology*, 26(2), pp.132--141.

17. Kang, S., Gordon, D., Brown, A., Ott, J. and Finch, S. (2003). Tradeoff between no-call reduction in genotyping error rate and loss of sample size for genetic case/control association studies. p.116.

18. Kim, W., Londono, D., Zhou, L., Xing, J., Nato, A., Musolf, A., Matise, T., Finch, S. and Gordon, D. (2013). Single-variant and multi-variant trend tests for genetic association with next-generation sequencing that are robust to sequencing error. *Human heredity*, 74(3-4), pp.172--183.

19. Luo, L., Boerwinkle, E. and Xiong, M. (2011). Association studies for next-generation sequencing. *Genome research*, 21(7), pp.1099--1108.

20. Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A. and others, (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747--753.

21. Mitra, S. (1958). On the limiting power function of the frequency chi-square test. *Annals of Mathematical Statistics*, 29(4), pp.1221--1233.

22. Mote, V. and Anderson, R. (1965). An Investigation of the Effect of Misclassification on the Properties of χ^2-tests in the Analysis of Categorical Data. *Biometrika*, pp.95--109.

23. Soon, W., Hariharan, M. and Snyder, M. (2013). High-throughput sequencing for biology and medicine. *Molecular systems biology*, 9(1).

24. R Development Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/.

25. Tung, L. (n.d.). *The impact of genotype misclassification errors on the power to detect a genetic association and gene-environment interaction with Cox proportional hazards*. 1st ed.

26. Wikipedia, (2014). *Genome-wide association study*. [online] Available at: http://en.wikipedia.org/wiki/Genome-wide_association_study.

Appendices

Code R

A. Non-centrality parameter from Fisher's information matrix of association test for a base pair

```
ncp_power_fisher<-function(pt10,pt11,q0,N,e,a){

q1<-1-q0

p01<-(1-e+(2*e-1)*pt10)*q0

p02<-(e-(2*e-1)*pt10)*q0

p11<-(1-e+(2*e-1)*pt11)*q1

p12<-(e-(2*e-1)*pt11)*q1

i11<-N*(2*e-1)^2*(q0^2*(1/p01+1/p02)+q1^2*(1/p11+1/p12))

i12<-N*(q0^2*(2*e-1)*(1/p01-1/p02)-q1^2*(2*e-1)^2*(1/p11+1/p12))

i21<-N*(2*e-1)*(q0^2*(1/p01-1/p02)-q1^2*(1/p11-1/p12))

i22<-N*(q0^2*(1/p01+1/p02)+q1^2*(2*e-1)*(1/p11-1/p12))

i33<-N*(1/q0+1/q1)

i13<-0

i23<-0

i31<-0

i32<-0

J<-i11-c(i12, i13)%*%solve(matrix(c(i22,i32,i23,i33),2,2))%*%c(i21,i31)

ncp<-(pt10-pt11)*J*(pt10-pt11)

power_ncp<-1-pchisq(qchisq(1-a,1),1,ncp)
```

```
return(c(ncp,power_ncp))

}
```

