Journal of Probability and Statistics

Advanced Designs and Statistical Methods for Genetic and Genomic Studies of Complex Diseases

Guest Editors: Yongzhao Shao, Wei Pan, and Xiaohua Douglas Zhang

Advanced Designs and Statistical Methods for Genetic and Genomic Studies of Complex Diseases

Journal of Probability and Statistics

Advanced Designs and Statistical Methods for Genetic and Genomic Studies of Complex Diseases

Guest Editors: Yongzhao Shao, Wei Pan, and Xiaohua Douglas Zhang

Editorial Board

Contents

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 805426, 3 pages
doi:10.1155/2012/805426

Editorial

Advanced Designs and Statistical Methods for Genetic and Genomic Studies of Complex Diseases

Yongzhao Shao,[1] Wei Pan,[2] and Xiaohua Douglas Zhang[3]

[1] *Division of Biostatistics, New York University School of Medicine, 650 First Avenue, No. 538, New York, NY 10016, USA*
[2] *Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303 Minneapolis, MN 55455, USA*
[3] *Biometrics Research, Merck Research Laboratories, WP53B-120, West Point, PA 19486, USA*

Correspondence should be addressed to Yongzhao Shao, shaoy01@nyumc.org

Received 21 August 2012; Accepted 21 August 2012

The completion of the Human Genome Project and the International HapMap Project, coupled with rapid advancement of high-throughput biotechnologies including next-generation sequencing (NGS), has facilitated the discovery of genetic and genomic variants linked to many human diseases. Massive amounts of data from genetic and genomic studies provide a great opportunity for researchers to investigate and propose novel statistical methods and algorithms that can effectively identify disease-associated or causal genetic/genomic markers while avoiding an abundance of false positive results.

Despite many recent advances in statistical designs and methods for the analysis of genetic and genomic data on complex diseases, numerous challenges remain. For example, complex diseases including many cancers are heterogeneous in both disease phenotypes and disease etiology. The specification of disease phenotype and measurement of risk factors or environmental exposures are often subject to missing data, measurement errors, or over-simplification. Disease susceptibility is often affected by heterogeneous genetic or genomic factors including rare variants and further altered by various environmental exposures. Therefore, novel study designs and analysis methods are essential for proper adjustment of latent heterogeneity, and for robust inferences using data with possible misspecification of disease phenotypes, incompletely measured exposures, or other complexities.

This special issue is devoted to original research articles as well as overview papers that propose and discuss innovative study designs, novel probabilistic and statistical models, and analysis methods and/or algorithms for genetic and genomic studies of complex diseases. In particular, genome-wide association studies (GWAS) provide an important screening approach to identify single nucleotide polymorphisms (SNPs) and pathways that

underlie complex diseases and traits without requiring prior knowledge about disease-associated chromosomal loci or genetic functions. There are several papers in this special issue that contribute novel and innovative statistical methods for the design, analysis, and prioritization of GWAS results. The paper by, J. Zhao and Z. Chen introduces a two-stage penalized logistic regression approach to case-control genome-wide association studies. While the common practice is to examine each SNP separately, ignoring correlation among the SNPs, the proposed method takes into account correlations among the vast number of SNPs to select etiologically important SNPs. The paper by J. H. Zhao and J. Luan provides an indepth review of mixed models with whole genome data which can deal with complex dependencies introduced by known or unknown familial relationships. Controlling false discovery rate (FDR) or false discovery proportions (FDPs) is one of the most fundamental statistical issues for GWAS and other genetic and genomic studies involving testing a large number of hypotheses. Controlling FDR, as suggested by Benjamini and Hochberg among many others, has been the most widely studied approach; however, the FDR is only the mathematical expectation of the false discovery proportion (FDP) which can be more directly relevant to specific studies. Direct control of the random variable FDP has recently attracted much attention. The paper by Y. Ge and X. Li proposed an upper bound to directly control the FDP under the assumption of independence among some test statistics. The paper by S. Shang et al. develops statistical designs including sample size calculations that can control the FDP to a prescribed level and achieve some desired overall power of making a desired percentage of true discoveries under some semiparametric assumptions of weak dependence between test statistics. For design studies involving testing a vast number of hypotheses, the paper by C.-H. Tseng and Y. Shao evaluates the growth rate of the required sample size as the number of hypotheses to be tested grows rapidly from 10 to 10 billion. The paper by P. Liu and C. Wang introduces a semiparametric optimal testing (SPOT) procedure for high-dimensional data with a small sample size as arising in microarray and RNA-seq experiments. The SPOT procedure is robust because it does not depend on any parametric assumption for the alternative means. The problem of high-dimensional data with a small sample size is also tackled by the paper of S. Kwon et al. where a multinomial ordinal probit model with singular value decomposition is developed for testing a large number of single nucleotide polymorphisms (SNPs) simultaneously for association with a multidisease status or multinomial trait. Indeed, several groups of researchers have been developing statistical methods that can effectively deal with multivariate outcomes, these novel methods and algorithms are important for genetic and genomic studies and are reviewed in the paper by Q. Yang and Y. Wang. Motivated by studying the genetic basis of Huntington's diseases, T. Chen et al. propose methods for the prediction of disease onset from mutation status using proband and relative data. Using prostate cancer as a prototype example, the paper by Pearlman A. et al. provides a case study of translational research, where genomic copy number alterations (CNA) are being clustered to build a metastatic potential score towards the development of statistical prediction models for the risk of metastasis at the time of primary tumor diagnosis. The paper of H. Jia and J. Li proposes a novel computational method that combines sequence overrepresentation and cross-species sequence conservation to detect transcriptional factor binding sites (TFBSs) in the upstream regions of a given set of coregulated genes. In modeling heterogeneity for many applications, testing homogeneity is an interesting and challenging question even in parametric context. The paper by Liu et al. develops an empirical likelihood-based method for the problem of testing homogeneity in a semi-parametric two-sample problem. Missing parental genotype data is quite common for linkage analysis

particularly for late-onset diseases, the paper by J. Han and Y. Shao introduces a method for reconstruction of parental genotypes and a transmission/disequilibrium heterogeneity (RC-TDH) test for fine mapping of complex diseases. The RC-TDH test extends the current classic transmission/disequilibrium test (TDT) or RC-TDT. The paper by J. Xu reviews the high-dimensional Cox regression analysis in genetic/genomic studies with censored survival outcomes. Gene-environmental interactions are important for studying complex diseases as evidenced by the well-known fact that about 80% of lung cancer patients are smokers. However, measurements on environmental factors are often misclassified or with measurement error. The paper by I. Lobach and R. Fan proposes methods for genotype-based Bayesian analysis of gene-environment interactions with multiple genetic markers and misclassifications in environmental factors. Next-generation sequencing (NGS) has been increasingly used in genetic/genomic studies to investigate roles of rare genetic variants. The paper by T. Wang et al. introduces some novel statistical designs and analysis methods for analyzing pooled sequencing data for rare variants.

The editors of this special issue would like to thank the large number of authors who have shared with them their research achievements. They sincerely thank the large number of professional and diligent referees whose great efforts resulted in rapid reviews and useful feedbacks incorporated into the revisions of the herein published papers. Without their generous contributions this special issue would not be possible.

Yongzhao Shao
Wei Pan
Xiaohua Douglas Zhang

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 256574, 14 pages
doi:10.1155/2012/256574

Research Article

The Transmission Disequilibrium/Heterogeneity Test with Parental-Genotype Reconstruction for Refined Genetic Mapping of Complex Diseases

Jing Han and Yongzhao Shao

Division of Biostatistics, NYU School of Medicine, New York University, 650 First Avenue, 5th Floor, New York, NY 10016, USA

Correspondence should be addressed to Yongzhao Shao, yongzhao.shao@nyumc.org

Received 2 March 2012; Accepted 1 May 2012

Academic Editor: Xiaohua Douglas Zhang

In linkage analysis for mapping genetic diseases, the transmission/disequilibrium test (TDT) uses the linkage disequilibrium (LD) between some marker and trait loci for precise genetic mapping while avoiding confounding due to population stratification. The sib-TDT (S-TDT) and combined-TDT (C-TDT) proposed by Spielman and Ewens can combine data from families with and without parental marker genotypes (PMGs). For some families with missing PMG, the reconstruction-combined TDT (RC-TDT) proposed by Knapp may be used to reconstruct missing parental genotypes from the genotypes of their offspring to increase power and to correct for potential bias. In this paper, we propose a further extension of the RC-TDT, called the reconstruction-combined transmission disequilibrium/heterogeneity (RC-TDH) test, to take into account the identical-by-descent (IBD) sharing information in addition to the LD information. It can effectively utilize families with missing or incomplete parental genetic marker information. An application of this proposed method to Genetic Analysis Workshop 14 (GAW14) data sets and extensive simulation studies suggest that this approach may further increase statistical power which is particularly valuable when LD is unknown and/or when some or all PMGs are not available.

1. Introduction

Genetic linkage analysis is an important step in localizing and identifying genes in the chromosomes that underlie many human diseases and other traits of interest. A brief overview of commonly used statistical methods for linkage analysis including recently developed model-free and model-based methods for mapping qualitative- and quantitative-trait loci, can be found in Shao [1]. For more extensive discussions on linkage analysis, readers can consult Ott [2].

Mapping genes that underlie complex diseases is of great current interest. The essence of linkage analysis is to identify statistical association between the inheritance of a complex genetic disease phenotype and inheritance of specific pieces of genetic material (called marker alleles). Many complex diseases including cancers have an inheritable component. For marker alleles that are associated with inheritance of complex diseases, it is common that the transmission probabilities of a marker allele of interest vary across heterozygous parents, due to locus heterogeneity, etiological heterogeneity, and many other complexities and/or combinations of them [3, 4]. Under such transmission heterogeneity, the transmission likelihood generally has the form of mixture models with many parameters [4, 5]. It can be shown that the efficient score test of such mixture likelihood includes two parts, one part related to transmission disequilibriums reflected by existence of linkage disequilibrium (LD) and the other related to transmission heterogeneity in the form of excessive dispersion in sharing of genetic markers as might be inferred from identical by descent (IBD) patterns (e.g., allele-sharing patterns among affected sib-pairs).

The transmission/disequilibrium test (TDT) developed by Spielman et al. [6] uses the LD information between some marker and disease loci for precise genetic mapping while avoiding confounding due to population stratification. It has been extended in multiple directions to meet the need for mapping complex traits, for example [7, 8]. In particular, missing parental genetic marker genotypes are very common for studying diseases with late onset. The sib-TDT (S-TDT) and combined-TDT (C-TDT) proposed by Spielman and Ewens [9] can deal with families without parental marker genotypes (PMGs) and can combine with data from families having PMG available. For some families with missing PMG, the reconstruction-combined TDT (RC-TDT) proposed by Knapp [10, 11] may be used to reconstruct missing PMG from the genotypes of their offspring to increase power of the C-TDT with a correction for potential bias in using reconstructed PMG [12].

An attractive feature of the RC-TDT is that it utilizes the missing PMG that can be uniquely determined from the genotypes of the children and corrects potential biases resulting from using reconstructed PMG by employing appropriate null expectation and variance, supplied in Tables 1 and 2 of Knapp [10]. Similar to the TDT and C-TDT, the RC-TDT is powerful only when there is strong LD. Usually LD is unknown, and it is difficult to measure, thus it is generally desirable to combine LD information with information on allele sharing obtained based on IBD patterns [5, 13].

For fine mapping of complex genetic disorders, Shao [4] derived a general mixture likelihood for allele transmission under various transmission disequilibrium and/or heterogeneity and further proposed a *transmission disequilibrium/heterogeneity* (TDH) test to efficiently combine the transmission disequilibrium and heterogeneity information to maximize the power for detecting linkage using genetic data from nuclear families. The TDH test was shown to be an efficient score test of the general mixture likelihood derived in Shao [4] which is a summation of two parts, a *transmission/disequilibrium test* (TDT) part which utilizes the LD information and a *transmission heterogeneity test* (THT) part that utilizes IBD-sharing information. To see that the THT utilizes IBD-sharing information, it should be pointed out that general mixture likelihood contains the mixture binomial likelihood discussed in Huang and Jiang [13] and Lo et al. [5], and the test statistic of the classical mean test for affected sib-pairs (ASPs) is a special case of the THT statistic with $n_{ai} = 2$ in Shao [4]. The classical mean test for affected sib-pairs is the most well-known IBD sharing-based linkage test [14]. The THT is applicable to general sibship and thus can be regarded as an extension of the classical mean test for affected sib-pairs.

Table 1: Moments of T_i under H_0.

PMG	ET_i	$\mathrm{Var}(T_i)$	$\mathrm{Var}[(T_i - ET_i)^2]$
AB/AA	$3n_{ai}/2$	$n_{ai}/4$	$n_{ai}(n_{ai}-1)/8$
AB/AB	n_{ai}	$n_{ai}/2$	$n_{ai}(2n_{ai}-1)/4$
AB/BB	$n_{ai}/2$	$n_{ai}/4$	$n_{ai}(n_{ai}-1)/8$

Table 2: Distribution of T_i when one PMG is missing but reconstructible.

PMG	Condition (R)	Range of T_i	$P_{H_0}(T_i = c \mid R)$
$AB \times AB$	$N^{AA} > 0$ and $N^{BB} > 0$	$0 \le c < n_a$	$\dfrac{\binom{2n_a}{c}(1/2)^{2n_a} - \binom{n_a}{c}(1/2)^{2n_a-c}(3/4)^{n_u}}{1-(1/2)^{n_c}[2(3/2)^{n_c}-1]}$
		$c = n_a$	$\dfrac{\binom{2n_a}{n_a}(1/2)^{2n_a} - (1/2)^{n_c}[2(3/2)^{n_u}-1]}{1-(1/2)^{n_c}[2(3/2)^{n_c}-1]}$
		$n_a < c \le 2n_a$	$\dfrac{\binom{2n_a}{c}(1/2)^{2n_a} - \binom{n_a}{c-n_a}(1/2)^{c}(3/4)^{n_u}}{1-(1/2)^{n_c}[2(3/2)^{n_c}-1]}$
$AA \times AB$	$N^{AA} > 0$ and $N^{AB} > 0$	$c = n_a$	$\dfrac{\binom{n_a}{c-n_a}(1/2)^{n_a} - (1/2)^{n_c}}{1-2(1/2)^{n_c}}$
		$n_a < c < 2n_a$	$\dfrac{\binom{n_a}{c-n_a}(1/2)^{n_a}}{1-2(1/2)^{n_c}}$
		$c = 2n_a$	$\dfrac{\binom{n_a}{c-n_a}(1/2)^{n_a} - (1/2)^{n_c}}{1-2(1/2)^{n_c}}$
$BB \times AB$	$N^{AB} > 0$ and $N^{BB} > 0$	$c = 0$	$\dfrac{\binom{n_a}{c}(1/2)^{n_a} - (1/2)^{n_c}}{1-2(1/2)^{n_c}}$
		$0 < c < n_a$	$\dfrac{\binom{n_a}{c}(1/2)^{n_a}}{1-2(1/2)^{n_c}}$
		$c = n_a$	$\dfrac{\binom{n_a}{c}(1/2)^{n_a} - (1/2)^{n_c}}{1-2(1/2)^{n_c}}$

In practice, parental marker genotypes are often incomplete for many genetic studies particularly for late onset diseases. Only using families with complete parental maker genotype information would lead to throwing away a large portion of the useful data and can also lead to biases. It is thus crucially important to make the TDH test applicable to families with missing or incomplete parental marker genotype information. In this paper, we develop a transmission disequilibrium/heterogeneity test with parental-genotype reconstruction, which utilizes both the LD information and the IBD-sharing information and can combine families with or without PMG information.

The transmission disequilibrium/heterogeneity test with parental-genotype reconstruction (RC-TDH) will be introduced in the next section. In Section 3, the RC-TDH test is applied to a data set from GAW14, and the results are compared with those of the RC-TDT. Finally, simulation studies that use common genetic models [5, 15] are carried out to compare the power and the true size of the RC-TDT and RC-TDH test. The numerical results suggest that RC-TDH test may greatly increase the statistical power which is particularly valuable whenever LD levels are unknown and/or whenever there is missing PMG information as in studying of a disease with late age of onset.

It should be pointed out that the main comparison made in this paper will be between RC-TDT and RC-TDH. We will not formally compare them with the classical IBD-based linkage tests such as those implemented in Genehunter and other softwares. The main rationale is as follows. We are mainly interested in fine mapping of genetic variants that underlie complex diseases, where the classical linkage tests are known to have low power because they do not utilize LD information effectively. With the rapid advancement of biotechnology, it is now feasible and affordable to use dense genetic markers, for example, the single nucleotide polymorphisms (SNPs), for genomewide linkage scan. With a large number of dense genetic markers (e.g., SNPs) some of the markers can be expected to fall into the LD block of the causal genetic variants; thus LD would generally exist to some degree for many markers. Thus the TDT and TDH tests would have power advantage over classical linkage tests which only effectively utilize the IBD information.

2. Method

2.1. Notation

It will be assumed that there are two alleles A and B at the marker locus, and allele A is of particular interest. Let n_{ai} denote the number of affected children, let n_{ui} denote the number of unaffected children, and let $n_{ci} = n_{ai} + n_{ui}$ denote the size of the sibship for family i. In each family, all children have been typed at the marker locus, but the PMG may or may not be available. Let $N_{ai}^{g}(N_{ui}^{g})$ be random variables, denoting the number of affected (or unaffected) children with genotype g in family i. Small letters (i.e., n_{ai}^{g} and n_{ui}^{g}) are used to denote the observed values of N_{ai}^{g} and N_{ui}^{g}. Further, let $N_{i}^{g} = N_{ai}^{g} + N_{ui}^{g}$ and $n_{i}^{g} = n_{ai}^{g} + n_{ui}^{g}$ denote the random variable and the observed number of children with genotype g in family i, respectively. T_i denotes the number of A alleles in affected children (i.e., $T_i = 2N_{ai}^{AA} + N_{ai}^{AB}$). The notation introduced here is consistent with Knapp [10, 11] and Han [16].

2.2. The TDH Test with Complete PMG

For completeness, we first consider the case when PMG are observed along with children's marker genotypes. Let x_i be the number of alleles A transmitted by the ith marker heterozygous parent to the affected children. When the exact number x_i of marker alleles A

transmitted to affected children cannot be determined as might happen in families with two heterozygous parents, then T_i can be used to replace x_i. Using T_i in families with ambiguous transmissions, the TDT statistic can be written as $T_D = T_d^2$ where

$$T_d = \sum_i \frac{T_i - ET_i}{\sqrt{\sum_i \mathrm{Var}(T_i)}}. \tag{2.1}$$

The transmission heterogeneity test (THT) statistic is denoted as $T_H = T_h^2$ where

$$T_h = \frac{\max\left\{\sum_i \left[(T_i - ET_i)^2 - \mathrm{Var}(T_i)\right], 0\right\}}{\sqrt{\sum_i \mathrm{Var}\left[(T_i - ET_i)^2\right]}}, \tag{2.2}$$

where the moments of T_i under H_0 given the parental marker genotypes (PMGs) are summarized in Table 1.

The transmission disequilibrium/heterogeneity (TDH) test is based on the following test statistic [4]:

$$T_{DH} = T_D + T_H. \tag{2.3}$$

In terms of statistical optimality, it can be shown that the TDH test is the efficient score test from the mixture likelihood function under transmission disequilibrium and heterogeneity [4]. In theory, the efficient score test is known to be locally most powerful.

2.3. The Reconstruction-Combined TDH (RC-TDH) Test

When at least one parent with missing PMG, Knapp [10] proposed a reconstruction-combined TDT (RC-TDT) to reconstruct PMG from the genotypes of their offspring and correct for the biases resulting from using reconstructed PMG. To improve the power to detect linkage, we propose the reconstruction-combined TDH test (RC-TDH) using the following test statistic:

$$\frac{[\sum(T_i - e_i)]^2}{\sum v_i} + \frac{\left[\max\left\{\sum\left((T_i - e_i)^2 - E_{H_0}\left[(T_i - e_i)^2 \mid R\right]\right), 0\right\}\right]^2}{\sum \mathrm{Var}_{H_0}\left[(T_i - e_i)^2 \mid R\right]}, \tag{2.4}$$

where T_i denotes the number of marker alleles A in affected children, and $e_i = E_{H_0}(T_i \mid R)$, $v_i = \mathrm{Var}_{H_0}(T_i \mid R)$ denote the appropriate null expectation and variance of T_i, respectively, as can be found in Tables 1 and 2 of Knapp [10]. In the RC-TDH statistic, the first term is the RC-TDT statistic of Knapp [10] and the second term is the RC-THT statistic with the restriction. To get the appropriate null expectation $\mathrm{Var}_{H_0}[(T_i - E_{H_0}(T_i \mid R))^2 | R]$, we need to derive the conditional distribution of T_i given the constraint for reconstruction R.

When one parental genotype is missing and reconstructible, the conditional probabilities of T_i are listed in Table 2. Note that the family index i has been dropped in the formula in Table 2. In the first column, the first parental genotype is typed and the second

one is reconstructed. The second column presents a necessary and sufficient condition, for the observed marker genotypes in the offspring, to allow reconstruction of the parental genotypes. The details of the derivation are provided in Han [16].

When both parental genotypes are missing, the reconstruction condition and the conditional probabilities of T_i are the same as that of one parental genotype is missing and the known parental genotype is AB.

When at least one parental genotype is missing and cannot be reconstructed, but the condition for the S-TDT is satisfied (i.e., there is at least one affected and at least one unaffected child in this family, not all of the children possess the same genotype), the distribution of T_i can be calculated using the affected and unaffected children genotypes by the hypergeometric distribution. The details are provided in the Appendix section.

As in C-TDT and RC-TDT, families not belonging to the previous categories will be ignored.

3. Application to Genetic Analysis Workshop 14 Data

The proposed RC-TDH test was applied to a Genetic Analysis Workshop 14 (GAW14) dataset to compare the power with that of RC-TDT. The GAW14 simulated data were generated by Dr. David Greenberg. A behavioral disorder has been simulated in multiple replicates of four different populations/groups. There are 100 families in the Aipotu, Karnagar, and Danacaa data sets. There are 100 replicates for each data set. The results of power comparison of RC-TDH with RC-TDT to analyze the linkage between the trait b disease allele and the marker B01T0561 are presented in Table 3. This trait has incomplete penetrance with f_{DD} = 30%. Application of the RC-TDH is illustrated in Table 3 with 50% and 100% missing parental genotypes. The power is based on type I error at 0.05 level.

4. Simulation

4.1. Simulation Set-Up

Simulation studies are conducted to compare the powers of the proposed RC-TDH test with the RC-TDT. To attain the correct type I error rates, we directly simulated the critical values under the null hypothesis of no linkage, in which θ (recombination frequency) = 0.5. In the simulations for the null distribution, 1,000,000 replicates of samples of nuclear families are generated and the empirical critical values are obtained. Based on 500 independent replicates and the empirical critical values, we estimate the power of the tests using the relative frequencies of the simulated test statistics which exceed the empirical critical values.

To generate the family-based data, as in earlier work [5], we consider two biallelic loci: one disease locus (with disease allele D and normal allele d) and one marker locus (with allele A and B). The frequency for disease allele D is p_D and for marker allele A is p_A. The linkage disequilibrium is the deviation of the frequency of DA haplotype from its equilibrium value (expected by chance). Define the LD parameter as

$$\Delta = \frac{p_{DA} - p_D \cdot p_A}{\min\left(p_D \cdot p_B,\ p_d \cdot p_A\right)}. \tag{4.1}$$

In our simulations, we assume A is the allele in LD with D. Thus, the range of the LD parameter Δ is in $[0, 1]$, in which 0 indicates linkage equilibrium. There are three penetrance parameters, f_{DD}, f_{Dd}, and f_{dd}, corresponding to three possible disease genotypes.

Table 3: Power comparison of the RC-TDH test with RC-TDT using GAW14 data.

Population	100% PMG Missing		50% PMG Missing	
	RC-TDT	RC-TDH	RC-TDT	RC-TDH
Aipotu	0.27	0.58	0.57	1.00
Karnagar	0.14	0.33	0.46	1.00
Danacaa	0.37	0.86	0.74	1.00

In the study of 100% PMG missing, we ignore all the parental marker genotypes. In the study of 50% PMG missing, we use 50% families with parental marker genotypes and 50% families without parental marker genotypes.

Simulation study 1 closely followed the approach used by Boehnke and Langefeld [15]. For each model, a disease prevalence K_p of 5% was assumed. The disease allele frequency p that resulted from each of the disease models can be calculated by $K_p = p^2 f_{DD} + 2p(1-p)f_{Dd} + (1-p)^2 f_{dd}$. Summary of the parameters used in this simulation study is in Table 4.

Summary of the parameters used in simulation study 2 is in Table 5. Four commonly used disease models are used here: dominant ($f_{Dd} = f_{DD}$), additive ($f_{Dd} = (f_{DD} + f_{dd})/2$), multiplicative ($f_{Dd} = \sqrt{f_{DD} \cdot f_{dd}}$), and recessive ($f_{Dd} = f_{dd}$) models.

4.2. Simulation Results

Table 6 presents estimates of the critical values for RC-TDH at significance levels of .05, .01, and .001. Table 7 presents the estimates of the true type I error rate, at nominal significance levels of .05, .01, and .001. The simulations support the validity of approximating the null distribution with a standard normal distribution for RC-TDT.

The results of simulation study 1 are shown in Table 8. The disease models are denoted by "D," "A," and "R" for the mode of inheritance (i.e., dominant, additive, and recessive); "1" and "2" for the value of f_{DD} (i.e., 1.0 and 0.5). The presented results come from the simulations with 4 sibs in each family, which have the same trend as those with 2 or 6 sibs in each family. In instances for which there is no parental genotype information available, application of the RC-TDH instead of the RC-TDT results in a consistent gain of power, especially when linkage disequilibrium is weak.

We conducted simulation study 2 to compare the power of the proposed RC-TDH test with that of RC-TDT according to linkage disequilibrium in different scenarios based on Table 5, such as tight linkage versus weak linkage, full penetrance versus incomplete penetrance. Each simulated sample consists of families with an identical number of sibs (n_c) in each family (with $n_c = 3$), which are ascertained on the basis of the presence of an affected child. Each sample consists of a total of 600 children. Half of the 200 families have complete PGM, and half of the families without PGM. To assess the power of the tests, 500 replicate samples are generated, under different simulation scenarios. For each replicate sample, the statistics obtained with the proposed RC-TDH and with the RC-TDT were calculated.

To compare power of the RC-TDH with that of the RC-TDT at different LD levels, we set the range of LD between 0 and 1, recombination fraction at 0.01, the frequency of allele D at 0.1, the frequency of allele A at 0.5, penetrance for genotype DD at full penetrance 1, penetrance for genotype dd at 0.01, and then the penetrance for genotype Dd can be determined by the modes of inheritance. The results in Table 9 and Figure 1 show that the power increases with LD, and the proposed RC-TDH is more powerful than RC-TDT, especially when LD is weak as in scenario 1 of Table 4.

Table 4: Parameters used in simulation study 1.

Scenario	Mode	p_D	p_A	f_{DD}	f_{dd}	f_{Dd}
1	Dominant	0.013	0.4	1.0	0.025	1.000
2		0.016	0.4	0.8	0.025	0.800
3		0.027	0.4	0.5	0.025	0.500
4		0.074	0.4	0.2	0.025	0.200
5	Additive	0.026	0.4	1.0	0.025	0.513
6		0.032	0.4	0.8	0.025	0.413
7		0.053	0.4	0.5	0.025	0.263
8		0.143	0.4	0.2	0.025	0.113
9	Recessive	0.160	0.4	1.0	0.025	0.025
10		0.180	0.4	0.8	0.025	0.025
11		0.229	0.4	0.5	0.025	0.025
12		0.378	0.4	0.2	0.025	0.025

Table 5: Parameters used in simulation study 2.

Scenario	θ	p_A	p_D	f_{dd}	f_{DD}	f_{Dd}			
						Dom	Rec	Add	Mul
S1	.01	.50	.10	.01	1.0	1.0	.01	.505	.100
S2	.10	.50	.10	.01	1.0	1.0	.01	.505	.100
S3	.01	.10	.10	.01	1.0	1.0	.01	.505	.100
S4	.10	.10	.10	.01	1.0	1.0	.01	.505	.100
S5	.01	.50	.10	.01	0.5	0.5	.01	.255	.071
S6	.10	.50	.10	.01	0.5	0.5	.01	.255	.071
S7	.01	.10	.10	.01	0.5	0.5	.01	.255	.071
S8	.10	.10	.10	.01	0.5	0.5	.01	.255	.071

Penetrance is the conditional probability of observing a phenotype given a specified disease genotype. In scenario 1, we set f_{DD} (the penetrance for a subject whose marker genotype is DD) at 1, which is an idealistic penetrance. To compare the power of the proposed RC-TDH with that of its competitor under different penetrance, f_{DD} is varied from full penetrance to incomplete penetrance 0.5, which is more realistic. The results in Table 9 and Figure 2 show that the proposed RC-TDH has better power than RC-TDT with half penetrance for genotype DD individuals as in scenario 5 of Table 5.

In summary, our simulation results show that the proposed RC-TDH is generally more powerful than RC-TDT for a broad range of LD, the tightness of the linkage, and across disease models.

5. Discussion

For mapping complex diseases, it is common that the transmission probabilities of a marker allele of interest vary across heterozygous parents, due to locus heterogeneity, etiological heterogeneity, and many other complexities and/or combinations of them [3, 4]. Under such transmission heterogeneity, the transmission likelihood generally has the form of mixture models with many parameters, and the efficient score test has two parts in the form of a TDH test [4]. This paper studies a TDH test which allows the inclusion of reconstructed

Table 6: Simulated critical values for RC-TDH.

Sibship size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
3	5.23	8.59	14.04
4	5.40	9.08	15.43
6	5.52	9.82	16.91

Note: determined on the basis of the dominant model with $f_{DD} = 0.2$ (Scenario 4 in Table 4).

Table 7: Simulated true type I error rates of the RC-TDT and of RC-TDH.

Sibship size	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.001$	
	RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH
3	0.0490	0.0502	0.0094	0.0100	0.0008	0.0010
4	0.0485	0.0499	0.0097	0.0099	0.0010	0.0010
6	0.0503	0.0497	0.0101	0.0100	0.0008	0.0010

Determined on the basis of the dominant model with $f_{DD} = 0.2$ (scenario 4 in Table 4).

Table 8: Powers of RC-TDT and RC-TDH in simulation study 1.

Model	$\Delta = 0.1$		$\Delta = 0.5$		$\Delta = 0.9$	
	RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH
$D1$	0.13	0.87	0.68	0.97	0.99	1.00
$D2$	0.08	0.41	0.61	0.73	0.97	0.98
$A1$	0.10	0.43	0.65	0.77	0.97	0.97
$A2$	0.09	0.16	0.56	0.59	0.98	0.96
$R1$	0.21	0.86	0.99	1.00	1.00	1.00
$R2$	0.15	0.40	0.98	0.99	1.00	1.00

D (dominant), R (recessive), A (additive); f_{DD}: 1 (1.0), 2 (0.5); with type-I error rate .05 based on 500 independent replicates of 150 nuclear families. Δ is the measurement for linkage disequilibrium. When $\Delta = 0$, there is no linkage disequilibrium. In this simulation study, all the parental marker genotypes are missing.

Table 9: Powers of the RC-TDT and RC-TDH in simulation study 2.

Scenario	Δ	Dominant		Recessive		Additive		Multiplicative	
		RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH
S1	0.0	0.00	0.65	0.00	0.37	0.00	0.19	0.00	0.13
	0.2	0.10	0.82	0.02	0.53	0.03	0.35	0.01	0.25
	0.4	0.57	0.96	0.37	0.85	0.33	0.72	0.25	0.63
	0.6	0.98	1.00	0.86	1.00	0.87	0.98	0.78	0.94
	0.8	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S5	0.0	0.00	0.17	0.00	0.04	0.00	0.01	0.00	0.01
	0.2	0.02	0.27	0.00	0.09	0.01	0.10	0.00	0.06
	0.4	0.25	0.62	0.04	0.29	0.16	0.43	0.06	0.24
	0.6	0.79	0.93	0.22	0.53	0.65	0.81	0.40	0.62
	0.8	0.99	1.00	0.61	0.89	0.96	0.99	0.81	0.94
	1.0	1.00	1.00	0.89	0.99	1.00	1.00	0.97	0.99

In this simulation, we used 50% families with available parental marker genotypes and 50% families without parental marker genotypes.

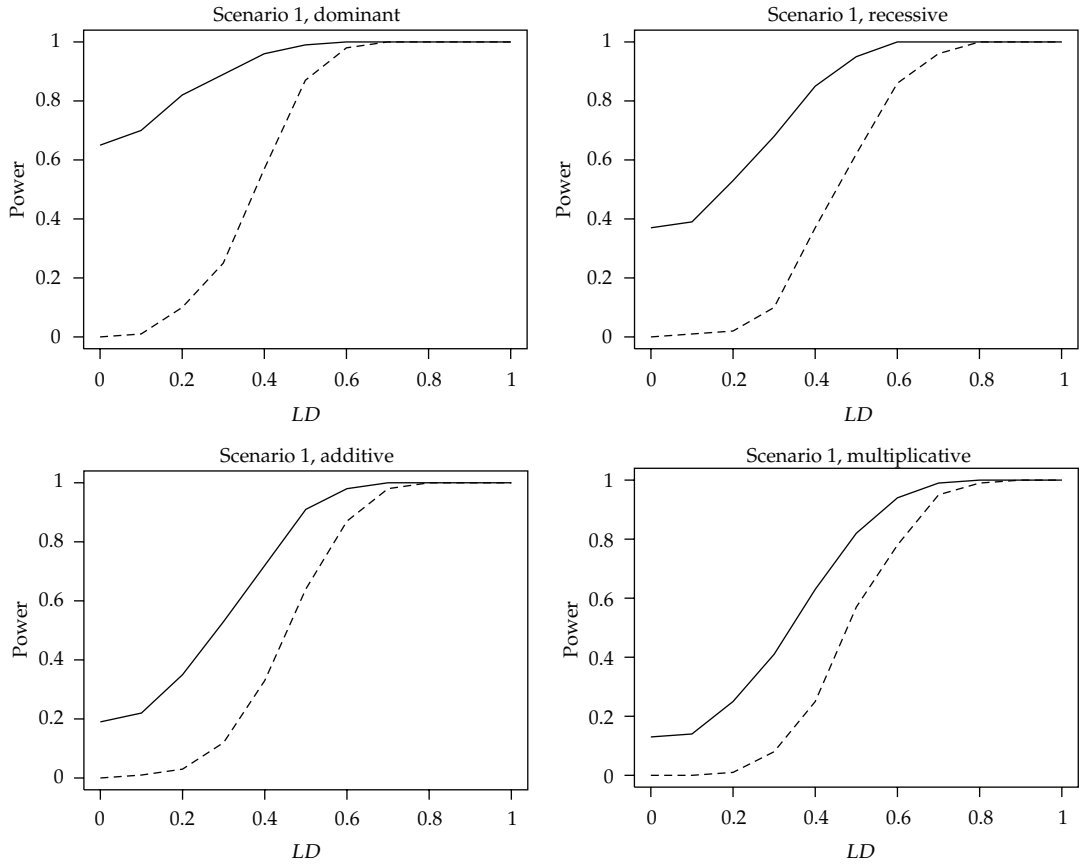


Figure 1: Power of RC-TDH (solid) and RC-TDT (dashed) in Table 5 (scenario 1). This figure is based on scenario 1: $\theta = 0.01$, $p_D = 0.1$, $p_A = 0.5$, $f_{DD} = 1$ and $f_{dd} = 0.01$. The type I error rate is 0.001 based on 500 independent replicates of 200 nuclear families, 50% of which without parental information. Every family contains 3 sibs and at least one is affected. *LD* is the measurement for linkage disequilibrium as defined by Δ in Section 4.1. When $LD = 0$, there is no linkage disequilibrium.

parental marker genotype data and extends the RC-TDT of Knapp [10, 11]. The proposed new approach was validated by simulation studies and GAW14 data sets, and the results indicate that the new approach might improve the power of family-based linkage analysis for a broad range of *LD*. Moreover, the simulation studies also indicate that the systematic power advantage of the RC-TDH test over the RC-TDT holds regardless of the underlying genetic models (e.g., recessive, dominant, additive, multiplicative).

Similar to RC-TDT, the new approach can utilize the missing parental information that can be reconstructed from the child genotypes, especially including some families with genotype-concordant or phenotype-concordant sibs. In addition, the proposed test is a sibship-oriented method which does not require specification of the underlying genetic model; it naturally uses the multiple siblings by considering the sibship as a whole. The second part of the RC-TDH statistic, the THT part of the test statistic, is based on information from IBD. This is quite obvious in the situation of affected sib-pairs, where the THT is essentially equivalent to the so-called mean test [4, 13].

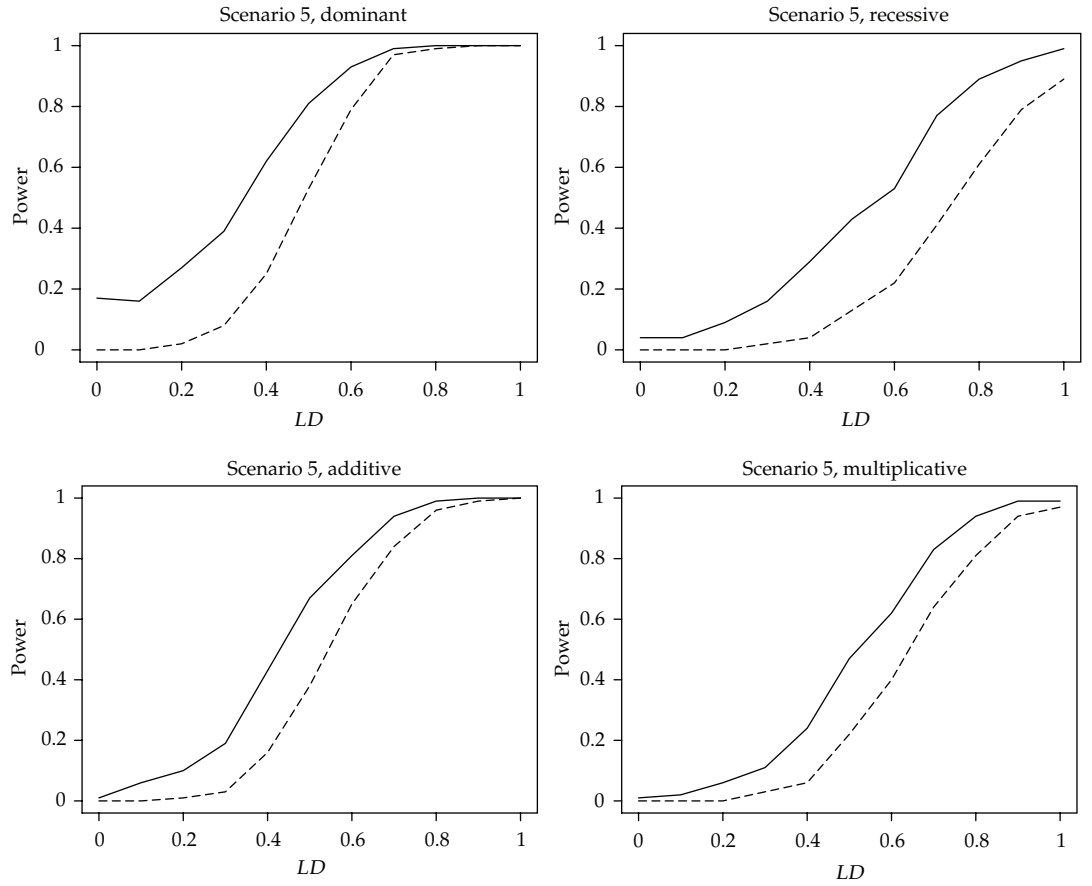


Figure 2: Power of RC-TDH (solid) and RC-TDT (dashed) in Table 5 (scenario 5). This figure is based on scenario 5: $\theta = 0.01$, $p_D = 0.1$, $p_A = 0.5$, $f_{DD} = 0.5$ and $f_{dd} = 0.01$. The type I error rate is 0.001 based on 500 independent replicates of 200 nuclear families, 50% of which without parental information. Every family contains 3 sibs and at least one is affected. *LD* is the measurement for linkage disequilibrium as defined by Δ in Section 4.1. When $LD = 0$, there is no linkage disequilibrium.

Many other linkage analysis tests such as the tests implemented by Genehunter have relatively low power with respect to TDT or TDH when *LD* is present. In reality, some degree of *LD* is often present particularly when we use dense genetic markers (e.g., SNPs) along the genome because they are available at increasingly cheaper cost, and these dense markers are already very affordable. With a large number of dense genetic markers, some markers may be expected to fall into the *LD* block of the causal variants. When using these affordable dense markers along the genome or candidate gene regions, we believe that RC-TDH will have better chance of success than the classical IBD-based linkage methods in detecting linkage signals along the genome.

As high density SNP arrays become increasingly affordable to researchers, genome-wide linkage studies are becoming common. Our TDH test has simple closed form test statistics which is computationally easy in addition to good overall power across a broad range of *LD*. Thus the proposed method would be potentially useful for genomewide linkage analysis. In contrast, likelihood ratio test for mixture likelihood is generally computationally

intensive [5, 17]. Many existing linkage tests and algorithms such as the likelihood ratio test discussed in Lo et al. [5] would be too computationally intensive for genomewide studies or when the number of genotyped markers is large.

It is possible to further extend the method to be applicable to markers with more than two alleles, which would be of great interest in studying haplotypes of multiple loci. However, our proposed tests are already applicable to the commonly used biallelic markers; for instance, the widely used single nucleotide polymorphisms (SNPs) are convenient biallelic markers.