B. Likelihood Ratio Test allowing for misclassification using NGS data by EM algorithm

```
library(MASS)

comb<-function (n,m){

        factorial(n)/(factorial(m)*factorial(n-m))

                                        }

#a: number of settings

EMLRT <- function (n1_cases, n1_control, S, qt0, N, V, p, a){

qt1 <- 1 - qt0

N0 <- N * qt0 # Number of Controls, remains constant through each EM algorithm update

N1 <- N - N0 # Number of Cases

EMLRT_H0 <- function(n1_cases,n1_control, S, p, a) {

e_matrix_H0 <- matrix(,S,2)

p_matrix_H0 <- matrix(,S,3)

lnL0 <- c()

r_H0_vec <- c()

r_control <- c()

r_case <- c()

control_geno <- c()

case_geno <- c()

control_geno_e <- c()

case_geno_e <- c()
```

```
t00 <- c()

t01 <- c()

t02 <- c()

t10 <- c()

t11 <- c()

t12 <- c()

for (i in 1: S) { # S: starting points

#cat ("The number of starting point:", S, "\n")

r_H0 <- 0

lnL0_new <- 0

lnL0_old <- 10

                                    at0_old <- runif(1,0,1) # starting point of at0

                                    #at1_ini <- runif(1,0,1) # starting point of at1

                                    #Under H0, at0 = at1

                                    at1_old <- at0_old

                                    e0_old <- runif(1,0,0.1) # starting point of e0

# Here, assuming the symmetric error rates so that e0_21 = e0_12

                                    #e0_old <- runif(1,0,0.1) # starting point of e1

                                    e0_old <- e0_old


                                    e1_old <- runif(1,0,0.1) # starting point of e1_12

                                    e1_old <- e1_old

                                    #e1_21_old <- runif(1,0,0.1) # starting point of e1_21
```

```
pt00_old <- (1-at0_old)^2

pt01_old <- 2*at0_old*(1-at0_old)

pt02_old <- (at0_old)^2

pt10_old <- (1-at1_old)^2

pt11_old <- 2*at1_old*(1-at1_old)

pt12_old <- (at1_old)^2

flag <- 0

while (flag == 0) {

if (abs(lnL0_new - lnL0_old) > 10^(-3)) {

lnL0_old <- lnL0_new

u00_old <- (2-0)/2*e0_old + (0/2)*(1-e0_old)

u01_old <- (2-1)/2*e0_old + (1/2)*(1-e0_old)

u02_old <- (2-2)/2*e0_old + (2/2)*(1-e0_old)

u10_old <- (2-0)/2*e1_old + (0/2)*(1-e1_old)

u11_old <- (2-1)/2*e1_old + (1/2)*(1-e1_old)

u12_old <- (2-2)/2*e1_old + (2/2)*(1-e1_old)

pt0_old <- pt00_old

pt1_old <- pt01_old

pt2_old <- pt02_old


lnL0_control_old <- 0

lnL0_case_old <- 0

        for (m in 1:N0) {
```

```
            lnL0_control_old <- lnL0_control_old +

                        log((comb(V, n1_control[m])*u00_old^n1_control[m]*(1-
u00_old)^(V-n1_control[m]))*pt0_old*qt0 +

                        (comb(V, n1_control[m])*u01_old^n1_control[m]*(1-
u01_old)^(V-n1_control[m])*pt1_old*qt0) +

                        (comb(V, n1_control[m])*u02_old^n1_control[m]*(1-
u02_old)^(V-n1_control[m])*pt2_old*qt0))

                        }

    for (m in 1:N1) {

            lnL0_case_old <- lnL0_case_old +

                        log((comb(V, n1_cases[m])*u10_old^n1_cases[m]*(1-
u10_old)^(V-n1_cases[m]))*pt0_old*qt1 +

                    (comb (V, n1_cases[m])*u11_old^n1_cases[m]*(1-u11_old)^(V-
n1_cases[m])*pt1_old*qt1) +

                      (comb(V,n1_cases[m])*u12_old^n1_cases[m]*(1-u12_old)^(V-
n1_cases[m])*pt2_old*qt1))

                        }

    lnL0_new <- lnL0_control_old + lnL0_case_old

#ti,j: tao_ij_m  postior probability

    for (m in 1:N0) # control group

    {

    t00[m] <- (u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt0_old)/

            ((u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt0_old)+

             (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt1_old)+

             (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt2_old))

    t01[m] <- (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt1_old)/
```

```
                ((u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt0_old)+
                 (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt1_old)+
                 (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt2_old))
t02[m] <- (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt2_old)/
                ((u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt0_old)+
                 (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt1_old)+
                 (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt2_old))
                }
for (m in 1:N1) # case group, i=1
{
# j = 0
t10[m] <- (u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt0_old)/
                ((u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt0_old)+
                 (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt1_old)+
                 (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt2_old))

# j = 1
t11[m] <- (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt1_old)/
                ((u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt0_old)+
                 (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt1_old)+
                 (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt2_old))
# j = 2
t12[m] <- (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt2_old)/
```

```
                ((u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt0_old)+
                 (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt1_old)+
                 (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt2_old))
        }
t00_sum_old <- 0
t01_sum_old <- 0
t02_sum_old <- 0
t10_sum_old <- 0
t11_sum_old <- 0
t12_sum_old <- 0
for (m in 1:N0) {
        t00_sum_old <- t00_sum_old + t00[m]
        t01_sum_old <- t01_sum_old + t01[m]
        t02_sum_old <- t02_sum_old + t02[m]
                        }
for (m in 1:N1) {
        t10_sum_old <- t10_sum_old + t10[m]
        t11_sum_old <- t11_sum_old + t11[m]
        t12_sum_old <- t12_sum_old + t12[m]
                        }
pt0_new <- (t00_sum_old + t10_sum_old) / N
pt1_new <- (t01_sum_old + t11_sum_old) / N
pt2_new <- (t02_sum_old + t12_sum_old) / N
```

```
M1_control <- 0

M2_control <- 0

M1_case <- 0

M2_case <- 0

        for (m in 1: N0) { # Control group

                M1_control <- M1_control + n1_control[m] * t00[m] + (V - n1_control[m]) *
t02[m]

                M2_control <- M2_control + V * (t00[m] + t02[m])

                                }

        for (m in 1:N1) { # Case Group

                M1_case <- M1_case + n1_cases[m] * t10[m] + (V - n1_cases[m]) * t12[m]

                M2_case <- M2_case + V * (t10[m] + t12[m])

                                }

                e0_new <- M1_control/M2_control

                e1_new <- M1_case/M2_case

e0_old <- e0_new

e1_old <- e1_new

pt00_old <- pt0_new

pt01_old <- pt1_new

pt02_old <- pt2_new

pt10_old <- pt0_new

pt11_old <- pt1_new

pt12_old <- pt2_new

pt_new <- c(pt0_new, pt1_new, pt2_new)
```

```
r_H0 <- r_H0 + 1

} #if

else {flag <- 1}

if (r_H0 >= 100) {flag <- 1}

} # while

e_matrix_H0[i,] <- c(e0_new, e1_new)

p_matrix_H0[i,] <- c(pt0_new,pt1_new,pt2_new)

lnL0[i] <- lnL0_new

r_H0_vec[i] <- r_H0

} # for(i in 1:S)

write.table(e_matrix_H0, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/e_matrix_H0_rep", sep =
""),p,".txt", sep = ""))

write.table(p_matrix_H0, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/p_matrix_H0_rep", sep =
""),p,".txt", sep = ""))

write.table(r_H0_vec, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/r_H0_rep", sep = ""),p,".txt", sep
= ""))

write.table(lnL0, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/lnL0_rep", sep = ""),p,".txt", sep =
""))

#return(lnL0)

#cat("LRT_H0 is: ", LRT_H0, "\n")

} # End of LRT_H0

EMLRT_H1 <- function(n1_cases,n1_control, S, p, a){

e_matrix_H1 <- matrix(,S,2)
```

```
p_matrix_H1 <- matrix(,S,6)
lnL1 <- c()
r_H1_vec <- c()
#r_H1 <- 0 # count of step
for (i in 1: S) { # S: starting points under H1
#cat ("The number of starting point:", S, "\n")
r_H1 <- 0
                                    at0_old <- runif(1,0,1) # starting point of at0 under H1
                                    at1_old <- runif(1,0,1) # starting point of at1 under H1
                                    #Under H1, at0 != at1
                                    e0_old <- runif(1,0,0.1) # starting point of e0 under H1
                                    # Here, assuming the symmetric error rates so that e0_21 = e0_12
                                    #e0_old <- runif(1,0,0.1) # starting point of e1 under H1

                                    e1_old <- runif(1,0,0.1) # starting point of e1 under H1
                                    #e1_old <- runif(1,0,0.1) # starting point of e1
pt00_old <- (1-at0_old)^2
pt01_old <- 2*at0_old*(1-at0_old)
pt02_old <- (at0_old)^2
pt10_old <- (1-at1_old)^2
pt11_old <- 2*at1_old*(1-at1_old)
pt12_old <- (at1_old)^2
lnL1_new <- 0
```

```
lnL1_old <- 10

# %%%%%%%%%% Loop starts

t00 <- c()