Appendix

A. Computational Details for the RC-TDH Test

When there are no parents who have been typed, the conditional probability has been derived in equation (A.6) of Knapp [10]. When only one parent has been typed as AB, the same constraint for reconstruction applies, thus (A.6) of Knapp [10] also works. Next we derive the the conditional probability when only one parent has been typed as AA. The case of when only one parent has been typed as BB is obvious due to symmetry between A and B.

A.1. One Parental Genotype Has Been Typed as AA

Note that the family index i has been dropped in the following formula.

Only one parental genotype has been typed, which is AA, but the genotype of the missing parent can be reconstructed as AB, if there is at least one child with genotype AB and at least one child with genotype AA. Here, the condition R is $N^{AB} > 0$ and $N^{AA} > 0$. To calculate the conditional distribution of T, we first calculate the probability of satisfying the constraint for reconstruction, R:

$$\begin{aligned} P_{H_0}(R) &= P_{H_0}\left(N^{AA} > 0 \text{ and } N^{AB} > 0\right) \\ &= 1 - P_{H_0}\left(N^{AA} = 0\right) - P_{H_0}\left(N^{AB} = 0\right) + P_{H_0}\left(N^{AA} = 0 \text{ and } N^{AB} = 0\right) \\ &= 1 - 2\left(\frac{1}{2}\right)^{n_c}. \end{aligned} \tag{A.1}$$

Then we calculate the joint probability of T and R:

$$\begin{aligned} P_{H_0}(\{T = c\} \cap R) &= P_{H_0}\left(T = c \cap N^{AA} > 0 \cap N^{AB} > 0\right) \\ &= P_{H_0}(T = c) - P_{H_0}\left(T = c \cap \left(N^{AA} = 0 \cup N^{AB} = 0\right)\right) \\ &= P_{H_0}(T = c) - P_{H_0}\left(T = c \cap N^{AA} = 0\right) - P_{H_0}\left(T = c \cap N^{AB} = 0\right) + 0 \\ &= \binom{n_a}{c - n_a}\left(\frac{1}{2}\right)^{n_a} - P_{H_0}\left(T = c \cap N^{AA} = 0\right) - P_{H_0}\left(T = c \cap N^{AB} = 0\right). \end{aligned} \tag{A.2}$$

There are three cases for the calculation:

case 1: $c = n_a$, $P_{H_0}(\{T = c\} \cap R) = \binom{n_a}{c-n_a}(1/2)^{n_a} - (1/2)^{n_c}$,

case 2: $n_a < c < 2n_a$, $P_{H_0}(\{T = c\} \cap R) = \binom{n_a}{c-n_a}(1/2)^{n_a}$,

case 3: $c = 2n_a$, $P_{H_0}(\{T = c\} \cap R) = \binom{n_a}{c-n_a}(1/2)^{n_a} - (1/2)^{n_c}$.

Therefore the distribution of T conditioned on R is

$$P_{H_0}(T = c \mid R) = \begin{cases} \dfrac{\binom{n_a}{c-n_a}(1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}, & c = n_a, \\ \dfrac{\binom{n_a}{c-n_a}(1/2)^{n_a}}{1 - 2(1/2)^{n_c}}, & n_a < c < 2n_a, \\ \dfrac{\binom{n_a}{c-n_a}(1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}, & c = 2n_a. \end{cases} \tag{A.3}$$

A.2. At Least One Parental Genotype Is Missing and Cannot Be Reconstructed, but the Condition for the S-TDT Is Satisfied

In a sibship with a affected and u unaffected sibs, the total number of sibs is $t = a+u$. Suppose that in this sibship the number of sibs who are of genotype AA is r and the number of sibs who are of genotype AB is s. Let x be the number of AA sibs and let y be the number of AB sibs who are classified as affected. As discussed in Spielman and Ewens [9], given the totals r, s, a, u, and t, the numbers x, y can be regarded as two entries in a 2×3 contingency table with marginal totals a, u, r, s, and $t - r - s$. Therefore, the distribution of $T = 2x + y$ can be obtained by the generalized hypergeometric distribution [18, page 47]. More specifically, we have

$$P(T = c) = \sum_{i=\max(c-2a,c-2r,0)}^{\min(s,a,c)} \frac{\binom{r}{(c-i)/2} \cdot \binom{s}{i} \cdot \binom{t-r-s}{a-((c+i)/2)}}{\binom{t}{a}}, \qquad 1 \le c \le \min(2r + s, 2a). \tag{A.4}$$

More formulas of parental marker genotype reconstruction probabilities under various missing genotypes types and constraints, as well as detailed derivations of these formulas, can be found in Han [16].

Acknowledgments

This research was partially supported by a Stony Wold-Herbert Foundation grant, the MPD Research Consortium Project Grant (1P01 CA108671), and the New York University Cancer Center Supporting Grant (2P30 CA16087) and by the NYU NIEHS Center Grant (5P30 ES00260). The research of JH was carried out as part of her Ph.D. dissertation work at New York University.

References

[1] Y. Shao, "Linkage Analysis," in *Encyclopedia of Quantitative Risk Analysis and Assessment*, John Wiley & Sons, Hoboken, NJ, USA, 2008.

[2] J. Ott, *Analysis of Human Genetic Linkage*, Johns Hopkins University, 3rd edition, 1999.

[3] E. S. Lander and N. J. Schork, "Genetic dissection of complex traits," *Science*, vol. 265, no. 5181, pp. 2037–2048, 1994.

[4] Y. Shao, "Adjustment for transmission heterogeneity in mapping complex genetic diseases using mixture models and score tests," *Proceeding of the American Statistical Association*, pp. 383–393, 2005.

[5] S. H. Lo, X. Liu, and Y. Shao, "A marginal likelihood model for family-based data," *Annals of Human Genetics*, vol. 67, no. 4, pp. 357–366, 2003.

[6] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *American Journal of Human Genetics*, vol. 52, no. 3, pp. 506–516, 1993.

[7] H. Zhao, "Family-based association studies," *Statistical Methods in Medical Research*, vol. 9, no. 6, pp. 563–587, 2000.

[8] W. J. Ewens and R. S. Spielman, "The transmission/disequilibrium test," in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop, and C. Cannings, Eds., John Wiley & Sons, 2nd edition, 2003.

[9] R. S. Spielman and W. J. Ewens, "A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test," *American Journal of Human Genetics*, vol. 62, no. 2, pp. 450–458, 1998.

[10] M. Knapp, "The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test," *American Journal of Human Genetics*, vol. 64, no. 3, pp. 861–870, 1999.

[11] M. Knapp, "Using exact P values to compare the power between the reconstruction- combined transmission/disequilibrium test and the sib transmission/disquilibrium test," *American Journal of Human Genetics*, vol. 65, no. 4, pp. 1208–1210, 1999.

[12] D. Curtis, "Use of siblings as controls in case-control association studies," *Annals of Human Genetics*, vol. 61, no. 4, pp. 319–333, 1997.

[13] J. Huang and Y. Jiang, "Linkage detection adaptive to linkage disequilibrium: the disequilibrium maximum-likelihood-binomial test for affected-sibship data," *American Journal of Human Genetics*, vol. 65, no. 6, pp. 1741–1759, 1999.

[14] W. C. Blackwelder and R. C. Elston, "A comparison of sib-pair linkage tests for disease susceptibility loci," *Genetic Epidemiology*, vol. 2, no. 1, pp. 85–97, 1985.

[15] M. Boehnke and C. D. Langefeld, "Genetic association mapping based on discordant sib pairs: the discordant-alleles test," *American Journal of Human Genetics*, vol. 62, no. 4, pp. 950–961, 1998.

[16] J. Han, *Family-based linkage analysis allowing for missing parental information [Ph.D. thesis]*, New York University, 2005.

[17] X. Liu and Y. Shao, "Asymptotics for likelihood ratio tests under loss of identifiability," *The Annals of Statistics*, vol. 31, no. 3, pp. 807–832, 2003.

[18] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley & Sons, New York, NY, USA, 3rd edition, 1968.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 524724, 19 pages
doi:10.1155/2012/524724

Research Article

Design and Statistical Analysis of Pooled Next Generation Sequencing for Rare Variants

Tao Wang,[1] Chang-Yun Lin,[2] Yuanhao Zhang,[3] Ruofeng Wen,[3] and Kenny Ye[1]

[1] *Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, NY 10461, USA*
[2] *Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan*
[3] *Department of Applied Mathematics and Statistics, Stony Brook University, New York, NY 11794, USA*

Correspondence should be addressed to Tao Wang, tao.wang@einstein.yu.edu

Received 23 March 2012; Accepted 6 June 2012

Academic Editor: Wei T. Pan

Next generation sequencing (NGS) is a revolutionary technology for biomedical research. One highly cost-efficient application of NGS is to detect disease association based on pooled DNA samples. However, several key issues need to be addressed for pooled NGS. One of them is the high sequencing error rate and its high variability across genomic positions and experiment runs, which, if not well considered in the experimental design and analysis, could lead to either inflated false positive rates or loss in statistical power. Another important issue is how to test association of a group of rare variants. To address the first issue, we proposed a new blocked pooling design in which multiple pools of DNA samples from cases and controls are sequenced together on same NGS functional units. To address the second issue, we proposed a testing procedure that does not require individual genotypes but by taking advantage of multiple DNA pools. Through a simulation study, we demonstrated that our approach provides a good control of the type I error rate, and yields satisfactory power compared to the test-based on individual genotypes. Our results also provide guidelines for designing an efficient pooled.

1. Introduction

An understanding of the role of genetic variants in human diseases provides valuable insights into the etiology of diseases. Next generation sequencing (NGS), also known as massively parallel sequencing, is a revolutionary technology for biomedical research [1]. The production of large numbers of low-cost reads makes NGS useful for many applications.

One of the most important applications is to identify DNA variants responsible for human diseases [2]. However, it is still extremely expensive and time consuming to sequence individual genomes of a large number of individuals required to achieve reasonable statistical power for identifying disease variants of common diseases [3, 4]. The yields of a typical single run of NGS are very high (e.g., more than 20 billion bases can be obtained routinely for illumina genome analyzer (GA) II)). Indeed, the throughput of the smallest functional unit, for example, a single "lane" of a sequencer can generate data amounting to many thousand-fold coverage for a small target region of interest, which is far greater than what is needed for genotyping one individual as the individual genotype at a specific position is expected to be accurately called at about 15–30 fold coverage. In this case, it is much more efficient to simultaneously sequence multiple targeted regions of many individuals.

To maximize the power of next generation sequencer, one technology that allows sequencing multiple DNA samples together is bar-coding. Bar-coding ligates the DNA fragments of each sample to a short, sample-specific DNA sequence, and then sequences these DNA fragments from multiple subjects in one single sequencing run [5]. However, the cost on sequencing template preparation could be high for bar-coding if the construction of libraries and bar-coding must be applied to each individual before sequencing. Another approach called "DNA sudoku" is designed to ascertain rare variants by assigning each individual to multiple DNA pools, and decoding the identity of rare variants based on a certain pooling scheme [6–8]. Because the number of pools required for "DNA sudoku" to infer the identity of a rare variant could be much smaller than the number of subjects, the cost on the sequencing and preparation of templates is reduced.

For screening disease association of genetic variants, an alternative approach simply sequences pooled DNA samples of cases and controls, respectively. The idea of this approach is based on comparing the estimated allele frequencies between cases and controls without actually inferring individual genotypes. Compared to bar-coding or "DNA Sudoku", the pooling approach is even more cost- and time-efficient and hence well-suited for screening disease variants. The savings on cost and time come from two sources. The first is that estimating the allele frequency requires much less depth of coverage per individual than that required for calling the genotype of each individual [7]. The second is the reduced efforts in library preparation for a large number of DNA samples. The pooling approach was proposed earlier for high throughput SNP arrays [9–11], but it was not widely accepted as SNP array technology does not provide the required accuracy of the estimate of the allele frequency in the pooled sample. NGS technology, however, has been demonstrated that it could provide an accurate estimate of the allele frequency, as shown by recent studies [12–14]. Another advantage of pooled sequencing is that it has the potential to detect rare variants which may explain the "missing" heritability that are not identified by current array technologies.

To identify disease association by pooled NGS, several key issues have to be addressed by the study design or the analytic approach. NGS has relatively high levels of base-calling errors, which are highly specific to genomic positions as well as experimental runs [15]. Although the average error rate is likely less than 1% after filtering out bases with low quality, the consequence of sequencing error is still not negligible in pooled sequencing, particularly because sequencing errors and the bases of a true variant could confound each other in a large DNA pool. As such, simply applying some filters to eliminate sequencing errors may lead to falsely remove bases of the true variant. Instead of trying to eliminate sequencing errors, this issue could be addressed by using efficient experiment designs and appropriate analytic approaches so that the locus-specific sequencing error rate can be accurately estimated and

incorporated into statistical models of testing disease association. It was shown by us and others that disease association can be validly and efficiently examined when the sequencing error parameters can be correctly specified [16, 17]. However, current statistical approaches often assume that the sequencing error parameters in the statistical models are known [16] or able to be estimated by using an internal control, such as a segment of plasmid DNA, in the pool [13, 17]. Because of the high variation in the error rates between genomic positions as well as different runs/lanes of NGS instruments, it is not adequate to use the average error rate for adjusting for the bias, potentially causing either inflated type I error rates or loss in statistical power [16, 17]. How to accurately estimate position- and lane-specific error rates relies on efficient pooled sequencing designs.

One major advantage of the sequencing technology over the SNP array is that it can ascertain novel rare variants that are not present on the array panels. However, it is well known that the power of testing association of rare variants individually is very limited due to the low occurrence of the rare alleles. To improve the statistical power, many statistical approaches have been proposed to simultaneously test a group of rare variants over recent years. Among them, the "collapsing" approach defines a score for each individual by the unweighted or weighted sum of the rare variant alleles of multiple positions in the targeted region. This approach essentially increases the "allele frequency" by pooling multiple variants, and hence improves the power, but the power of this approach relies on the assumption that all rare variant alleles have effects in the same direction [18–23]. To avoid such an assumption that is often not realistic, other approaches such as the test statistic based on the genomic distance and C-alpha test were also proposed [24, 25]. Nevertheless, these approaches all require individual genotypes for accounting for the linkage disequilibrium (LD) among multiple variants. Because LD information is largely lost in pooled sequencing, how to test disease association of a group of rare variants is still an open question.

In this paper, we proposed blocked pooling design combining bar-coding and pooling sequencing, along with a new multivariant testing procedure, for testing disease association of rare variants. We conducted a simulation study to examine the performance of the new approach under various situations.

2. Methods

2.1. Blocked Pooling Design

Sequencing error is the major concern of pooled sequencing because it has a significant impact on the validity and efficiency of testing disease association [17]. Because sequencing errors and bases of true variant alleles confound each other in a single DNA pool, it is often too difficult to differentiate them to obtain an accurate estimate of the sequencing error rate and the allele frequency. Nevertheless, if the sequencing error rates across multiple pools are consistent, a more accurate estimate may be obtained by combining data of multiple DNA pools. To understand sequencing error rates across multiple pools, we have conducted a study that used the GA II system to sequence the pooled mitochondria DNA (mtDNA) from 20 subjects, whose mtDNA had been sequenced previously using Sanger dideoxy sequencing on an ABI3730XL [15]. The pooled mtDNA samples were multiplexed by bar-coding at 2 pools per lane and replicated in another lane on a different flow cells. Using the results of Sanger sequencing as the reference, the data suggested that locus-specific base-calling error rates are quite consistent between two pools multiplexed in one lane, but vary between two

lanes in different flow cells. In addition, sequencing error rates across genomic positions have a significant variation. Although the majority of positions have an error rate lower than 1%, it can be as high as 20%, suggesting that the use of the average error rate of all genomic positions to account for sequencing error is not adequate in testing disease association, even when such an error rate is estimated from a segment of plasmid DNA as the internal control [13, 16, 17].

Based on data from the pooled mtDNA sequencing study, we propose to combine pooling and bar-coding approaches to sequence multiple DNA pools of cases and controls in one lane with each pool indexed. This experiment design can be looked upon as blocked design, which is known to improve the statistical validity as well as power, in particular, when a large variability between blocks (here lanes) is present [26]. With multiple pools indexed in one lane, the sequencing errors are largely consistent across multiple indexed pools, while bases of the true variant alleles may vary because different numbers of alleles are likely sampled in different pools. The idea of blocked pooling design is that each pool can serve as the control of other pools in the same block to eliminate effects of sequencing errors, and eventually improve the validity and efficiency of testing disease association. Furthermore, an unbalanced pooling design (different sizes of pools) could be considered to obtain an even more accurate estimation of the sequencing error rate. For example, a pool with a single individual and another pool with a large number of individuals can be multiplexed in one lane. In this design, the pool with one individual serves as the control for accurately estimating the sequencing error rate, while the pool with a large number of individuals provides the data for accurately estimating the allele frequency. The pool with a small number of individuals could provide a more accurate estimate of the sequencing error rate because of the large difference between the allele frequency (e.g., 0, 0.5, or 1 for one individual) and the sequencing error rate. In the ideal situation in which there are no sequencing errors, the balanced pooling design provides the most efficient estimate of the allele frequency because of the consistent depth of coverage for each individual. However, in the presence of sequencing errors it is necessary to balance between estimating the allele frequency and estimating the sequencing error rate to obtain an optimal association result. We empirically evaluated the importance of parameters of blocked designs in terms of the bias and standard error (SE) of the estimate of the sequencing error rate and the allele frequency.

2.2. Estimating the Sequencing Error Rate and Testing Association of Single Variants

For a case-control study, let the phenotype of a subject be denoted by $i = 1, 0$ for cases or controls, respectively. We are interested in the question of whether the variant allele is associated with disease. Let θ_i be the allele frequency of the group i. The statistical hypothesis of association can be tested by examining if cases have a different frequency of the variant allele from controls, which could be written as $H_0 : \theta_1 = \theta_0$ versus $H_1 : \theta_1 \neq \theta_0$.

Let n_{ij} be the total number of chromosomes and let v_{ij} be the number of the variant alleles at a locus of interest for the jth pool of ith group. For a pooled sequencing, v_{ij} is unknown and has to be estimated from sequencing reads. We assume that cases and controls are assigned in L_1 and L_0 pools, respectively, indexed in a single sequencing lane. After sequencing, m_{ij} sequencing bases at the locus are observed, and x_{ij} out of m_{ij} bases report the variant allele for the jth pool. To estimate the sequencing error rate (e) and the allele

frequency (θ), we consider a simple EM algorithm, given by

(0) Initial $\theta^{(0)}$ and $e^{(0)}$,

(1) E step

$$w_j = p(v_j \mid x_j) = \frac{\binom{n_j}{v_j}\theta^{(0)v_i}(1-\theta^{(0)})^{n_j-v_j}\binom{m_j}{x_j}\mathcal{A}}{\sum_0^{n_j}\binom{n_j}{v_j}\theta^{(0)v_j}(1-\theta^{(0)})^{n_j-v_j}\binom{m_j}{x_j}\mathcal{A}}, \tag{2.1}$$

where $\mathcal{A}$ donates
$[v_j/n_j(1-e^{(0)}) + (1-v_j/n_j)e^{(0)}]^{x_j}\{1-[v_j/n_j(1-e^{(0)}) + (1-v_j/n_j)e^{(0)}]\}^{m_j-x_j}$,

(2) M step

$$\theta^{(1)} = \frac{\left[\sum_{j=1}^{L}\sum_{v_j=0}^{n_j}(w_j v_j)\right]}{\sum_{j=1}^{L} n_j}, \tag{2.2}$$

$$e^{(1)} = \frac{\sum_{j=1}^{L}\sum_{v_j=0}^{n_j} w_j(x_j - v_j m_j/n_j)}{\sum_{j=1}^{L} m_j}, \tag{2.3}$$

(3) Iteratively update θ and e until converge.

For testing disease association of a rare variant, we have proposed a simple testing procedure based on a parametric bootstrap (PB), which is defined by the following steps:

(1) estimating the sequencing error rate ($\hat{e}$) and allele frequency ($\hat{\theta}$) of DNA pools under the null hypothesis by the above EM algorithm;

(2) calculating the test statistic $T = \sum(n_{1j}/\sum n_{1j})x_{1j}/m_{1j} - \sum(n_{0j}/\sum n_{0j})x_{0j}/m_{0j}$;

(3) sampling $\tilde{x}_i = (\tilde{x}_{01}, \ldots, \tilde{x}_{11}, \ldots)$ and calculating the test statistic $\tilde{T}$. First, the number of the variant alleles for each pool is sampled from $Binom(n_{ij}, \hat{\theta})$; $\tilde{x}_{ij}$ is then sampled from $Binom(m_{ij}, v_{ij}/n_{ij}(1-\hat{e}) + (1-v_{ij}/n_{ij})\hat{e})$; and lastly $\tilde{T}$ is calculated based on $\tilde{x}_i$;

(4) replicating (3) many times and estimating the P value by the proportion of $|\tilde{T}| > |T|$.

2.3. *Testing Association of Multiple Rare Variants*

Because the statistical power to detect disease association of rare variants individually is often limited, it is useful to jointly test association of a group of rare variants, for example, rare variants in an exon or a gene. Our test statistic is based on P values of individual variants. Let $p_r (r = 1, \ldots, R)$ be the P value for variant r. The test statistic is defined by

$$z = \frac{a^T Z}{\sqrt{a^T a}}, \tag{2.4}$$

where $Z = (Z_1, \ldots, Z_R)^T$ in which the element $Z_r = \Phi^{-1}(p_r)$ is the corresponding upper-tail Z score transformed from the P value, and $a = (a_1, \ldots, a_R)^T$ in which a_r is the weight given to variant r. Because the functional information of each variant is usually not available, a reasonable approach is to give equal weights to all variants because it is not prejudiced about which variants are expected to be more relevant to disease. Of note, this test statistic is in spirit close to many test statistics based on individual genotypes, such as the VEGAS statistic [27], Empirical Bayesian score statistic proposed by Goeman (2006) [28], the statistic based on the genomic distance [29], the logistic kernel machine based test statistic [30, 31], as well as C-alpha test [24].

When multiple rare variants are in linkage equilibrium (no correlation), the statistic follows a standard normal distribution. The question is that, when multiple rare variants are in LD, the P value cannot be obtained based on a standard distribution. The permutation procedure randomly shuffling the disease status is often used to account for the correlation among genetic variants. However, such a procedure requires individual genotypes that are not available in pooled sequencing. Instead, we can take a Monte Carlo approach by simulating the test statistics of individual variants under the null hypothesis from multivariate normal distribution to evaluate the P value. In this approach, we simulate the multivariate normally distributed vector with mean 0 and covariance Σ, the $R \times R$ matrix of pair-wise correlations. To do this, we use the Cholesky decomposition: a vector of R independent, standard normally distributed random variables is first generated; then it is multiplied by the Cholesky decomposition matrix of Σ. The simulated test statistic is calculated based on the multivariate normally distributed vector. A large number of multivariate normal vectors are simulated, and the empirical P value is defined by the proportion of simulated test statistics that exceed the observed test statistic .

The statistical challenge is how to estimate the covariance matrix Σ without individual genotypes available. By treating single pools as the sample unit, we estimate the covariance matrix based on the number of variant alleles of pools, instead of the number of alleles of individuals. One option is the standard unbiased empirical covariance matrix $\widehat{S}$ with entries defined as $s_{rr'} = (1/(L_0 + L_1 - 1)) \sum_{j=1}^{L_0+L_1} (\widehat{v}_{jr} - \overline{v}_r)(\widehat{v}_{jr'} - \overline{v}_{r'})$, which $\widehat{v}_{jr}$ is the estimated number of variant alleles of the rth variant in the jth pool. However, this unbiased estimate is known to be inefficient, particularly because the number of the pools is often relatively small. Because rare mutations usually occur on different haplotypes within a target region [32], and therefore their correlations are often low. This motivated us to use an empirical Bayesian shrinkage estimate of the covariance, which may provide better balance between efficiency and bias [33]. The proposed shrinkage estimate is in the following form:

$$S^* = \lambda I_{R\times R} + (1 - \lambda)\widehat{S}, \tag{2.5}$$

where $\lambda = (\sum_{r \neq r'} \widehat{var}(s_{rr'}) + \sum_r \widehat{var}(s_{rr'}))/(\sum_{r \neq r'} s_{rr'}^2 + \sum_r (s_{rr'} - 1)^2)$ is the shrinkage intensity. The idea of this empirical Bayesian estimate is that, when the data do not provide evidence of correlation of variants, the estimate is shrunk toward an identity matrix, the possibly efficient estimator under the assumption of independency of variants. Of note, this estimate is essentially equivalent to that proposed by Schafer [34].

2.4. Simulations

We conducted a simulation study to examine the impact of varied parameters of pooled designs on the estimation of the sequencing error rate and the allele frequency as well as the

test of disease association in terms of validity and efficiency. For each replicate, the pooled sequencing reads of each pool were simulated in the following two-steps: the individual genotypes were first generated under Hardy-Weinberg equilibrium; the sequencing reads of each pool were then generated independently. The sample size was set at 500 cases and 500 controls; individuals were included in different numbers of DNA pools under either the balanced or the unbalanced designs. For the unbalanced designs, one half of the pools included single subjects, and the remaining individuals were evenly assigned to the other half of pools. We set the numbers of reads were consistent cross pools. The type I error rate and power will be evaluated by the proportion of replicates having a P value that is less than a significant level of 0.05. For each simulated situation, the process was repeated for 1,000 replicates.

The performance of the PB test was examined for testing single variants for different types of designs under different allele frequencies (1% and 5%), sequencing error rates (0.5% and 1%), depths of coverage per chromosome (5, 10, and 20×), and numbers of pools (2, 10 and 40). To evaluate the type I error rate, we simulated the sequencing reads under the null hypothesis of no association, in which circumstance the cases and controls have the same allele frequency. For comparison, we also considered a Naïve Fisher's (FN) exact test that is based on the estimated allele frequency without taking sequencing errors into account, Fisher's exact test based on the estimated allele frequency with taking sequencing error into account (FE), and Fisher's exact test based on the true individual genotypes (FT). For the FN test, the number of variant alleles is directly estimated by the proportion of reads that report the variant allele. For the FE test, the number of the variant alleles is based on the allele frequency estimated by the EM algorithm; and the FT test assumes that the genotype of each individual is known and hence the number of the variant alleles can be simply counted. To evaluate the power, we fixed the allele frequency in controls, but allowed the allele frequency in cases to vary in order to yield different effect sizes.

The performance of the PB test was then examined for testing multiple variants. Different numbers of variants and varied correlations were considered. To simulate correlated variants, a set of variables was sampled from multivariate normal distribution with mean 0 and covariance Σ, which had equal pair-wise correlations ($\rho = 0$ or 0.5). The haplotype was generated by dichotomizing the normal variables based on the allele frequencies of cases and controls. The genotypes in each DNA pool were randomly sampled from a large number of haplotypes, and reads for each variant were then sampled independently. We examined multivariant tests based on three different estimates of covariance matrix [the unbiased empirical covariance estimate (E), the independent matrix (I), and the shrinkage estimate (S)] and compared them to the single-variant test with Bonferroni correction (min$P$).

3. Results

3.1. Estimating the Sequencing Error Rate and the Allele Frequency

Table 1 presents results for estimating the sequencing error rate and the allele frequency for balanced and unbalanced pooled sequencing designs under different sequencing depths of coverage, numbers of pools, allele frequencies, and sequencing error rates. As expected, the unbalanced pooling design had smaller bias of the estimate of the sequencing error rate than the balanced design. For example, when the allele frequency and the sequencing error rate were both 1% and the number of pools was 10, the bias of the unbalanced design was <0.0001,

Table 1: Estimated bias and standard error (SE) of the sequencing error rate and the allele frequency for the unbalanced and balanced designs under different sequencing depths of coverage, numbers of pools, allele frequencies, and sequencing error rates.

Depths of coverage	Pool number	θ	e	Unbalanced design				Balanced design			
				Bias (e)	SE (e)	Bias (θ)	SE (θ)	Bias (e)	SE (e)	Bias (θ)	SE (θ)
5×	2	0.01	0.005	0.0000	0.0015	0.0002	0.0029	−0.0011	0.0007	0.0011	0.0019
			0.01	−0.0001	0.0020	0.0001	0.0034	−0.0063	0.0014	0.0062	0.0025
		0.05	0.005	−0.0002	0.0014	0.0002	0.0046	−0.0011	0.0056	0.0010	0.0062
			0.01	−0.0002	0.0019	0.0002	0.0051	−0.0066	0.0054	0.0061	0.0060
	10	0.01	0.005	−0.0001	0.0014	0.0003	0.0028	−0.0009	0.0024	0.0011	0.0029
			0.01	0.0000	0.0021	0.0003	0.0034	−0.0074	0.0027	0.0078	0.0033
		0.05	0.005	0.0002	0.0014	−0.0004	0.0047	0.0084	0.0116	−0.0076	0.0108
			0.01	0.0000	0.0020	−0.0001	0.0051	0.0031	0.0127	−0.0032	0.0119
	40	0.01	0.005	0.0002	0.0014	−0.0001	0.0027	0.0007	0.0017	−0.0003	0.0017
			0.01	0.0001	0.0020	0.0001	0.0034	−0.0005	0.0033	0.0010	0.0038
		0.05	0.005	0.0019	0.0015	−0.0018	0.0046	0.0098	0.0063	−0.0078	0.0058
			0.01	0.0009	0.0020	−0.0005	0.0048	0.0090	0.0079	−0.0073	0.0072
10×	2	0.01	0.005	−0.0001	0.0010	0.0002	0.0020	−0.0006	0.0002	0.0006	0.0012
			0.01	−0.0001	0.0014	0.0001	0.0025	−0.0056	0.0001	0.0055	0.0014
		0.05	0.005	−0.0001	0.0010	0.0000	0.0033	0.0002	0.0035	−0.0002	0.0038
			0.01	−0.0001	0.0014	0.0004	0.0035	−0.0049	0.0033	0.0045	0.0039
	10	0.01	0.005	0.0000	0.0010	0.0000	0.0020	−0.0003	0.0017	0.0004	0.0019
			0.01	0.0000	0.0014	0.0000	0.0024	−0.0074	0.0023	0.0076	0.0029
		0.05	0.005	0.0000	0.0010	0.0001	0.0032	0.0041	0.0093	−0.0037	0.0087
			0.01	−0.0001	0.0014	0.0001	0.0037	−0.0010	0.0101	0.0009	0.0093
	40	0.01	0.005	0.0000	0.0009	0.0001	0.0017	0.0006	0.0010	−0.0001	0.0008
			0.01	0.0000	0.0014	0.0002	0.0023	0.0002	0.0015	0.0001	0.0011
		0.05	0.005	0.0006	0.0011	−0.0004	0.0032	0.0035	0.0021	−0.0021	0.0024
			0.01	0.0003	0.0016	0.0000	0.0035	0.0025	0.0032	−0.0015	0.0032
20×	2	0.01	0.005	−0.0001	0.0006	0.0001	0.0014	−0.0003	0.0001	0.0003	0.0009
			0.01	−0.0001	0.0010	0.0001	0.0017	−0.0053	0.0000	0.0052	0.0010
		0.05	0.005	0.0000	0.0007	0.0001	0.0024	0.0000	0.0023	0.0000	0.0027
			0.01	−0.0001	0.0010	0.0001	0.0025	−0.0051	0.0014	0.0047	0.0021
	10	0.01	0.005	−0.0001	0.0007	0.0001	0.0014	−0.0002	0.0011	0.0003	0.0011
			0.01	0.0000	0.0010	0.0000	0.0017	−0.0074	0.0026	0.0076	0.0032
		0.05	0.005	0.0000	0.0007	0.0000	0.0023	0.0003	0.0034	−0.0002	0.0035
			0.01	0.0000	0.0010	0.0001	0.0026	−0.0047	0.0042	0.0043	0.0042
	40	0.01	0.005	0.0001	0.0006	0.0000	0.0010	0.0004	0.0007	0.0000	0.0004
			0.01	0.0000	0.0009	0.0001	0.0013	0.0002	0.0009	0.0000	0.0005
		0.05	0.005	0.0002	0.0008	−0.0002	0.0022	0.0019	0.0013	−0.0006	0.0014
			0.01	0.0002	0.0011	−0.0001	0.0024	0.0009	0.0019	−0.0003	0.0017

while the bias of the balanced design was −0.0074. In addition, the SE of the sequencing error rate of the unbalanced design was comparable to that of the balanced design. Interestingly, the bias of the allele frequency of the unbalanced design was also smaller than the balanced design and their SEs were comparable. Surprisingly, the bias and SE of both the sequencing error rate and the allele frequency were not significantly improved by increasing the number of pools from 2 to 40. As expected, the bias and SE of the sequencing error rate and the allele frequency tended to decrease with an increasing sequencing coverage.

3.2. Testing for Single Variants

3.2.1. Type I Error Rate

The empirical type I error rate at a significance level of 0.05 is shown in Table 2. In general, the FT test tended to be overconservative when the allele frequency was low. When the depth of coverage is relatively low (5×), the FE test often had a very poor control of type I error rate for both the unbalanced and balanced designs, partially because the variance of the estimate of the number of variant alleles is not negligible due to the low depth of coverage. The FN test was either overliberal or overconservative because it ignores both the sequencing error and the variation of the estimate of the number of variant alleles. Table 2 indicates that the type I error rate of the PB test was consistently close to the nominal level of 0.05 for the unbalanced design, while it could be either liberal or conservative for the balanced design, which is likely due to that the balanced design could not provide an accurate estimate of the sequencing error rate and the allele frequency under low depths of coverage. With an increased depth of sequencing coverage (10× and 20×), the FE test had an improved control of the Type I error rate for the unbalanced design, while it was still a little conservative for the balanced design. The FN test tended to be more conservative for both the balanced and unbalanced designs with an increasing depth of coverage. The PB test consistently kept a good control of the type I error for the unbalanced design.

3.2.2. Power

We only evaluated the power of the PB test for the unbalanced design because the balanced design did not provide a good control of the type I error rate. As the reference, the FT test that assumes individual genotypes are observed was compared.

Figure 1 shows the empirical power of the PB test for testing association of single variants. Because of the confounding effect of the sequencing error as well as the uncertainty of the estimate of the number of variants allele in a pool, the PB test was generally less powerful than the FT test. However, the loss in power was reduced with a decreasing sequencing error rate or an increasing sequencing depth of coverage. The power of the PB test was not significantly different between various numbers of pools, in particular for the numbers of pools were 10 and 40. The difference in power between the PB test and the FT test seemed more obvious for a more common variant, which could be due to the conservativeness of the FT test for testing relatively rare variants. The results of two versions of Fisher's exact test based on the estimated number of the variant alleles were not presented here, because they generally have a poor control of the type I error rate. Nevertheless, after adjusting for the inflated type I rate they tend to be less powerful than the proposed PB test,

Table 2: Type I error rates at a level of 5% for the PB test and Fisher's exact tests under various depths of coverage, numbers of pools, allele frequencies, and the error rates for testing association. Sample size was set at 500 cases and 500 controls.

Depth of coverage	Pool number	θ	e	Unbalanced design				Balanced design			
				PB	FT	FN	FE	PB	FT	FN	FE
5×	2	0.01	0.005	0.046	0.033	0.041	0.096	0.041	0.036	0.027	0.017
			0.01	0.044	0.041	0.024	0.16	0.013	0.028	0.008	0.004
		0.05	0.005	0.053	0.041	0.072	0.086	0.059	0.037	0.049	0.113
			0.01	0.047	0.044	0.062	0.091	0.059	0.042	0.052	0.092
	10	0.01	0.005	0.048	0.026	0.04	0.111	0.056	0.039	0.025	0.065
			0.01	0.057	0.039	0.028	0.152	0.02	0.026	0.012	0.033
		0.05	0.005	0.051	0.037	0.066	0.083	0.09	0.035	0.052	0.302
			0.01	0.046	0.033	0.06	0.093	0.076	0.05	0.048	0.317
	40	0.01	0.005	0.05	0.029	0.032	0.082	0.049	0.028	0.019	0.044
			0.01	0.049	0.03	0.02	0.122	0.072	0.05	0.017	0.133
		0.05	0.005	0.054	0.032	0.065	0.073	0.07	0.044	0.047	0.043
			0.01	0.055	0.032	0.065	0.098	0.083	0.033	0.05	0.088
10×	2	0.01	0.005	0.045	0.033	0.02	0.066	0.051	0.034	0.01	0.018
			0.01	0.04	0.037	0.012	0.107	0.032	0.037	0.011	0.013
		0.05	0.005	0.055	0.042	0.056	0.063	0.057	0.053	0.03	0.055
			0.01	0.039	0.042	0.034	0.056	0.04	0.045	0.033	0.051
	10	0.01	0.005	0.05	0.032	0.016	0.072	0.061	0.045	0.021	0.062
			0.01	0.045	0.037	0.016	0.098	0.012	0.031	0.006	0.032
		0.05	0.005	0.051	0.046	0.052	0.067	0.066	0.042	0.044	0.168
			0.01	0.048	0.039	0.04	0.071	0.059	0.042	0.03	0.181
	40	0.01	0.005	0.043	0.036	0.023	0.051	0.048	0.026	0.012	0.034
			0.01	0.05	0.032	0.01	0.079	0.053	0.026	0.003	0.035
		0.05	0.005	0.041	0.042	0.04	0.049	0.06	0.049	0.039	0.045
			0.01	0.051	0.047	0.041	0.065	0.053	0.033	0.031	0.046
20×	2	0.01	0.005	0.06	0.024	0.018	0.054	0.058	0.033	0.015	0.037
			0.01	0.036	0.034	0.008	0.049	0.014	0.031	0.003	0.006
		0.05	0.005	0.043	0.036	0.039	0.043	0.045	0.035	0.032	0.036
			0.01	0.054	0.039	0.043	0.054	0.037	0.04	0.026	0.035
	10	0.01	0.005	0.045	0.025	0.014	0.047	0.061	0.03	0.015	0.051
			0.01	0.05	0.04	0.004	0.058	0.011	0.026	0.004	0.034
		0.05	0.005	0.043	0.035	0.034	0.041	0.055	0.038	0.036	0.05
			0.01	0.06	0.041	0.036	0.064	0.048	0.038	0.029	0.051
	40	0.01	0.005	0.051	0.023	0.012	0.035	0.042	0.036	0.013	0.034
			0.01	0.054	0.033	0.002	0.052	0.039	0.022	0.002	0.023
		0.05	0.005	0.052	0.033	0.045	0.049	0.061	0.046	0.038	0.036
			0.01	0.052	0.04	0.034	0.056	0.052	0.048	0.039	0.043

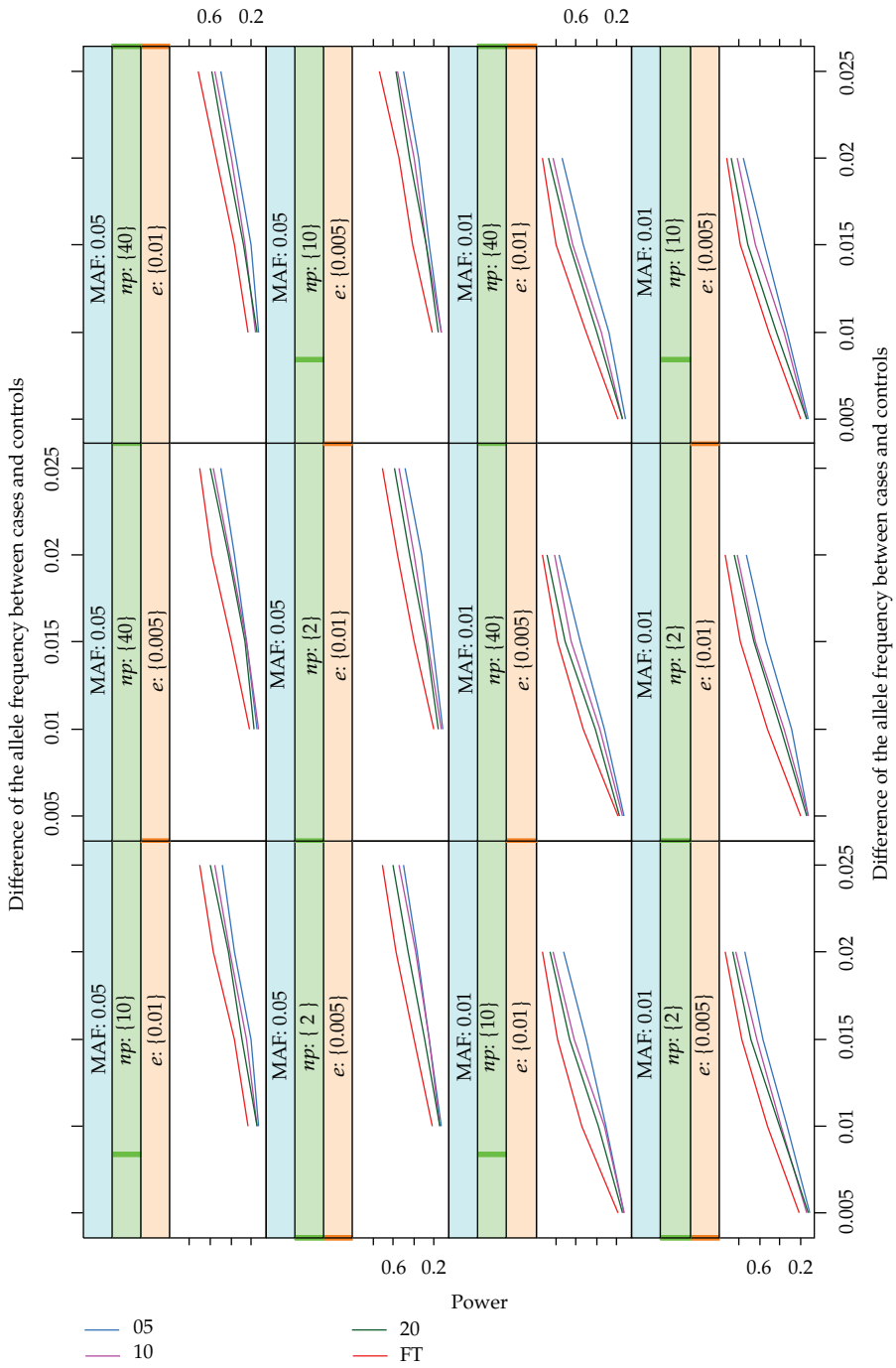


Figure 1: Empirical power at the a level of 5% for the parametric bootstrap (PB) test as a function of the difference of the allele frequency between cases and controls under various sequencing error rates, numbers of pools, and depths of sequencing coverage for testing association. Sample size was set at 500 cases and 500 controls. The minor allele frequencies (MAF) of controls were set at 0.01 and 0.5, sequencing error rates (e) were set at 0.005 and 0.01; and numbers of pools were set at 2, 10, and 40. Lines with different colors indicate the power of the PB test under different depths of coverage, which are compared to that of the Fisher's exact test (read line) based on the true individual genotypes (FT).

in particular for rare variants, because of the tendency of conservativeness of Fisher's exact test itself in particular for rare variants (data not shown).

3.3. Testing for Multiple Variants

3.3.1. Type I Error Rate

The empirical type I error rate at a significant level of 0.05 for testing association of multiple rare variants is shown in Table 3. The multi-variant PB test based on the empirical unbiased estimate of the covariance had the worst performance, it was too liberal when multiple rare variants were in linkage equilibrium, while it was overconservative when variants were in LD. This was more obvious when the sequencing error rate was high (1%). As expected, the test based on an *identity* covariance matrix had a good control of the type I error rate when multiple variants were uncorrelated, but it tended to be liberal when variants were in LD. The single-variant test based on Bonferroni correction was consistently conservative when variants were in either LD or linkage equilibrium. Compared to other tests, the multivariants PB test based on a shrinkage estimate had the best performance. The results were similar for different numbers of pools for an unbalanced design. As expected, the type I error rate was improved for the test based on the empirical estimate of the covariance with an increasing number of pools. The PB test based on the shrinkage estimate kept a good control of the type I error rate.

3.3.2. Power

Figure 2 shows the empirical power of different tests for testing association of multiple variants under various numbers of pools, numbers of variants, sequencing error rates, depths of sequencing coverage, and correlation structures. In general, the single-variant test with Bonferroni correction had the worst performance in terms of power, which may be due to two reasons: first, it does not make use of the accumulated effects from all variants; second, it has a conservative type I error rate. Among different multi-variant tests, the test based on the unbiased estimate of the covariance was consistently less powerful than the other two tests, even though it had a liberal type I error rate when variants were in LD (data not shown). The power of the tests based on a shrinkage estimate and an identity covariance matrix was comparable when variants are in linkage equilibrium (Figure 2(a)), but the identity covariance matrix seemed slightly more powerful than the shrinkage estimate in particular when the variants were in LD, which may be due to the fact that the test based on an identity covariance matrix had a liberal type I error rate in this case (data not shown).

4. Discussion

In this paper, we addressed two important questions of testing disease association of rare variants by pooled sequencing. One critical issue is that the sequencing error rate is high and has a significant variability across genomic positions. Ignoring the position-specific sequencing error could lead to a biased estimate of the allele frequency, and eventually a biased association result that can be either conservative or liberal, which was shown in our simulations. Another important issue is that the pooling procedure introduces an extra

Table 3: Type I error rates at a level of 5% for multivariant tests under various allele frequencies, error rates, depths of coverage, and numbers of variants for testing association. Sample size was set at 500 cases and 500 controls. Simulations were based on the unbalanced design with 10 and 20 pools.

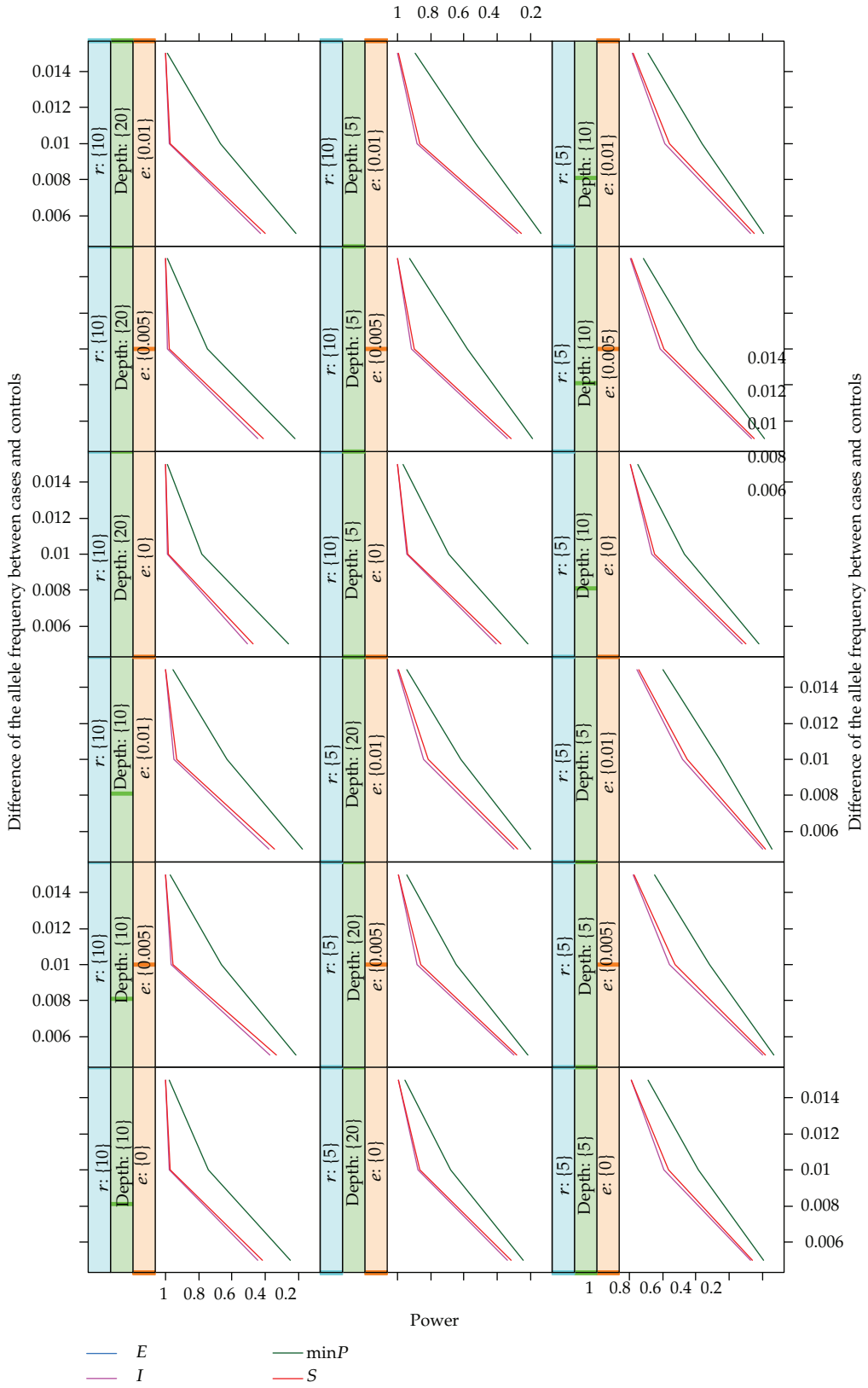
Number of pool	Number of variant	Depth	θ	e	Uncorrelated				Correlated ($\rho = 0.5$)			
					E	I	S	min	E	I	S	min
10	5	5×	0.01	0.005	0.058	0.038	0.039	0.037	0.051	0.067	0.062	0.036
				0.01	0.051	0.045	0.045	0.034	0.046	0.05	0.046	0.035
			0.05	0.005	0.062	0.052	0.058	0.048	0.028	0.046	0.038	0.04
				0.01	0.072	0.042	0.044	0.042	0.062	0.034	0.07	0.056
		10×	0.01	0.005	0.065	0.053	0.056	0.035	0.033	0.038	0.035	0.038
				0.01	0.057	0.043	0.044	0.042	0.037	0.053	0.043	0.04
			0.05	0.005	0.066	0.048	0.048	0.046	0.03	0.06	0.04	0.046
				0.01	0.07	0.052	0.052	0.06	0.032	0.072	0.052	0.052
		20×	0.01	0.005	0.063	0.049	0.052	0.039	0.039	0.058	0.045	0.035
				0.01	0.069	0.049	0.049	0.036	0.032	0.045	0.042	0.041
			0.05	0.005	0.054	0.042	0.042	0.042	0.038	0.066	0.048	0.064
				0.01	0.076	0.06	0.06	0.042	0.018	0.05	0.03	0.032
	10	5×	0.01	0.005	0.071	0.05	0.049	0.03	0.027	0.049	0.045	0.036
				0.01	0.064	0.051	0.052	0.023	0.03	0.049	0.043	0.036
			0.05	0.005	0.066	0.048	0.048	0.066	0.008	0.046	0.028	0.058
				0.01	0.068	0.044	0.048	0.036	0.02	0.082	0.062	0.05
		10×	0.01	0.005	0.063	0.058	0.058	0.04	0.024	0.063	0.05	0.045
				0.01	0.062	0.05	0.049	0.036	0.02	0.049	0.039	0.043
			0.05	0.005	0.064	0.054	0.052	0.048	0.014	0.056	0.044	0.044
				0.01	0.082	0.068	0.066	0.05	0.016	0.056	0.032	0.04
		20×	0.01	0.005	0.067	0.048	0.049	0.041	0.021	0.06	0.041	0.049
				0.01	0.067	0.046	0.045	0.038	0.022	0.053	0.04	0.026
			0.05	0.005	0.066	0.044	0.046	0.046	0.008	0.06	0.04	0.05
				0.01	0.098	0.078	0.084	0.044	0.008	0.068	0.038	0.044
20	5	5×	0.01	0.005	0.058	0.053	0.052	0.048	0.037	0.055	0.05	0.039
				0.01	0.052	0.048	0.05	0.049	0.028	0.044	0.037	0.041
			0.05	0.005	0.057	0.047	0.048	0.061	0.027	0.058	0.046	0.053
				0.01	0.058	0.051	0.047	0.044	0.026	0.055	0.044	0.047
		10×	0.01	0.005	0.052	0.043	0.045	0.031	0.039	0.06	0.052	0.058
				0.01	0.05	0.041	0.041	0.054	0.029	0.046	0.043	0.037
			0.05	0.005	0.048	0.038	0.04	0.044	0.025	0.055	0.045	0.054
				0.01	0.051	0.042	0.043	0.044	0.025	0.063	0.042	0.051
		20×	0.01	0.005	0.058	0.051	0.05	0.035	0.034	0.053	0.046	0.042
				0.01	0.051	0.04	0.038	0.035	0.026	0.043	0.041	0.041
			0.05	0.005	0.051	0.049	0.05	0.042	0.022	0.056	0.037	0.047
				0.01	0.054	0.045	0.041	0.042	0.02	0.055	0.034	0.044
	10	5×	0.01	0.005	0.068	0.06	0.058	0.024	0.03	0.058	0.046	0.03
				0.01	0.07	0.046	0.048	0.044	0.024	0.048	0.044	0.034
			0.05	0.005	0.08	0.06	0.65	0.052	0.012	0.062	0.04	0.038
				0.01	0.054	0.048	0.046	0.036	0.025	0.066	0.046	0.038
		10×	0.01	0.005	0.064	0.048	0.05	0.038	0.02	0.046	0.032	0.036
				0.01	0.076	0.05	0.054	0.048	0.02	0.068	0.054	0.034
			0.05	0.005	0.068	0.052	0.05	0.044	0.008	0.072	0.038	0.032
				0.01	0.08	0.048	0.052	0.05	0.01	0.07	0.038	0.05

Table 3: Continued.

Number of pool	Number of variant	Depth	θ	e	Uncorrelated				Correlated ($\rho = 0.5$)			
					E	I	S	min	E	I	S	min
			0.01	0.005	0.056	0.054	0.052	0.042	0.022	0.04	0.034	0.036
		20×		0.01	0.072	0.06	0.06	0.04	0.028	0.064	0.052	0.028
			0.05	0.005	0.074	0.054	0.06	0.038	0.026	0.09	0.054	0.038
				0.01	0.044	0.038	0.038	0.046	0.014	0.074	0.034	0.034

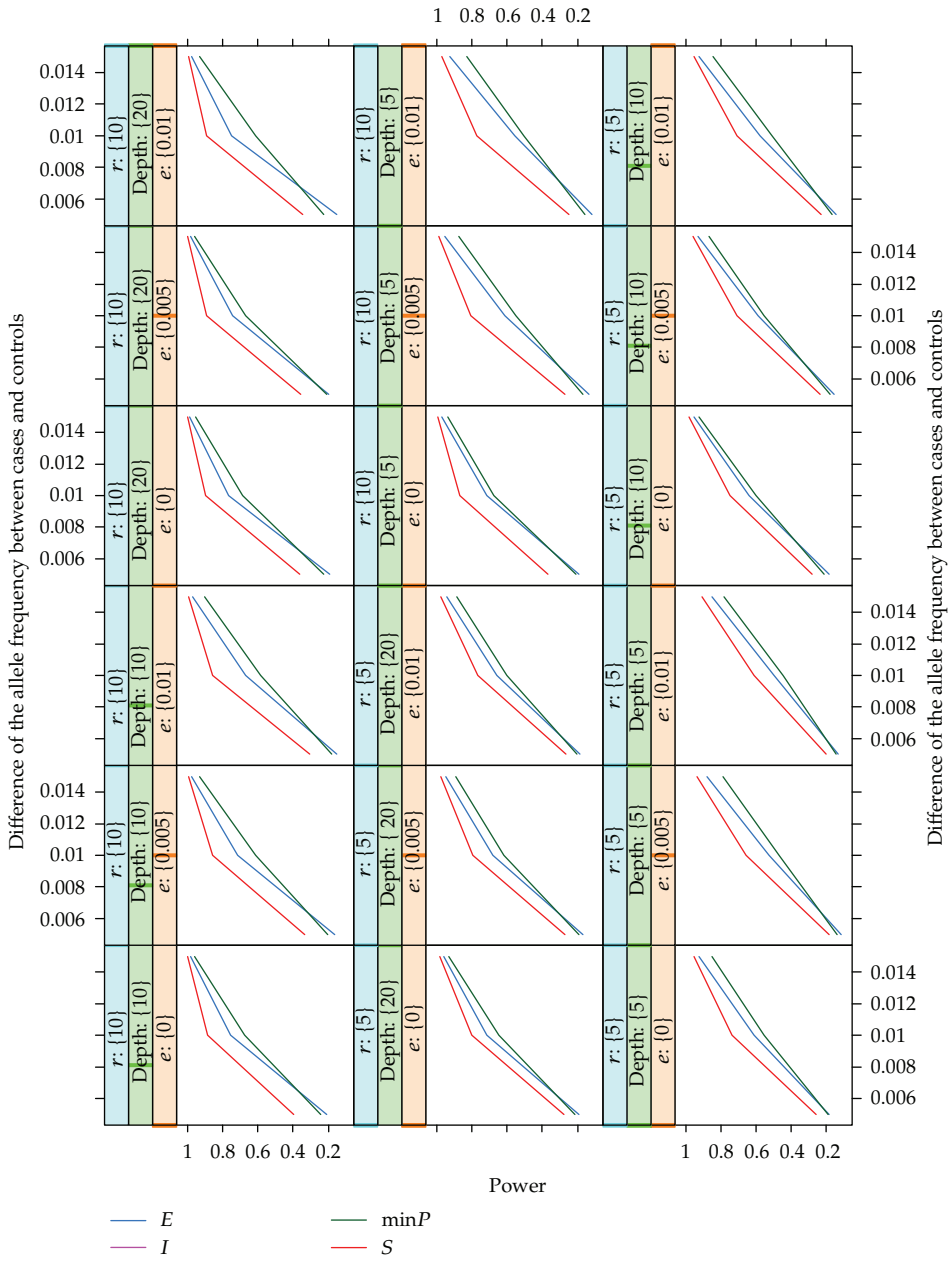
variance of the estimated number of variant alleles in a pool. Ignoring the uncertainty of the number of variant alleles could result in an inflated type I error rate, in particular in the case that the sequencing depth of coverage is low. This problem was indicated by the simulation results of the FE test which is directly based on the estimated number of variant alleles. To tackle these two questions, we proposed to use blocked pooling design to efficiently estimate the position-specific sequencing error rate and the allele frequency, along with a parameter bootstrap testing procedure to account for the extra variance of the estimate of the number of variant alleles in a pool.

We have proposed blocked pooling design to address the above two questions. Although blocked design in this paper was discussed based on lanes of flow cell, the similar idea could be extended to flow cells to take into account two sources of variation: variation between lanes within a flow cell and variation between flowcells. Based on blocked pooling design, an EM algorithm was used for estimating the position-specific sequencing error rate by making use of data from multiple pools. We examined the bias and standard error of the estimate of the sequencing error rates of different pooling designs under various situations through simulations. Intuitively, the EM algorithm should have a better performance when the number of pools is large and the number of individuals in a single pool is small because of the large difference between the minimal allele frequency of a pool and the sequencing error rate. As the result, we found the unbalanced design in which one half of pools included single individuals could provide a much more accurate estimate of the sequencing error rate as well as the allele frequency, while it does not sacrifice much on the variance of these estimates. Previously, we found that misspecification sequencing error has much more important impact on the statistical power than other parameters of pooled sequencing, for example, the depth of coverage and the number of pools [17]. Because the unbalanced design could provide more accurate estimates of the error rate and the allele frequency, the proposed PB test based on the unbalanced design not only consistently maintained a good control of the Type I error rate, but also provided higher power than the balanced design under various situations, even when the depth of coverage was low (5×). For balanced design, however, the proposed PB test tended to be anticonservative for low coverage data. As such, we suggest that the unbalanced blocked design, rather than the more commonly-used balance design, should be used in practice. Before a pooled sequencing study, it may be a good strategy to perform a simulation study to obtain the optimal unbalanced design based on the size of sequencing region and total depth of coverage. Under our simulated situations, for the given number of subjects, depth of coverage and type of design, the number of pools ranging from 10 to 40 did not significantly improve the estimate of the sequencing error rate and the allele frequency, and hence it was not a significant parameter for the statistical power. This result could be important, because it suggested that the pooled sequencing can be very cost-effective by including a small number of large pools with many individuals and small



(a)

Figure 2: Continued.



(b)

Figure 2: Empirical power at a level of 5% as a function of the difference of the allele frequency between cases and controls for the proposed PB test based on various estimates of the covariance matrix for testing multiple rare variants under the unbalanced design. The allele frequency of controls was 0.01; the sample size was set at 500 cases and 500 controls; the error rates (e) were set at 0, 0.005, and 0.01; the depths of coverage were set at 10× and 20×, and the numbers of pools (r) were set at 5 and 10. Figure (a) shows the power of the PB test based on an identity covariance matrix (I), the PB test based on the shrinkage estimate of the covariance matrix (S) and the single-variant test with Bonferroni correction the number of variants for independent variants (min$P$). Figure (b) shows the power of the PB test based on the empirical estimate of the covariance matrix (E), the PB test based on the shrinkage estimate of the covariance matrix (S) and the single-variant test with Bonferroni correction (min$P$) the number of variants for independent variants.

pools with single individuals in an unbalanced design, which is able to achieve adequate power.

As a single rare variant is likely to have a low marginal effect on disease risk, particularly in the presence of genetic heterogeneity, it is beneficial to jointly test a group of rare variants in a functional unit, such as genes or pathways. We extended the PB method for multiple rare variants. As with other multivariant tests based on individual genotypes, the multivariant PB test is designed for situations in which many rare variants present in the target region. Because our multivariant test is defined by the sum of Z scores transformed from single P values, it does not rely on the assumption on the direction of effects. Even if the effects of rare allele are uniformly in one direction, such as increasing risk, the proposed test can easily incorporate such information by using one-sided single P values to define the test statistic. Its another advantage is that the power is not primarily driven by more common variants when variants with different allele frequencies present in the target region. Because individual genotypes are not available in pooled sequencing, permutation testing is not an option for accurate significance estimation in scenarios where LD is present. We proposed a Monte Carlo approach by simulating the null distribution of the test statistic based on the estimate covariance between variants. The validity and efficiency of this approach rely on how well the covariance can be estimated. Because of the limit number of pools, the test based on the empirical unbiased covariance estimate did not have a good control of the type I error rate and often led to loss in power. However, the test based on the shrinkage estimate could provide a more satisfactory control on the type I error rate. Yet, it maintained comparable power to the test based on the unknown true covariance. One concern of the proposed approach is that the simulation procedure may lead to significant computational time for large-scale sequencing-based studies. To reduce computational burden, more effective approaches could also be obtained based on the shrinkage estimate of the covariance matrix [35].

The test procedure relies on several assumptions for the different steps of resequencing. The first step of resequencing is typically pulldown of the target genomic region and amplification. We assumed that the targeted genomic regions of subjects in a pool are amplified independently with an equal probability. One concern about this is the presence of heterogeneity in DNA amount in a pool. In this case, individuals are not evenly represented in the pool, and hence the assumption of the resampling approach that alleles of different subjects are drawn with the same probability is not valid. Indeed, the presence of heterogeneity in DNA amount was found to inflate the variance of the test statistic and hence lead to an inflated type I error rate (data not shown). However, if multiple independent markers ($\geq$30) are sequenced, it may be possible to use an approach similar to the genomic control to adjust for the inflated variance [36, 37].

In summary, our results suggest that pooled next-generation sequencing with the unbalance blocked design and the appropriate analytic approach could be a valid and cost-effective tool for screening the association of rare variants with diseases. Compared with individual sequencing, it is beneficial in terms of the reduction in cost and time but does not sacrifice much in statistical efficiency.

Acknowledgment

T. Wang was supported in part by the CTSA Grant UL1 RR025750 and KL2 RR025749 and TL1 RR025748 from the National Center for Research Resources (NCRR), a component of the

National Institutes of Health (NIH) and NIH roadmap for Medical Research, R21HG006150 from National Human Genome Research Institute (NHGRI). The codes written in R for the proposed PB test is available by email to Dr. Tao Wang (tao.wang@einstein.yu.edu).

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[2] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[3] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases," *Nature Genetics*, vol. 40, no. 6, pp. 695–701, 2008.

[4] D. R. Bentley, "Whole-genome re-sequencing," *Current Opinion in Genetics & Development*, vol. 16, pp. 545–552, 2006.

[5] D. W. Craig, J. V. Pearson, S. Szelinger et al., "Identification of genetic variants using bar-coded multiplexed sequencing," *Nature Methods*, vol. 5, no. 10, pp. 887–893, 2008.

[6] Y. Erlich, K. Chang, A. Gordon et al., "DNA Sudoku - Harnessing high-throughput sequencing for multiplexed specimen analysis," *Genome Research*, vol. 19, no. 7, pp. 1243–1253, 2009.

[7] A. Futschik and C. Schlötterer, "The next generation of molecular markers from massively parallel sequencing of pooled DNA samples," *Genetics*, vol. 186, no. 1, pp. 207–218, 2010.

[8] N. Shental, A. Amir, and O. Zuk, "Identification of rare alleles and their carriers using compressed se(que)nsing," *Nucleic Acids Research*, vol. 38, no. 19, Article ID gkq675, p. e179, 2010.

[9] T. Ito, S. Chiku, E. Inoue et al., "Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data," *The American Journal of Human Genetics*, vol. 72, no. 2, pp. 384–398, 2003.

[10] S. H. Shaw, M. M. Carrasquillo, C. Kashuk, E. G. Puffenberger, and A. Chakravarti, "Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes," *Genome Research*, vol. 8, no. 2, pp. 111–123, 1998.

[11] D. Zeng and D. Y. Lin, "Estimating Haplotype-disease associations with pooled genotype data," *Genetic Epidemiology*, vol. 28, no. 1, pp. 70–82, 2005.

[12] S. E. Calvo, E. J. Tucker, A. G. Compton et al., "High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex i deficiency," *Nature Genetics*, vol. 42, no. 10, pp. 851–858, 2010.

[13] T. E. Druley, F. L. M. Vallania, D. J. Wegner et al., "Quantification of rare allelic variants from pooled genomic DNA," *Nature Methods*, vol. 6, no. 4, pp. 263–265, 2009.

[14] S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd, "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes," *Science*, vol. 324, no. 5925, pp. 387–389, 2009.

[15] T. Wang, K. Pradhan, L. J. Wong, K. Ye, and T. E. Rohan, "Estimating allele frequency from next-generation sequencing of pooled mitochondrial DNA samples," *Frontiers in Genetics*, vol. 2, article 51, 2011.

[16] S. Y. Kim, Y. Li, Y. Guo et al., "Design of association studies with pooled or un-pooled next-generation sequencing data," *Genetic Epidemiology*, vol. 34, no. 5, pp. 479–491, 2010.

[17] T. Wang, C. Y. Lin, T. E. Rohan, and K. Ye, "Resequencing of pooled DNA for detecting disease associations with rare variants," *Genetic Epidemiology*, vol. 34, no. 5, pp. 492–501, 2010.

[18] J. C. Cohen, R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson, and H. H. Hobbs, "Multiple rare alleles contribute to low plasma levels of HDL cholesterol," *Science*, vol. 305, no. 5685, pp. 869–872, 2004.

[19] S. Morgenthaler and W. G. Thilly, "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)," *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 615, no. 1-2, pp. 28–56, 2007.

[20] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 311–321, 2008.

[21] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS Genetics*, vol. 5, no. 2, Article ID e1000384, 2009.

[22] A. P. Morris and E. Zeggini, "An evaluation of statistical approaches to rare variant analysis in genetic association studies," *Genetic Epidemiology*, vol. 34, no. 2, pp. 188–193, 2010.
[23] A. L. Price, G. V. Kryukov, P. I. W. de Bakker et al., "Pooled association tests for rare variants in exon-resequencing studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 832–838, 2010.
[24] B. M. Neale, M. A. Rivas, B. F. Voight et al., "Testing for an unusual distribution of rare variants," *PLoS Genetics*, vol. 7, no. 3, Article ID e1001322, 2011.
[25] W. Pan and X. Shen, "Adaptive tests for association analysis of rare variants," *Genetic Epidemiology*, vol. 35, no. 5, pp. 381–388, 2011.
[26] B. Hunter, Ed., *Statistics for Experimenters*, Wiley, 1987.
[27] J. Z. Liu, A. F. McRae, D. R. Nyholt et al., "A versatile gene-based test for genome-wide association studies," *The American Journal of Human Genetics*, vol. 87, no. 1, pp. 139–145, 2010.
[28] J. J. Goeman, S. A. van de Geer, and H. C. van Houwelingen, "Testing against a high dimensional alternative," *Journal of the Royal Statistical Society B*, vol. 68, no. 3, pp. 477–493, 2006.
[29] J. Wessel and N. J. Schork, "Generalized genomic distance-based regression methodology for multilocus association analysis," *The American Journal of Human Genetics*, vol. 79, no. 5, pp. 792–806, 2006.
[30] L. C. Kwee, D. Liu, X. Lin, D. Ghosh, and M. P. Epstein, "A powerful and flexible multilocus association test for quantitative traits," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 386–397, 2008.
[31] M. C. Wu, P. Kraft, M. P. Epstein et al., "Powerful SNP-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010.
[32] J. K. Pritchard, "Are rare variants responsible for susceptibility to complex diseases?" *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 124–137, 2001.
[33] S. Greenland, "Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and Empirical-Bayes regression," *Statistics in Medicine*, vol. 12, no. 8, pp. 717–736, 1993.
[34] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, article 32, 2005.
[35] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests," *The American Journal of Human Genetics*, vol. 81, pp. 1158–1168, 2007.
[36] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.
[37] B. Devlin, S. A. Bacanu, and K. Roeder, "Genomic control to the extreme," *Nature Genetics*, vol. 36, no. 11, pp. 1129–1131, 2004, Author reply p. 31.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 817948, 13 pages
doi:10.1155/2012/817948

Research Article

Sample Size Calculation for Controlling False Discovery Proportion

Shulian Shang,[1] Qianhe Zhou,[2] Mengling Liu,[1] and Yongzhao Shao[1]

[1] *Division of Biostatistics, New York University School of Medicine, New York, NY 10016, USA*
[2] *Novartis Institutes for Biomedical Research, Cambridge, MA 02139, USA*

Correspondence should be addressed to Yongzhao Shao, yongzhao.shao@nyumc.org

Received 31 March 2012; Accepted 5 June 2012

Academic Editor: Xiaohua Douglas Zhang

The false discovery proportion (FDP), the proportion of incorrect rejections among all rejections, is a direct measure of abundance of false positive findings in multiple testing. Many methods have been proposed to control FDP, but they are too conservative to be useful for power analysis. Study designs for controlling the mean of FDP, which is false discovery rate, have been commonly used. However, there has been little attempt to design study with direct FDP control to achieve certain level of efficiency. We provide a sample size calculation method using the variance formula of the FDP under weak-dependence assumptions to achieve the desired overall power. The relationship between design parameters and sample size is explored. The adequacy of the procedure is assessed by simulation. We illustrate the method using estimated correlations from a prostate cancer dataset.

1. Introduction

Modern biomedical research frequently involves parallel measurements of a large number of quantities of interest, such as gene expression levels, single nucleotide polymorphism (SNP) and DNA copy number variations. The scientific question can often be formulated as a multiple testing problem. In order to address the multiplicity issue, many methods have been proposed to control the family-wise error rate (FWER), false discovery rate (FDR) or false discovery proportion (FDP). Controlling FDR has been widely used in high-dimensional data analysis [1–3]. FDR is the expected value of the FDP, which is the proportion of incorrect rejections among all rejections. Controlling FDR ensures that the average of FDP from many independently repeated experiments is under control. However, the variability of FDP is ignored, and the actual FDP could be much greater than FDR with high probability, especially

when test statistics are correlated. Consequently, researchers have proposed many procedures to control FDP directly [4–11]:

$$\mathrm{P}\ (\ \mathrm{FDP}\ \leq r_1) \geq c_1 \tag{1.1}$$

for given r_1 and c_1. This is a more stringent criterion than FDR because the proportion of false rejections is bounded above by r_1 with high probability. The FDP controlling procedures are generally too conservative to be useful for the purpose of study design or power analysis.

When we design studies involving multiple testing, it is important to determine sample size to ensure adequate statistical power. Methods for calculating sample size have been proposed to control various criteria, for example, FWER [12–14], FDR [15–20], the number of false discoveries [19, 21] and FDP [22]. For controlling FDP, Oura et al. [22] provided a method to calculate sample size using the beta-binomial model for the sum of rejection status of true alternative hypotheses. It is assumed that only test statistics of true alternative hypotheses are dependent, with a parametric correlation structure. This assumption is restrictive because null test statistics can also be correlated and the dependence structure can be more complicated than the assumed parametric correlation structure. Furthermore, the computation is intensive because computation of the beta-binomial distribution is required. However, to our knowledge this is the only paper that directly deals with this important design problem.

In this paper, we provide a more general method of sample size calculation for controlling FDP under weak-dependence assumptions. Under some assumptions on dependence among test statistics, explicit formulas for the mean and variance of FDP have been derived for each fixed effect size [23]. The formulas elucidate the effects of various design parameters on the variance of FDP. Moreover, the formulas provide a convenient tool to calculate sample size for controlling the FDP. As in [13, 18, 19, 24], we consider the probability of detecting at least a specified proportion of true alternative hypotheses as the power criterion. An iterative computation algorithm for calculating sample size is provided. Simulation experiments indicate that studies with the resultant sample sizes satisfy the power criterion at the given rejection threshold. We illustrate the sample size calculation procedure using a prostate cancer dataset.

2. Methods

2.1. Notation

Suppose that m hypotheses are tested simultaneously. Let M_0 denote the index set of m_0 tests for which null hypotheses are true and M_1 the index set of $m_1 = m - m_0$ tests for which alternative hypotheses are true. Denote the proportion of true null hypotheses by $\pi_0 = m_0/m$. We reject a hypothesis if the P value is less than some threshold α, and denote the rejection status of the ith test by $R_i(\alpha) = I(p_i < \alpha)$, where p_i denotes the P value of the ith test and $I(\cdot)$ is an indicator function. The number of rejections is $R = \sum_{i=1}^{m} R_i(\alpha)$. Let the comparison-wise type II error of the ith test be β_i and the average type II error be

$$\overline{\beta} = \frac{1}{m_1} \sum_{i \in M_1} \beta_i. \tag{2.1}$$

Table 1 summarizes the outcomes of m tests and their expected values.

Table 1: Outcomes and expected outcomes of testing m hypotheses.

	Outcomes		
	Reject H_0	Accept H_0	Total
H_0 is true	V	$m_0 - V$	m_0
H_1 is true	U	$m_1 - U$	m_1
Total	R	$m - R$	m
	Expected outcomes		
	Reject H_0	Accept H_0	Total
H_0 is true	$m_0\alpha$	$m_0(1-\alpha)$	m_0
H_1 is true	$m_1(1-\overline{\beta})$	$m_1\overline{\beta}$	m_1
Total	$m_0\alpha + m_1(1-\overline{\beta})$	$m_0(1-\alpha) + m_1\overline{\beta}$	m

Denote the Pearson correlation coefficient of two rejection indicators by

$$\theta^{ij} = \operatorname{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}. \tag{2.2}$$

Furthermore, for $i, j \in M_0$, define

$$\theta_V^{ij} = \operatorname{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}. \tag{2.3}$$

Let the average correlation be denoted as

$$\overline{\theta}_V = \frac{\sum_{i,j\in M_0, i\neq j}\theta_V^{ij}}{m_0(m_0-1)}. \tag{2.4}$$

Similarly, for $i, j \in M_1$, we define

$$\theta_U^{ij} = \operatorname{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}. \tag{2.5}$$

The average correlation is

$$\overline{\theta}_U = \frac{\sum_{i,j\in M_1, i\neq j}\theta_U^{ij}}{m_1(m_1-1)}. \tag{2.6}$$

In addition, for $i \in M_1$, $j \in M_0$, denote

$$\theta_{UV}^{ij} = \operatorname{corr}\left\{R_i(\alpha), R_j(\alpha)\right\}. \tag{2.7}$$

Denote the average correlation by

$$\overline{\theta}_{UV} = \frac{\sum_{i\in M_1}\sum_{j\in M_0}\theta_{UV}^{ij}}{m_0 m_1}. \tag{2.8}$$

2.2. The Effect of Design Parameters on the Variance of FDP

It has been shown via numerical studies that the variability of FDP increases when test statistics are dependent [25, 26]. But the relationship between design parameters and the variance of FDP has not been examined through analytical formulas. Under the assumptions of common effect size and weak dependence among test statistics, explicit formulas for the mean (μ_Q) and variance (σ_Q^2) of the FDP have been derived [23]:

$$\mu_Q \approx \frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)\left(1-\overline{\beta}\right)}, \tag{2.9}$$

$$\sigma_Q^2 \approx \frac{\pi_0(1-\pi_0)^2\alpha(1-\alpha)\left(1-\overline{\beta}\right)}{\left\{\pi_0\alpha + (1-\pi_0)\left(1-\overline{\beta}\right)\right\}^4}\Sigma, \tag{2.10}$$

where

$$\begin{aligned}\Sigma = {} & \frac{1}{m}\left(1-\overline{\beta}+\frac{\pi_0}{1-\pi_0}\omega\overline{\beta}\right)+\left(\pi_0-\frac{1}{m}\right)\left(1-\overline{\beta}\right)\overline{\theta}_V \\ & + \pi_0\omega\overline{\beta}\overline{\theta}_U - 2\pi_0\sqrt{\omega\overline{\beta}\left(1-\overline{\beta}\right)}\overline{\theta}_{UV}\end{aligned} \tag{2.11}$$

and $\omega = \alpha/(1-\alpha)$.

The variance formula (2.10) elucidates the effects of various design parameters on the variance of FDP. To explore the effects, in Figure 1 we calculated σ_Q using (2.10) and plotted it against m for different correlations $\overline{\theta}_V$. We set $\pi_0 = 0.7$ and m in the range of 1000 to 10000. The average correlations $\overline{\theta}_U$ and $\overline{\theta}_{UV}$ are fixed to be 0.001 and 0, respectively. The levels of α and $\overline{\beta}$ are chosen such that FDR is 3% or 5%. At each value of $\overline{\theta}_V$, σ_Q decreases as the number of tests m increases. The solid line shows the standard deviation of the FDP when $\overline{\theta}_V$ is 0. When $\overline{\theta}_V$ is not 0, σ_Q increases evidently. If test statistics are highly correlated, FDP can be much greater than its mean FDR at a given rejection threshold due to its large variability.

In Figure 2, the relationship between σ_Q and π_0 was investigated. When other parameters are fixed, σ_Q increases as π_0 increases.

Figure 3 shows that σ_Q increases as $\overline{\beta}$ increases. When other factors are fixed, the variability of FDP is smaller when the comparison-wise type II error is smaller.

2.3. Power and Sample Size Analysis

Under some general regularity conditions including weak dependence among test statistics, the FDP follows an asymptotic normal distribution $N(\mu_Q, \sigma_Q^2)$ [23, 27]. As was pointed out by Shang et al. [23], $Y = \log(\text{FDP})$ also has an asymptotic normal distribution by the delta

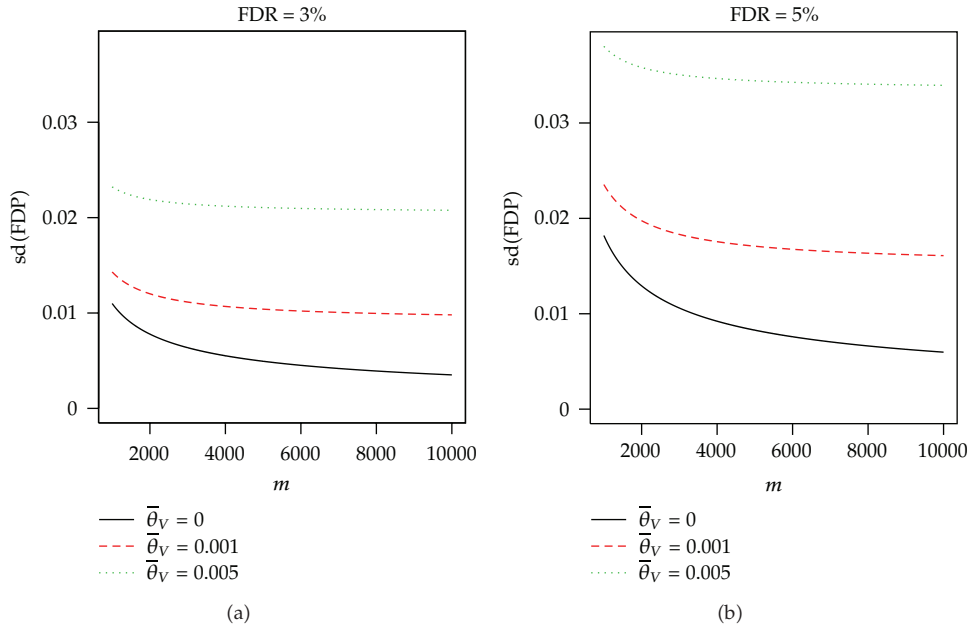


Figure 1: Plots of sd(FDP) at different m and $\overline{\theta}_V$. $\pi_0 = 0.7$, $\alpha = 0.01$, $\overline{\theta}_U = 0.001$ and $\overline{\theta}_{UV} = 0$. (a) $\overline{\beta} = 0.25$, FDR = 3%; (b) $\overline{\beta} = 0.56$, FDR = 5%.

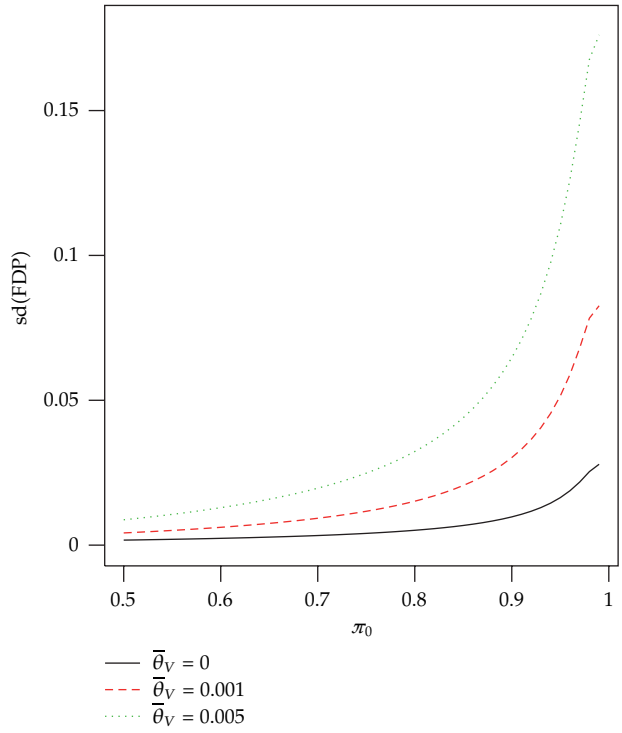


Figure 2: Plots of sd(FDP) at different π_0 and $\overline{\theta}_V$. $m = 10000$, $\alpha = 0.01$, $\overline{\beta} = 0.20$, $\overline{\theta}_U = 0.001$ and $\overline{\theta}_{UV} = 0$.