t01 <- c()

t02 <- c()

t10 <- c()

t11 <- c()

t12 <- c()

flag <- 0

while (flag == 0) {

if (abs(lnL1_new - lnL1_old) > 10^(-3)) { # EM under H1

lnL1_old <- lnL1_new

u00_old <- (2-0)/2*e0_old + (0/2)*(1-e0_old)

u01_old <- (2-1)/2*e0_old + (1/2)*(1-e0_old)

u02_old <- (2-2)/2*e0_old + (2/2)*(1-e0_old)

u10_old <- (2-0)/2*e1_old + (0/2)*(1-e1_old)

u11_old <- (2-1)/2*e1_old + (1/2)*(1-e1_old)

u12_old <- (2-2)/2*e1_old + (2/2)*(1-e1_old)

lnL1_control_old <- 0

lnL1_case_old <- 0

        for (m in 1:N0) {

                lnL1_control_old <- lnL1_control_old +

                                log((comb(V, n1_control[m])*u00_old^n1_control[m]*(1-
u00_old)^(V-n1_control[m]))*pt00_old*qt0 +
```

```
                                (comb(V, n1_control[m])*u01_old^n1_control[m]*(1-
u01_old)^(V-n1_control[m])*pt01_old*qt0) +

                                (comb(V, n1_control[m])*u02_old^n1_control[m]*(1-
u02_old)^(V-n1_control[m])*pt02_old*qt0))

                                }

        for (m in 1:N1) {

                lnL1_case_old <- lnL1_case_old +

                                log((comb(V, n1_cases[m])*u10_old^n1_cases[m]*(1-
u10_old)^(V-n1_cases[m]))*pt10_old*qt1 +

                              (comb (V, n1_cases[m])*u11_old^n1_cases[m]*(1-u11_old)^(V-
n1_cases[m])*pt11_old*qt1) +

                                 (comb(V,n1_cases[m])*u12_old^n1_cases[m]*(1-u12_old)^(V-
n1_cases[m])*pt12_old*qt1))

                                }

        lnL1_new <- lnL1_control_old + lnL1_case_old

        for (m in 1:N0) # control group

        {

        t00[m] <- (u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt00_old)/

                ((u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt00_old)+

                 (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt01_old)+

                 (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt02_old))

        t01[m] <- (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt01_old)/

                        ((u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt00_old)+

                         (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt01_old)+

                         (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt02_old))

        t02[m] <- (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt02_old)/
```

```
                ((u00_old^n1_control[m]*(1-u00_old)^(V-n1_control[m])*pt00_old)+
                 (u01_old^n1_control[m]*(1-u01_old)^(V-n1_control[m])*pt01_old)+
                 (u02_old^n1_control[m]*(1-u02_old)^(V-n1_control[m])*pt02_old))
                }
for (m in 1:N1) # case group, i=1
{
# j = 0
t10[m] <- (u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt10_old)/
                ((u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt10_old)+
                 (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt11_old)+
                 (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt12_old))
# j = 1
t11[m] <- (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt11_old)/
                ((u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt10_old)+
                 (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt11_old)+
                 (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt12_old))
# j = 2
t12[m] <- (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt12_old)/
                ((u10_old^n1_cases[m]*(1-u10_old)^(V-n1_cases[m])*pt10_old)+
                 (u11_old^n1_cases[m]*(1-u11_old)^(V-n1_cases[m])*pt11_old)+
                 (u12_old^n1_cases[m]*(1-u12_old)^(V-n1_cases[m])*pt12_old))
        }
t00_sum_old <- 0
```

```
        t01_sum_old <- 0
        t02_sum_old <- 0
        t10_sum_old <- 0
        t11_sum_old <- 0
        t12_sum_old <- 0
        for (m in 1:N0) {
                t00_sum_old <- t00_sum_old + t00[m]
                t01_sum_old <- t01_sum_old + t01[m]
                t02_sum_old <- t02_sum_old + t02[m]
                                }
        for (m in 1:N1) {
                t10_sum_old <- t10_sum_old + t10[m]
                t11_sum_old <- t11_sum_old + t11[m]
                t12_sum_old <- t12_sum_old + t12[m]
                                }
        pt00_new <- t00_sum_old / N0
        pt01_new <- t01_sum_old / N0
        pt02_new <- t02_sum_old / N0
        pt10_new <- t10_sum_old / N1
        pt11_new <- t11_sum_old / N1