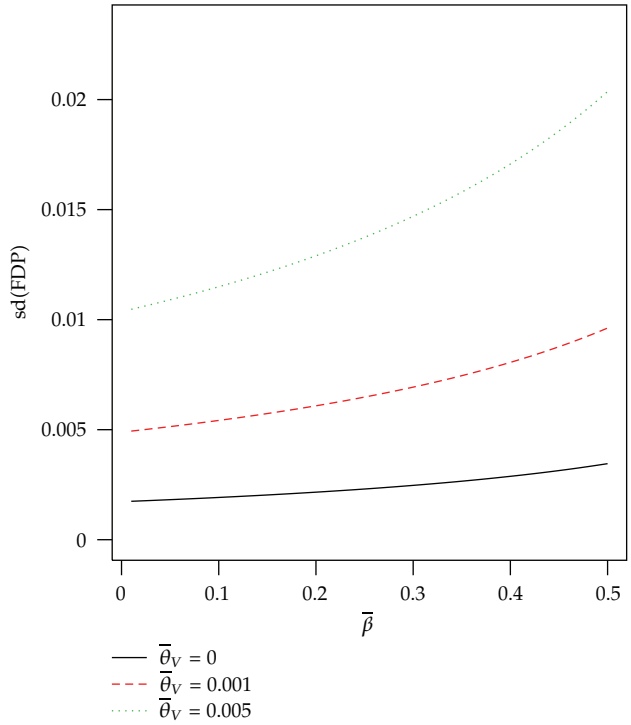


Figure 3: Plots of sd(FDP) at different $\overline{\beta}$ and $\overline{\theta}_V$. $m = 10000$, $\pi_0 = 0.7$, $\alpha = 0.004$, $\overline{\theta}_U = 0.001$ and $\overline{\theta}_{UV} = 0$.

method, and under weak dependence log(FDP) is closer to normal than the FDP itself. The approximate mean and variance of $Y = \log(\text{FDP})$ are [23]

$$\mu_Y \approx \log(\mu_Q) \approx \log\left\{\frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)\left(1-\overline{\beta}\right)}\right\}, \tag{2.12}$$

$$\sigma_Y^2 \approx \frac{(1-\pi_0)^2(1-\alpha)\left(1-\overline{\beta}\right)}{\pi_0\alpha\left\{\pi_0\alpha + (1-\pi_0)\left(1-\overline{\beta}\right)\right\}^2}\Sigma, \tag{2.13}$$

where Σ is in (2.11).

To control FDP with desired power, criterion (1.1) has to be satisfied. Asymptotic normality of log(FDP) implies that

$$\Phi\left(\frac{\log r_1 - \mu_Y}{\sigma_Y}\right) \geq c_1, \tag{2.14}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of standard normal distribution, μ_Y is in (2.12), and $\sigma^2{}_Y$ is in (2.13).

There are two commonly used power criteria in multiple testing: the average power, defined as $\mathrm{E}\,(U/m_1)$, and the overall power, defined as $\mathrm{P}\,(U/m_1 \geq r_2)$ for given r_2. When

a study is designed using the average power criterion, the proportion of true alternative hypotheses rejected will be greater than a prespecified number on average. However, under dependence among test statistics the variability of U/m_1 increases [18], and the study can be underpowered with high probability. Consequently, the overall power has been used in [13, 18, 19, 24] and we also use this power criterion,

$$\mathrm{P}\left(\frac{U}{m_1} \geq r_2\right) \geq c_2 \tag{2.15}$$

for given r_2 and c_2.

Under the weak-dependence assumptions in [18, 23], U/m_1 has an asymptotic normal distribution:

$$\frac{U}{m_1} \sim N\left(1-\overline{\beta}, \frac{\overline{\beta}\left(1-\overline{\beta}\right)}{m_1}\left\{1+\overline{\theta}_U(m_1-1)\right\}\right). \tag{2.16}$$

Setting the inequality in (2.15) to equality, the following equation for $\overline{\beta}$ can be obtained as in [18]:

$$\overline{\beta} = 1 - r_2 - \frac{1-2r_2+\sqrt{4m_1^* r_2(1-r_2)+1}}{2m_1^*+2}, \tag{2.17}$$

where $m_1^* = m_1/\{1+\overline{\theta}_U(m_1-1)\}z_{1-c_2}^2$ and $\Phi(z_{1-c_2}) = c_2$.

For illustration, consider that a two-sample one-sided t-test is performed. Let δ denote the effect size (mean difference divided by the common standard deviation), and a_1 and $1-a_1$ denote the allocation proportion for two groups. We first find α and $\overline{\beta}$ which fulfill criteria (1.1) and (2.15). The required sample size n is the smallest integer satisfying the following inequality:

$$1-\overline{\beta} \leq 1-\Gamma_{n-2}\left(-t_{n-2,\alpha}\middle|\delta\sqrt{a_1(1-a_1)n}\right), \tag{2.18}$$

where $\Gamma_{n-2}(\cdot \mid b)$ is the CDF of a noncentral t-distribution with $n-2$ degrees of freedom and noncentrality parameter b and $t_{n-2,\alpha}$ is the upper α critical value of the central t-distribution with $n-2$ degrees of freedom.

Following the notation defined in Section 2.1, the correlations can be calculated as:

$$\theta_V^{ij} = \frac{\Psi_{n-2}\left(-t_{n-2,\alpha}, -t_{n-2,\alpha}; \rho^{ij}\right) - \alpha^2}{\alpha(1-\alpha)}, \tag{2.19}$$

$$\theta_U^{ij} = \frac{\Psi_{n-2}\left(t_{n-2,\overline{\beta}}, t_{n-2,\overline{\beta}}; \rho^{ij}\right) - \left(1-\overline{\beta}\right)^2}{\overline{\beta}\left(1-\overline{\beta}\right)}, \tag{2.20}$$

$$\theta_{UV}^{ij} = \frac{\Psi_{n-2}\left(-t_{n-2,\alpha}, t_{n-2,\overline{\beta}}; \rho^{ij}\right) - \alpha\left(1-\overline{\beta}\right)}{\sqrt{\alpha(1-\alpha)\overline{\beta}\left(1-\overline{\beta}\right)}}, \tag{2.21}$$

where Ψ_{n-2} is the CDF of a bivariate t-distribution with $n-2$ degrees of freedom and ρ^{ij} denotes the Pearson correlation between the ith and jth test statistics. As can be seen from these formulas, the correlations depend on α and $\overline{\beta}$. No analytical solutions can be found for these two parameters. We use the following iterative computation algorithm to calculate sample size.

Algorithm.

(1) Input design parameters r_1, c_1, r_2, c_2, m, π_0, δ, a_1, ρ_U, ρ_V and ρ_{UV}.

(2) Start from $\overline{\theta}_U = 0$, $\overline{\theta}_V = 0$ and $\overline{\theta}_{UV} = 0$.

(3) Calculate $\overline{\beta}$ from (2.17).

(4) Using the current values of $\overline{\theta}_U$, $\overline{\theta}_V$, $\overline{\theta}_{UV}$ and $\overline{\beta}$, solve for α from equation $\Phi((\log r_1 - \mu_Y)/\sigma_Y) = c_1$.

(5) Using the current estimates of $\overline{\beta}$ and α, calculate θ_V, θ_U and θ_{UV} from (2.19), (2.20) and (2.21), respectively. Obtain the average correlations $\overline{\theta}_V$, $\overline{\theta}_U$ and $\overline{\theta}_{UV}$.

(6) With updated estimates of $\overline{\theta}_V$, $\overline{\theta}_U$ and $\overline{\theta}_{UV}$, repeat steps 3 to 5 until the estimates of $\overline{\beta}$ and α converge.

(7) Plug the estimated $\overline{\beta}$ and α into (2.18) to solve the sample size.

The estimates of rejection threshold α and comparison-wise type II error $\overline{\beta}$ can also be obtained.

3. Numerical Studies

3.1. Simulation

The proposed sample size calculation procedure was illustrated for one-sided t-test comparing the mean of two groups. The effect size $\delta = 1$ and allocation proportion $a_1 = 0.5$. Two types of correlation structures were used: blockwise correlation and autoregressive correlation structure. In the blockwise correlation structure, a proportion of test statistics were correlated in units of blocks. The correlation coefficient within block was a constant, and test statistics were independent across blocks. True null test statistics and true alternative

test statistics were independent. In the autoregressive correlation structure, the correlation matrix (σ_{ij}) for dependent test statistics was parameterized by $\sigma_{ij}(\rho) = \rho^{|i-j|}$, where $\sigma_{ij}(\rho)$ is the Pearson correlation coefficient for the ith and jth test statistics and ρ is a correlation parameter.

Oura et al. [22] provided a sample size calculation method for controlling FDP using the beta-binomial model. Only test statistics of true alternative hypotheses are allowed to be dependent, with blockwise correlation structure. For comparison, this method and the sample size calculation procedure for controlling FDR with dependence adjustment in [18] were also assessed. Specifically, the criteria for controlling FDP are

$$\mathrm{P}\ (\mathrm{FDP}\ \leq 0.05) \geq 0.95, \qquad \mathrm{P}\left(\frac{U}{m_1} \geq 0.9\right) \geq 0.8. \tag{3.1}$$

The criteria for controlling FDR are

$$\mathrm{FDR}\ \leq 0.05, \qquad \mathrm{P}\left(\frac{U}{m_1} \geq 0.9\right) \geq 0.8. \tag{3.2}$$

Table 2 presents the sample size estimates for the blockwise correlation structure. Several parameter configurations were used. The block size is 20 or 100, for $m = 2000$ or 10000, respectively. We observe that the sample size increases as the correlation between test statistics gets stronger, represented by a greater correlation parameter or a larger proportion of correlated test statistics. When the correlation is fixed, as the number of tests m increases, the required sample size decreases. With the other parameters fixed, when the number of true alternative hypotheses increases (π_0 decreases), the required sample size decreases.

The sample sizes for controlling FDP are greater than those for controlling FDR because controlling FDP is in general more stringent. In the case that $\pi_0 = 0.9$, $p_v = 0.3$, $\rho_v = 0.6$ and $m = 2000$ (see Table 2), the sample size for controlling FDP is 81, which is 23% greater than the sample size for controlling FDR. The sample sizes using the method in [22] are in parentheses and are slightly smaller than ours. In terms of computational efficiency, our algorithm converges very fast and generally within 10 steps. The computation is not heavy, and in fact, very similar and comparable to that in [18] for controlling FDR with dependence adjustment. The method of Oura et al. [22] is more computationally intensive. It becomes not feasible when the number of tests or the number of blocks of dependent test statistics is large. Simulation studies show that FDP is controlled and the power is achievable with the sample size given by our procedure at the calculated rejection threshold α (results not shown).

Table 3 presents the sample sizes for the autoregressive correlation structure. Similar trends for sample size are observed as the design parameters vary. The method in [22] is not applicable to this dependence structure.

3.2. Sample Size Calculation Based on a Prostate Cancer Dataset

We use a prostate cancer dataset as source of correlation structure to illustrate the proposed sample size calculation method while ensuring overall power. The study by Wang et al. [28] investigated the association between mRNA gene expression levels and the aggressive phenotype of prostate cancer. The dataset contains 13935 mRNA measured from 62 patients with aggressive prostate cancer and 63 patients with nonaggressive disease. The method in

Table 2: Sample size calculation for controlling FDP and FDR to achieve overall power, blockwise correlation structure.

π_0	ρ_u	ρ_v	p_u	p_v	$m = 2000$		$m = 10000$	
					FDP	FDR	FDP	FDR
0.9	0.2	0.2	0.1	0.1	75	66	68	65
	0.5	0.5	0.1	0.1	77	66	70	65
	0.8	0.8	0.1	0.1	81	67	74	66
	0	0.2	0	0.3	76	66	68	64
	0	0.5	0	0.3	78	66	71	64
	0	0.6	0	0.3	81	66	73	64
	0.2	0	1	0	77 (74)	67	70	67
	0.5	0	1	0	79 (75)	69	72	69
	0.8	0	1	0	82 (77)	72	75	72
0.7	0.2	0.2	0.1	0.1	53	49	50	48
	0.5	0.5	0.1	0.1	54	49	51	48
	0.8	0.8	0.1	0.1	55	49	53	49
	0	0.2	0	0.3	53	48	50	48
	0	0.5	0	0.3	55	48	52	48
	0	0.8	0	0.3	58	48	56	48
	0.2	0	1	0	54 (53)	49	50	49
	0.5	0	1	0	55 (54)	50	50	50
	0.8	0	1	0	56 (55)	52	50	52

ρ_u: correlation between test statistics for which the alternative hypotheses are true; ρ_v: correlation between test statistics for which the null hypotheses are true; p_u: proportion of correlated test statistics for which the alternative hypotheses are true; p_v: proportion of correlated test statistics for which the null hypotheses are true.

[23] was used to estimate the correlation between gene expression levels. The estimated average correlation of expression levels of null genes, alternative genes and between null genes and alternative genes are 0.0040, 0.0043 and −0.0005, respectively.

Sample size was calculated for one-sided t-test. The total number of genes is $m = 10000$, and the following criteria are to be satisfied: $\mathrm{P}(\mathrm{FDP} \leq 0.10) \geq 0.7$ and $\mathrm{P}(U/m_1 \geq 0.9) \geq 0.8$. Table 4 presents the sample sizes for various values of m_1. We performed simulation studies to confirm that these sample sizes provided adequate power at the rejection threshold given by our algorithm. Simulation data were generated with blockwise dependence structure such that the average correlation was close to the estimated correlation from the real dataset.

4. Discussion

In practice, when planning a study one typically needs to make some assumptions. For designing multiple testing studies, a common assumption is that the dependence between test statistics is weak. In this paper, we provide a computationally effective method of sample size calculation for controlling FDP under weak dependence while achieving the desired overall power. This approach uses semiparametric assumptions on dependence structure. We only need to estimate the Pearson correlation between test statistics, and thus this method is applicable to many realistic settings where weak dependence can be assumed. The variance formula of FDP provides a convenient tool to uncover the relationship between the design

Table 3: Sample size calculation for controlling FDP and FDR to achieve overall power, autoregressive correlation structure.

π_0	ρ_u	ρ_v	p_u	p_v	$m = 2000$		$m = 10000$	
					FDP	FDR	FDP	FDR
0.9	0.2	0.2	0.1	0.1	75	66	67	64
	0.5	0.5	0.1	0.1	75	66	67	64
	0.8	0.8	0.1	0.1	77	66	68	64
	0.2	0.2	0.4	0.4	75	66	67	64
	0.5	0.5	0.4	0.4	76	66	68	64
	0.8	0.8	0.4	0.4	81	66	69	64
0.7	0.2	0.2	0.1	0.1	53	48	49	48
	0.5	0.5	0.1	0.1	53	48	49	48
	0.8	0.8	0.1	0.1	54	48	50	48
	0.2	0.2	0.4	0.4	53	49	49	48
	0.5	0.5	0.4	0.4	54	49	50	48
	0.8	0.8	0.4	0.4	56	49	51	48

ρ_u: correlation parameter for test statistics for which the alternative hypotheses are true; ρ_v: correlation parameter for test statistics for which the null hypotheses are true; p_u: proportion of correlated test statistics for which the alternative hypotheses are true; p_v: proportion of correlated test statistics for which the null hypotheses are true.

Table 4: Sample size calculation using the prostate cancer dataset.

π_0	m_1	n	$\widehat{\mathrm{P}}(\mathrm{FDP} \leq 0.10)$	$\widehat{\mathrm{P}}((U/m_1) \geq 0.9)$	$\widehat{\alpha}$
0.95	500	74	0.830	0.915	0.003
0.9	1000	63	0.855	0.970	0.007
0.85	1500	57	0.885	0.955	0.012
0.8	2000	52	0.875	0.945	0.018
0.75	2500	48	0.870	0.955	0.026
0.7	3000	44	0.915	0.925	0.034

$\widehat{\mathrm{P}}(\mathrm{FDP} \leq 0.10)$ and $\widehat{\mathrm{P}}((U/m_1) \geq 0.9)$: empirical probability from 200 simulation runs.

parameters and the variability of FDP under the assumption of weak dependence. Simulation studies indicate that the algorithm is computationally efficient and stable.

We have used one-sided t-test to illustrate the method, and the procedure can be easily extended to two-sided t-test and other tests. Common effect size is assumed in the sample size calculation algorithm. In practice, one can try different effect sizes, see the range of sample sizes and the variability and then make a decision. Effects of variations on other parameters such as π_0 can be examined similarly.

Acknowledgments

We would like to thank the reviewers and the Editor for their careful reading of our paper and the constructive suggestions. This research was partially supported by a Stony Wold-Herbert Foundation grant and NIH grants 2P30 CA16087, 5P30 ES00260, UL1 TR000038 and R03 CA153083.

References

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[2] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society B*, vol. 64, no. 3, pp. 479–498, 2002.

[3] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society B*, vol. 66, no. 1, pp. 187–205, 2004.

[4] A. Farcomeni, "Generalized augmentation to control the false discovery exceedance in multiple testing," *Scandinavian Journal of Statistics*, vol. 36, no. 3, pp. 501–517, 2009.

[5] Y. Ge, S. C. Sealfon, and T. P. Speed, "Multiple testing and its applications to microarrays," *Statistical Methods in Medical Research*, vol. 18, no. 6, pp. 543–563, 2009.

[6] C. R. Genovese and L. Wasserman, "Exceedance control of the false discovery proportion," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1408–1417, 2006.

[7] C. Genovese and L. Wasserman, "A stochastic process approach to false discovery control," *The Annals of Statistics*, vol. 32, no. 3, pp. 1035–1061, 2004.

[8] E. L. Korn, M.-C. Li, L. M. McShane, and R. Simon, "An investigation of two multivariate permutation methods for controlling the false discovery proportion," *Statistics in Medicine*, vol. 26, no. 24, pp. 4428–4440, 2007.

[9] E. L. Korn, J. F. Troendle, L. M. McShane, and R. Simon, "Controlling the number of false discoveries: application to high-dimensional genomic data," *Journal of Statistical Planning and Inference*, vol. 124, no. 2, pp. 379–398, 2004.

[10] N. Meinshausen, "False discovery control for multiple tests of association under general dependence," *Scandinavian Journal of Statistics*, vol. 33, no. 2, pp. 227–237, 2006.

[11] M. J. van der Laan, S. Dudoit, and K. S. Pollard, "Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, article 15, 2004.

[12] S. H. Jung, H. Bang, and S. Young, "Sample size calculation for multiple testing in microarray data analysis," *Biostatistics*, vol. 6, no. 1, pp. 157–169, 2005.

[13] S. J. Wang and J. J. Chen, "Sample size for identifying differentially expressed genes in microarray experiments," *Journal of Computational Biology*, vol. 11, no. 4, pp. 714–726, 2004.

[14] M. C. K. Yang, J. J. Yang, R. A. McIndoe, and J. X. She, "Microarray experimental design: power and sample size considerations," *Physiological Genomics*, vol. 16, pp. 24–28, 2003.

[15] S. H. Jung, "Sample size for fdr-control in microarray data analysis," *Bioinformatics*, vol. 21, no. 14, pp. 3097–3104, 2005.

[16] S. S. Li, J. Bigler, J. W. Lampe, J. D. Potter, and Z. Feng, "FDR-controlling testing procedures and sample size determination for microarrays," *Statistics in Medicine*, vol. 24, no. 15, pp. 2267–2280, 2005.

[17] S. Pounds and C. Cheng, "Sample size determination for the false discovery rate," *Bioinformatics*, vol. 21, no. 23, pp. 4263–4271, 2005.

[18] Y. Shao and C.-H. Tseng, "Sample size calculation with dependence adjustment for FDR-control in microarray studies," *Statistics in Medicine*, vol. 26, no. 23, pp. 4219–4237, 2007.

[19] C. A. Tsai, S. J. Wang, D. T. Chen, and J. J. Chen, "Sample size for gene expression microarray experiments," *Bioinformatics*, vol. 21, no. 8, pp. 1502–1508, 2005.

[20] X. D. Zhang, *Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale Rnai Research*, Cambridge University Press, 2011.

[21] M. L. T. Lee and G. A. Whitmore, "Power and sample size for dna microarray studies," *Statistics in Medicine*, vol. 21, no. 23, pp. 3543–3570, 2002.

[22] T. Oura, S. Matsui, and K. Kawakami, "Sample size calculations for controlling the distribution of false discovery proportion in microarray experiments," *Biostatistics*, vol. 10, no. 4, pp. 694–705, 2009.

[23] S. Shang, M. Liu, and Y. Shao, "A tight prediction interval for false discovery proportion under dependence," *Open Journal of Statistics*, vol. 2, no. 2, pp. 163–171, 2012.

[24] W.-J. Lin, H.-M. Hsueh, and J. J. Chen, "Power and sample size estimation in microarray studies," *Bmc Bioinformatics*, vol. 11, article 48, 2010.

[25] R. Heller, "Correlated z-values and the accuracy of large-scale statistical estimates," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 1057–1059, 2010.

[26] Y. Pawitan, S. Calza, and A. Ploner, "Estimation of false discovery proportion under general dependence," *Bioinformatics*, vol. 22, no. 24, pp. 3025–3031, 2006.

[27] A. Farcomeni, "Some results on the control of the false discovery rate under dependence," *Scandinavian Journal of Statistics*, vol. 34, no. 2, pp. 275–297, 2007.
[28] L. Wang, H. Tang, V. Thayanithy et al., "Gene networks and microRNAs implicated in aggressive prostate cancer," *Cancer Research*, vol. 69, no. 24, pp. 9490–9497, 2009.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 935621, 10 pages
doi:10.1155/2012/935621

Research Article

Sample Size Growth with an Increasing Number of Comparisons

Chi-Hong Tseng[1] and Yongzhao Shao[2]

[1] *Department of Medicine, UCLA School of Medicine, Los Angeles, CA 90095, USA*
[2] *Division of Biostatistics, NYU School of Medicine, New York, NY 10016, USA*

Correspondence should be addressed to Chi-Hong Tseng, tseng.ch@gmail.com

Received 30 March 2012; Accepted 8 June 2012

Academic Editor: Wei T. Pan

An appropriate sample size is crucial for the success of many studies that involve a large number of comparisons. Sample size formulas for testing multiple hypotheses are provided in this paper. They can be used to determine the sample sizes required to provide adequate power while controlling familywise error rate or false discovery rate, to derive the growth rate of sample size with respect to an increasing number of comparisons or decrease in effect size, and to assess reliability of study designs. It is demonstrated that practical sample sizes can often be achieved even when adjustments for a large number of comparisons are made as in many genomewide studies.

1. Introduction

With the recent advancement in high-throughput technologies, simultaneous testing of a large number of hypotheses has become a common practice for many types of genomewide studies. Examples include genetic association studies and DNA microarray studies. In a genomewide association analysis, a large number of genetic markers are tested for association with the disease [1]. In DNA microarray studies, the interest is typically to identify differentially expressed genes between patient groups among a large number of candidate genes [2].

The challenges for designing such large-scale studies include the selection of features of scientific importance to be investigated, selection of appropriate sample size to provide adequate power, and choices of methods appropriate for the adjustment of multiple testing [3–7]. There exist recent methodological breakthroughs on multiple comparisons, such as in the frontier of controlling the false discovery rate (FDR) [8, 9], which is particularly useful for the study of DNA microarray and protein arrays. It is also increasingly used in

genomewide association studies [10]. On the other hand, the Bonferroni type adjustment is still surprisingly useful. For example, Klein et al. [1] successfully identified two SNPs which are associated with the age-related macular degeneration disease (AMD) using a Bonferroni adjustment. Witte et al. [11] provided an interesting observation that the relative sample size, based on Bonferroni adjustment, is approximately in a linear relationship to the logarithm of the number of comparisons.

An appropriate sample size is crucial for the success of studies involving a large number of comparisons. However, optimal and reliable sample size is extremely challenging to identify, as it typically depends on other design parameters that often have to be estimated based on preliminary data. Preliminary data are often limited at the design stage of studies, which lead to unreliable estimates of design parameters and create extra uncertainty in sample size estimation. Thus, it is of great practical interest to examine the relationship between sample size and other design parameters, such as the number of comparisons to be made. In this paper, we analyze this problem beyond witte et al.'s [11] observation by providing explicit sample size formulas, examining various genomic analyses, and deriving sample size formula for FDR control. The explicit sample size formulas are desirable because they elucidates how the change in other design parameters would affect sample size. This is of fundamental importance for understanding the reliability of study designs.

2. Sample Size Formulas

For testing a single hypothesis, the sample size problem is typically formulated as finding the number of subjects needed to ensure desired power $1-\beta$ for detecting an effect size Δ at a prespecified significance level α. Consider an one-sided test for equality of two normal means assuming known variances σ_1^2 and σ_2^2, respectively. The sample size per group (n) is as follows [12]:

$$n = \frac{\left(z_\alpha + Cz_\beta\right)^2}{\Delta^2}, \tag{2.1}$$

where $\Delta = |\mu_1 - \mu_2|/\sqrt{\sigma_1^2 + \sigma_2^2}$, $C = 1$, $\Phi(z_t) = 1-t$, and $\Phi(z)$ is the distribution function (CDF) of the standard normal distribution.

Many of the most widely used statistical tests have similar sample size formulas as in (2.1). For example, the commonly used Mann-Whitney test for comparing two continuous distributions without normality assumption has the same form of sample size formula as in (2.1). Similarly, for testing equality of two binomial proportions, using independent samples or using correlated samples as in McNemar's test, the sample size formulas are also of form (2.1) as discussed in Rosner [12].

For testing a single hypothesis, the influences of α, β, and Δ on the sample size n can be inferred easily from the above sample size formula (2.1), and are well known. When testing multiple hypotheses, one must guard against an abundance of false-positive results. The traditional criterion for error control in such situations is the familywise error rate (FWER), which is the probability of rejecting one or more true null hypotheses. The simplest and most commonly used method for controlling FWER is the Bonferroni correction, which is discussed in the next subsection.

2.1. FWER Control

In this section, we present sample size formulas for multiple comparisons in the context of controlling the familywise error rate (FWER). Suppose we make multiple comparisons with Δ being the same. If we wish to retain a familywise error rate α, and power $(1-\beta)$, then with the Bonferroni adjustment, $\alpha_{\text{bon}} = \alpha/M$, the sample size corresponding to (2.1) becomes

$$n_M = \frac{\left(z_{\alpha/M} + Cz_\beta\right)^2}{\Delta^2}. \tag{2.2}$$

To see how n_M changes as M increases, we can use the following well-known fact: when $\alpha < 0.5$, $\phi(z_\alpha)(1/z_\alpha - 1/z_\alpha^3) \le 1 - \Phi(z_\alpha) \le \phi(z_\alpha)/z_\alpha$. Since $\alpha/M = 1 - \Phi(z_{\alpha/M})$, we can approximate $z_{\alpha/M}$ by $z^*_{\alpha/M}$, where

$$z^{*2}_{\alpha/M} \equiv 2\log\left(\frac{M}{\alpha}\right) - \log(2\pi)\log\log\left(\frac{M}{\alpha}\right). \tag{2.3}$$

The explicit approximation of $z^2_{\alpha/M}$ in (2.3) works extremely well for M ranging from 10 to 10^{10}. Putting (2.3) into (2.2) yields the following approximation of the required sample size n_M:

$$n^*_M = \frac{\left(z^*_{\alpha/M} + Cz_\beta\right)^2}{\Delta^2}. \tag{2.4}$$

Then, for fixed (α, β, Δ), from (2.3) and (2.4), we have

$$n_M \approx n^*_M \approx \frac{2}{\Delta^2}\log\frac{M}{\alpha}, \quad \text{as } M \longrightarrow +\infty. \tag{2.5}$$

A few facts are self-evident from the above approximation. First, n_M is an approximately linear function of $\log M$ (base 10) with slope $2/\Delta^2$. Second, the impact of β on n_M (or n^*_M) is negligible when M is large. Third, a decrease in α is equivalent to an increase in M on n_M (or n^*_M). The impact of Δ on n_M (or n^*_M) is demonstrated in Figure 1 with $\alpha = 0.05$, $1-\beta = 0.90$, and $\Delta = 0.5, 1$, and 2, respectively. It shows that n_M (open circles) can indeed be approximated well by a linear function of $\log M$. The lines are calculated based on approximate normal quantiles (2.4) for n^*_M. Moreover, when Δ is large (e.g., $\Delta = 2$), the slope is very small.

The simple Bonferroni correction is very useful, when the number of true alternatives is small. This often occurs, for example, in candidate gene association studies. The Bonferroni approach is easy to apply, for example, it is convenient when the hypotheses involve many covariates and nuisance parameters, whereas the permutation approaches may not be applicable, because they require some symmetry or exchangeability on the null hypotheses [13, 14]. Next, we give two practical examples to illustrate the growth rate of sample size relative to the number of tests M to be performed.

Table 1: An SNP from Klein et al. [1].

Attribute	rs1329428 (C/T)
Risk allele	C
OR (dominant)	4.7
Freq in HapMAP CEU	82%
OR (recessive)	6.2
Freq in HapMAP CEU	41%

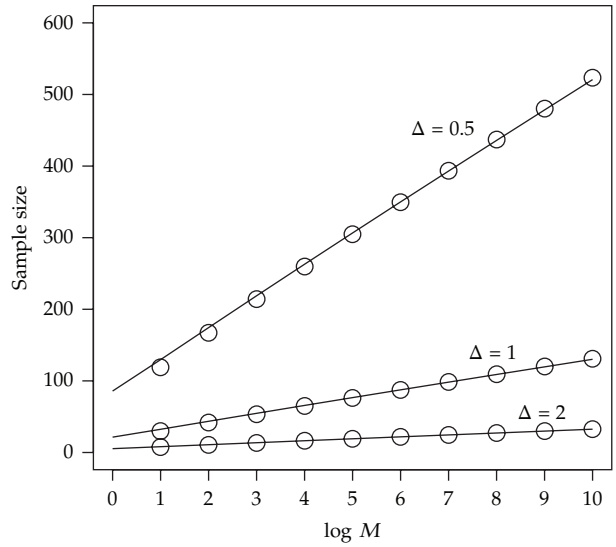


Figure 1: Sample size versus log M (base 10) to detect effect sizes $\Delta = 0.5, 1$ or 2 with $1 - \beta = 90\%$ power at the familywise significance level $\alpha = 5\%$, when Bonferroni adjustment is used. The open circles represent the sample sizes calculated based on exact normal quantiles (2.2).

The AMD Example

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. Klein et al. [1] reported a genomewide screen of 96 cases and 50 controls for polymorphisms associated with AMD. They examined 116,204 single-nucleotide polymorphisms (SNPs). Two of the SNPs are found to be strongly associated with the disease phenotype. This is an example to test equality of two binomial proportions of two independent groups (cases and controls). The required sample size for each marker is given in (2.2) or (2.4) with $\Delta^2 = 2(p_1 - p_2)^2 \overline{p}\,\overline{q}$, $C = \sqrt{(p_1\, q_1 + p_2 q_2)/(2\overline{p}\,\overline{q})}$, and $\overline{p} = (p_1 + p_2)/2$. Illustration for sample size growth with the Bonferroni correction is plotted in Figure 2 against log M using the SNP rs1329428 (Table 1) identified in Klein et al. [1]. Using Bonferroni adjustment, the sample sizes are calculated to provide 90% power to detect the association at the familywise significance level $\alpha = 5\%$. The open circles and plus signs are sample sizes n_M using (2.2) according to the dominant and recessive odds ratios, respectively. The corresponding lines are sample sizes n_M^* based on (2.4).

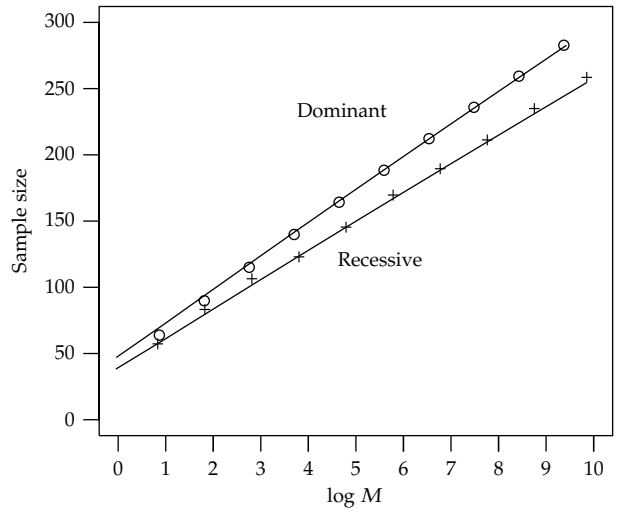


Figure 2: Sample sizes to detect the association at rs1329428 versus numbers of SNPS in genome wide screen of the AMD study.

The TDT Example

To test for linkage or association in family-based studies, the transmission/disequilibrium test (TDT) of Spielman et al. [15] examines the transmission of an allele from heterozygous parents to their affected offspring. If an allele is associated with the disease risk, its transmission may occur more than 50% of the times. Risch and Merikangas [16] studied the required sample size for TDT in affected sib pairs. TDT is equivalent to McNemar's test for two correlated proportions with the hypothesis $H_0 : p = 0.5$ versus $H_1 : p > 0.5$, for the specified alternative $p = p_A$, where p_A is the probability that an A/B parent transmits allele A to an affected offspring. The sample size (matched pairs) needed is given in (2.1) with $C = 2\sqrt{p_A(1-p_A)}$, $\Delta^2 = 2(p_A - 0.5)^2 p_D$, and p_D is the projected proportion of discordant pairs among all matched pairs. If we assume that each family used in the analysis has only one marker heterozygous parent, then n is the number of families required. Demonstration of sample sizes for TDT is plotted in Figure 3 using the setup given in Risch and Merikangas [16]. Using Bonferroni adjustment, the sample sizes are calculated to provide $1 - \beta = 90\%$ power to identify a disease gene at the familywise significance level $\alpha = 5\%$. The plus signs and open triangles are the sample size n_M calculated based on (2.2) corresponding to disease frequencies equal to 0.1 and 0.5, respectively. The corresponding lines are for n_M^* based on (2.4).

2.2. FDR Control

For the test of multiple hypotheses, such as the analysis of many genes using microarray, the outcomes can be described in Table 2.

It is likely that many genes are differentially expressed in a microarray study [7]. A natural way to control the overall false positives is to control the expected proportion of false

Table 2: Possible outcome for testing M hypotheses.

Truth	Test decision		Total
	Reject H_0	Accept H_0	
H_0	V	$m_0 - V$	m_0
H_1	U	$m_1 - U$	m_1
Total	R	$M - R$	M

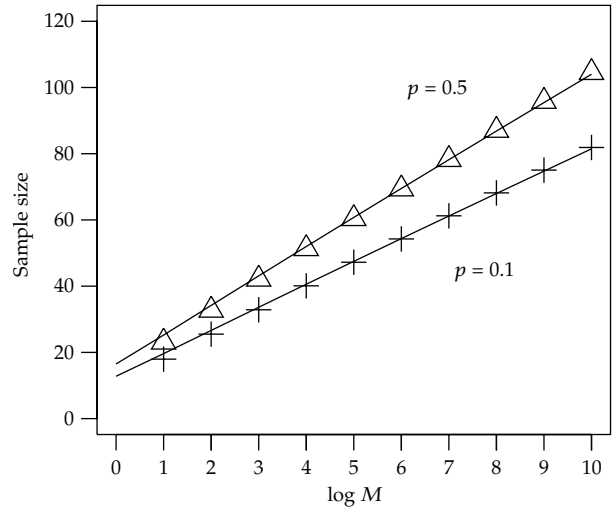


Figure 3: Number of families needed versus $\log M$ (base 10). Sample size for the TDT in the example of Risch and Merikangas [16], with disease frequencies of 0.1 (plus signs) and 0.5 (open triangles).

positives. Benjamini and Hochberg [8] defined the false discovery rate (FDR), using Table 2, as

$$\text{FDR} = P(R > 0)E\left[\frac{V}{U} \mid R > 0\right], \quad \text{FDR} = 0 \text{ for } R = 0. \tag{2.6}$$

Storey [9] defines positive FDR (pFDR) as $\text{pFDR} = \text{FDR}/P(R > 0)$. When M is large as assumed next, $P(R > 0) \approx 1$, unless the power $1 - \beta$ is too small, then $\text{FDR} \approx \text{pFDR}$.

The required sample size for multiple testing depends on α, $(1 - \beta)$, M, and Δ of each individual gene. For easy exposition, we assume an equal effect size Δ for all differentially expressed genes, say m_1 genes; thus, the power $(1 - \beta)$ of detecting any individual differentially expressed gene is the same for all of the m_1 genes between samples of two conditions of sizes n_1 and n_2. The expected outcomes in multiple testing can be expressed as functions of α, β, m_0, and m_1 and are summarized in Table 3.

By law of large numbers, from Table 3, $\text{FDR} = E(V/R) = m_0\alpha/(m_0\alpha + m_1(1 - \beta))$. Denote the desired FDR level by f. Then from the above equation, we have

$$\alpha_{\text{fdr}} = \frac{f}{1 - f}\left[\left(1 - \frac{m_1}{M}\right)^{-1} - 1\right](1 - \beta). \tag{2.7}$$

Table 3: Expected outcome for testing M hypotheses.

Truth	Test decision		Total
	Reject H_0	Reject H_a	
H_0	αm_0	$(1-\alpha)m_0$	m_0
H_1	$(1-\beta)m_1$	βm_1	m_1
Total	$\alpha m_0 + (1-\beta)m_1$	$(1-\alpha)m_0 + \beta m_1$	M

To account for the dependence among tests, we follow Shao and Tseng [17]. Let T_i be the test statistic of an one-sided two sample z-test for the ith alternative hypothesis, let p_i be its P value, and let $u_i = I(p_i < \alpha)$ be the rejection status at the level α; $u_i = 1$ if the ith test result is a rejection and 0 otherwise. Furthermore, if we denote the pairwise correlation coefficient between two tests by $\rho_U^{ij} = \mathrm{Corr}(T_i, T_j)$, then it can be shown that the correlation between u_i and u_j, $\theta_U^{ij} = \mathrm{Corr}(u_i, u_j)$ can be derived from the correlations of test statistics as follows:

$$\theta_U^{ij} = \frac{F\left(\tilde{z}_\alpha, \tilde{z}_\alpha; \rho_U^{ij}\right) - (1-\beta)^2}{\beta(1-\beta)}, \tag{2.8}$$

where F is the CDF of the standard bivariate normal distribution, and $\tilde{z}_\alpha = -z_\alpha + \Delta/\sqrt{n_1^{-1} + n_2^{-1}}$ [18]. Under local dependence assumptions, the total number of true discoveries, $U = \sum_{i=1}^{m_1} u_i$, has an approximately normal distribution: $U \sim N(m_1(1-\beta), \sigma_U^2)$, where $\sigma_U^2 = m_1\beta(1-\beta)[1 + \overline{\theta}_U(m_1 - 1)]$, and $\overline{\theta}_U = (m_1(m_1-1))^{-1}\sum_{i\neq j}\theta_U^{ij}$ is the average correlation among true discoveries. The local dependence assumption can be viewed in a simplified formulation of the central limit theorem under the "strong mixing" given in Theorem 27.4 of Billingsely [19]. "Mixing" means, roughly, that random variables temporally far apart from one another are nearly independent. We think that the local dependence assumption is reasonable in many genetic studies. For example, linkage disequilibrium can result in local dependence of genetic markers. In biomarkers study, biomarkers of the same pathway are often correlated and result in local dependence.

It is often desirable to find sample size to ensure a familywise power Ψ of identifying at least a given fraction $r \in (0,1)$ out of m_1 true discoveries: $\Psi = P(U \geq [m_1 r])$. The above normal approximation of U allows a closed form solution for the comparison-wise β:

$$\beta_{\mathrm{fdr}} = 1 - r - \frac{1 - 2r + \sqrt{4m_1^* r(1-r) + 1}}{2m_1^* + 2}, \tag{2.9}$$

where $m_1^* = m_1/\{[1 + \overline{\theta}_U(m_1 - 1)]z_{1-\Psi}^2\}$. When m_1 is large, to have a family-wise power Ψ in detecting at least $100r\%$ out of m_1 true alternatives, and with an FDR f, the sample size needed for a one-sided z-test is given by (2.1), with α and β determined by (2.7) and (2.9) iteratively.

A Microarray Example.

We now consider a well-known dataset from a study of leukemia in Gloub et al. [2] to demonstrate the relationship between sample size and number of multiple comparisons when

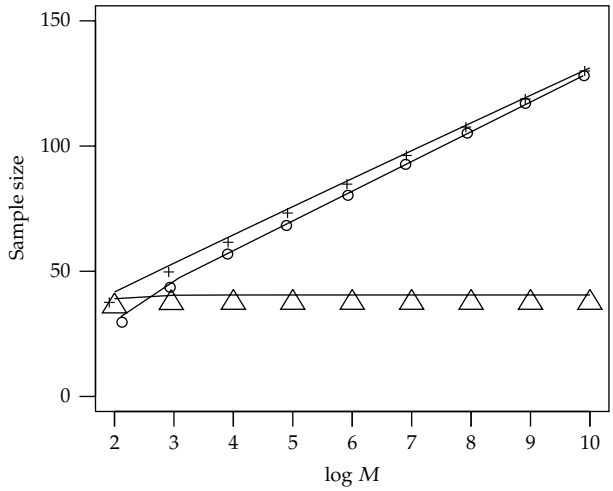


Figure 4: Sample size versus $\log M$ (base 10) for controlling FDR $f = 5\%$ with $\Psi = 90\%$. The open circles represent the sample sizes needed when the number of true alternatives m_1 stays as constant ($m_1 = 40$), the plus signs give the sample sizes when $m_1 = 2\log M$, and the triangles are the sample sizes when the proportion of true alternatives is constant ($m_1 = M/10$).

controlling FDR. The original purpose of the experiment described in Gloub et al. [2] is to identify the susceptible genes related to clinical heterogeneity in two subclass of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset contains 7129 attributes from 47 patients with ALL and 25 patients with AML. We can apply (2.1), (2.7), and (2.9) iteratively to obtain the required sample size when controlling FDR. Figure 4 provides 3 different settings for controlling FDR $f = 5\%$ with $\Psi = 90\%$. Based on the top 100 most differentially expressed genes in Gloub et al. [2], $\overline{\theta}_U = 0.07$ (see (2.9)). The open circles represent the sample sizes n_M needed when the number of true alternatives m_1 stays constant ($m_1 = 40$). In this case, we observe that the sample size is a linear function of $\log M$ as M increases. The "plus" signs denote the sample sizes n_M when the number of true alternatives increases in a slower pace than M ($m_1 = 2\log M$); the sample size is also approximately a linear function of $\log M$. The triangles denote the sample sizes n_M when the proportion of true alternatives is constant ($m_1/M = 10\%$), and the sample sizes roughly remain constant as the number of tests increases which is expected from (2.7). The lines in Figure 4 represent sample sizes n_M^* based on (2.4).

3. Discussion

In this short paper, we have shown that a large increase in the number of comparisons often only requires a small increase in the sample size. We further demonstrated that when controlling FDR, the sample size may even sometimes stay constant as the number of comparisons increases (Figure 4). The sample size required for testing M hypotheses is generally not growing faster than a linear function of $\log M$, even when a simple Bonferroni adjustment is used, and the slope of the linear growth rate (in $\log M$) is small when detecting a large effect size. These results have important implications in practice due to the wide use of multiple comparisons.

In this paper, we discuss the sample size formulas based on fixed effect size in alternative hypotheses. In reality, the effect sizes may follow a distribution, and simulation method may be useful in determining the sample size. We used z-test to derive the sample size formula, because large sample size is usually required for studies with multiple comparisons. If the effect size is large and sample size is small, t-test may be more appropriate. However, we expect the relationship between sample size and the logarithm of number of comparisons made is still linear.

In practice, if feasible, using a conservative sample size can reduce the chance of obtaining false-positive results and ensure reproducibility [6]. The simple sample size formulas provided in this paper might be used to select a suitable sample size by varying other design parameters and by taking into consideration the reliability of the proposed designs. While FDR is very useful and is increasingly used in multiple comparisons, our experience in helping biomedical investigators and the analysis in this paper indicate that the simple Bonferroni approach can often provide conservative but useful sample sizes in many situations.

References

[1] R. J. Klein, C. Zeiss, E. Y. Chew et al., "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] P. H. Westfall and S. S. Young, *Resample-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, New York, NY, USA, 1993.

[4] J. C. Hsu, *Multiple Comparisons: Theory and Methods*, Chapman & Hall, London, UK, 1996.

[5] P. H. Westfall and R. D. Wolfinger, "Multiple tests with discrete distributions," *American Statistician*, vol. 51, no. 1, pp. 3–8, 1997.

[6] N. J. Risch, "Searching for genetic determinants in the new millennium," *Nature*, vol. 405, no. 6788, pp. 847–856, 2000.

[7] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[8] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[9] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, vol. 64, no. 3, pp. 479–498, 2002.

[10] Q. Yang, J. Cui, I. Chazaro, L. A. Cupples, and S. Demissie, "Power and type I error rate of fales discovery rate approaches in genome-wide association studies," *BMC Genetics*, vol. 6, supplement 1, article S134, 2005.

[11] J. S. Witte, R. C. Elston, and L. R. Cardon, "On the relative sample size required for multiple comparisons," *Statistics in Medicine*, vol. 19, no. 3, pp. 369–372, 2000.

[12] B. Rosner, *Fundamentals of Biostatistics*, Duxbury, Los Angeles, Calif, USA, 2006.

[13] Y. Ge, S. Dudoit, and T. P. Speed, "Resampling-based multiple testing for microarray data analysis," *Test*, vol. 12, no. 1, pp. 1–77, 2003.

[14] Y. Huang, H. Xu, V. Calian, and J. C. Hsu, "To permute or not to permute," *Bioinformatics*, vol. 22, no. 18, pp. 2244–2248, 2006.

[15] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *American Journal of Human Genetics*, vol. 52, no. 3, pp. 506–516, 1993.

[16] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[17] Y. Shao and C. H. Tseng, "Sample size calculation with dependence adjustment for FDR-control in microarray studies," *Statistics in Medicine*, vol. 26, no. 23, pp. 4219–4237, 2007.

[18] H. Ahn and J. J. Chen, "Generation of over-dispersed and under-dispersed binomial variates," *Journal of Computational and Graphical Statistics*, vol. 4, no. 1, pp. 55–64, 1995.
[19] P. Billingsley, *Probability and Measure*, John Wiley & Sons, New York, NY, USA, 1995.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 151259, 15 pages
doi:10.1155/2012/151259

Research Article

Genotype-Based Bayesian Analysis of Gene-Environment Interactions with Multiple Genetic Markers and Misclassification in Environmental Factors

Iryna Lobach[1] and Ruzong Fan[2]

[1] *Department of Population Health, Division of Biostatistics, School of Medicine, New York University, New York, NY 10016, USA*
[2] *Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD 20852, USA*

Correspondence should be addressed to Iryna Lobach, iryna.lobach@nyumc.org

Received 1 March 2012; Revised 23 May 2012; Accepted 25 May 2012

Academic Editor: Wei T. Pan

A key component to understanding etiology of complex diseases, such as cancer, diabetes, alcohol dependence, is to investigate gene-environment interactions. This work is motivated by the following two concerns in the analysis of gene-environment interactions. First, multiple genetic markers in moderate linkage disequilibrium may be involved in susceptibility to a complex disease. Second, environmental factors may be subject to misclassification. We develop a genotype based Bayesian pseudolikelihood approach that accommodates linkage disequilibrium in genetic markers and misclassification in environmental factors. Since our approach is genotype based, it allows the observed genetic information to enter the model directly thus eliminating the need to infer haplotype phase and simplifying computations. Bayesian approach allows shrinking parameter estimates towards prior distribution to improve estimation and inference when environmental factors are subject to misclassification. Simulation experiments demonstrated that our method produced parameter estimates that are nearly unbiased even for small sample sizes. An application of our method is illustrated using a case-control study of interaction between early onset of drinking and genes involved in dopamine pathway.

1. Introduction

A key component to prevention and control of complex diseases, such as cancer, hypertension, diabetes, and alcoholism, is to study the independent, cumulative, and interactive effects of genetic and environmental factors. This analysis has the potential to impact the

understanding of the role of genetic influences under various environmental exposures, thus providing valuable information to (1) better understand the biological pathways involved in the disease and its progression, thus providing major clues to the underlying causes of alcohol dependence; (2) design personalized interventions targeted to individuals with enhanced vulnerability to the disease (the risk genes may help identify patients at higher risk long before any symptoms occur); (3) gain critical understanding for drug discovery.

This work is motivated by the following two concerns in the analysis of gene-environment interactions. **First**, complex diseases are caused by multiple variants with small-to-moderate effect sizes working in concert [1]. Most of the results of published genome-wide association studies are based on single nucleotide polymorphism (SNP) analysis [2]. This approach may suffer from low power due to a large number of tests and small effect sizes of individual SNPs. Furthermore, the true causal genetic marker is often not genotyped, rather is captured through linkage disequilibrium (LD) with the typed markers. Since each SNP has only partial linkage disequilibrium with the causal SNP, the observed effect size of the typed SNP is lower than the effect size of the causal SNP. In light of this concern, we propose to use a risk function that allows the genetic markers in linkage disequilibrium to enter the model directly [3]. This model eliminates the need to estimate haplotype phase and hence protects against bias due to the uncertainty that may arise due to the haplotype phase ambiguity [4–8]. In addition, the computation burden can be significantly reduced since the proposed approach uses genotype data directly. **Second**, many variables that are of interest to biomedical researchers are subject to misclassification, for example, due to uncertainty associated with a recall or a measurement at an individual level. Misclassification may result in bias and loss of power to detect gene-environment interactions [9]. Oftentimes uncertainty associated with these variables may not be avoided in practice. The loss of power prevents the ability to discover gene-environment interactions in small studies or studies involving analysis of subtypes of complex diseases.

An example of biomedical problem of gene-environment interactions is the analysis of role of age when first got drunk in the etiology of alcohol dependence. The age at which a person gets drunk for the first time may influence genes linked to alcoholism, making the youngest drinkers most susceptible to severe problems [10]. Twin study found that when twins started drinking early (age < 13 years old), genetic factors contributed greatly to risk for alcohol dependence, at rates as high as 90 percent in the youngest drinkers [10]. Some early-onset drinkers do not develop alcohol problems and some late-onset drinkers do, hence it is important to investigate genetic and environmental influences that predispose for or protect against alcohol dependence in these two groups. However, the definition of early age of getting drunk is subject to misclassification due to uncertainty associated with the recall.

In light of these concerns, we develop a Bayesian methodology for analysis of gene-environment interactions in case-controls studies. Estimation and inference are based on a pseudolikelihood function [3, 11, 12]. This pseudolikelihood function offers the following advantages. One is that environmental variables measured exactly are modeled completely nonparametrically. Furthermore, *a priori* information about the probability of disease can be incorporated directly. The pseudolikelihood function exploits gene-environment independence assumption which is a reasonable assumption in many practical applications. If the gene-environment interaction is not significantly present in the population, then the distribution of genotype can be specified within strata defined by an environmental covariate. The proposed analysis is based on a pseudolikelihood function hence conventional Bayesian techniques may not be applied directly. Validity of Bayesian techniques need to be examined when the likelihood function is not a proper likelihood [13]. We followed Monahan and Boos

[13] and Lobach et al. [3] to validate our Bayesian approach under this pseudolikelihood function. Our Bayesian approach has the ability to shrink the parameter estimates towards prior and hence reduce variability in parameter estimates. This property is essential when environmental exposure is subject to misclassification, especially in studies with smaller sample sizes, for example, of subtypes of complex disease. On the other hand, if sample size is large enough, estimation and inference can be based on the asymptotic posterior distribution that we derived which will ease the computational burden.

An outline of this paper is as follows. In Section 2 we introduce notation and formally state the problem. In Section 2 we present the Bayesian model under various scenarios. Section 3 describes asymptotic posterior distribution. Section 4 describes simulation experiment. Section 5 describes application of the Bayesian model to the analysis of alcoholism study. Section 6 gives concluding remarks.

2. Bayesian Model Based on Pseudolikelihood

2.1. Notation and Risk Function

Consider a sample consisting of n_0 controls and n_d cases at disease stage or type $d = 1, \ldots, K$. Define D as the disease status. Following Lobach et al. [11], we pretend that this case-control sample is collected using a simple Bernoulli scheme, where the selection probability of a subject given disease status is proportional to $n_d/\mathrm{pr}(D = d)$, $d = 0, 1, \ldots, K$. Let $R = 1$ denote the indicator of whether or not a subject is selected into the case-control sample. All participants of the study will have this selection status $R = 1$. The observed genetic data consist of unphased genotypes $G = (G_1, \ldots, G_I)$ at I loci. Let $Q(G; \theta)$ be a model describing Hardy-Weinberg equilibrium (HWE).

Let (T, Z) denote all nongenetic variables of interest. Suppose T is the set of factors subject to misclassification, and Z is the set of variables observed exactly. We assume that the observed genetic data does not contain any additional information on disease status and the true environmental covariate given the genetic variable of interest. Let X denote the error-prone version of T. Suppose the misclassification process is defined by the following parametric structure $p_{\mathrm{miss}}(x \mid T, G, Z, D, \xi)$. This model is general enough to capture differential misclassification. The joint distribution of the environmental factors in the underlying population can be specified in the following form $p_{T|Z}(t \mid z, \xi) f_Z(z)$. While T may be a vector of factors, for simplicity of presentation in what follows we suppose that T is a factor.

Given the environmental covariates T and Z, genotype data G, the risk of disease in the underlying population is given by the following polytomous logistic model:

$$\mathrm{pr}(D = k \geq 1 \mid \mathbf{G}, T, Z) = \frac{\exp\{\beta_{k0} + m_k(\mathbf{G}, T, Z; \beta)\}}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{G}, T, Z; \beta)\}}, \tag{2.1}$$

where $m(\bullet)$ is a function of known form parameterizing the risk of disease in terms of parameters β. For the ith marker, denote the two alleles by M_i and m_i, with frequencies P_{M_i} and P_{m_i}, respectively. Following Lobach et al. [3], we define the following dummy variables and two risk models: genotype effect model and additive effect model.

Define the following dummy variables:

$$A_i = \begin{cases} 1, & \text{if } G_i = M_iM_i, \\ 0, & \text{if } G_i = M_im_i, \\ -1, & \text{if } G_i = m_im_i, \end{cases} \qquad B_i = \begin{cases} -P_{m_i}^2, & \text{if } G_i = M_iM_i, \\ P_{M_i}P_{m_i}, & \text{if } G_i = M_im_i, \\ -P_{M_i}^2, & \text{if } G_i = m_im_i. \end{cases} \tag{2.2}$$

Notice that $A_i + 1$ is the number of allele M_i at the ith marker, and hence A_i can be used to model the allele or additive effect of M_i. Let $\mathrm{pr}(g;\theta)$ be a parametric form of the joint distribution of the observed genetic markers. In the following, we provide two examples of function $m_k(\cdot)$ using the genotype information $\mathbf{G} = (G_1, G_2, \ldots, G_I)$.

2.1.1. Genotype Effect Model (GEM)

The following specification of the risk function incorporates both additive and dominance effects of genotype, as well as the multiplicative gene-environment interactions

$$\begin{aligned} m_k(\mathbf{G}, T, Z; \beta) = m_k(\mathcal{A}, \mathcal{B}, T, Z; \beta) &= T\beta_{kT} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} \\ &+ \sum_{i=1}^{I} TA_i\beta_{kATi} + \sum_{i=1}^{I} ZA_i\beta_{kAZi} + \sum_{i=1}^{I} B_i\beta_{kDi} + \sum_{i=1}^{I} TB_i\beta_{kDTi} + \sum_{i=1}^{I} ZB_i\beta_{kDZi}. \end{aligned} \tag{2.3}$$

In this formulation, the regression coefficients β_{kAi} and β_{kDi} model risk due to the additive and dominance effect, respectively [14, 15]. The remaining terms capture the multiplicative gene-environmental interaction.

2.1.2. Additive Effect Model (AEM)

Suppose that the dominance effect is not significantly present in the model (2.3). In this situation, the risk function takes the following form:

$$m_k(\mathbf{G}, T, Z; \beta) = m_k(\mathcal{A}, T, Z; \beta) = T\beta_{kT} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} + \sum_{i=1}^{I} TA_i\beta_{kATi} + \sum_{i=1}^{I} ZA_i\beta_{kAZi}. \tag{2.4}$$

2.2. Pseudolikelihood

Let us denote $\kappa_k = \beta_{k0} + \log(n_k/n_0) - \log(\pi_k/\pi_0)$, $k = 1, 2, \ldots, K$, and $\tilde{\kappa} = (\kappa_1, \ldots, \kappa_K)^T$. In addition, let $\tilde{\beta}_0 = (\beta_{10}, \ldots, \beta_{K0})^T$, $\Omega = (\tilde{\beta}_0^T, \beta^T, \Theta^T, \tilde{\kappa}^T)^T$, $\mathcal{B} = (\Omega^T, \eta^T)^T$, and $\upsilon = (\eta^T, \xi^T)^T$. Define

$$S(k, \mathbf{g}, t, z; \Omega) = \frac{\exp\left[1_{(k\geq 1)}(k)\{\kappa_k + m_k(\mathbf{g}, t, z; \beta)\}\right]}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{g}, t, z; \beta)\}} \mathrm{pr}(\mathbf{g}; \Theta). \tag{2.5}$$

We assume that G and (X, Z) are independently distributed in the underlying population. Only changes in notation are needed to model genotype and environment within strata thus relaxing gene-environment independence assumption. An example of gene-environment dependence is polymorphisms in nicotine metabolism pathway that may regulate the degree of addiction to nicotine, thus creating gene-environment interaction. Furthermore, these polymorphisms may interact with smoking status while being involved in lung cancer [16]. We suppose that the type of genetic covariate measured does not depend on the individual's true genetic covariate, given disease status, environmental covariates and the measured genetic information. Furthermore, we suppose that the observed genetic variable does not contain any additional information on disease status and true environmental covariate given the genetic variable of interest.

Similarly to Lobach et al. [11], we propose to use the following pseudolikelihood function in place of the likelihood function to estimate the parameters:

$$\begin{aligned} L_{\text{Pseudo}}(k, \mathbf{g}, x, z; \Omega, \eta, \xi) &\equiv \text{pr}(D = k,\ \mathbf{G} = \mathbf{g},\ X = x \mid Z = z,\ R = 1) \\ &= \frac{\sum_{t^*} S(k, \mathbf{g}, t^*, z; \Omega) p_{\text{miss}}(x \mid k, \mathbf{g}, t^*, z; \xi) f_T(t \mid z; \eta)}{\sum_{k^*=0}^{K} \sum_{t^*} \sum_{\mathbf{g} \in \mathcal{G}} \int S(k^*, \mathbf{g}, t^*, z; \Omega) f_T(t^* \mid z; \eta)}, \end{aligned} \tag{2.6}$$

where $\mathcal{G}$ is the set of all possible genotypes in the population. Lobach et al. [12] proved that maximization of L_{Pseudo}, although not the actual retrospective likelihood for case-control data, leads to consistent and asymptotically normal parameter estimates. Observe that conditioning on Z in L_{Pseudo} allows it to be free of the nonparametric density function $f_Z(z)$, thus avoiding the difficulty of estimating potentially high-dimensional nuisance parameters.

2.3. Bayesian Analysis Based on Pseudolikelihood

Since in our setting the retrospectively collected data is analyzed as if they were coming from a random sample, function (2.6) is not a real likelihood function and hence the traditional Bayesian analysis is not technically correct. Conventional approaches to validity of posterior probability statements follow from the definition of the likelihood as the joint density of observations.

For simplicity of presentation we introduce new notation for this section only.

Monahan and Boos [13] introduced a definition based on coverage of posterior sets that are constructed to contain the correct probability of including a parameter τ, if the underlying distribution of τ is the prior $p(\tau)$ and the model $f(Y \mid \tau)$ of data Y is correct. This approach has been used in gene-environment interaction setting [3]. For example, in the one-dimensional case, the natural posterior coverage set functions are the one-sided intervals $I_\alpha^* = R_\alpha(Y) = (-\infty, \tau_\alpha^*)$, where τ_α^* is α-percentile of the posterior $f(Y \mid \tau)$. Validity for such a posterior means that all these intervals I_α^* have the correct coverage α. In practice, it is often challenging to verify the required probability analytically. Monahan and Boos [13] proposed a convenient numerical method. Briefly, define τ_k, $k = 1, \ldots, m$ to be a sample generated independently from a continuous prior $p(\tau)$. For each τ_k, let Y^k denote a value generated from $f(Y \mid \tau_k)$. In addition, for each k, define H_k to be a variable in the following form:

$$H_k = \int_{-\infty}^{\tau_k} f\left(\tau \mid Y^k\right) d\tau. \tag{2.7}$$

This corresponds to posterior coverage set functions of the form $(-\infty, \tau_\alpha^k)$, where τ_α^k is the αth percentile point of posterior density $f(\tau \mid Y^k)$. Monahan and Boos [13] argued that if the distribution of H_k fails to follow the uniform distribution for any prior, then the likelihood function cannot be a coverage proper Bayesian likelihood.

We propose to use the methodology described above to validate the likelihood function and apply conventional MCMC techniques to estimate parameters. We note that the method developed by Monahan and Boos is devised to invalidate a pseudolikelihood. Therefore to validate a pseudolikelihood, we propose to consider a comprehensive set of scenarios to examine coverage probabilities of posterior sets, and if these scenarios fail to invalidate a pseudolikelihood, we suppose that it is valid.

2.4. Fully Bayesian Model

We consider the case when the environmental covariates (T, X), genetic variant G, and disease status D are binary. Let $\text{pr}(G = 1) = \theta$, $\text{pr}(T = 1) = \eta$. For simplicity of presentation, consider an additive model. Define the vector of risk parameters $\mathcal{B} = (\beta_t, \beta_A, \beta_B, \beta_{tA}, \beta_{tB})^T$. Consider a multiplicative interaction and let $m(t, g, \mathcal{B}) = \beta_t t + \beta_A A + \beta_{tA} tA + \beta_B B + \beta_{tB} tB$. Make the following definition:

$$S(d, g, t, \mathcal{B}, \theta) = \frac{\exp\left[I_{(d\geq 1)}(d)\left\{\kappa_d + m(t, g, \mathcal{B})\right\}\right]}{1 + \exp\left\{\beta_0 + m(t, g, \mathcal{B})\right\}} \theta^g (1-\theta)^{1-g}. \tag{2.8}$$

If X is an observed environmental covariate prone to misclassification, denote the misclassification probabilities as $\text{pr}(X = 1 \mid T = 0) = \xi_1$ and $\text{pr}(X = 0 \mid T = 1) = \xi_0$. Hence, the distribution of misclassification process is $f_{\text{mem}}(x \mid t, \xi_0, \xi_1) = \{x\xi_1 + (1-x)(1-\xi_1)\}(1-t) + \{x(1-\xi_0) + (1-x)\xi_0\}t$.

On the risk parameters, we impose a normal prior with mean $\mu_{\mathcal{B}}$ and covariance matrix $\Sigma_{\mathcal{B}}$.

Similarly to the appendix in Fan and Xiong [14] and Lobach et al. [3], the following expectations, variances, and covariances can be derived. $E(A_i) = P_{M_i} - P_{m_i}$, $E(B_i) = 0$, $\text{Var}(A_i) = 2P_{M_i}P_{m_i}$, $\text{Var}(B_i) = P_{M_i}^2 P_{m_i}^2$, $\text{Cov}(A_i, A_j) = 2\Delta_{M_iM_j}$, $\text{Var}(B_i, B_j) = \Delta_{M_iM_j}^2$, $i \neq j$. And $\text{Cov}(A_i, B_i) = 0$ for all i and j;

$$\mathbf{V_A} = 2\begin{pmatrix} P_{M_1}P_{m_1} & \Delta_{M_1M_2} & \cdots & \Delta_{M_1M_I} \\ \Delta_{M_1M_2} & P_{M_2}P_{m_2} & \cdots & \Delta_{M_2M_I} \\ \vdots & \vdots & \cdots & \vdots \\ \Delta_{M_1M_I} & \Delta_{M_2M_I} & \cdots & P_{M_I}P_{m_I} \end{pmatrix}, \qquad \mathbf{V_D} = \begin{pmatrix} P_{M_1}^2 P_{m_1}^2 & \Delta_{M_1M_2}^2 & \cdots & \Delta_{M_1M_I}^2 \\ \Delta_{M_1M_2}^2 & P_{M_2}^2 P_{m_2}^2 & \cdots & \Delta_{M_2M_I}^2 \\ \vdots & \vdots & \cdots & \vdots \\ \Delta_{M_1M_I}^2 & \Delta_{M_2M_I}^2 & \cdots & P_{M_I}^2 P_{m_I}^2 \end{pmatrix}. \tag{2.9}$$

Define $\mathcal{A} = (A_1, \ldots, A_I)$ and $\mathcal{B} = (B_1, \ldots, B_I)$. Let $\mathbf{O_I}$ be a $I \times I$ matrix with zero elements. Based on the expectations and covariances described above, we have $\text{Cov}(\mathcal{A}, \mathcal{B}) = \begin{pmatrix} \mathbf{V_A} & \mathbf{O_I} \\ \mathbf{O_I} & \mathbf{V_D} \end{pmatrix}$.

In the case when misclassification is large, the sampling distribution of risk parameter estimates is likely to be skewed [11, 17]. However, because the shape of the normal distribution is symmetric, this prior is likely to bring the sampling distribution of the risk parameter estimates closer to normal. For the frequency parameters η and θ, we use

noninformative uniform (0,1) priors. In this setting, the prior information imposed on θ is noninformative. If *a priori* information is available about the genotype frequencies, it can be specified using a corresponding distribution or HWE.

Then, the joint posterior distribution for the model unknowns is proportional to

$$\prod_{i=1}^{n} \frac{\sum_{t^*=0}^{1} S(d_i, g_i, t^*, \mathcal{B}, \theta) p_{\text{miss}}(x_i \mid t^*, \xi_0, \xi_1) \eta^{t^*} (1-\eta)^{1-t^*}}{\sum_{t^*=0}^{1} \sum_{d=0}^{1} \sum_{g=0}^{1} S(d_i, g_i, t^*, \mathcal{B}, \theta) \eta^{t^*} (1-\eta)^{1-t^*}} \times |\Sigma_{\mathcal{B}}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathcal{B} - \mu_{\mathcal{B}})^T \Sigma_{\mathcal{B}}^{-1} (\mathcal{B} - \mu_{\mathcal{B}}) \right\} \times I_{(0,1)}(\eta) \times I_{(0,1)}(\theta). \tag{2.10}$$

Note that in this formulation, we specify a known misclassification process. We recommend performing sensitivity analysis to see whether parameter estimates change when misclassification probabilities are specified slightly differently. Furthermore, we recommend conservative setting when LD is set to be zero as *a priori*.

3. Asymptotic Posterior Distribution

We now consider properties of an asymptotic posterior distribution based on the pseudolikelihood (2.6). MCMC techniques can be computationally challenging. Knowing the form of an asymptotic posterior distribution would ease the computational burden.

For simplicity, we suppose that the parameter ξ that controls misclassification error distribution is known, although this is not required. Denote Θ_0 and $\widehat{\Theta}_n$ to be values that maximize prior and pseudolikelihood, respectively. Let $\Psi(d, g, x, z, \Theta, \xi)$ be the derivative of $\log\{L_i(d, g, x, z, \Theta, \xi)\}$ with respect to Θ and

$$\Lambda = \sum_{d} \frac{n_d}{n} E\{\Psi(D, G, X, Z, \Omega, \eta, \xi) \mid D = d\} \times E\{\Psi(D, G, X, Z, \Omega, \eta, \xi) \mid D = d\}^T. \tag{3.1}$$

Furthermore, if $p(\Theta)$ is the prior distribution of the vector of parameters, define $l(\Theta)$ to be the derivative of $\log\{p(\Theta)\}$ with respect to Θ. Then define

$$\mathcal{L}_n(\Theta, \xi) = \sum_{i=1}^{n} \Psi(D_i, G_i, X_i, Z_i, \Theta, \xi) \tag{3.2}$$

and matrices

$$\mathcal{I}(\Theta) = -E\left[\frac{\partial\{\mathcal{L}_n(\Theta, \xi)\}}{\partial(\Theta)}\right]; \qquad \mathcal{J}(\Theta) = -E\left[\frac{\partial\{l(\Theta)\}}{\partial(\Theta)}\right]. \tag{3.3}$$

Bernardo and Smith [18] showed that under suitable regularity conditions the posterior distribution of vector of parameters $\widehat{\Theta}_n$ converges to normal $\mathcal{N}(\mathcal{M}, \Sigma)$ distribution. Mean vector and covariance matrix can be consistently estimated as follows:

$$\begin{aligned} \widehat{\mathcal{M}_n} &= \widehat{\Sigma}_n^{-1}\left\{\mathcal{J}\left(\widehat{\Theta}_n\right)\widehat{\Theta}_n + \mathcal{J}(\Theta_0)\Theta_0\right\}, \\ \widehat{\Sigma}_n &= \left\{\mathcal{J}(\widehat{\Theta}_n) + J(\Theta_0)\right\}^{-1}. \end{aligned} \tag{3.4}$$

It can be easily seen that $n^{-1}\partial\{\mathcal{L}_n(\widehat{\mathcal{B}}, \xi)\}/\partial\mathcal{B}^T$ is a consistent estimate of $\mathcal{J}(\Theta)$. Alternatively, if $\widehat{\Sigma}$ is the sample covariance matrix of the terms $\Psi(D_i, G_i, X_i, Z_i, \widehat{\mathcal{B}}, \xi)$, then $\widehat{\Sigma} + \widehat{\Lambda}$ consistently estimates $\mathcal{J}(\Theta)$.

Note that the posterior distribution has precision equal to the sum of precision provided by the observed data and the prior precision matrix. This formulation suggests an approximation, namely, that for large n, prior is small compared to the one provided by the observed data. Hence, with a large sample size, one can reduce computational burden by using the asymptotic distribution and using precision provided by the observed data while specifying the posterior distribution.

4. Simulation Experiments

We investigated the case of small $n_0 = n_1 = 350$ and large ($n_0 = n_1 = 1{,}500$) sample sizes.

We validated the pseudolikelihood function using methodology described by Monahan and Boss [13] in a few scenarios by varying sample size, effect size, and misclassification probabilities. In 96% of cases that we considered, the Kolmagorov-Smirnov test failed to reject the null hypothesis that the sample of H_k (2.7) comes from the uniform (0,1) distribution at the 0.05 significance level. Hence, we concluded that the pseudolikelihood is valid for subsequent analysis. Hence, we proceeded to estimating parameters.

We implemented Metropolis-Hastings algorithm in the following setting. On the risk parameters $\mathcal{B}$, we imposed a normal $\mathcal{N}(\mathcal{B}^{\text{mean}}, \Sigma_{\mathcal{B}})$ prior, where $\mathcal{B}^{\text{mean}}$ is equal to the pseudo-MLE estimates. To examine sensitivity of the estimates to this specification, we considered a case when $\mathcal{B}^{\text{mean}}$ is a vector of zero values. Covariance matrix was specified as a diagonal matrix with diagonal elements equal to 3^2. Alternatively, we specified the corresponding matrix according to the known structure that is a function of LD. In all of these scenarios, the results we obtained were comparable. Table 1 presents results based on $\mathcal{B}^{\text{mean}} = (0, 0, 0)$ and covariance matrix with diagonal elements equal to 3^2.

To examine performance of our approach, we performed two simulation experiments. In the first experiment, we investigated performance of Bayesian method compared to pseudo-MLE. The goal of this experiment was to examine the ability of Bayesian approach to shrink the parameter estimates towards prior when misclassification causes the estimates to have skewed distribution. In the second experiment, we examined performance of the asymptotic posterior distribution.

Experiment 1. We generated the true environmental variables T from a binomial distribution with $\text{pr}(T = 1) = 0.5$. The misclassification probabilities are $\text{pr}(X = 0 \mid T = 1) = 0.20$ and $\text{pr}(X = 1 \mid T = 0) = 0.25$. We simulated three genetic markers in LD corresponding to $\Delta = 0.03$

Table 1: Biases and root mean squared errors (RMSEs) of risk parameters for the naive approach that ignores existence of misclassification and the proposed method in the case when $\mathrm{pr}(D = 1)$ is known and when it is estimated. The results are based on 500 samples of 1,500 cases and 1,500 controls. Genotype is simulated at the three marker loci with $P_{M_i} = 0.25$, $i = 1, 2, 3$, with linkage disequilibrium corresponding to $\Delta_{M_iM_j} = 0.03$. The environmental covariate (X) is binary and measured with error with misclassification probabilities being 0.20 for exposed and 0.25 for nonexposed subjects. The data is simulated and analyzed under the genotype effect model.

Parameter	True value	Naive analysis		Pseudo-MLE		MCMC	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
κ	0.484	0.481	0.231	−0.054	0.020	−0.032	0.013
β_X	0.693	−0.351	0.132	0.014	0.039	0.008	0.021
β_{A1}	0.406	0.257	0.073	−0.011	0.016	−0.003	0.009
β_{A2}	0.789	0.194	0.046	−0.003	0.015	−0.002	0.006
β_{A3}	0.693	0.283	0.089	−0.005	0.016	−0.003	0.008
β_{AX1}	0.916	−0.425	0.193	0.039	0.046	0.017	0.025
β_{AX2}	0.693	−0.317	0.113	0.038	0.041	0.023	0.021
β_{AX3}	1.099	−0.515	0.282	0.039	0.058	0.019	0.032
β_{D1}	0.262	0.299	0.133	0.026	0.152	0.009	0.368
β_{D2}	0.095	0.258	0.105	0.005	0.099	0.003	0.039
β_{D3}	0.693	0.231	0.092	0.018	0.128	0.008	0.087
β_{DX1}	1.099	−0.495	0.326	0.018	0.301	0.006	0.093
β_{DX2}	0.916	−0.413	0.235	0.006	0.208	0.005	0.121
β_{DX3}	1.099	−0.486	0.313	0.023	0.286	0.017	0.138
P_{M_i}	0.250	<0.001	<0.001	−0.001	<0.001	<0.001	<0.001
$\mathrm{pr}(X = 1)$	0.500			0.003	0.001	<0.001	<0.001
$\mathrm{pr}(D = 1)$	0.005			0.003	<0.001	<0.001	<0.001

and $P_{M_i} = 0.25$. In the study with $1,500$ cases and $1,500$ controls, we generated a binary disease status according to the following logistic model:

$$\mathrm{logit}\{\mathrm{pr}(D = 1 \mid G, X)\} = \beta_0 + \beta_t t + \sum_{j=1}^{3} \beta_{A_j} A_j + \sum_{j=1}^{3} \beta_{B_j} B_j + \sum_{j=1}^{3} \beta_{ATj} A_j T + \sum_{j=1}^{3} \beta_{TBj} B_j T. \quad (4.1)$$

To examine the case when genetic data is missing, we simulated a similar set of 1,500 cases and 1,500 controls with 50% of genetic information missing completely at random. To investigate a smaller study, we simulated 350 cases and 350 controls with the disease status defined by the risk model with all β_{B_j} and β_{BTj} set to 0. Results presented in Tables 1 and 2 illustrate that the proposed Bayesian approach produced parameter estimates that are less variable and less biased. We examine the empirical distribution of parameter estimates based on a small sample and found that it is skewed, which may be due to small sample size and presence of misclassification. We observed this phenomena in our previous work [3, 11]. The Bayesian solution brings the advantage, that is, a symmetric prior can shrink parameter estimates towards normal distribution. Furthermore, we presented performance of the naive approach that ignores existence of misclassification.

Table 2: Biases and root mean squared errors (RMSEs) of risk parameters obtained based on pseudo-MLE and the proposed MCMC. The results are based on 500 samples of 350 cases and 350 controls. Genotype is simulated at the two marker loci with $P_{M_i} = 0.25$, $i = 1, 2$. The environmental covariate (X) is binary and measured with error misclassification probabilities being 0.20 for exposed and 0.25 for nonexposed subjects. The data is simulated and analyzed under the additive effect model and the LD measure $\Delta_{M1M2} = 0.03$.

Parameter	True value	Pseudo-MLE		MCMC	
		Bias	RMSE	Bias	RMSE
β_X	1.099	0.035	0.392	0.013	0.236
β_{A1}	0.406	−0.268	1.035	−0.079	0.397
β_{A2}	0.789	−0.319	1.062	−0.085	0.372
β_{A3}	0.693	−0.293	1.043	−0.092	0.365
β_{AX1}	0.916	0.432	1.135	0.103	0.432
β_{AX2}	0.693	0.391	1.047	0.085	0.481
β_{AX3}	1.099	0.293	1.113	0.097	0.427

Table 3: Biases and root mean squared errors (RMSEs) of risk parameters obtained based on asymptotic posterior distribution. The results are based on 500 samples of 1,500 cases and 1,500 controls. Genotype is simulated at the two marker loci with $P_{M_i} = 0.25$, $i = 1, 2$. The environmental covariate (X) is binary and measured with error with misclassification probabilities being 0.20 for exposed and 0.25 for nonexposed subjects. The data is simulated and analyzed under the additive effect model and the LD measure $\Delta_{M1M2} = 0.03$.

Parameter	True value	Bias	Estimated SE	SE
β_X	0.693	0.010	0.032	0.039
β_{A1}	0.406	−0.005	0.012	0.015
β_{A2}	0.789	−0.004	0.011	0.014
β_{A3}	0.693	−0.004	0.016	0.016
β_{AX1}	0.916	0.023	0.045	0.044
β_{AX2}	0.693	0.019	0.061	0.058
β_{AX3}	1.099	0.020	0.052	0.054
β_{D1}	0.262	0.016	0.431	0.410
β_{D2}	0.095	0.009	0.052	0.063
β_{D3}	0.693	0.013	0.099	0.100
β_{DX1}	1.099	0.011	0.013	0.015
β_{DX2}	0.916	0.013	0.025	0.027
β_{DX3}	1.099	0.016	0.027	0.030

Experiment 2. We examined performance of estimation based on the derived asymptotic posterior in the simulation setup described in Experiment 1 corresponding to $n_1 = n_2 = 1,500$. Results presented in Table 3 illustrate that the parameter estimates are nearly unbiased. Moreover, estimated variances of parameter estimates are very close to the observed variability with one exception, namely, β_x. Variability of β_x may be inflated due to the misclassification in environmental exposure.

5. Analysis of Alcohol Dependence

The Collaborative Studies on the Genetics of Alcoholism (COGA) is a nine-center nationwide study that was initiated in 1989 and has had as its primary aim the identification of genes that contribute to alcoholism susceptibility and related characteristics [19–21]. COGA is funded through the National Institute on Alcohol Abuse and Alcoholism (NIAAA). The focus of this study is a case-control design of unrelated individuals for a genetic association analysis of addiction. Analyses that include incorporation of important demographic and environmental factors such as age when first got drunk, sex, income, and education into association studies are pursued. Our project involves analysis of 40 SNPs residing in genes involved in dopamine pathways. Specifically, we consider D2 dopamine receptor gene (DRD2) encoding a protein which plays a central role in reward-mediating mesocorticolimbic pathways; a member of the immunoglobulin gene superfamily NCAM1 encoding protein involved in various neural functions; tetratricopeptide repeat domain 12 gene (TTC12); CHRNA3 gene shown to be involved in higher craving after quitting and increased withdrawal symptoms over time. Cases are defined as individuals with DSM-IV alcohol dependence (lifetime). Controls are defined as individuals who have been exposed to alcohol, but have never met lifetime diagnosis for alcohol dependence or dependence on other illicit substances. The sample consists of 50.7% of male and 49.3% female participants; 60% report their race as Caucasian and 40% are non-Caucasian. We categorized age when first got drunk as "Early" if it is less or equal to 13 (EAD = 1, 45.2% of all participants) and people with low income are the ones who make less than 30 K per year (LI = 1, 45% of all participants).

Define T to be the true unobserved indicator of early drinking, that is, $T = 1$ corresponds to the early onset of drinking, $T = 0$ to the late onset. Let X be the observed value of the early onset of drinking. Because we do not have external data or internal replicates to estimate misclassification probability, we performed sensitivity analysis for various values of misclassification.

We used the following risk model:

$$\operatorname{logit}\{\operatorname{pr}(D = 1 \mid G = (A, B), T)\} = \beta_0 + \beta_t T + \beta_A A + \beta_B B + \beta_{AT} AT + \beta_{BT} BT. \tag{5.1}$$

The results of sensitivity analysis (not shown) suggest that when $\operatorname{pr}(X = 0 \mid T = 1)$ is ignored or underestimated, the interaction effect is not significant. The setting corresponds to the case when exposed subjects are defined as nonexposed, thus reducing the association signal. However, the estimation procedure appears to be robust to underestimation of $\operatorname{pr}(X = 1 \mid T = 0)$. This scenario corresponds to the case when a nonexposed subject is considered to be exposed.

Parameter estimates obtained using our method corresponding to $\operatorname{pr}(X = 0 \mid T = t) = 0.25$ and $\operatorname{pr}(X = 1 \mid T = 0) = 0.25$ are presented in Table 4 demonstrating significant interaction between various genetic markers and early onset of drinking.

6. Discussion

Motivated by concerns in the analysis of gene-environment interactions, we proposed a genotype-based Bayesian approach for the analysis of case-control studies when environmental exposure cannot be observed directly and is subject to misclassification. The formulation of risk functions and the estimation procedure are along the lines of our previous work:

Table 4: Risk parameter estimates and standard errors in the alcohol dependence data.

Gene, SNP	Estimate of log(OR)	Standard error
NCAM1, rs586903	1.78	0.06
NCAM1, rs2303377	2.58	0.11
NCAM1, rs2156485	1.87	0.07
TTC12, rs7103866	2.21	0.03
TTC12, rs723077	1.92	0.03
TTC12, rs2288159	2.21	0.01
CHRNA3, rs1051730	1.77	0.03
CHRNA3, rs8192475	1.62	0.02

genotype and additive effect models [14, 15] and pseudolikelihood approach [3, 11, 12]. The risk function of genotype effect model involves both the additive and dominance effect while taking into account possible interactions between genes expressed in terms of interaction between their additive and dominance components, while the additive effect model only involves the additive effect and possible interactions. The additive effect model contains less parameters than the genotype effect model. In applications, the additive effect models should be used in analyzing data as the first step. If the dominance effect is strong enough to compensate the increase of the number of the parameters in the genotype effect models, one may use the genotype effect models.

The proposed method has several unique advantages. First, the observed genetic information enters the model directly and the LD structure is captured in the regression coefficients. This aspect offers advantages from the practical point of view, the computational burden is less demanding because haplotype phase need not to be estimated. In the cases when LD is moderate, which is the focus of our work, the computational demands can be substantial even with the current state of technology. Furthermore, the risk due to uncertainty associated with the haplotype phase estimation can be avoided. Second, the estimating procedure is based on a pseudolikelihood model, similarly to the method investigated previously, that allows efficient estimation of parameters, models environmental covariates completely nonparametrically, and incorporates information about the probability of disease [3, 11, 12]. In epidemiologic studies, the vector of environmental covariates measured exactly is often, high dimensional, and a good estimate about probability of disease in a population is known. Additionally, the Bayesian formulation of the proposed method allows shrinking parameter estimates towards prior which offers advantage in cases when misclassification is present.

Because of the Bayesian formulation and the need to validate posterior sets obtained using a pseudolikelihood, the proposed method can be highly computationally intensive. Moreover, the validation of pseudolikelihood requires evaluation of ratio of two likelihood functions. For example, in our simulation experiments and data analysis, this part required us to obtain a precise value of ratios similar to $\exp(3000)/\exp(2908)$. Hence, we employed GNU Multiple Precision Arithmetic Library (http://gmplib.org/).