        pt12_new <- t12_sum_old / N1
e0_new <- 0
e1_new <- 0
```

```
M1_control <- 0

M2_control <- 0

M1_case <- 0

M2_case <- 0

        for (m in 1: N0) { # Control group

                M1_control <- M1_control + n1_control[m] * t00[m] + (V - n1_control[m]) *
t02[m]

                M2_control <- M2_control + V * (t00[m] + t02[m])

                                }

        for (m in 1:N1) { # Case Group

                M1_case <- M1_case + n1_cases[m] * t10[m] + (V - n1_cases[m]) * t12[m]

                M2_case <- M2_case + V * (t10[m] + t12[m])

                                }

                e0_new <- M1_control/M2_control

                e1_new <- M1_case/M2_case
e0_old <- e0_new

e1_old <- e1_new

pt00_old <- pt00_new

pt01_old <- pt01_new

pt02_old <- pt02_new

pt10_old <- pt10_new

pt11_old <- pt11_new

pt12_old <- pt12_new

pt_new <- c(pt00_new, pt01_new, pt02_new, pt10_new, pt11_new, pt12_new)
```

```
r_H1 <- r_H1 + 1}

else {flag <- 1} #if

if (r_H1 >= 100) {flag <- 1}

} # while

e_matrix_H1[i,] <- c(e0_new, e1_new)

p_matrix_H1[i,] <- c(pt00_new,pt01_new,pt02_new,pt10_new,pt11_new,pt12_new)

lnL1[i] <- lnL1_new

r_H1_vec[i] <- r_H1

} # for(i in 1:S)

write.table(e_matrix_H1, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/e_matrix_H1_rep", sep =
""),p,".txt", sep = ""))

write.table(p_matrix_H1, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/p_matrix_H1_rep", sep =
""),p,".txt", sep = ""))

write.table(r_H1_vec, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/r_H1_rep", sep = ""),p,".txt", sep
= ""))

write.table(lnL1, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/lnL1_rep", sep = ""),p,".txt", sep =
""))

} # End of the LRT_H1 function

EMLRT_H0(n1_cases,n1_control, S, p, a)

EMLRT_H1(n1_cases,n1_control, S, p, a)

} # End of EMLRT function

EMLRT_all <- function (V, N , qt0, at0_ini, at1_ini, e0_ini, e1_ini, S, R, a) {

for (p in 1:R) {
```

```
qt1 <- 1 - qt0

N0 <- N * qt0 # Number of Controls, remains constant through each EM algorithm update

N1 <- N - N0 # Number of Cases

pt00 <- (1-at0_ini)^2

pt01 <- 2*at0_ini*(1-at0_ini)

pt02 <- (at0_ini)^2

pt10 <- (1-at1_ini)^2

pt11 <- 2*at1_ini*(1-at1_ini)

pt12 <- (at1_ini)^2

e0 <- e0_ini

e1 <- e1_ini

r_control <- c()

r_case <- c()

control_geno <- c()

case_geno <- c()

control_geno_e <- c()

case_geno_e <- c()

t00 <- c()

t01 <- c()

t02 <- c()

t10 <- c()

t11 <- c()

t12 <- c()
```

```
r_control <- runif (N0,0,1)

for (j in 1:N0) { # [1

        if (r_control[j] <= pt00)          {control_geno[j] <- "00"}

        else {if ((r_control[j] > pt00) & (r_control[j] <= pt00 + pt01))         {control_geno[j] <-
"01"}

                else      {control_geno[j] <- "11"}

                }

                        }  # 1]

r_case <-  runif (N1,0,1)

for (j in 1:N1) { #[2

        if (r_case[j] <= pt10)             {case_geno[j] <- "00"}

        else {if ((r_case[j] > pt10)&(r_case[j] <= pt10 + pt11))        {case_geno[j] <- "01"}

                else {case_geno[j] <- "11"}

                }

                        } # 2]

n1_control <- c()

n1_cases <- c()

for (j in 1:N0) { #[3

        if (control_geno[j] == "00") { # no allele "1" for individual i at a single base pair position

                                                xless_controls <- rbinom(1,V,e0) # number of
misclassified "2" to "1" alleles

                                                if (xless_controls == 0) {control_geno_e[j] <- "00"}

                                                        else {if ((xless_controls > 0) &
(xless_controls < V)) {control_geno_e[j] <- "01"}

                                                                else {control_geno_e[j] <- "11"}
```

```
                                                    }

                                                n1_control[j] <- xless_controls

                                                }

        else {if (control_geno[j] == "11"){ # all allele "1" for individual i at a single base pair position #[4

                                                    xless_controls <- rbinom(1,V,1-e0) # numbers of observed less common alleles

                                                    if (xless_controls == 0) {control_geno_e[j] <- "11"}

                                                    else {if ((xless_controls > 0)&(xless_controls < V)) {control_geno_e[j] <- "01"}