The form of our pseudolikelihood function is complex and it does not seem feasible to validate a pseudolikelihood function algebraically. Instead, we propose to apply Monahan and Boos method to examine coverage probabilities of posterior sets. If a comprehensive set of scenarios fails to invalidate a pseudolikelihood function, we suppose that the pseudolikelihood is valid. This reasoning may be similar to the conventional hypothesis testing where the null hypothesis is assumed to be true (pseudolikelihood is valid), and the observed

data is used to quantify evidence in favor of the alternative hypothesis (pseudolikelihood is not valid). Of course, a strong basis for validity of a pseudolikelihood is needed. We employ the following arguments. Our previous research approach [3, 11, 12] demonstrated validity of this pseudolikelihood in frequentist sense, that is, we have shown that estimation and inferences are correct when this pseudolikelihood is used in place of a real likelihood function. Hence, posterior distribution based on a pseudolikelihood may be invalid only for certain prior distributions. Therefore, to invalidate a pseudolikelihood, one should find a prior distribution for which the posterior is not valid. However, in our setting, the number of possible prior settings is narrow, because what we advocate is the use of symmetry of prior distribution as a way to improve precision of estimation and inference. We are restricting the prior of regression coefficients to be Gaussian and advocate mean zero and large variance. While one can try other priors for other parameters, the number of possible prior settings is still reasonable and it is practically feasible to look at their performance in terms of probability of coverage sets.

While the major motivation of the proposed work is dictated by the need of a symmetric prior on risk coefficients, other types of *a priori* information can enter our model. For example, if *a priori* information about the LD structure is available, it can be modeled in the *a priori* distribution. Furthermore, if misclassification probabilities are not known precisely, one can specify uncertainty associated with values of misclassification.

A major practical advantage of this proposed work is that it allows the model to exploit recent advances in genotyping technology. Specifically, with the recent advances genetic markers become more and more densely typed and multiple markers are likely to be observed in a functional unit of interest. These units of interest may be defined in terms of LD blocks using information available in linkage maps. While in situations when linkage disequilibrium is strong, the haplotype-based analysis is advantageous; in more common scenarios when linkage disequilibrium is moderate, our approach provides advantages.

However, in the context when the number of genetic markers in a functional unit of interest is large our methodology may require model averaging and model selection component. Hence, behavior of this pseudolikelihood needs to be examined in this setting. A practical strategy can be that one starts with screening analysis first to get interesting genetic variants and SNPs using traditional methods which is computationally less demanding. Then, one may apply the proposed approaches for possible gene-environment interactions and further investigations by focusing on these important and interesting genetic variants and SNPs.

Acknowledgments

R. Fan was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Maryland, USA. Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI) (U01 HG004422). SAGE is one of the genomewide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning as well as with general study coordination was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and

the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1. This work has utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at New York University Langone Medical Center.

References

[1] D. Thomas, "Gene-environment-wide association studies: emerging approaches," *Nature Reviews Genetics*, vol. 11, no. 4, pp. 259–272, 2010.

[2] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, "A comprehensive review of genetic association studies," *Genetics in Medicine*, vol. 4, no. 2, pp. 45–61, 2002.

[3] I. Lobach, B. Mallick, and R. J. Carroll, "Semiparametric Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data," vol. 4, no. 3, pp. 305–316, 2011.

[4] S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti, "Haplotype inference in random population samples," *American Journal of Human Genetics*, vol. 71, no. 5, pp. 1129–1137, 2002.

[5] J. Marchini, D. Cutler, N. Patterson et al., "A comparison of phasing algorithms for trios and unrelated individuals," *American Journal of Human Genetics*, vol. 78, no. 3, pp. 437–450, 2006.

[6] Z. S. Qin, T. Niu, and J. S. Liu, "Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms," *American Journal of Human Genetics*, vol. 71, no. 5, pp. 1242–1247, 2002.

[7] M. Stephens and P. Donnelly, "A comparison of bayesian methods for haplotype reconstruction from population genotype data," *American Journal of Human Genetics*, vol. 73, no. 5, pp. 1162–1169, 2003.

[8] M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics*, vol. 68, no. 4, pp. 978–989, 2001.

[9] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models Edition*, Chapman & Hall CRC Press, 2nd edition, 2006.

[10] A. Agrawal, C. E. Sartor, M. T. Lynskey et al., "Evidence for an interaction between age at first drink and genetic influences on DSM-IV alcohol dependence symptoms," *Alcoholism*, vol. 33, no. 12, pp. 2047–2056, 2009.

[11] I. Lobach, R. J. Carroll, C. Spinka, M. H. Gail, and N. Chatterjee, "Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement errors in environmental exposures," *Biometrics*, vol. 64, no. 3, pp. 673–684, 2008.

[12] I. Lobach, R. Fan, and R. J. Carroll, "Genotype-based association mapping of complex diseases: gene-environment interactions with multiple genetic markers and measurement error in environmental exposures," *Genetic Epidemiology*, vol. 34, no. 8, pp. 792–802, 2010.

[13] J. F. Monahan and D. D. Boos, "Proper likelihoods for Bayesian analysis," *Biometrika*, vol. 79, no. 2, pp. 271–278, 1992.

[14] R. Fan and M. Xiong, "High resolution mapping of quantitative trait loci by linkage disequilibrium analysis," *European Journal of Human Genetics*, vol. 10, no. 10, pp. 607–615, 2002.

[15] R. Fan, J. Jung, and L. Jin, "High-resolution association mapping of quantitative trait loci: a population-based approach," *Genetics*, vol. 172, no. 1, pp. 663–686, 2006.

[16] M. K. Ho and R. F. Tyndale, "Overview of the pharmacogenomics of cigarette smoking," *Pharmacogenomics Journal*, vol. 7, no. 2, pp. 81–98, 2007.

[17] D. W. Schafer and K. G. Purdy, "Likelihood analysis for errors-in-variables regression with replicate measurements," *Biometrika*, vol. 83, no. 4, pp. 813–824, 1996.

[18] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, Chichester, UK, 1994.

[19] H. J. Edenberg, "The collaborative study on the genetics of alcoholism: an update," *Alcohol Research and Health*, vol. 26, no. 3, pp. 214–218, 2002.

[20] L. J. Bierut, N. L. Saccone, J. P. Rice et al., "Defining alcohol-related phenotypes in humans: the collaborative study on the genetics of alcoholism," *Alcohol Research and Health*, vol. 26, no. 3, pp. 208–213, 2002.

[21] H. J. Edenberg and T. Foroud, "The genetics of alcoholism: identifying specific genes through family studies," *Addiction Biology*, vol. 11, no. 3-4, pp. 386–396, 2006.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 873570, 19 pages
doi:10.1155/2012/873570

Research Article

Clustering-Based Method for Developing a Genomic Copy Number Alteration Signature for Predicting the Metastatic Potential of Prostate Cancer

Alexander Pearlman,[1] Christopher Campbell,[1] Eric Brooks,[2] Alex Genshaft,[2] Shahin Shajahan,[2] Michael Ittman,[3] G. Steven Bova,[4] Jonathan Melamed,[5] Ilona Holcomb,[6] Robert J. Schneider,[7] and Harry Ostrer[1]

[1] *Department of Pathology, Albert Einstein College of Medicine, Bronx, NY 10461, USA*
[2] *Human Genetics Program, Department of Pediatrics, NYU Langone Medical Center, New York, NY 10016, USA*
[3] *Department of Pathology, Baylor College of Medicine, Houston, TX 77030, USA*
[4] *Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA*
[5] *Department of Pathology, NYU Langone Medical Center, New York, NY 10016, USA*
[6] *Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA*
[7] *NYU Cancer Institute and Department of Microbiology, NYU Langone Medical Center, New York, NY 10016, USA*

Correspondence should be addressed to Alexander Pearlman, apearlman@gmail.com

Received 1 March 2012; Revised 17 May 2012; Accepted 31 May 2012

Academic Editor: Xiaohua Douglas Zhang

The transition of cancer from a localized tumor to a distant metastasis is not well understood for prostate and many other cancers, partly, because of the scarcity of tumor samples, especially metastases, from cancer patients with long-term clinical follow-up. To overcome this limitation, we developed a semi-supervised clustering method using the tumor genomic DNA copy number alterations to classify each patient into inferred clinical outcome groups of metastatic potential. Our data set was comprised of 294 primary tumors and 49 metastases from 5 independent cohorts of prostate cancer patients. The alterations were modeled based on Darwin's evolutionary selection theory and the genes overlapping these altered genomic regions were used to develop a metastatic potential score for a prostate cancer primary tumor. The function of the proteins encoded by some of the predictor genes promote escape from anoikis, a pathway of apoptosis, deregulated in metastases. We evaluated the metastatic potential score with other clinical predictors available at diagnosis using a Cox proportional hazards model and show our proposed score was the only significant predictor of metastasis free survival. The metastasis gene signature and associated score could be applied directly to copy number alteration profiles from patient biopsies positive for prostate cancer.

1. Introduction

Prostate cancer is a common public health problem. In 2012, this disease was expected to be diagnosed in an estimated 241,740 men (29% of all male cancers) and to result in 28,170 deaths (9% of male cancer deaths) [1]. If left untreated, around 70% of prostate cancers remain asymptomatic and indolent for decades [2]. If treated with radical prostatectomy or radiation therapy, the risk of metastasis is reduced, but erectile dysfunction, urinary incontinence, and rectal bleeding may occur, affecting the patient's quality of life. Because it is currently difficult to determine accurately which patients will develop metastatic disease, physicians treat patients with mid-to-late stage local disease aggressively, even when such treatment may not be required. Clinical parameters, such as, serum concentration of prostate-specific antigen (PSA), extension beyond surgical margins, invasion of seminal vesicles, extension beyond the capsule, surgical Gleason score, prostate weight, race, and year of surgery, are employed in existing nomograms for prediction of local recurrences after surgery [3], but, many of these parameters are not available at diagnosis and cannot be used for guiding therapeutic decisions. Development of a robust risk model from a biopsy that accurately predicts the potential of a local prostate cancer to metastasize would justify aggressive treatment in high-risk cases and improve the quality of life for men with indolent disease by allowing them to avoid treatment-related side effects. Thus, the goal of this study was to develop a method to identify tumor genomic biomarkers that could be applied to prediction models that help guide clinical treatment decisions.

The method chosen for developing the predictive model was the analysis of genomic DNA copy number alterations (CNAs) in prostate cancers, because these cancers have long been known to harbor multiple genomic imbalances that result from CNAs [4, 5]. High-resolution measurements of CNAs have functional value, in some cases providing evidence for alterations in the quantity of normal, mutant, or hybrid-fusion transcripts and proteins in the cancer cells. The resulting changes in abundance or altered structure of RNA transcripts and proteins (e.g., truncating dominant negative mutations) may impact the fitness of the cell and provide some of the mechanisms necessary for distant site migration, invasion, and growth. From the multiple CNAs identified in tumors, CNA-based gene signatures were developed into a score that suggested the ability to predict metastasis free survival.

2. Methods

2.1. Cohorts and Samples

We studied four publically available prostate cancer cohorts and a fifth cohort reported here: (1) 294 primary tumors and matched normal tissue samples from NYU School of Medicine (NYU $n = 29$), Baylor College of Medicine (Baylor $n = 20$) [6], Memorial Sloan-Kettering Cancer Center (MSK $n = 181$) [7], and Stanford University (SU $n = 64$ (single reference used for each tumor)) [8]. (2) 49 metastatic tumors and matched normal samples from Johns Hopkins School of Medicine (Hopkins $n = 13$) [9] and MSK ($n = 36$) [7]. The 13 patients in the Hopkins cohort had multiple metastases dissected at autopsy, totaling 55 samples for the study. We also studied a sixth, publically available cohort of 337 cell lines originating from varying tumor cell types (ArrayExpress ID: E-MTAB-38).

2.2. Sample Processing

Genomic DNA (gDNA) from the NYU cohort was extracted from fresh-frozen prostate tumors using a Gentra DNA extraction kit (Qiagen). Purified gDNA was hydrated in reduced TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0). The gDNA concentration was measured using the NanoDrop 2000 spectrophotometer at optical density (OD) wavelength of 260 nm. Protein and organic contaminations were measured at OD 280 nm and 230 nm, respectively. Samples that passed OD quality control thresholds were then run on a 1% agarose gel to assess the integrity of the gDNA. 500 ng of gDNA samples was run on the Affymetrix Human SNP Array 6.0 at the Rockefeller University Genomics Resource Center using standard operating procedures. Samples that were obtained from public sources were processed according to the methods outlined in their respective publications. Affymetrix .cel files were processed using the Birdseed v2 algorithm [10].

2.3. Study Design

The case samples in this study were either metastatic tumors (METS) or primary tumors from men treated with radical prostatectomy that were clinically followed up and reported to develop distant metastases (mPTs). METS and mPTs are clearly discernible phenotypes that can be classified unequivocally as cases. The control samples were defined as primary tumors that had not progressed to form distant metastases following radical prostatectomy either because clinical followup was not available or because the treatment rendered the patient not informative for this outcome. Radical prostatectomy treats both indolent primary tumors (iPTs) that would not metastasize and primary tumors that would otherwise progress to form metastases, if left untreated. Thus, the control primary tumors actually represent a mixture of iPTs and unrealized mPTs. Assuming a randomly sampled cohort, it is expected that about 30% of the control group of primary tumors would be unrealized mPTs [2]. Considering the scarcity of clinically informative mPTs and iPTs for study, our strategy for identifying CNA biomarkers from tumors with inferred metastatic outcomes allowed a greater number of individual genomes to be used. Accordingly, all of the clinically informative mPTs available to us were not used to identify the biomarkers and only tested in a Cox proportional hazard model to assess the clinical usefulness of these predictors. Future tumor cohort study design using the method presented in this paper should consider the prevalence of metastatic progression to assure a large enough representation of both mPTs and iPTs. The natural history of prostate cancer, without medical intervention, (e.g., watchful waiting or active surveillance) is well documented [2]. Assuming a randomly sampled cohort, this information allowed us to estimate the prevalence of mPTs to be 30%.

2.4. Cancer Genomics Copy Number Algorithm

A genomic DNA copy number analysis pipeline (Figure 1) was designed using the R-statistical software [11] (R) to process the raw intensity data through a series of computational steps resulting in ranked lists of genes and associated significance that could be used for functional mining and prediction model development. The R-package will be provided upon request and raw and processed data can be obtained from Gene Expression Omnibus accession# GSE27105.

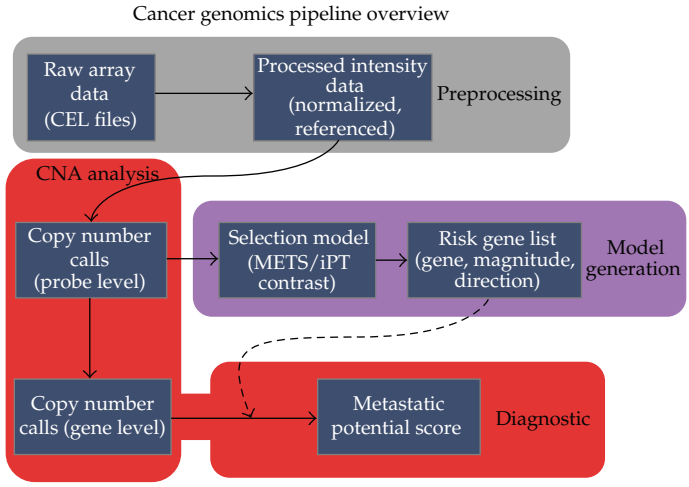


Figure 1: Array CGH analysis pipeline for processing pixel image data from Affymetrix SNP arrays to produce genotype and signal intensity measures for copy number analysis used for developing bioclinical models and diagnostics.

2.5. *Raw Data Processing*

Signal intensity files (.cel) for the Affymetrix SNP Array 6.0 or 500 k mapping arrays were processed using the Affymetrix Power Tools, Birdseed V2 [10], and BRLMM [12] algorithms, respectively, resulting in genotype allele calls and signal intensity measures for each SNP and copy number probe. After the first stage, the genotype calls were prepared for downstream principal component analysis for ethnic identification and quality control testing, especially important when investigating racially driven health disparities (Figure 2). Men of African descent have an increased incidence, earlier onset, and more aggressive form of the disease than those of European origin. Even when adjusted for the increased level of incidence in African Americans, the mortality rate of African American men is more than twice that of Caucasian men [1]. Although not presented in the current work, sophisticated CNA models of metastatic disease may provide a biological explanation for the epidemiological observations of racial health disparity of metastasis.

The probe-summarized intensity signals were log transformed and standardized (mean centered, standard deviation scaled) on an individual array basis and the relative copy number was calculated by subtracting the normal from the tumor intensity for each patient on a probe basis. The resulting copy number profile (CN) represented the amplification and deletion events that accumulated in each cancer sample tested.

Next, the probes were ordered as they appear in the genome and the copy number signal data (CN) was smoothed. The smoothing was conducted using a running median function (runmed in *R*, with endrule parameter equal to "median"). The smoothing function was termed $S(\mathrm{CN})_k$, where k represents the probe width of the smoothing window. The values of k usually range from 5 to 151, depending on the array's probe density and were chosen not to exceed a biologically meaningful span of total genetic distance. Considerations for k should include the average alteration size (estimated empirically from each data set) and distance between probes as determined by the array probe density. As an extreme example, smoothing the entire arm of a chromosome will remove all local variation that

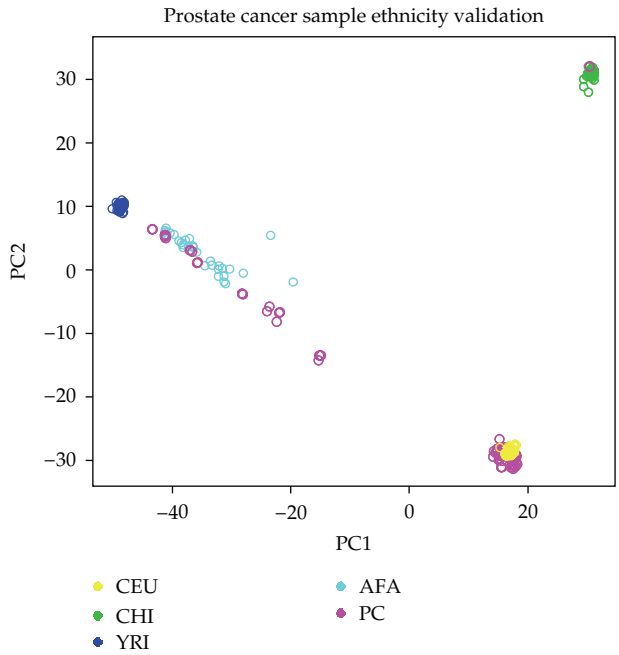


Figure 2: Principal component analysis identity testing of a variety of normal SNP profiles from the germline DNA of prostate cancer patients (PC) used in the study compared to a set of HapMap normal reference populations from Nigeria (YRI), Europe (CEU), China (CHB), or, of African American (AFA) decent. The x and y axes represent the 1st and 2nd eigenvectors.

exists on that arm. The function $S(\mathrm{CN})_k$ thus yielded n smoothing profiles per sample, with n representing the number of different values used for k. An example of the multiple n values used for chromosome 1 of a particular sample is shown in Figure 3.

2.6. Copy Number Alteration Calling Algorithm

The next part of this stage involved assigning copy number events to each probe. The reason we developed a CNA caller from scratch was because the standard calling algorithms required parameter inputs that were dependent on the signal-to-noise distribution of the copy number measures. Because cancer samples' signal-to-noise are notoriously variable, both on a chromosome basis (within a sample profile) and across samples, this made the standard CNA calling approaches inefficient without significant reconfiguration. Therefore, we developed a method that was dynamic to the signal-to-noise variation observed in cancer genomes. We validate the effectiveness our approach (Figure 4) using a benchmark simulation data set used to test a variety of algorithms [13]. Given that SNP arrays are not designed to provide quantitative measures of copy number (but do respond linearly to CNAs), we restrict our calls to three categories: amplifications (1), deletions (−1), and neutral events (0). To determine the "center" of the genome so that thresholds can be drawn, we assume that a majority of the intensity values reflect a 2-copy state for the referenced sample, that is, the majority of the referenced tumor sample exists in a 2-copy state (manual calling is used for those samples in which this assumption is not valid). To accomplish this, we sample

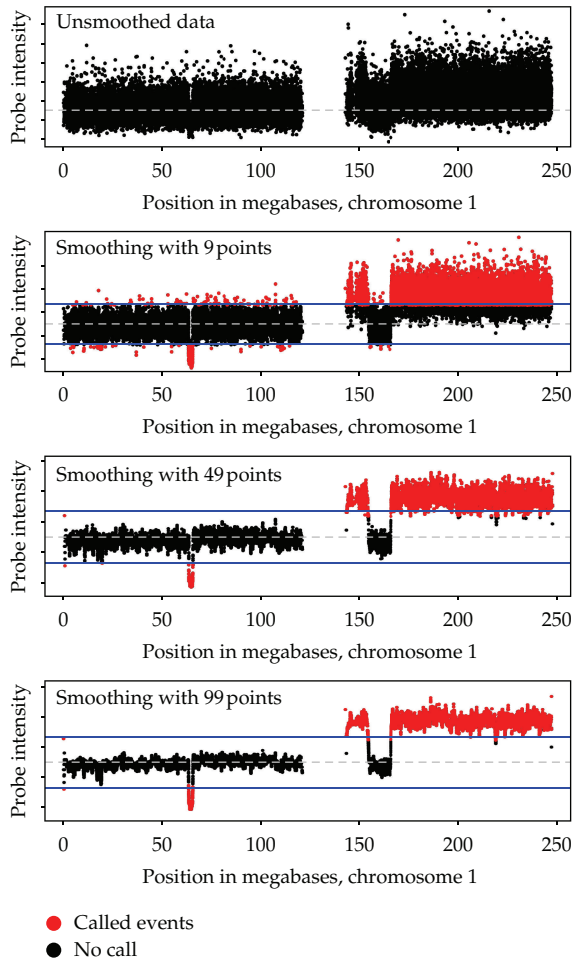


Figure 3: A representative primary tumor chromosome 1 copy number profile (top panel) and corresponding $S(\mathrm{CN})n$ $[k = \{9, 49, 99\}]$ in the bottom panels. Therefore, $n = 3$ because three different smoothing lengths are used. Black probes represent probes that are not called while red probes are the called events that exceed the amplification and deletion thresholds.

10,000 random stretches of probes covering approximately 500 kilobases from the autosomes, calculate the median of each, and use the most frequently occurring value to scale the sample appropriately. Following scaling of the genome, thresholds were drawn based on quantile values and copy number states were assigned to each probe. Since this thresholding scheme was applied to every smoothing, there were n event calls per probe. These calls result in a "ρ" profile, where $T()$ represents the function of trinary binning:

$$\rho_k = T(S(\mathrm{CN})_k). \tag{2.1}$$

The $n\rho$ calls for each probe were then combined by summation, resulting in a composite profile (ρ') that ranged from $-n$ (signifying that a deletion was called at every smoothing

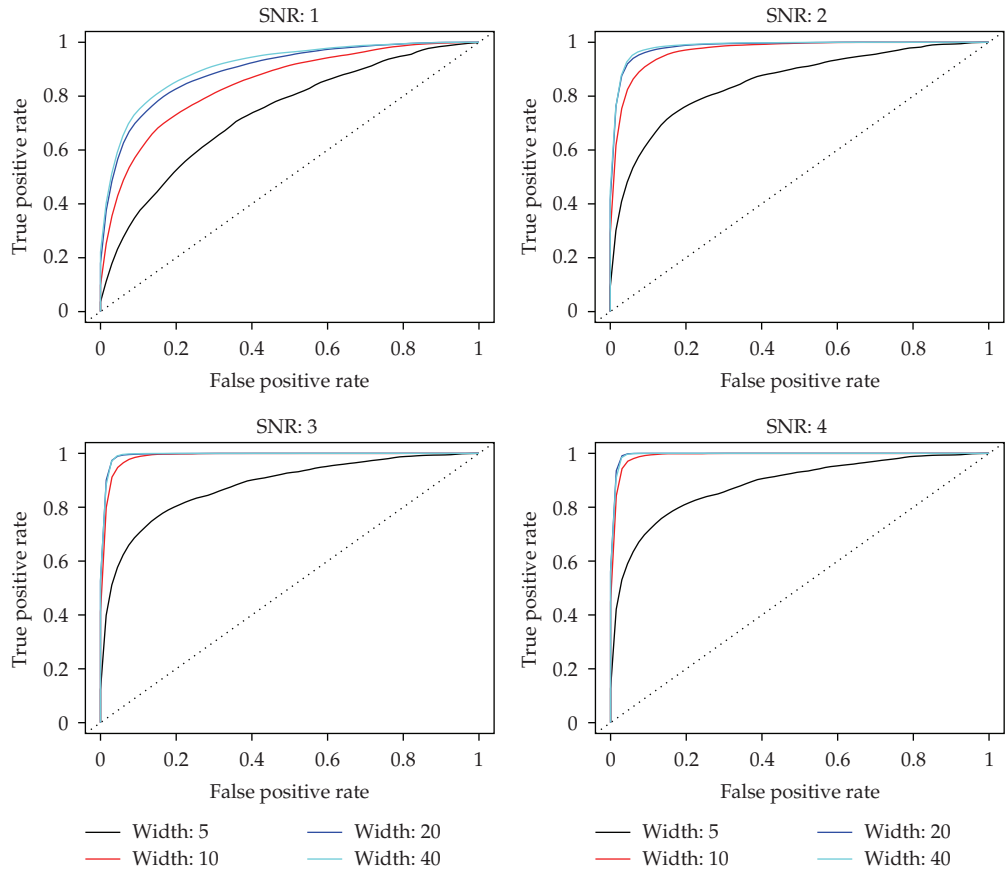


Figure 4: Receiver-operating characteristic curves showing the performance of our CNA-calling algorithm on the simulated data [13]. Each panel represents a different signal-to-noise ratio and the curves represent varying event widths of the simulated data. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. Each curve is generated by testing varying thresholds on 100 simulated chromosomes for the condition specified. The curves are combined using vertical averaging. The dashed line represents the random model.

for that probe) to $+n$ (signifying that an amplification was called at every smoothing for that probe):

$$\rho' = \sum_{i=1}^{n} \rho_i. \tag{2.2}$$

One ρ' profile was thus generated per sample, representing a composite of n smoothings, and this metric was used for the rest of the primary analysis. We benchmarked our copy number calling method using a published simulation data set [13] comprised of randomly generated artificial chromosomes. Each chromosome was generated with an aberration flanking the center probe with Gaussian noise $N(0, 0.25^2)$ superimposed. All combinations of signal to noise (SN = $4, 3, 2,$ and 1) and aberration widths ($W = 40, 20, 10,$ and 5) were produced for a total of 160,000 analysis runs. Receiver-operating characteristics (ROC) were computed from the benchmark simulation dataset [13];

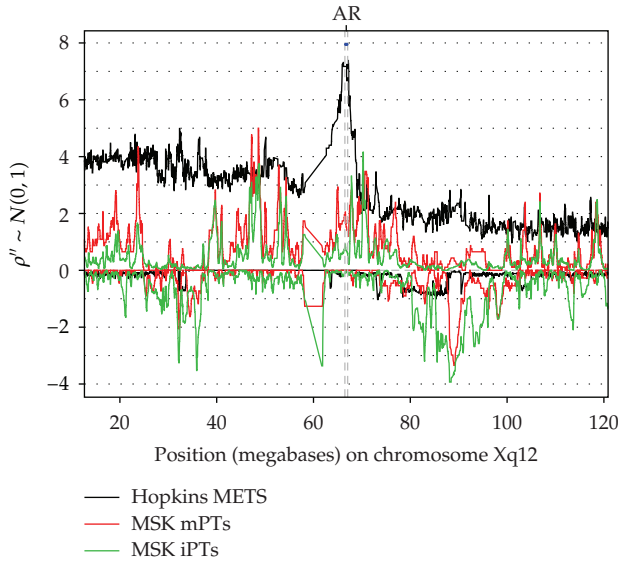


Figure 5: Copy number profile, ρ'' shows an amplification of a region on chromosome X harboring the androgen receptor (AR) locus. The x-axis represents the ordered chromosome position and the y-axis represents standardized population frequencies exhibiting amplifications (above 0) and deletions (below 0). The three populations of tumors are represented as red, black, or green lines for mPTs, androgen ablation treated metastases (METS), and iPTs, respectively.

where ROC is defined as a pair, ROC = (TPR, FPR), TPR = (the number of probes within the aberration width that is above a threshold)/(the total number of probes within the aberration width). FPR = (the number of probes outside the aberration width that is above a threshold)/(the total number of probes outside the aberration width). The threshold values are selected to continuously range over the values of the data points, and since ROC is piecewise constant, only changing when a threshold is equal to the value of a data point, we only need to consider values of the data points in their sorted order. The area under the curve (AUC) of each ROC curve was used to gauge performances.

To examine the frequency of amplification and deletions for subgroups of samples or populations and evaluate the sensitivity of our CNA-calling method, we further combined the ρ' data to create ρ'' by summing across the ρ' profiles on a probe basis across multiple samples. Two values of ρ'' were calculated for population or subpopulation. The first value represented the sum of all positive ρ' values in the population at any probe and was thus called ρ''_{amp}. Likewise, the second value representing the sum of all negative ρ' values in the population at any probe was called ρ''_{del}:

$$\rho''_{\text{amp|del}} = \sum_{i=1}^{n \text{ samples}} \rho'_{[\text{amp|del}]}. \tag{2.3}$$

An example of copy number ρ'' plot (Figure 5) is observed in a select region on chromosome X from metastases of men treated with androgen ablation therapy and primary tumors of iPTs and mPTs from other men not treated. Furthermore, differential analysis of the ρ'' values can be used to identify probes or regions of probes that comprise genes that may contribute to

the phenotype being tested (e.g., iPT versus mPT or response to therapy versus no response to therapy).

2.7. Semisupervised Clustering Algorithm

Since sufficient labels were not available to train a model from primary tumors alone, we first created from a cohort of men that developed distant metastases a simplified summary metastasis profile to capture the high-frequency events, that are in part, assumed to correlate to the outcome. This clustering approach is not unsupervised, class-less clustering because we know some information about one of the components which is the summary profile from known metastasis samples. To reflect the frequency of events observed for individual metastasis CNA profiles in the summary metastasis profile, the average number of ρ' events calculated for the group of metastases was used to set a threshold for the number of total ρ' events used to build the summary metastasis profile. The actual probes chosen for the metastasis summary profile were based on their ranked frequency which resulted in a threshold of at least 25% of the samples exhibiting the event. Although not tested here, the theoretical specificity of the summary profile is expected to decrease as the threshold for minimum number of events called decreases, while the sensitivity of the profile decreases as the threshold of minimum number of events called increases. In the case of the MSK cohort, clustering of the 36 metastases ρ' profiles independently yielded two well-separated clusters from which we built two metastasis summary profiles to perform semisupervised clustering with the primary tumors. Alternatively, the 13-patient Hopkins cohort made up of 55 metastases yielded only one homogeneous cluster and associated summary metastasis profile. To overcome the inherent variability with clustering algorithms, we employed a resampling hierarchical clustering method to infer an initial grouping for the unclassified primary tumors. For each iteration, a subset of the individual ρ' profiles from the unknown primary tumors were randomly chosen with replacement and clustered with the summary copy number profile derived from the metastasis samples (one metastasis summary profile from the Hopkins cohort and two from MSK cohort). Therefore, the semisupervised clustering analysis presented here was developed to classify prostate primary tumors into subgroups with different metastatic potential (mPT and iPT) based on their CNA profiles. Distance was calculated using a binary metric, and the samples were joined using hierarchical clustering (complete-linkage method). The cluster tree was divided into two groups at the final join, and the primary tumor samples were scored 1 if they fell in the same cluster as the metastasis profile, and 0 if they were in the other cluster. Using the results from 20,000 resampling iterations of the clustering, a proximity score was generated for each sample, representing the number of times it fell in a cluster with the metastasis profile. A sample with a high score was considered to be more metastatic (mPT), while lower scoring tumors were more indolent (iPT). The similarity scores distributed throughout the possible range of values (0 to 1), allowing us to form distinct groups of tumors with significant contrast between high- and low-metastatic distance to MSK metastasis signature 1 (Figure 6). The group of samples with scores closer to the center of the distribution were omitted to further define the contrast between high- and low-scoring samples.

2.8. Metastasis Genes Inferred through Evolutionary Selection Modelling

Genomic DNA copy number alterations in local and metastatic prostate tumors are typically numerous, systematic in their genomic placement and varied in size from point mutations

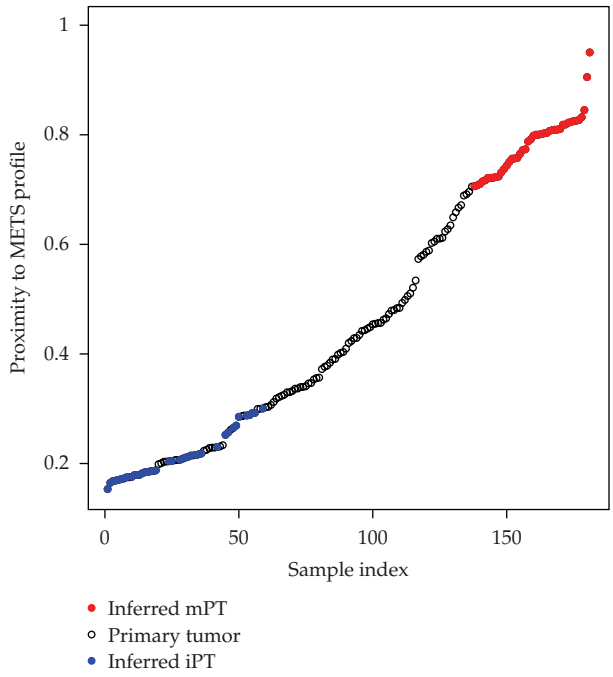


Figure 6: Plot of ranked proximity score for MSK signature 2. Proximity represents the number of times a particular sample clustered with the MSK metastasis profile 2. The samples with higher scores (red points) are classified as inferred mPTs and the samples with lower scores (blue points) are classified as inferred iPTs. Primary tumors (hollow points) interspersed with the blue iPT tumors were excluded as iPTs for MSK signature 2 because they did not consistently classify as iPTs in the proximity analysis using MSK signature 1.

to duplications or deletions of entire chromosomes. Given these observations, geneticists have postulated that Darwinian selection may operate on the genomic instability in tumors [14]. High-resolution measurements of CNAs in somatic tumors have informative value, in some cases reflecting the direction in which the biochemistry of the cell controls the quantity of normal, mutant, or hybrid-fusion transcripts and proteins. During this genomic transformation, the resulting modified transcripts and proteins may impact the fitness of the cell. Guided by these principles of evolutionary selection, our analyses sought to identify the CNA landscape that reflects selection mechanisms of metastasis. Genomic selection towards a metastatic cancer phenotype can be both positive and negative and be observed in CNAs exhibiting both amplifications and deletions. For example, genes that promote metastasis and amplified in metastatic tumors would reflect positive selection, while metastasis suppressor genes that are deleted in metastases reflect negative selection. The genes associated with these regions, altered at high frequency in metastatic tumors and enriched in mPTs more so than iPTs, lead to enhanced metastatic potential. We identified specific CNAs that selected positively for metastatic potential, exhibiting amplifications in metastases and mPTs and deletions in iPTs. CNAs identified to exhibit negative selection for metastatic potential were observed to be deleted in metastases and mPTs and amplified in the iPTs. Therefore, we designed models based on Darwin's evolutionary selection theory to score positive and negative selection based on the mPT and iPT classifications derived through semisupervised clustering using the ρ' data. For each probe on the array, we calculated an enrichment score,

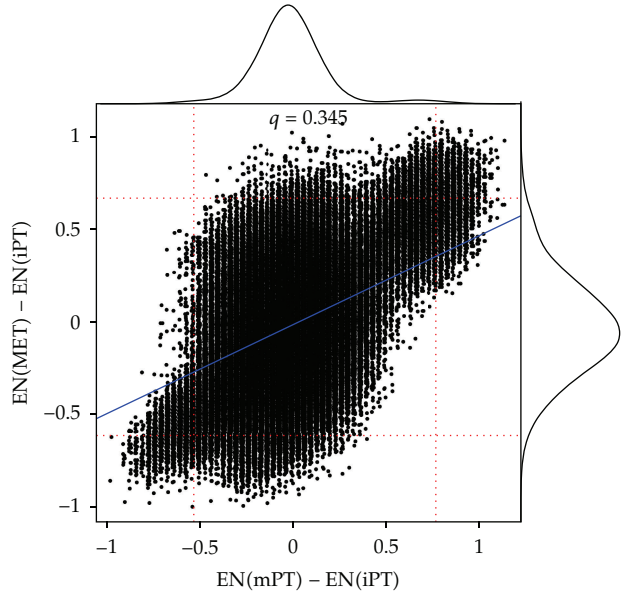


Figure 7: Scatterplot of the enrichment scores of the METS versus those of mPTs, normalized by the enrichment scores of iPTs. Kernel density estimation curves are shown protruding from the x and y axes. The horizontal and vertical dashed red lines denote the trim points (quantiles 0.99 and 0.01). A linear regression line, based on the trimmed values, is shown in blue. The value of q is the Pearson correlation coefficient for the trimmed x and y values.

EN(x), which represented the relative number of amplifications versus deletions, observed in each subgroup (metastasis, mPT and iPT):

$$\mathrm{EN}(x) = \frac{(\#\mathrm{Amp} - \#\mathrm{Del})}{\#\mathrm{Samples}}. \tag{2.4}$$

Next, we modeled the relative enrichment by contrasting the metastasis and mPT copy number alterations with those observed in the iPT group:

$$\mathrm{SM} = e^{[\mathrm{EN(METS)}+q^{*}\mathrm{EN(mPT)}-\mathrm{EN(iPT)}]}. \tag{2.5}$$

The first two enrichment terms (for metastatic and metastatic-like samples) being summed were designed to assign a higher score when the METS and mPT samples had more amplifications than deletions. Greater amplification enrichment in the METS and mPTs resulted in higher scores. The third term, EN(iPT), was higher when the iPT samples exhibit the opposite effect (enrichment for deletions over amplifications). The middle term, EN(mPT), was multiplied by a data-driven coefficient, q, representing the average contribution of mPT on a probe basis (Figure 7).

For example, probes that were amplified in all metastases and mPTs but deleted in all iPTs (positive selection driving the metastasis cells) would yield the highest possible score. Likewise, probes that were deleted in all metastases and mPT samples, but amplified in all iPT samples (negatively select or inhibit the promotion of the metastasis cells), would

reach the minimum possible score. Therefore, regions of the genome that enhance and inhibit metastasis formation will be captured by our evolutionary selection model.

Following this probe scoring method we developed a Z-score model in order to extend this analysis to the gene level. We assign each probe to a gene, provided it falls within 10,000 bp up- or downstream of the transcription start or stop site. The SM scores for the probes within a gene are averaged and compared to the mean and standard deviation of a background distribution, which was calculated by sampling the top 5th percentile of amplified or deleted probes from all genes on the array with the same number of probes as the gene in question. The result is a Z-score for each gene in the genome that is represented on the array.

2.9. Metastatic Potential Score and Survival Analysis

We developed an algorithm based on genomic CNAs to calculate a metastatic potential score (MPS), with a higher score indicating a greater likelihood of metastasis. The MPS score for a new individual patient only depends on the CNA profile of this new patient. It can be calculated without requirement for other samples, since it's simply based on the concordance/discordance relationship to the CNA metastasis gene signature previously identified as selecting for the metastatic phenotype through our selection model. The MPS was calculated using a weighted Z score from the top set of CNAs overlapping metastasis genes determined by the significance of their selection model Z scores. We used $Z \geq 1.7$ as a cutoff point because for standard normal distribution, the tail of 1.7 is about 5%. The metastatic potential score was defined as the following:

$$\mathrm{MPS} = \sum_{i=1}^{n} Z_i' * \mathrm{Dir}_{\mathrm{sig}}(i) * \mathrm{Dir}_{\mathrm{samp}}(i). \tag{2.6}$$

For each tumor profile, logistic adjusted Z scores (Z') from genes ($i \ldots n$) that match the direction of the metastasis gene signature (a vector of −1s and +1s representing whether the gene was deleted or amplified in the signature, resp.) were added, whereas Z' from genes that mismatch the direction of the signature were subtracted. As the direction component of the risk model score (Dir) reflects, if the CNAs of the metastasis signature ($\mathrm{Dir}_{\mathrm{sig}}$) and the unknown sample profile ($\mathrm{Dir}_{\mathrm{samp}}$) are in the same direction, the coefficient will be 1; if they are in opposing directions, the coefficient will be −1; and if $\mathrm{Dir}_{\mathrm{samp}}(i) = 0$, then the entire term will not count towards the score. For example, if a gene i, that is typically amplified in metastases ($\mathrm{Dir}_{\mathrm{sig}}(i)$) and mPTs, is also amplified in the unknown profile ($\mathrm{Dir}_{\mathrm{samp}}(i)$) that Z score is added, whereas if gene i in the profile is deleted, as expected in iPTs, the Z score is subtracted. Neutral genes that are neither amplified nor deleted in the unknown profile are not scored in this model.

Three metastasis signatures, derived from a combination of five cohorts were used to develop the MPS. The first signature was identified using 49 primary tumors of unknown clinical outcome from NYU ($n = 29$) and Baylor ($n = 20$) and a metastasis cohort from Hopkins ($n = 13$). The other two signatures were identified using 75% of the MSK cohort of primary tumors of unknown outcome ($n = 126$) along with a set of metastatic tumors ($n = 36$) from the same MSK cohort. The CNA-based gene signatures from these 2 sets of cohorts were concatenated and derived into the MPS which we assessed in a Cox proportional hazard model with samples set aside for testing purposes only. The test cases were comprised

of bona fide mPTs (primary tumors that later developed into distant metastasis), whereas the test controls were derived from a random sample of tumors with unknown outcome not used to build the MPS. All presurgery predictors (PSA, clinical stage, biopsy Gleason) and other demographic variables (age at diagnosis and race) were tested independently and in combination with the MPS in Cox proportional hazards survival analysis with the time variable represented by progression to metastasis.

3. Results

3.1. Prediction Models

Our selection models resulted in three hundred and sixty-eight genes (from 3 metastasis signatures) with a CNA status that was concordant among METS and mPTs and contrasted with iPTs ($Z \geq 1.7$) (Supplemental Table 1, see Supplementary Materials available online at doi:10.1155/2012/873570). With these genes, we developed the MPS and tested the accuracy as an independent predictor of metastasis, with a subset of primary tumors ($n = 52$) not used to develop the signatures ($n = 13$ mPTs and $n = 39$ control primary tumors, Table 1). As a continuous predictor, applying the MPS to a Cox proportional hazards model resulted in a significant association to the endpoint of metastasis-free survival (2.88; 95% CI = 1.15 – 7.2; $P = 0.02$) (Table 2).

Patients diagnosed with prostate cancer have several pretreatment variables, such as, clinical stage (combination of digital rectum exam, PSA, and ultrasound/MRI), biopsy Gleason score and other demographic measures (e.g., age or race) to guide the decision to undergo surgery. These variables have marginal clinical utility and, in our cohorts, none of these clinical variables were statistically significant in univariate or multivariate logistic regression models. In multivariate Cox regression models (Table 2), only the MPS score reached statistical significance, indicating, that the MPS score was the only reproducible predictor of metastasis-free survival.

Notably, the clinical stage was specific when palpable tumor was detected (T2 or greater); however, it lacked sensitivity, because 47% (9/19) of pathological stage-4 cases that evaluated *ex-vivo* were diagnosed as T1C before surgery [7]. Twenty-seven percent (13 out of 49) of clinical stage T1C tumors that were upstaged following prostatectomy resulted in distant metastasis formation. Therefore, staging at the time of biopsy can seriously underestimate the severity of disease. Similarly, the biopsy Gleason score versus the postsurgery Gleason score was underestimated in 38% of cases and overestimated in 8% [7] (Figure 8).

3.2. Metastatic Potential Score Distributions

Significant differences as measured by Mann-Whitney test of the MPS were observed for the metastasis ($P < 0.001$) and mPT ($P = 0.001$) groups, compared to the control primary tumors (Figure 9). The MPS in the lymph-node-positive primary tumors (derived from the MSK ($n = 9$) and Stanford ($n = 9$) cohorts did not differ significantly from the control tumor group ($P_{\text{MSK}} = 0.34, P_{\text{Stanford}} = 0.13, P_{\text{Combined}} = 0.08$), which reflected the marginal ability of this clinical parameter to predict distant metastasis in previous reports [15].

Consistent with our assumption that the control cohorts contained a fraction of mPTs, their MPS overlapped the MPS range of the cases. Furthermore, control primary tumors

Table 1: Clinical And histological characteristics of samples used to validate the metastatic potential score model.

	Case	Control
n	13	39
Age		
Mean	59.5	59.1
Median	61	58
Standard deviation	7.1	7.3
Range	46–67	46–73
Race		
Asian	0 (0%)	1 (1.9%)
Black	1 (1.9%)	4 (7.7%)
Unknown	0 (0%)	2 (3.8%)
White Non-Hispanic	12 (23.1%)	32 (61.5%)
Clinical stage		
T1C	4 (7.7%)	23 (44.2%)
T2	5 (9.6%)	16 (30.8%)
T3	4 (7.7%)	0 (0%)
T4	0 (0%)	0 (0%)
Biopsy Gleason score		
5	0 (0%)	0 (0%)
6	4 (7.7%)	26 (50%)
7	7 (13.5%)	10 (19.2%)
8	2 (3.8%)	2 (3.8%)
9	0 (0%)	1 (1.9%)
Prediagnosis biopsy PSA (ng/mL)		
Median	6.9	5.6
<4	2 (3.8%)	6 (11.5%)
4–10	6 (11.5%)	24 (46.2%)
>10	4 (7.7%)	7 (13.5%)
Pretreatment PSA (ng/mL)		
Median	12.8	5.6
<4	2 (3.8%)	7 (13.5%)
4–10	4 (7.7%)	26 (50%)
>10	7 (13.5%)	6 (11.5%)

Table 2: Cox proportional hazards model analysis of the metastatic potential score and clinical predictors.

Component	Hazard ratio	*P*	95% CI
Univariate			
MPS	2.87	0.02	1.2–7.2
Pretreatment PSA	1.00	0.04	1.0–1.1
Clinical stage T2-T3	1.27	0.70	0.4–4.2
Multivariate			
MPS	2.61	0.05	1.0–6.8
Clinical stage T2-T3	0.90	0.87	0.3–3.1
Pretreatment PSA	1.00	0.18	1.0-1.0

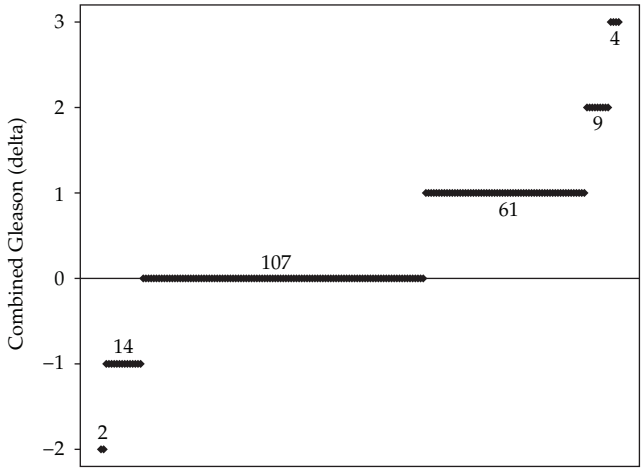


Figure 8: Biopsy versus pathology Gleason score. The difference between the Gleason score as measured from a biopsy of the tumor relative to the pathological assessment of the score using the radical prostatectomy surgical specimen (y-axis). The x-axis represents the sample index.

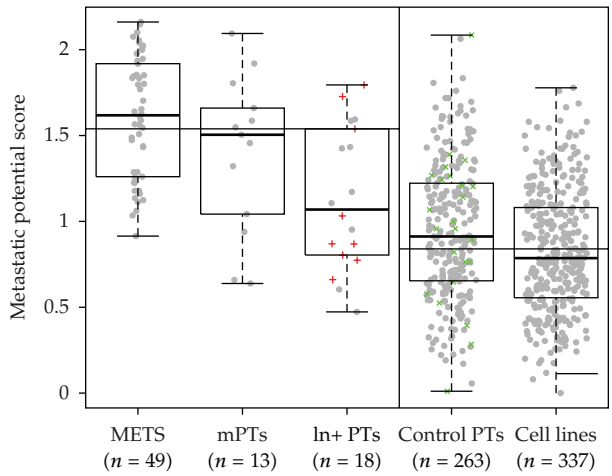


Figure 9: Boxplot showing the metastatic potential scores for all samples involved in the analysis. All high-risk tumors are shown in the left three boxes (metastases, progressors, and lymph-node-positive samples), while unknown control primary tumors and the publically available cell line data are shown in the right boxes. The red "+" symbols in the lymph-node-positive box represent those samples from the MSK dataset, distinguishing them from the SU cohort lymph-node-positive samples. The green "x" symbols in the control primary tumors plot represent selected low-risk primary tumors (individuals with no biochemical recurrence (PSA) for at least 80 months).

(from MSK cohort) that did not recur biochemically (as measured by PSA) after 80 months of followup, (represented by green Xs in Figure 9) were not significantly correlated with the MPS. To determine whether other cancer types exhibited a similar metastatic landscape of CNAs to that observed in prostate cancer, we calculated the metastatic potential score for 337 cancer cell lines. We observed an overall distribution that overlapped with low-risk prostate primary tumors (Figure 9). However, 22 of the 337 cell lines ranked by MPS were above the

75th percentile of the prostate primary tumors and metastases. These cell lines originated from tumors of the lung ($n = 10$), breast ($n = 3$), colon ($n = 2$), and melanoma ($n = 2$). Other singletons in this group of 22 cell lines originated from thyroid, rectum, pharynx, pancreas, and kidney.

3.3. *Biomarker Functional Significance*

Another way to validate our algorithms is by data mining the functional attributes of the metastasis genes identified by the selection model. As expected, many of the top-ranking metastasis genes identified have molecular functions related to alteration of nuclear and extracellular matrix structure and metabolic modification that enhance processes characteristic of escape from anoikis (a key metastasis specific process). A heat map of the CNA events of signature genes for all prostate tumors is suggestive of a path toward the different high frequency amplification versus deletion events that contrast the high-risk and low-risk tumors (Figure 10). The mid-risk region with its relative paucity of signature events may represent the starting point of two alternative pathways of subsequent copy number alteration, one leading to metastasis and the other to an indolent state. The locking in of these "antimetastasis" events in indolent tumors may explain why they failed to metastasize despite extended periods of watchful waiting.

One of the top predictor genes, the solute carrier family SLC7A5 gene, deleted on chromosome 16q24.2, encodes a neutral aminoacid transporter protein (LAT1) that has been implicated in multiple cancers (prostate [16], breast [17], ovarian [18], lung [19], and brain [20]) and has been shown to have utility as a diagnostic [21–23] and drug target in cell line [24–26] and preclinical animal models [27]. The normal function of LAT1 is to regulate cellular aminoacid concentration, L-glutamine (efflux) and L-leucine (influx). Reduced activity of LAT1 results in increased concentrations of L-glutamine which has been shown to constitutively fuel mTOR activity [28]. Seven other solute carrier superfamily members (SLCO5A1, SLC7A2, SLC10A5, SLC26A7, SLC25A37, SLC38A8, and SLC39A14) were predictive of metastatic potential in our models, likely creating a cellular environment conducive to metastasis.

A second subset of signature genes included 6 Cadherin family members encoding calcium dependent cell adhesion glycoproteins (CDH2, CDH8, CDH13, CDH15, CDH17, and PCDH9). Many of the Cadherin family proteins have putative functions associated with metastasis progression [29] and have been included in diagnostic panels [30, 31].

A third subset of 5 genes predicted to contribute to metastatic potential were potassium channels, KCNB2, KCNQ3, KCNAB1, KCTD8, and KCNH4. Notably, 3 other potassium channels reside in the highly amplified region between 8q13 and 8q24 (KCNS2, KCNV1 and KCNK9) that did not rank high in our analysis but may have weak or modifier effects. High levels of cytoplasmic potassium ion concentrations have been shown to inhibit the hallmark mitochondrial apoptotic cascade of membrane disruption and ensuing release of cytochrome C, caspase, and nuclease degradation of cellular components [32]. Furthermore, another study showed that the methylation status of potassium channel, KCNMA1 (10q22.3), was predictive of prostate cancer recurrence [33]. The activity of voltage-gated potassium channels in prostate cancer cell lines, LNCaP (low metastatic potential) and PC3 (high metastatic potential), were observed to be markedly different [34]. The complete set of metastasis signature genes likely represents various subsets of functions. Representation of different gene family members suggests that each tumor may have a unique profile to

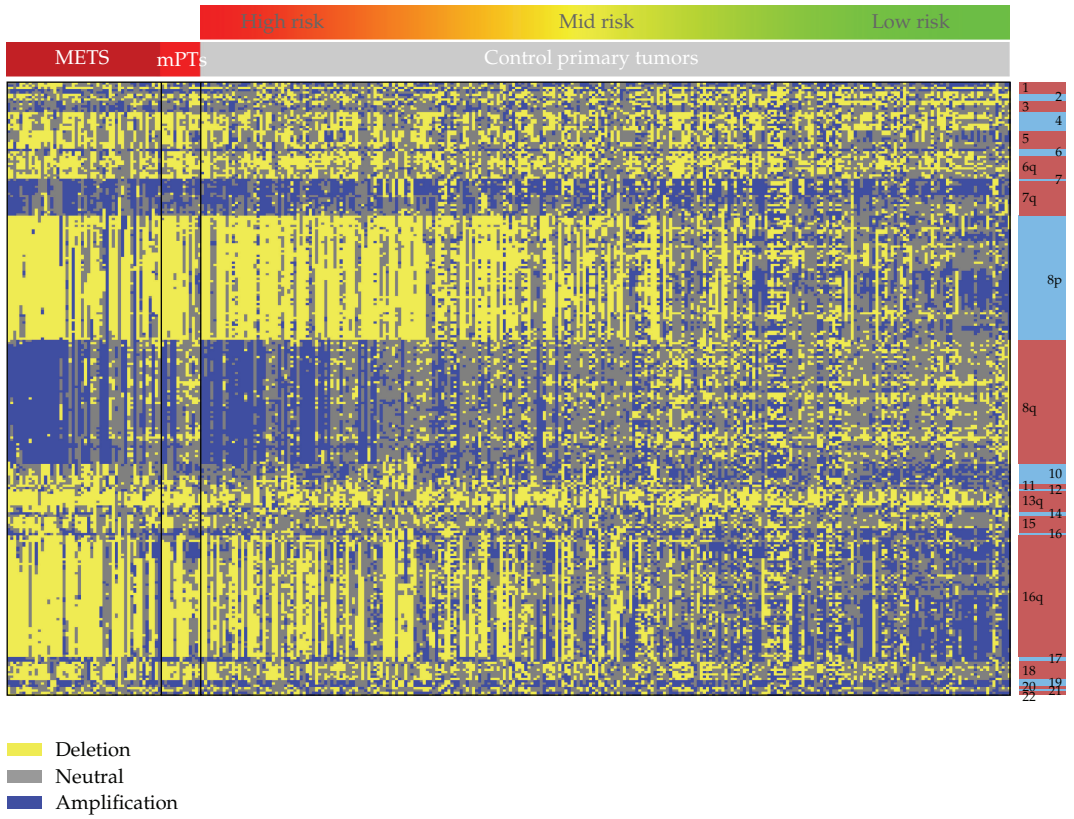


Figure 10: Heatmap showing copy number amplifications and deletions for tumor samples in the gene signature. The genes are arranged in genomic order; position is indicated by the colored bar on the right. The tumor samples (x-axis) are arranged by subtype (metastatic (METS), progressors (mPTs), and control primary tumors) and further sorted by their metastatic potential score. A strong pattern emerges in the metastasis samples on the left and is shared by the progressors and high-risk primary tumors. Further towards the right, the metastatic pattern diminishes and even shows a reversal in copy number pattern in some chromosomal areas.

progress to metastasis, yet different members of a gene family may contribute to a functional redundancy. Notably, the genomic DNA landscape around the androgen receptor locus on chromosome X represents a compelling observation linking CNAs to a functional cause and effect response of androgen ablation therapy (Figure 5).

4. Summary

In this study, we developed a semisupervised clustering algorithm that can infer the classification of a primary tumor based on metastatic risk. This was essential to overcome the limitations inherent to prostate cancer cohorts for collecting long-term clinical outcome data. Our novel approach to modeling the CNA data based on Darwin's evolutionary selection theory allowed us to identify genes associated with the specific metastatic processes of anoikis. Current clinical models for assessing risk are aimed at predicting biochemical recurrence, rather than metastasis, and do not include genomic information. This limitation

was underscored in a study with a large cohort of greater than 10,000 men who had undergone radical prostatectomy [35]. Within that cohort, about 20% of men developed biochemical recurrence within 5 years of the procedure, but subsequently only 10% of the men with biochemical recurrence developed distant metastases after 12 years.

This proposed new classification method and selection model allowed us to develop a metastatic potential score that could be used for predicting an individual's metastasis-free survival at the time of diagnosis. With validation in additional cohorts and statistical models with known metastasis outcome, this approach may lead to a significant advancement in determining whether aggressive treatment of prostate cancer is necessary. This predictor might be important for correctly categorizing men at the time of diagnosis and could predict whether surgery, radiation therapy, or watchful waiting was warranted. Because the proposed tool, tumor genomic analysis, is comprehensive for identifying the genetic changes that are associated with the pathogenesis of metastasis, there is a greater likelihood of selecting a sufficient number of markers that are both sensitive and specific predictors. This method could be applied to other cancers (e.g., breast) that exhibit variation in the metastatic potential of the primary tumor and have similar difficulties in collecting tumor samples with long-term clinical outcome data.

Acknowledgments

The authors would like to thank Dr. Kelly Maxwell for extracting the genomic DNA for the NYU cohort and all of the reviewers for their thoughtful suggestions.

References

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA—A Cancer Journal for Clinicians*, vol. 59, no. 4, pp. 225–249.

[2] L. Klotz, L. Zhang, A. Lam, R. Nam, A. Mamedov, and A. Loblaw, "Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer," *Journal of Clinical Oncology*, vol. 28, no. 1, pp. 126–131, 2010.

[3] M. Ohori, M. Kattan, P. T. Scardino, and T. M. Wheeler, "Radical prostatectomy for carcinoma of the prostate," *Modern Pathology*, vol. 17, no. 3, pp. 349–359, 2004.

[4] R. Beroukhim et al., "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, pp. 899–905, 2010.

[5] J. Sun, W. Liu, T. S. Adams et al., "DNA copy number alterations in prostate cancers: a combined analysis of published CGH studies," *Prostate*, vol. 67, no. 7, pp. 692–700, 2007.

[6] P. Castro, C. J. Creighton, M. Ozen, D. Bcrel, M. P. Mims, and M. Ittmann, "Genomic profiling of prostate cancers from African American men," *Neoplasia*, vol. 11, no. 3, pp. 305–312, 2009.

[7] B. S. Taylor et al., "Integrative genomic profiling of human prostate cancer," *Cancer Cell*, vol. 18, pp. 11–22, 2010.

[8] J. Lapointe, C. Li, C. P. Giacomini et al., "Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis," *Cancer Research*, vol. 67, no. 18, pp. 8504–8510, 2007.

[9] W. Liu, S. Laitinen, S. Khan et al., "Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer," *Nature Medicine*, vol. 15, no. 5, pp. 559–565, 2009.

[10] J. M. Korn, F. G. Kuruvilla, S. A. McCarroll et al., "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs," *Nature Genetics*, vol. 40, no. 10, pp. 1253–1260, 2008.

[11] R. D. C. Team, *R Foundation for Statistical Computing*, Vienna, Austria, 2009.

[12] N. Rabbee and T. P. Speed, "A genotype calling algorithm for affymetrix SNP arrays," *Bioinformatics*, vol. 22, no. 1, pp. 7–12, 2006.

[13] W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, vol. 21, no. 19, pp. 3763–3770, 2005.

[14] D. P. Cahill, K. W. Kinzler, B. Vogelstein, and C. Lengauer, "Genetic instability and darwinian selection in tumours," *Trends in Cell Biology*, vol. 9, no. 12, pp. M57–M60, 1999.
[15] S. A. Boorjian, R. H. Thompson, S. Siddiqui et al., "Long-term outcome after radical prostatectomy for patients with lymph node positive prostate cancer in the prostate specific antigen era," *Journal of Urology*, vol. 178, no. 3, pp. 864–871, 2007.
[16] T. Sakata, G. Ferdous, T. Tsuruta et al., "L-type amino-acid transporter 1 as a novel biomarker for high-grade malignancy in prostate cancer," *Pathology International*, vol. 59, no. 1, pp. 7–18, 2009.
[17] K. Kaira, N. Oriuchi, H. Imai et al., "L-type amino acid transporter 1 and CD98 expression in primary and metastatic sites of human neoplasms," *Cancer Science*, vol. 99, no. 12, pp. 2380–2386, 2008.
[18] M. Kaji, M. Kabir-Salmani, N. Anzai et al., "Properties of L-type amino acid transporter 1 in epidermal ovarian cancer," *International Journal of Gynecological Cancer*, vol. 20, no. 3, pp. 329–336, 2010.
[19] H. Imai, K. Kaira, N. Oriuchi et al., "L-type amino acid transporter 1 expression is a prognostic marker in patients with surgically resected stage I non-small cell lung cancer," *Histopathology*, vol. 54, no. 7, pp. 804–813, 2009.
[20] K. Kobayashi, A. Ohnishi, J. Promsuk et al., "Enhanced tumor growth elicited by L-type amino acid transporter 1 in human malignant glioma cells," *Neurosurgery*, vol. 62, no. 2, pp. 493–503, 2008.
[21] J. M. S. Bartlett, J. Thomas, D. T. Ross et al., "Mammostrat as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy," *Breast Cancer Research*, vol. 12, no. 4, article no. R47, 2010.
[22] B. Z. Ring, R. S. Seitz, R. A. Beck et al., "A novel five-antibody immunohistochemical test for subclassification of lung carcinoma," *Modern Pathology*, vol. 22, no. 8, pp. 1032–1043, 2009.
[23] B. Z. Ring, R. S. Seitz, R. Beck et al., "Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer," *Journal of Clinical Oncology*, vol. 24, no. 19, pp. 3039–3047, 2006.
[24] X. Fan, D. D. Ross, H. Arakawa, V. Ganapathy, I. Tamai, and T. Nakanishi, "Impact of system L amino acid transporter 1 (LAT1) on proliferation of human ovarian cancer cells: a possible target for combination therapy with anti-proliferative aminopeptidase inhibitors," *Biochemical Pharmacology*, vol. 80, no. 6, pp. 811–818, 2010.
[25] K. Yamauchi, H. Sakurai, T. Kimura et al., "System L amino acid transporter inhibitor enhances anti-tumor activity of cisplatin in a head and neck squamous cell carcinoma cell line," *Cancer Letters*, vol. 276, no. 1, pp. 95–101, 2009.
[26] C. S. Kim, S. H. Cho, H. S. Chun et al., "BCH, an inhibitor of system L amino acid transporters, induces apoptosis in cancer cells," *Biological and Pharmaceutical Bulletin*, vol. 31, no. 6, pp. 1096–1100, 2008.
[27] K. Oda, N. Hosoda, H. Endo et al., "L-Type amino acid transporter 1 inhibitors inhibit tumor cell growth," *Cancer Science*, vol. 101, no. 1, pp. 173–179, 2010.
[28] P. Nicklin, P. Bergman, B. Zhang et al., "Bidirectional transport of amino acids regulates mTOR and autophagy," *Cell*, vol. 136, no. 3, pp. 521–534, 2009.
[29] M. Yilmaz and G. Christofori, "Mechanisms of motility in metastasizing cells," *Molecular Cancer Research*, vol. 8, no. 5, pp. 629–642, 2010.
[30] A. Celebiler Cavusoglu, Y. Kilic, S. Saydam et al., "Predicting invasive phenotype with CDH1, CDH13, CD44, and TIMP3 gene expression in primary breast cancer," *Cancer Science*, vol. 100, no. 12, pp. 2341–2345, 2009.
[31] Y. Lu, W. Lemon, P.-Y. Liu et al., "A gene expression signature predicts survival of patients with stage I non-small cell lung cancer," *PLoS Medicine*, vol. 3, no. 12, article e467, pp. 2229–2243, 2006.
[32] D. Ekhterae, O. Platoshyn, S. Krick, Y. Yu, S. S. McDaniel, and J. X. J. Yuan, "Bcl–2 decreases voltage-gated K^+ channel activity and enhances survival in vascular smooth muscle cells," *American Journal of Physiology, Cell Physiology*, vol. 281, no. 1, pp. C157–C165, 2001.
[33] D. K. Vanaja, M. Ehrich, D. Van Den Boom et al., "Hypermethylation of genes for diagnosis and risk stratification of prostate cancer," *Cancer Investigation*, vol. 27, no. 5, pp. 549–560, 2009.
[34] M. E. Laniado, S. P. Fraser, and M. B. A. Djamgoz, "Voltage-gated K^+ channel activity in human prostate cancer cell lines of markedly different metastatic potential: distinguishing characteristics of PC-3 and LNCaP cells," *Prostate*, vol. 46, no. 4, pp. 262–274, 2001.
[35] T. Nakagawa, T. M. Kollmeyer, B. W. Morlan et al., "A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy," *PLoS One*, vol. 3, no. 5, Article ID e2318, 2008.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 913560, 14 pages
doi:10.1155/2012/913560

Research Article

Robust Semiparametric Optimal Testing Procedure for Multiple Normal Means

Peng Liu[1] and Chong Wang[1,2]

[1] *Department of Statistics, Iowa State University, Ames, IA 50011, USA*
[2] *Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University, Ames, IA 50011, USA*

Correspondence should be addressed to Peng Liu, pliu@iastate.edu

Received 27 March 2012; Accepted 10 May 2012

Academic Editor: Yongzhao Shao

In high-dimensional gene expression experiments such as microarray and RNA-seq experiments, the number of measured variables is huge while the number of replicates is small. As a consequence, hypothesis testing is challenging because the power of tests can be very low after controlling multiple testing error. Optimal testing procedures with high average power while controlling false discovery rate are preferred. Many methods were constructed to achieve high power through borrowing information across genes. Some of these methods can be shown to achieve the optimal average power across genes, but only under a normal assumption of alternative means. However, the assumption of a normal distribution is likely violated in practice. In this paper, we propose a novel semiparametric optimal testing (SPOT) procedure for high-dimensional data with small sample size. Our procedure is more robust because it does not depend on any parametric assumption for the alternative means. We show that the proposed test achieves the maximum average power asymptotically as the number of tests goes to infinity. Both simulation study and the analysis of a real microarray data with spike-in probes show that the proposed SPOT procedure performs better when compared to other popularly applied procedures.

1. Introduction

The problem of statistically testing mean difference for each of thousands of variables is commonly encountered in genomic studies. For example, the popularly applied microarray technology allows the gene expression study of tens of thousands of genes simultaneously. The recent advance of next-generation sequencing technology allows the measurement of gene expression in an even higher dimension. These high-throughput technologies have revolutionized the way genomic studies progress and provided rich data to explore. However, these experiments are expensive, and as a consequence, such experiments typically involve only a few samples for each treatment group. This results in the "large p, small n"

problem for hypothesis testing, and the power of the statistical tests can be very low after controlling the multiple testing error, such as the false discovery rate (FDR).

The normalized signal intensities from microarray experiments are generally assumed to follow normal distributions [1–4]. The recently emerging next-generation sequencing data may also be modeled approximately using normal distributions, when the number of reads are large or under certain transformation [5]. Thus multiple testing problem for normal means has wide applications in genetic and genomic studies, and it is also a general statistical question of interest.

Several testing procedures have been proposed in the context of microarray study, including the SAM test [6], Efron's t-test [7], the regularized t-test [8], the B-statistic [1] and its multivariate counterpart, the MB-statistic [9], the test of Wright and Simon [10], the moderated t-test [2], the F_S test [3] and the test of [11] which is similar to the F_S test, the F_{SS} test [4], and the LEMMA test [12]. Although numerous procedures have been proposed, very few can be justified to achieve the optimal power. Among these procedures, Hwang and Liu [4] proposed a framework and showed that an optimal testing procedure can be derived within such a framework. They also proposed a test with maximum average power (the MAP test) and an approximated version, the F_{SS} test. Here the optimality was defined in terms of maximizing the power averaged across all tests for which the null hypotheses are false while controlling FDR. This method provides theoretical guide for developing optimal multiple testing procedures. The popularly applied moderated t-statistic developed by Smyth [2] can also be shown to achieve optimal power asymptotically under different distributional assumptions from the F_{SS} test. Both the moderated t-statistic and the F_{SS} test assume that the mean expression levels (or the mean of interesting contrasts) of all genes follow a normal distribution although the parameters for this distribution vary between the two tests. Yet in practice such distribution depends on the population of genes selected in a particular study and often does not follow the prespecified parametric distribution. This raises concerns about the robustness of the moderated t and the F_{SS} tests.

The objective of this paper is to develop an optimal and robust multiple testing procedure without any distributional assumptions on the mean. As in Hwang and Liu [4], the optimality is defined in terms of maximizing the power averaged across all tests for which the null hypotheses are false while controlling FDR. We develop a semiparametric optimal testing procedure which we abbreviate as the SPOT procedure. The distribution of the mean expression across genes is not assumed to follow a parametric model which makes our method robust to violations to normal assumptions. We find that the SPOT procedure works very well in simulation studies and in an analysis of real microarray data with spike-in probes.

The remaining of this paper is organized as follows. We first introduce necessary notations in Section 2. Then, in Section 3, we describe the general concepts of optimal testing procedures. We propose our semiparametric optimal testing (SPOT) procedure in Section 4 and describe its implementation in Section 5. Section 6 presents simulation studies. Section 7 shows the analysis result of a real microarray dataset. Section 8 provides a summary of this paper.

2. Notations

An appropriate linear model is typically fitted for each gene based on the design of a microarray experiment. Section 2 of Smyth [2] provides a nice description of this topic.

Given the linear model, suppose that we have an interesting contrast to test for each gene. This contrast may be the difference between the means of two treatment groups or linear combination of means from several treatment groups. For the simplicity of description, we call the genes whose contrast means are not zero as the differentially expressed (DE) genes and the genes whose contrast means equal to zero as equivalently expressed (EE) genes. After fitting the linear model for each gene, we obtain an estimate for the contrast for each gene, X_g. In addition, we get the estimate of the sample residual variance, s_g^2, for each gene. For each $g = 1, \ldots, G$, X_g and s_g^2 are related to true parameters, μ_g and σ_g^2, by $X_g \mid \mu_g, \sigma_g^2 \sim N(\mu_g, v_g\sigma_g^2)$ and $s_g^2 \mid \sigma_g^2 \sim (\sigma_g^2/d_g)\chi_{d_g}^2$, where μ_g is the contrast mean for gene g, σ_g^2 is the true residual variance for gene g, and the coefficients v_g and d_g are determined by the design of the experiment. Two examples are given as follows.

Example 2.1. Two-channel microarray experiment to compare two treatments. Assume that each sample from treatment A is paired randomly with a sample from treatment B and each pair of samples is cohybridized onto one slide. After normalization and appropriate transformation, the difference of normalized expression measurements between the two samples on each slide is analyzed for each gene. Hence, this is a paired sample case and the number of data points for each gene is n, the number of slides. We are interested in identifying DE genes. In this case, X_g is the mean difference of the paired samples for gene g. s_g^2 is the sample variance for gene g. So $v_g = 1/n$ and $d_g = n - 1$.

Example 2.2. Affymetrix microarray experiment with two independent samples. Assume sample sizes are n_1 and n_2 for treatment A and treatment B, respectively. The statistic X_g is the difference in sample means of normalized expression measurements between two groups for gene g. s_g^2 is the pooled sample variance. Then $v_g = 1/n_1 + 1/n_2$ and $d_g = n_1 + n_2 - 2$.

Given the data X_g and s_g^2, an ordinary t-test with statistic $t_g = X_g/\sqrt{v_g}s_g$ may be used to test the null hypothesis H_g^0: $\mu_g = 0$. However, the power of such tests is low after controlling multiple testing error. So statistical methods with higher power are in demand for such high-dimensional testing problem as encountered in gene expression studies.

3. Optimal Testing Procedures

In the analysis of high-dimensional gene expression data such as microarray data, we are more interested in the average behavior of the tests across all genes rather than the performance of an individual test. Because the dimension of tests is huge, multiple testing errors should be controlled to avoid too many type I errors. Controlling FDR is an important method for controlling multiple testing errors and is widely used for genomic studies. Although many testing procedures have been developed as reviewed in Section 1, the paper by Hwang and Liu [4] provides some theoretical guide on how to derive optimal testing procedures within an empirical Bayes framework. The optimal tests are defined to be the ones that maximize the power averaged across all genes for which the null hypotheses are false while controlling FDR. Such optimal tests have been called MAP tests, where MAP stands for maximum average power [4].

In a Bayesian framework, we assume model parameters like μ_g and σ_g^2 follow some distributions. The residual variances of genes, σ_g^2, have been modeled by prior distribution like inverse gamma [2, 10] or log-normal [4] distribution independent of whether the null hypothesis is true or false. For EE genes, the mean of contrast X_g, μ_g, is equal to 0. For DE

genes, the mean μ_g is not 0 almost surely. Denote the alternative distribution of μ_g by $\pi_1(\cdot)$. Based on the Neyman-Pearson fundamental lemma, for a randomly selected gene g, the most powerful test statistic for testing $H_g^0 : \mu_g = 0$ versus $H_g^1 : \mu_g \sim \pi_1(\mu_g)$ is given by

$$T_g^{\mathrm{NP}} = \frac{\int\int f\left(X_g, s_g^2 \mid \mu_g, \sigma_g^2\right)\pi_1(\mu_g)\pi\left(\sigma_g^2\right)d\mu_g d\sigma_g^2}{\int f(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2)\pi(\sigma_g^2)d\sigma_g^2}, \tag{3.1}$$

where $\pi(\cdot)$ denote the prior distributions of σ_g^2. And the test rejects the null hypothesis H_g^0 when T_g^{NP} is large. The simultaneous testing procedure where all genes are tested using the most powerful statistics T_g^{NP}, $g = 1, 2, \ldots, G$, achieves the highest average power while controlling FDR, as proved in Hwang and Liu [4].

One popular multiple-testing method for microarray data is the moderated t-test proposed by Smyth [2]. Smyth proposed to model the residual variance σ_g^2 with the prior distribution:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2, \tag{3.2}$$

where $\chi_{d_0}^2$ denotes a chi-square distribution with degrees of freedom d_0 and s_0^2 is another hyperparameter. This prior distribution is equivalent to an inverse-gamma distribution and has been shown to fit real data well. Compared to a standard t-test statistic $t_g = x_g/\sqrt{v_g}s_g$, Smyth's moderated t-statistic takes the form of

$$\tilde{t}_g = \frac{x_g}{\sqrt{v_g}\tilde{s}_g}, \tag{3.3}$$

where

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0} \tag{3.4}$$

is a shrinkage estimator of σ_g^2 by shrinking s_g^2 toward s_0^2.

In practice, the unknown hyperparameters d_0 and s_0^2 for the distribution of the variance σ_g^2 can be estimated consistently by the method of moments, that is, equating the empirical and expected first two moments of $\log s_g^2$ [2]. Smyth [2] showed that the moderated t-test is equivalent to the B statistic proposed in Lönnstedt and Speed [1] which was derived as the posterior odds under the assumption that the distribution of μ_g under the alternative hypothesis follows $N(0, v_0\sigma_g^2)$, where v_0 is a constant. In fact, we can prove the claim that the moderated t-test achieves the optimal average power asymptotically under their assumptions for μ_g and σ_g^2. The proof is in the appendix.

However, note that the assumption that μ_g of DE genes follows a normal distribution with mean zero is restrictive. It is likely that, for example, there are more upregulated genes than downregulated genes for some studies which suggests that the mean of μ_g should be

positive. Hwang and Liu [4] have proposed a more general normal prior distribution of μ_g for DE genes:

$$\pi_1(\mu_g) \sim N\left(\theta, \tau_g^2\right), \tag{3.5}$$

where the mean of this distribution is not necessarily zero but to be estimated based on data. In addition, the variance for this distribution does not depend on the residual variance. Under this model, they have derived an optimal test and an approximated version of the test statistic (F_{SS} test) that is computationally faster. The F_{SS} statistic shrinks both the estimate of mean μ_g and the estimate of variance σ_g^2.

Both the moderated t-test and the F_{SS} test have been shown to achieve optimal power asymptotically under the assumption of normal distribution for the alternative means. Simulation studies also confirm that the power of the tests is superior under the model assumptions. However, a single normal distribution assumption on μ_g for DE genes may not be appropriate for all cases and the distribution of $\pi_1(\mu_g)$ may consist of a mixture of different subgroup distributions, for example, a mixture of two normal distributions with one having a negative mean and the other having a positive mean. If the parametric distributional assumptions of $\pi_1(\mu_g)$ are violated, the power of an optimal test built under those assumptions will suffer.

4. Semiparametric Optimal Testing (SPOT) Procedure

To obtain a more robust procedure, we propose to model the distribution of the mean μ_g nonparametrically while still deriving the optimal procedure. For the variance σ_g^2, the inverse gamma distributional assumption is reasonable and works well in practice, so we still keep this assumption. Hence, we will derive a semiparametric optimal testing procedure that we call the SPOT procedure.

Note that the numerator and denominator of the most powerful test statistic (3.1) are the joint marginal distributions of (X_g, s_g^2), under the alternative and null hypothesis, respectively. By denoting the marginal distributions by

$$\begin{aligned} m_1\left(X_g, s_g^2\right) &= \iint f\left(X_g, s_g^2 \mid \mu_g, \sigma_g^2\right)\pi_1(\mu_g)\pi\left(\sigma_g^2\right)d\mu_g d\sigma_g^2, \\ m_0\left(X_g, s_g^2\right) &= \int f\left(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2\right)\pi\left(\sigma_g^2\right)d\sigma_g^2, \end{aligned} \tag{4.1}$$

statistic (3.1) becomes

$$T_g^{\mathrm{NP}} = \frac{m_1\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)}. \tag{4.2}$$

The null marginal distribution $m_0(X_g, s_g^2)$ only involves integration with respect to variance σ_g^2. With consistent estimators of hyperparameters as proposed in Smyth [2], we can estimate $m_0(X_g, s_g^2)$ consistently. For the alternative marginal distribution $m_1(X_g, s_g^2)$, it is hard to find

a consistent estimator without any distributional assumption on μ_g. If we were to know which genes are DE, then we could estimate $m_1(X_g, s_g^2)$ nonparametrically with observed values of (X_g, s_g^2) from the DE gene population. Many nonparametric density estimators are consistent, for example, the histogram estimators and the kernel density estimators with proper choices of bandwidths [13]. However, the knowledge of differential expression is the research question of the study and of course is not available for all genes. Considering all genes without separating those that are differentially expressed from those that are not, we have a mixture distribution of differentially expressed and nondifferentially expressed genes. The mixture density of the marginal distributions, denoted by $m_m(X_g, s_g^2)$, can be estimated consistently by nonparametric density estimators with observed (X_g, s_g^2) for all genes $g = 1, \ldots, G$. Can this consistent estimator of $m_m(X_g, s_g^2)$ help us construct a most powerful test statistic, together with a consistent estimator of $m_0(X_g, s_g^2)$?

Suppose that p_0 and p_1 are proportions of EE and DE genes, respectively, with $0 \leq p_0, p_1 \leq 1$ and $p_0 + p_1 = 1$, then the mixture marginal density is

$$m_m\left(X_g, s_g^2\right) = p_0 m_0\left(X_g, s_g^2\right) + p_1 m_1\left(X_g, s_g^2\right). \tag{4.3}$$

The ratio of mixture marginal density $m_m(X_g, s_g^2)$ and the null marginal density $m_0(X_g, s_g^2)$ is a monotonic function of the statistic T_g^{NP} expressed in formula (4.2) because

$$\begin{aligned} \frac{m_m\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)} &= \frac{p_0 m_0\left(X_g, s_g^2\right) + p_1 m_1\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)}, \\ &= p_0 + p_1 \frac{m_1\left(X_g, s_g^2\right)}{m_0\left(X_g, s_g^2\right)}. \end{aligned} \tag{4.4}$$

Thus the test that rejects the null hypothesis when $m_m(X_g, s_g^2)/m_0(X_g, s_g^2)$ is large is also a most powerful test. Note that to calculate this statistic, we only need to estimate $m_m(X_g, s_g^2)$ and $m_0(X_g, s_g^2)$ but do not have to estimate the proportions p_0 and p_1.

Let $\widehat{m}_m(X_g, s_g^2)$ denote any consistent density estimator of $m_m(X_g, s_g^2)$, and let $\widehat{m}_0(X_g, s_g^2)$ denote any consistent estimator of $m_0(X_g, s_g^2)$, such that

$$\begin{aligned} \widehat{m}_m\left(X_g, s_g^2\right) &\xrightarrow{P} m_m\left(X_g, s_g^2\right) \quad \text{as } G \nearrow \infty, \\ \widehat{m}_0\left(X_g, s_g^2\right) &\xrightarrow{P} m_0\left(X_g, s_g^2\right) \quad \text{as } G \nearrow \infty, \end{aligned} \tag{4.5}$$

where $\xrightarrow{P}$ denotes convergence in probability. Then the statistic $\widehat{m}_m(X_g, s_g^2)/\widehat{m}_0(X_g, s_g^2)$ has the optimal testing power asymptotically. Notice the convergence with respect to G, which is usually huge in the microarray and RNA-seq studies.

We have already discussed the availability of a parametric consistent estimator of $m_0(X_g, s_g^2)$ through estimating the hyperparameters d_0 and s_0^2 of σ_g^2 in Section 3. For $m_m(X_g, s_g^2)$, any theoretically consistent density estimator $\widehat{m}_m(X_g, s_g^2)$ of joint data (X_g, s_g^2)

can be used to construct the test statistic (4.4) with asymptotically optimal average power. For example, nonparametric estimators such as histograms, kernel density estimates, and local polynomial estimators can all be utilized. As our test statistic $\widehat{m}_m(X_g, s_g^2)/\widehat{m}_0(X_g, s_g^2)$ involves both parametric and nonparametric parts, we name it the semiparametric optimal test (SPOT).

5. Implementation of SPOT

In this section, we discuss details in implementation of the proposed SPOT procedure.