                                                        else {control_geno_e[j] <- "00"}

                                                        }

                                                n1_control[j] <- xless_controls

                                                    }

                else {

                        x1_controls_e <- rbinom(1,V,(1/2)*e0 + (1/2)*(1-e0)) # Number of observed "1" alleles in control groups

                        if (x1_controls_e == 0) {control_geno_e[j] <- "00"}

                        else {if ((x1_controls_e > 0)&(x1_controls_e < V)) {control_geno_e[j] <- "01"}

                            else {control_geno_e[j] <- "11"}

                            }

                        n1_control[j] <- x1_controls_e

                        }

                } # 4]
```

```
                    } # 3]

for (j in 1:N1) { #[3

        if (case_geno[j] == "00") { # no allele "1" for individual i at a single base pair position

                                        xless_cases <- rbinom(1,V,e1) # number of misclassified "2" to "1" alleles

                                        if (xless_cases == 0) {case_geno_e[j] <- "00"}

                                                else {if ((xless_cases > 0) & (xless_cases < V)) {case_geno_e[j] <- "01"}

                                                        else {case_geno_e[j] <- "11"}

                                                        }

                                        n1_cases[j] <- xless_cases

                                        }

        else {if (case_geno[j] == "11"){ # all allele "1" for individual i at a single base pair position #[4

                                        xless_cases <- rbinom(1,V,1-e1) # number of misclassified "1" to "2" alleles

                                        if (xless_cases == 0) {case_geno_e[j] <- "11"}

                                        else {if ((xless_cases > 0)&(xless_cases < V)) {case_geno_e[j] <- "01"}

                                                else {case_geno_e[j] <- "00"}

                                                }

                                        n1_cases[j] <- xless_cases

                                        }

                else {

                        x1_cases_e <- rbinom(1,V,(1/2)*e1 + (1/2)*(1-e1))
```

```
                    if (x1_cases_e == 0) {case_geno_e[j] <- "00"}

                    else {if ((x1_cases_e > 0)&(x1_cases_e < V))
        {case_geno_e[j] <- "01"}

                        else {case_geno_e[j] <- "11"}

                        }

                    n1_cases[j] <- x1_cases_e

                    }

            } # 4]

                    } # 3]

write.table(case_geno_e, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/case_geno_e_rep", sep = ""), p,
".txt", sep = ""))

write.table(control_geno_e, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/control_geno_e_rep", sep = ""), p,
".txt", sep = ""))

u00 <- (2-0)/2*e0 + (0/2)*(1-e0)

u01 <- (2-1)/2*e0 + (1/2)*(1-e0)

u02 <- (2-2)/2*e0 + (2/2)*(1-e0)

u10 <- (2-0)/2*e1 + (0/2)*(1-e1)

u11 <- (2-1)/2*e1 + (1/2)*(1-e1)

u12 <- (2-2)/2*e1 + (2/2)*(1-e1)

# The initial log-likelihood

lnL0_control_ini <- 0

lnL0_case_ini <- 0

        for (m in 1:N0) {

            lnL0_control_ini <- lnL0_control_ini +
```

```
                                log((comb(V, n1_control[m])*u00^n1_control[m]*(1-u00)^(V-
n1_control[m]))*pt00*qt0 +

                                (comb(V, n1_control[m])*u01^n1_control[m]*(1-u01)^(V-
n1_control[m])*pt01*qt0) +

                                (comb(V, n1_control[m])*u02^n1_control[m]*(1-u02)^(V-
n1_control[m])*pt02*qt0))

                                }

        for (m in 1:N1) {

                lnL0_case_ini <- lnL0_case_ini +

                                log((comb(V, n1_cases[m])*u10^n1_cases[m]*(1-u10)^(V-
n1_cases[m]))*pt10*qt1 +

                              (comb (V, n1_cases[m])*u11^n1_cases[m]*(1-u11)^(V-
n1_cases[m])*pt11*qt1) +

                                 (comb(V,n1_cases[m])*u12^n1_cases[m]*(1-u12)^(V-
n1_cases[m])*pt12*qt1))

                                }

        lnL0_ini <- lnL0_control_ini + lnL0_case_ini

write.table(lnL0_ini, paste(paste("C:/Users/Ruiqi
Zhang/Dropbox/Research/2014.3/IncreaseSampleSize/S", a, "/lnL_ini_rep", sep = ""), p , ".txt",
sep = ""))

EMLRT(n1_cases, n1_control, S, qt0, N, V, p, a)

} # Replicates

} # End of EMLRT_all function
```