5.1. Estimation of $m_0(X_g, s_g^2)$

The null marginal density

$$\begin{aligned} m_0\left(X_g, s_g^2\right) &= \int f\left(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2\right)\pi\left(\sigma_g^2\right)d\sigma_g^2 \\ &= \int \frac{e^{-x_g^2/(2v_g\sigma_g^2)}}{\left(2\pi v_g\sigma_g^2\right)^{1/2}}\left(\frac{d_g}{2\sigma_g^2}\right)^{d_g/2}\frac{s^{2(d_g/2-1)}e^{-d_g s_g^2/(2\sigma_g^2)}}{\Gamma\left(d_g/2\right)}\left(\frac{d_0 s_0^2}{2}\right)^{d_0/2}\frac{\sigma_g^{-2(d_0/2+1)}e^{-d_0 s_0^2/2\sigma_g^2}}{\Gamma(d_0/2)}d\sigma_g^2 \\ &= C_2 \cdot s_g^{2(d/2-1)}\left(\frac{x_g^2/v_g + d_0 s_0^2 + d_g s_g^2}{2}\right)^{-(1+d_0+d_g)/2}, \end{aligned} \tag{5.1}$$

where C_2 is a constant. As in Smyth [2] and Hwang and Liu [4], we assume that the distribution of σ_g^2 does not depend on whether a gene is DE or EE. Then, all genes are used to estimate the parameters d_0 and s_0^2. We apply the method of moments proposed in Smyth [2] to get estimates of d_0 and s_0^2. Replacing unknown parameters d_0 and s_0^2 in $m_0(X_g, s_g^2)$ by their consistent method of moments estimates leads to a consistent estimator $\widehat{m}_0(X_g, s_g^2)$ of $m_0(X_g, s_g^2)$.

5.2. A Hybrid Method for Estimation of $m_m(X_g, s_g^2)$

Although any consistent estimator $\widehat{m}_m(X_g, s_g^2)$ can be used to construct a SPOT statistic of the form $\widehat{m}_m(X_g, s_g^2)/\widehat{m}_0(X_g, s_g^2)$, in practice, a density estimator that converges fast would be always preferred. It is known that the accuracy of the density estimators goes down quickly as the dimension increases [13]. We have tried a few two-dimensional density estimators for $\widehat{m}_m(X_g, s_g^2)$, including the kernel estimators. Due to the curse of dimensionality, the direct two-dimensional density estimators do not perform as satisfactory as a hybrid estimator that we develop and would suggest to use. This hybrid estimator has a component that is similar to kernel estimators, whereas it also utilizes the prior information on variances σ_g^2 to help improving the accuracy.

In constructing this estimator, we first estimate the marginal density of X_g by the typical kernel density estimate:

$$\widehat{f}(x_g) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{h}K\left(\frac{x_g - x_i}{h}\right), \tag{5.2}$$

where h is a positive value known as bandwidth. We estimate the conditional density of $f(s_g^2 \mid x_g)$ by using

$$f\left(s_g^2 \mid x_g\right) = \int f\left(s_g^2 \mid \sigma_g^2, x_g\right) f\left(\sigma_g^2 \mid x_g\right) d\sigma_g^2 = \int f\left(s_g^2 \mid \sigma_g^2\right) f\left(\sigma_g^2 \mid x_g\right) d\sigma_g^2, \tag{5.3}$$

where the second equality is a result of the independence between s_g^2 and x_g given the parameter σ_g^2. The distribution of $s_g^2 \mid \sigma_g^2$ is $(\sigma_g^2/d_g)\chi_{d_g}^2$ for normal-distributed observations. Now we need to estimate $f(\sigma_g^2 \mid x_g)$. Denote the set of genes that lie within bandwidth distance to gene g as $\{A_g : i \in A_g$ if and only if $|x_i - x_g| < h\}$. We estimate $f(\sigma_g^2 \mid x_g)$ by the following approximation that is based on the neighborhood of x_g, A_g:

$$\frac{1}{\#\{A_g\}}\sum_{i\in A_g}\frac{f(s_i^2 \mid \sigma^2)\pi(\sigma^2)}{\int f(s_i^2 \mid \sigma^2)\pi(\sigma^2)d\sigma^2}. \tag{5.4}$$

The $\#\{A_g\}$ in formula denotes the number of genes in set A_g. Substituting exact parametric form of $f(s_g^2 \mid \sigma^2)$ and $\pi(\sigma^2)$ into above formulas leads to the explicit form

$$\widehat{f}\left(s_g^2 \mid x_g\right) = C_3 \cdot \frac{1}{\#\{A_g\}}\sum_{i\in A_g} s_g^{2(d/2-1)}\left(\frac{d_g s_i^2 + d_0 s_0^2 + d_g s_g^2}{2}\right)^{-(d_0+d_g)/2}, \tag{5.5}$$

where C_3 is a constant. The product between the kernel estimate $\widehat{f}(x_g)$ and the conditional estimate $\widehat{f}(s_g^2 \mid x_g)$ provides us a joint density estimator of the mixture

$$\widehat{m}_m\left(X_g, s_g^2\right) = \widehat{f}(x_g) \cdot \widehat{f}\left(s_g^2 \mid x_g\right). \tag{5.6}$$

With this approximation, we cannot theoretically show that the resulting estimator, $\widehat{m}_m(X_g, s_g^2)$, is consistent but it works better in practice than the consistent kernel density estimator of the joint density $m_m(X_g, s_g^2)$.

6. Simulation Study

In order to evaluate the performance of our proposed SPOT procedure, we performed three simulation studies. The gene expression data were simulated from Normal (μ_{gi}, σ_g^2) for observations of gene g in treatment group i. The way to sample μ_{gi} and σ_g^2 differs across

simulation studies. We assume that there are two treatment groups and 3 replicates per treatment group. For each simulation setting, one hundred sets of gene expression data were independently simulated, and each dataset included 10,000 genes. The performances of the SPOT, moderated t, F_{SS}, and ordinary t-test statistics were evaluated for by comparing their average behavior averaged across the 100 datasets.

6.1. Simulation Study I

In the first simulation study, we have two settings that differ in the number of DE genes. For the first setting, $G_1 = 2,500$ are DE genes whereas the other $G_0 = 7,500$ are EE. In the second setting, only $G_1 = 1,800$ are DE while the other $G_0 = 8,200$ are EE. Gene expression means μ_{gi} and variances σ_g^2 were simulated as follows:

$$\begin{aligned}
&\mu_{g1} = 0 \quad \forall g;\\
&\mu_{g2} \sim \text{Normal}\left(0.5, 0.3^2\right) \quad \text{for } g = 1 \text{ to } 0.3G_1;\\
&\mu_{g2} \sim \text{Normal}\left(1, 0.3^2\right) \quad \text{for } g = (0.3G_1 + 1) \text{ to } 0.9G_1;\\
&\mu_{g2} \sim t_1\ (0.5) \quad \text{for } g = (0.9G_1 + 1) \text{ to } G_1;\\
&\mu_{g2} = 0 \quad \text{for } g = (G_1 + 1) \text{ to } 10000;\\
&\sigma_g^2 \sim \text{Gamma}\ (2, 4) \quad \forall g.
\end{aligned} \tag{6.1}$$

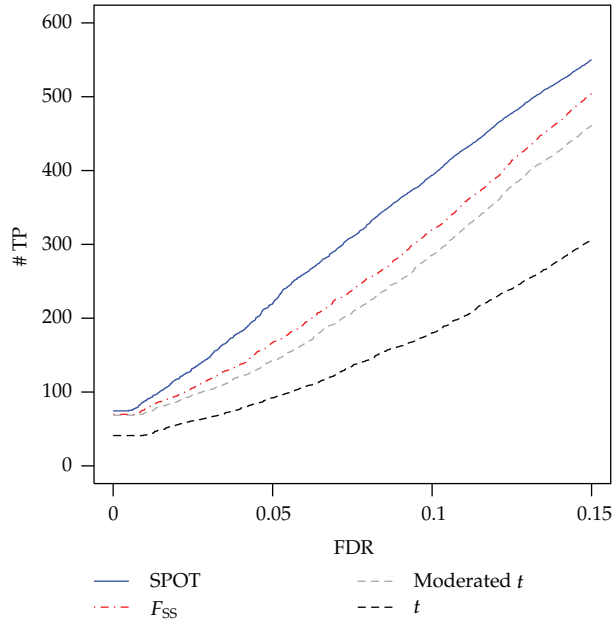
For each simulated data, SPOT, moderated t, F_{SS}, and ordinary t-test statistics were calculated and evaluated using the number of selected true positives at various FDR levels. The plots of number of true positives versus FDR for SPOT, moderated t, F_{SS}, and ordinary t-test statistics are shown in Figure 1. Simulation settings 1 and 2 generated similar results. The ordinary t-test is the poorest method under comparison. The moderated t-test is considerably better than the ordinary t-test although it is worse than F_{SS} test. Our proposed SPOT test is superior to all other three methods, with the largest number of true positive findings than the other three statistics at the same FDR levels.

6.2. Simulation Study II

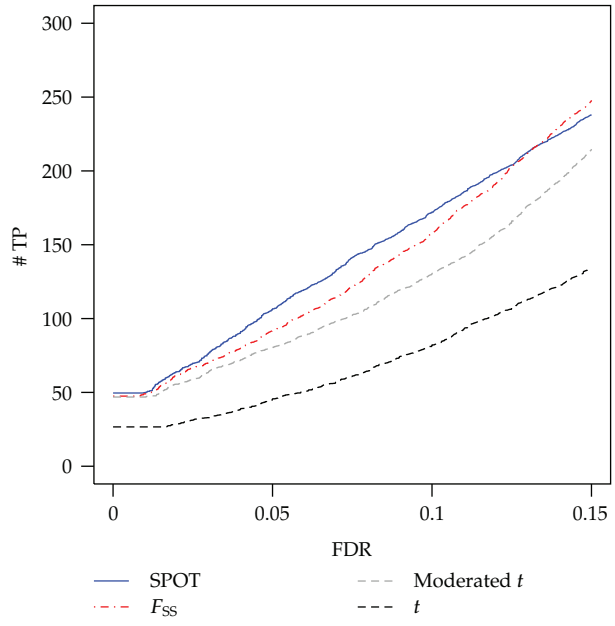
To check how the variance distribution affects the relative ranking of the SPOT procedure, we did another simulation study the same as the setting 1 of simulation study I except that the variances were simulated from a log-normal distribution, which is the assumption under which the F_{SS} test was derived. As Figure 2 shows, the results are similar to those from simulation I. The SPOT procedure still performs much better than all the other three methods.

6.3. Simulation Study III

Typically, the parametric test achieves higher power than the nonparametric test if the parametric assumption is appropriate. To check the robustness of the SPOT procedure, we



(a) Simulation I, setting 1



(b) Simulation I, setting 2

Figure 1: Simulation study I: plots of number of true positives (# TP) versus false discovery rate (FDR) from analyses using SPOT, moderated t, and F_{SS} methods.

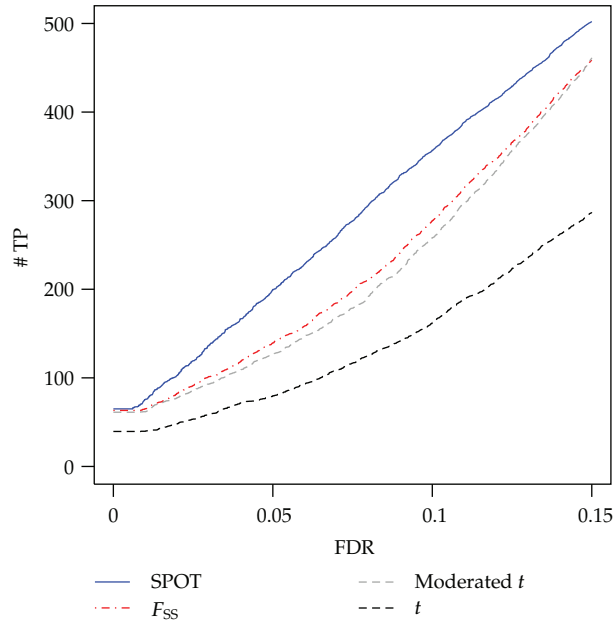


Figure 2: Simulation study II: plots of number of true positives (# TP) versus false discovery rate (FDR) from analyses using SPOT, moderated t, and F_{SS} methods.

simulated data under the parametric assumption for both μ_{gi} and σ_g^2 under which the F_{SS} test was derived. Specifically, for the 2,500 differentially expressed genes, μ_{gi} were drawn from a normal distribution with mean 1.2 and standard deviation 0.3, σ_g^2 were sampled from a log-normal distribution with parameters −0.96 and 0.8. Figure 3 shows that the SPOT procedure and the F_{SS} test are comparable to each other when FDR is small (less than 0.05) and they are both much better than the moderated t-test and the ordinary t-test. When FDR is between 0.05 and 0.15, the F_{SS} test is the best while the SPOT procedure is the next best performing procedure, which is still much better than the moderated t-test and the ordinary t-test.

7. Evaluation Using the Golden Spike Microarray Data

In this section, we compare the performances of different methods using a real microarray dataset from experiments conducted using Affymetrix GeneChip in the Golden Spike Project. The Golden Spike Project generated microarray datasets comparing two replicated groups in which the relative concentrations of a large number of genes are known. The two groups are the spike-in group and the control group, each with three chips. Data and information related to this project are available through the website http://www2.ccr.buffalo.edu/halfon/spike/. More specifically, the Golden Spike dataset included 1309 individual cRNAs "spiked in" at known relative concentrations between the two groups. The fold-changes between the spike-in and control group were assigned at different levels for different cRNAs, and the levels ranged from 1.2 to 4. Hence, these cRNAs were truly "differentially expressed" between groups and we consider them as DE genes. In addition, a background sample of 2551 RNA species was present at identical concentrations in both samples. So these 2551 RNA species were not differentially expressed between the two

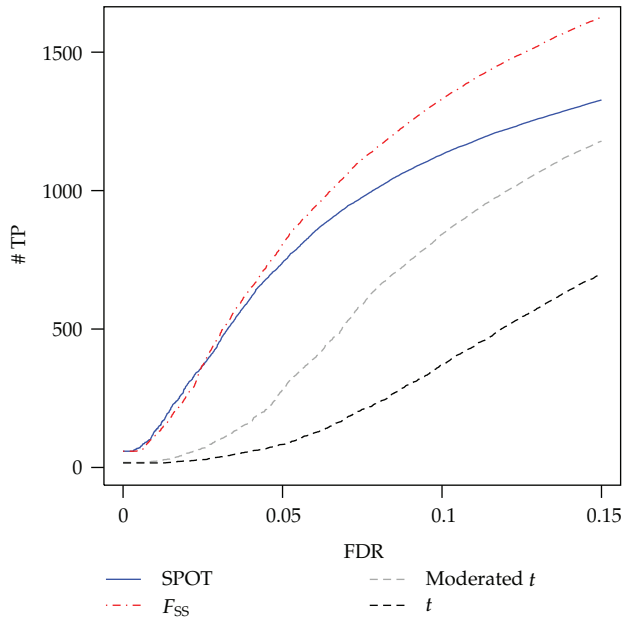


Figure 3: Simulation study III: plots of number of true positives (# TP) versus false discovery rate (FDR) from analyses using SPOT, moderated t, and F_{SS} methods.

Table 1: Golden Spike data: number of true positives selected by three testing procedures at critical FDR levels.

Method	FDR					
	0.01	0.02	0.05	0.1	0.15	0.2
SPOT	754	847	947	986	1015	1051
Moderated t	466	588	821	911	969	1018
F_{SS}	442	563	824	908	975	1016

groups. With the knowledge of the true differential expression status, this real microarray dataset provides an ideal case to evaluate the performances of different methods without imposing any distributional assumption for variances and means as usually is done in simulation studies.

With the summary dataset downloaded from the Golden Spike Project website, we calculated the SPOT, the moderated t, the ordinary t, and the F_{SS} statistics and evaluated their performances using the true statuses of RNA based on the design. Figure 4 shows the plots of number of true positives versus FDR for the ordinary t, moderated t, F_{SS}, and SPOT procedures over a range of FDR $\in [0, 0.15]$ which is of most practical interest. It can be observed that the performance of the SPOT procedure improves over the performances of the other three methods throughout the whole range of FDR in these plots. In addition, the improvement is substantial at lower FDR levels. For example, the SPOT procedure detects 754 true positives at the FDR level of 1% while the moderated t-test only detects 466 and the F_{SS} test only detects 442 true positives (Table 1).

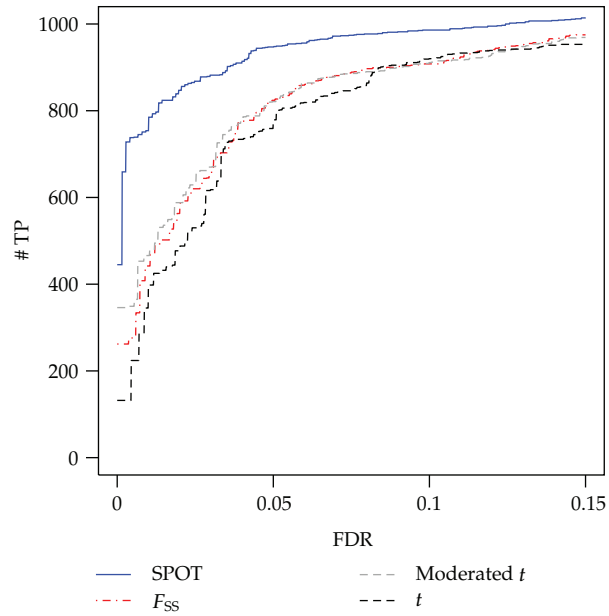


Figure 4: Golden spike data: plots of number of true positive (TP) genes versus false discovery rate (FDR) from analysis using SPOT, moderated t, and F_{SS} tests.

8. Summary

In this paper, we have derived a semiparametric optimal testing (SPOT) procedure for high-dimensional gene expression data analysis. Although the method is illustrated for analyzing microarray data, it can be applied to any high-dimensional testing problem with normal model. Our test statistic is justified to be asymptotically most powerful, without any assumption on the mean parameter of differential expression. The asymptotic property is derived when the number of genes is large, which is reasonable for high-dimensional gene expression studies. We also provided an approximate version to implement the SPOT procedure in practice and evaluated the performance of our proposed test statistic using both simulation studies and real microarray data analysis. The proposed SPOT method is shown to outperform the popularly applied moderated t and the F_{SS} statistics, which are optimal only under certain normality conditions of the mean. There is still potential in improving the performance of SPOT procedure if better density estimates can be found for the marginal distributions $m_m(X_g, s_g^2)$ and $m_0(X_g, s_g^2)$.

Appendix

Proof of the Claim That the Moderated t-Test Achieves the Optimal Average Power Asymptotically under the Assumptions That $\mu_g \sim N(0, v_0\sigma_g^2)$ and $1/\sigma_g^2 \sim 1/d_0 s_0^2 \chi_{d_0}^2$

Under Smyth's [2] model assumptions, the most power test statistic formula (3.1) derived under the Neyman-Pearson lemma becomes

$$T_g^{\mathrm{NP}} = \frac{\iint f\left(X_g, s_g^2 \mid \mu_g, \sigma_g^2\right)\pi_1(\mu_g)\pi_1\left(\sigma_g^2\right)d\mu_g d\sigma_g^2}{\int f(X_g, s_g^2 \mid \mu_g = 0, \sigma_g^2)\pi_0(\sigma_g^2)d\sigma_g^2}$$

$$= \frac{\int\int\left(e^{-(x_g-\mu_g)^2/2v_g\sigma_g^2}/\left(2\pi v_g\sigma_g^2\right)^{1/2}\right)\left(d_g/2\sigma_g^2\right)^{d_g/2}\left(s^{d_g-2}e^{-d_gs_g^2/2\sigma_g^2}/\Gamma\left(d_g/2\right)\right)\mathcal{A}}{\int\left(e^{-x_g^2/2v_g\sigma_g^2}/\left(2\pi v_g\sigma_g^2\right)^{1/2}\right)\left(d_g/2\sigma_g^2\right)^{d_g/2}\left(s^{2(d_g/2-1)}/\Gamma\left(d_g/2\right)\right)e^{-d_gs_g^2/2\sigma_g^2}\cdot\mathcal{B}}$$

$$= C_1\cdot\left(\frac{x_g^2/\left(v_0+v_g\right)+d_0s_0^2+d_gs_g^2}{x_g^2/v_g+d_0s_0^2+d_gs_g^2}\right)^{-(1+d_0+d_g)/2}, \tag{A.1}$$

where $\mathcal{A}$ denotes $(e^{-\mu_g^2/(2v_0\sigma_g^2)}/(2\pi v_0\sigma_g^2)^{1/2})(d_0s_0^2/2)^{d_0/2}(\sigma_g^{-d_0+2}/\Gamma(d_0/2))e^{-d_0s_0^2/(2\sigma_g^2)}d\mu_g d\sigma_g^2$, and $\mathcal{B}$ denotes $(d_0s_0^2/2)^{d_0/2}(\sigma_g^{-2(d_0/2-1)}/\Gamma(d_0/2))e^{-d_0s_0^2/(2\sigma_g^2)}d\sigma_g^2$, which is a monotonic function of Smyth's [2] moderated t-statistic, with C_1 being some constant. Thus the claim follows with existence of consistent estimates of d_0 and s_0^2, which has been shown in Smyth [2].

References

[1] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, no. 1, pp. 31–46, 2002.

[2] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, article 3, 2004.

[3] X. Cui, J. Hwang, J. Qiu, N. J. Blades, and A. Churchill, "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, vol. 6, no. 1, pp. 59–75, 2005.

[4] J. T. G. Hwang and P. Liu, "Optimal tests shrinking both means and variances applicable to microarray data analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 9, article 36, 2010.

[5] T. Cai, J. Jeng, and H. Li, "Robust detection and identification of sparse segments in ultra-high dimensional data analysis," *Journal of the Royal Statistical Society: Series B*, vol. 14, part 4, 2012.

[6] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[7] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayes analysis of a microarray experiment," *The Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1151–1160, 2001.

[8] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[9] Y. C. Tai and T. P. Speed, "A multivariate empirical Bayes statistic for replicated microarray time course data," *The Annals of Statistics*, vol. 34, no. 5, pp. 2387–2412, 2006.

[10] G. W. Wright and R. M. Simon, "A random variance model for detection of differential gene expression in small microarray experiments," *Bioinformatics*, vol. 19, no. 18, pp. 2448–2455, 2003.

[11] T. Tong and Y. Wang, "Optimal shrinkage estimation of variances with applications to microarray data analysis," *The Journal of the American Statistical Association*, vol. 102, no. 477, pp. 113–122, 2007.

[12] H. Bar, J. Booth, E. Schifano, and M. T. Wells, "Laplace approximated EM microarray analysis: an empirical Bayes approach for comparative microarray experiments," *Statistical Science*, vol. 25, no. 3, pp. 388–407, 2010.

[13] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer, New York, NY, USA, 2006.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 478680, 14 pages
doi:10.1155/2012/478680

Review Article

High-Dimensional Cox Regression Analysis in Genetic Studies with Censored Survival Outcomes

Jinfeng Xu

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546

Correspondence should be addressed to Jinfeng Xu, staxj@nus.edu.sg

Received 22 February 2012; Revised 21 May 2012; Accepted 26 May 2012

Academic Editor: Yongzhao Shao

With the advancement of high-throughput technologies, nowadays high-dimensional genomic and proteomic data are easy to obtain and have become ever increasingly important in unveiling the complex etiology of many diseases. While relating a large number of factors to a survival outcome through the Cox relative risk model, various techniques have been proposed in the literature. We review some recently developed methods for such analysis. For high-dimensional variable selection in the Cox model with parametric relative risk, we consider the univariate shrinkage method (US) using the lasso penalty and the penalized partial likelihood method using the folded penalties (PPL). The penalization methods are not restricted to the finite-dimensional case. For the high-dimensional ($p \to \infty$, $p \ll n$) or ultrahigh-dimensional case ($n \to \infty$, $n \ll p$), both the sure independence screening (SIS) method and the extended Bayesian information criterion (EBIC) can be further incorporated into the penalization methods for variable selection. We also consider the penalization method for the Cox model with semiparametric relative risk, and the modified partial least squares method for the Cox model. The comparison of different methods is discussed and numerical examples are provided for the illustration. Finally, areas of further research are presented.

1. Introduction

The modern high-throughput technologies offer the possibility of a powerful, genome-wide search for the genetic and environmental factors that have influential effects on diseases. The identification of such factors and the discernment of such a relationship can lead to better understanding of the causation of diseases and better predictive models. In the presence of a large number of covariates, it is very challenging to build a model which fully utilize all the information and excels in both parsimony and prediction accuracy. In classical settings where the number of covariates p is fixed and the sample size n is large, subset selection coupled with model selection criteria such as Akaike's information criterion (AIC) and

Bayesian information criterion (BIC) can be used to identify relevant variables or choose the best model with the optimal prediction accuracy. However, subset selection is inherently unstable because of its discreteness [1]. To overcome this drawback of subject selection, Tibshirani [2] proposed the least absolute shrinkage and selection operator (LASSO) for simultaneous coefficient estimation and variable selection. Fan and Li [3] further proposed the penalization method with the smoothly-clipped absolute deviation (SCAD) penalty and rigorously established its oracle properties. The optimal properties of the lasso or SCAD-based penalization methods are not restricted to the finite-dimensional case. In the high-dimensional case ($p \to \infty,\ p \ll n$), Fan and Peng [4] proved that the oracle properties are well retained. In the ultra high-dimensional case ($n \to \infty,\ n \ll p$), Fan and Lv [5] proposed the sure independence screening method (SIS) which first reduces dimensionality from high to a moderate scale that is below the sample size and then apply a penalization method. In a general asymptotic framework, the sure independence screening method is shown to fare well for even exponentially growing dimensionality. In high-dimensional or ultra high-dimensional situations, J. Chen and Z. Chen [6] proposed the extended Bayesian information criterion (EBIC) and established its selection consistency under mild conditions. The EBIC is further extended to the generalized linear model [7].

When the clinical outcome involves time to an event such as age at disease onset or time to cancer recurrence, the regression analysis is often conducted by the Cox relative risk model. The classical Cox model is only applicable to the situation where the number of subjects is much larger than the number of covariates. Thus, to accommodate the large p and small n scenario, some variable selection and dimension reduction techniques have to be implemented in a regression analysis. Recently, for variable selection in the Cox model, a number of approaches based on the efficient shrinkage method have been proposed and gained increased popularity. See, for example, LASSO [8], SCAD [9], and adaptive lasso [10, 11].

For high-dimensional variable selection in the Cox model with parametric relative risk, we review the univariate shrinkage method (US) [12] and the penalized partial likelihood approach [13]. The univariate shrinkage method [12] assumes the independence of the covariates in each risk set and the partial likelihood factors into a product. This leads to an attractive procedure which is univariate in its operation and most suitable for a high-dimensional variable selection setting. The variables are entered into the model based on the size of their Cox score statistics, and in nature the method is similar to univariate thresholding in linear regression and nearest shrunken centroids in classification. The univariate shrinakge method is applicable to the setting with an arbitrary number of variables but is less informative in identifying joint effects from multiple variables. The penalized partial likelihood approach [13] employs a class of folded-concave penalties to the Cox parametric relative risk model and strong oracle properties of non-concave penalized methods are established for nonpolynomial (NP) dimensional data. A coordinate-wise algorithm is used for finding the grid of solution paths. The penalized partial likelihood approach investigates joint effects from multiple variables and is applicable to both the finite-dimensional and high-dimensional cases. For the ultra high-dimensional case, some preliminary procedures such as the sure independence screening (SIS) method and the extended Bayesian information criterion (EBIC) can be used to reduce the number of variables to be moderately below the sample size before the penalized partial likelihood approach is formally adopted.

The aforementioned two methods both adopt the Cox parametric relative risk model for the covariance analysis. In practice, the parametric form of the relative risk model is quite

restrictive and may not be tenable. In Section 3, we review a penalization method in the Cox model with semiparametric relative risk approach [14]. The relative risk is assumed to partially linear with one parametric component and one nonparametric component. Two penalties are applied sequentially to simultaneously estimate the parameters and select variables for both the parametric and the nonparametric parts. The semiparametric relative risk model greatly relaxes the restrictive assumption of the classical Cox model and facilitate its use in exploratory data analysis. Although the method is proposed for the finite-dimensional setting, it is straightforward to be extended to the high dimensional and ultra high-dimensional situations the same as the penalization method for the Cox model with parametric relative risk.

In Section 4, we review a modified partial least squares method for dimension reduction in the Cox regression approach [15] which provides another alternative approach to dealing with the problem of high-dimensionality. By mimicking the partial least squares in the linear model, it first constructs the components which are linear combinations of original covariates. By sequentially determining the components and using the cross-validation to select the number of components, a parsimonious model with good predictive accuracy can be obtained.

In Section 5, we discuss the comparison of different methods and numerical examples are provided for the illustration. Finally, several important problems for future research are also presented in Section 6.

2. The Penalization Methods for the Cox Model with Parametric Relative Risk

2.1. The Cox Model with Parametric Relative Risk

We consider the setting where the time to event is subject to right censoring and the observations consist of $\{Y_i = T_i \wedge C_i,\ \delta_i = I(T_i \leq C_i),\ Z_i,\ i = 1, \ldots, n\}$, where T_i is the survival time, C_i the censoring time, and Z_i is the p-dimensional vector of covariates. The Cox relative risk model assumes that the conditional hazard function of T given the covariates $Z = z$ takes the following form:

$$\lambda(t \mid Z = z) = \lambda_0(t) \exp\left(\beta_0^T Z\right), \tag{2.1}$$

where $\lambda_0(t)$ is the unknown baseline hazard function and β_0 is the unknown vector of coefficients. The influential effects that the covariates might have on the time T_i are examined by the relative risk. The unknown coefficient vector β_0 is estimated by maximizing the partial likelihood function

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta^T Z_i)}{\sum_{j \in R_i} \exp(\beta^T Z_j)} \right\}^{\delta_i}, \tag{2.2}$$

or equivalently, the log partial likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta^T Z_i - \log \left[\sum_{j \in R_i} \exp\left(\beta^T Z_j\right) \right] \right\}, \tag{2.3}$$

where $R_i = \{j : Y_j \geq Y_i\}$. As in the least squares estimation, the estimation of β from the partial likelihood function requires the sample size n is much larger than the dimension of the covariate vector p. In practice, a marginal approach is often adopted which includes one covariate at a time and maximizes

$$L_k(\beta_k) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta_k Z_{ik})}{\sum_{j \in R_i} \exp(\beta_k Z_{jk})} \right\}^{\delta_i}, \tag{2.4}$$

or

$$\ell_k(\beta_k) = \sum_{i=1}^{n} \delta_i \left\{ \beta_k Z_{ik} - \log \left[\sum_{j \in R_i} \exp(\beta_k Z_{jk}) \right] \right\}, \tag{2.5}$$

for $k = 1, \ldots, p$.

2.2. The Univariate Shrinkage Method

To identify the variables which are associated with T, multiple testing procedures will be used to make valid statistical inferences. However, Tibshirani [12] looks at the problem from another perspective. Since the maximizer of the partial likelihood is not unique when $n \ll p$, he proposes the regularized partial likelihood approach by using the lasso penalty as follows:

$$J(\beta) = \ell(\beta) - \lambda \sum_{k=1}^{p} |\beta_k|. \tag{2.6}$$

By assuming that both conditionally on each risk set, and marginally, the covariates are independent of one another, and using Bayes's theorem, Tibshirani [12] shows that the log partial likelihood function

$$\ell(\beta) = \text{constant} + \sum_{k=1}^{p} \ell_k(\beta_k). \tag{2.7}$$

The regularized partial likelihood function is

$$J(\beta) = \text{constant} + \sum_{k=1}^{p} \ell_k(\beta_k) - \lambda \sum_{k=1}^{p} |\beta_k|, \tag{2.8}$$

Table 1: Simulation results for variable selection in the Cox model with parametric relative risk.

p	Method	MSE	FDR	PSR	MMS
250	US	1.62	0.06	0.58	3.97
	PS	0.92	0.16	0.76	6.55
	EBIC	0.96	0.10	0.69	4.70
500	US	1.71	0.07	0.55	3.95
	PS	1.01	0.18	0.74	6.63
	EBIC	1.06	0.11	0.66	4.73
1000	US	1.84	0.08	0.52	3.91
	PS	1.12	0.20	0.70	6.69
	EBIC	1.15	0.13	0.63	4.78

Table 2: Results for the microarray lung cancer dataset.

Method	Number of selected genes	Median P value ($\times 10^{-4}$)
US	5	10.06
PS	13	0.064
EBIC	8	0.082

and results in the the Cox univariate shrinkage (CUS) estimator which maximizes the penalized function. Since the maximization is a set of one-dimensional maximization $\ell_k(\beta_k) - \lambda|\beta_k|,\ k = 1,\ldots,p$, for a range of λ, we can fairly easily get the penalized estimates $\widehat{\beta}_k$. Actually, the entire paths of the regularization estimates can be obtained. It can also be shown that

$$\widehat{\beta}_k \neq 0 \Longleftrightarrow \frac{|U_k|}{\sqrt{V_k}} > \lambda, \tag{2.9}$$

where U_k and V_k are the gradient of the (unpenalized) log-partial likelihood and the (negative) observed Fisher information. This is similar to soft/hard thresholding. Hence, the Cox univariate shrinkage method ranks all the covariates based on the Cox score statistic. As the Cox score is often used for determining the univariate significance of covariates, the results have easy interpretation. The tuning parameter λ can be selected by cross-validation as in Verweij and van Houwelingen [16] or directly determined as in Donoho and Johnstone [17]. The Cox univariate shrinkage method presents a numerically convenient approach for high-dimensional variable selection in the Cox model. In the literature, the modified shooting algorithm [10] and the least squares approximation based algorithm [11] both yield the entire solution paths, but only when n is much larger than p.

One drawback of the Cox univariate shrinkage procedure is that the variables enter into the model based on their univariate Cox scores. Thus, when two predictors are both strongly predictive and highly correlated with each other, both will appear in the model. In that case, it may be more desirable to just include one of them for parsimony. This can be done using preconditioning [18] as is demonstrated by Tibshirani [12].

2.3. The Penalized Partial Likelihood Method

The penalized partial likelihood estimation with noncave penalties has been extensively studied by Fan and Li [9] for the case where the sample size n is much larger than the dimension of Z Bradic et al. [13] considered the folded penalties for the penalized partial likelihood estimation when the dimension of Z is nonpolynomial (NP). The folded penalties include the smoothly clipped absolute deviation (SCAD) and the minimax concave penalty (MCP) as special cases. The penalized log partial likelihood becomes

$$\ell(\beta) - \lambda_n \sum_{k=1}^{p} p_{\lambda_n}(|\beta_k|), \tag{2.10}$$

where $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a nonnegative tuning parameter. For a class of folded penalties, by clarifying the identification problem of the penalized partial likelihood estimates and deriving a large deviation result for divergence of a martingale from its compensator, Bradic et al. [13] establish the strong oracle properties for the penalized estimates. Note that their results also hold for the lasso penalty. The strong oracle property indicates that as both n and p goes to ∞, with probability tending to 1, the penalized estimator behaves as if the true relevant variables in the model were known. This is different from the classical notion of oracle which just requires that the estimator behaves like the oracle rather than an actual oracle itself. The strong oracle property implies the classical oracle property of Fan and Li [9] and sign consistency of Bickel et al. [19]. This tighter notion of an oracle property was first mentioned in Kim et al. [20] for the SCAD estimator of the linear model with polynomial dimensionality and then extended by Bradic et al. [21] to the penalized M-estimators under the ultrahigh dimensionality setting. Bradic et al. [13] further extended it to the Cox model by employing sophisticated techniques dealing with martingale and censoring structures.

Analogous to the Cox univariate shrinkage method, the penalized Cox relative risk method [13] proposes a coordinate wise algorithm which is especially attractive for the situation of $p \gg n$ and have been previously studied for linear and generalized linear models [5, 22, 23]. Since the coordinate-wise maximization algorithm in each iteration provides limits that are stationary points of the overall optimization, each output of the iterative coordinate ascent algorithm (ICA), Bradic et al. [13] propose gives a stationary point.

For each iteration, sequentially for $k = 1, \ldots, p$, by the partial quadratic approximation of $\ell(\beta)$ at the current estimate along the k-th coordinate while fixing the other coordinates, the k-th coordinate of the estimate is updated by maximizing the univariate penalized likelihood. Due to the univariate nature, the problem can be solved analytically, avoiding the challenges of nonconcave optimization. It updates each coordinate if the maximizer of the penalized univariate optimization increases the penalized objective function as well. The algorithm stops when two values of the penalized objective function are not different by more than a small threshold value.

Although both the iterative coordinate ascent algorithm (ICA) and the univariate shrinkage method exploit the convenient of univariate optimization, the univariate shrinkage method separates the coefficient estimates while in the iterative coordinate ascent algorithm, the coefficient estimates, are still related to one another in the iterative updating. These two methods also require different conditions. The univariate shrinkage method assumes that both conditionally on each risk set, and marginally, the covariates are independent of

one another while the penalized Cox relative risk method assumes the conditions on the folded-concave penalties, the sparsity level, the dimensionality of the covariate vector, and the magnitude of the tuning parameter λ_n.

2.4. The Penalized Partial Likelihood Approach for the Ultra High-Dimensional Case

While the univariate shrinkage method is applicable to an arbitrary dimensionality, the penalized partial likelihood requires that the sample size is larger than the number of variables. Thus, to apply the penalized partial likelihood approach to the ultra high-dimensional case, a preliminary screening procedure is needed. Fan and Lv [5] proposed the sure independence screening procedure which first shrinks the full model $1, \ldots, p$ straightforwardly and accurately down to a submodel with size $d = o(n)$. Thus, the original problem of estimating the sparse p-vector β reduces to estimating a sparse d-vector that is based on the now much smaller submodel. The penalized partial likelihood method in Section 2.3 can then be applied to the submodel. Fan and Lv [5] proved the sure independence screening method has optimal theoretical properties for even exponentially growing dimensionality.

For small n large P problems, the traditional model selection criteria such as AIC, BIC, and cross-validation choose too many features. To overcome the difficulties, J. Chen and Z. Chen [6] developed a family of extended Bayes' information criteria (EBIC). The EBIC is shown to be consistent with nice finite sample properties in both the linear model [6] and the generalized linear model [7]. For any subset model $s \subset \{1, 2, \ldots, p\}$, denote its size by $\nu(s)$. Let $\widehat{\beta}(s)$ be the maximum partial likelihood estimate corresponding to the subset model s. The extended Bayesian information criterion is defined as

$$-2\ell\left(\widehat{\beta}(s)\right) + \nu(s) \log n + 2\nu(s)\gamma \log p, \tag{2.11}$$

where γ is a prespecified constant and can be chosen to be 0.5 as suggested by J. Chen and Z. Chen [7]. Optimal theoretical properties such as selection consistency of the EBIC have been rigorously obtained by J. Chen and Z. Chen [6] for the linear model and by J. Chen and Z. Chen [7] for the generalized linear model. The EBIC can be appealingly applied to the Cox model and it is worthwhile to further investigate its theoretical properties in the Cox model which has not yet been addressed in the literature.

3. The Penalization Method for the Cox Model with Semiparametric Relative Risk

The Cox relative risk model is sometimes too restrictive in examining the covariate effects. It seems implausible that the linearity assumption holds in the presence of a large number of predictors. Intuitively, at least for some of them, the linearity assumption might be violated and the modeling of covariate effects via the parametric relative risk model might lead to erroneous results. On the other hand, there are two objectives in the high-dimensional regression analysis of genetic studies with censored survival outcomes, we not only want to identify the predictor variables which are associated with the time but also to discern

such a relationship if there does exist an association. Therefore, it is worth looking at other alternative survival models in examining the covariate effects.

Du et al. [14] proposed the penalized method for the Cox model with semiparametric relative risk model. Let $Z^T = (U^T, W^T)$, where U and W are the subvectors of Z with dimensions $d = p - q$ and q, respectively. Instead of (2.1), they assume that

$$\lambda(t \mid Z = z) = \lambda_0(t) \exp\left[\beta_0^T U + \eta(W)\right], \tag{3.1}$$

where $\eta(w) = \eta(w_1, \ldots, w_q)$ is an unknown multivariate smooth function. The model assumes the additivity of the effects of U and W and only the effect of U is postulated to be linear. The effect of W can be of any form. This greatly enhances the flexibility and facilitates more robust investigation of the covariate effects across a large number of genetic and environment factors. Similarly, the log partial likelihood is

$$\ell(\beta, \eta) = \sum_{i=1}^{n} \delta_i \left\{ \beta^T U_i + \eta(W_i) - \log\left[\sum_{j \in R_i} \exp\left(\beta^T U_j + \eta(W_j)\right) \right] \right\}. \tag{3.2}$$

Du et al. [14] proposed two penalties for the model (3.1), one penalty for the roughness of the function η and the other penalty for simultaneous coefficient estimation and variable selection. The estimation iterates between the estimation of η given an initial estimator of β and the estimation of β given an initial estimator of η. Given an estimate $\widehat{\beta}$ of β, η is estimated by maximizing

$$\ell\left(\widehat{\beta}, \eta\right) - \lambda J(\eta), \tag{3.3}$$

where J is a roughness penalty specifying the smoothness of η, and $\lambda > 0$ is a smoothing parameter controlling the tradeoff. A popular choice for J is the L_2-penalty which yields tensor product cubic splines for multivariate W. Given an estimate $\widehat{\eta}$ of η, β can be estimated by

$$\ell(\beta, \widehat{\eta}) - \sum_{k=1}^{d} p_{\theta_n}(|\beta_k|), \tag{3.4}$$

where $p_{\theta_n}(\cdot)$ is the SCAD penalty function and θ_n is the tuning parameter. In its numerical implementation, the SCAD penalty is approximated by a one-step approximation which transforms the SCAD penalty problem to a LASSO-type optimization, where the celebrated LARS algorithm [24] can be readily used to yield the entire solution path.

The algorithm converges quickly within a few iterations. In this approach, the SCAD penalty facilitates the simultaneous coefficient estimation and variable selection in the parametric component of relative risk model. As the multivariate smooth function W also involves multiple predictor variables, it is therefore necessary to identify the correct structure of η and relevant variables in W too. Taking care of variable selection for the parametric components, we still need an approach to assess the structure of the nonparametric components. By transforming the profile partial likelihood to a density estimation problem

with biased sampling, Du et al. [14] further derive a model selection tool based on the Kullback-Leibler geometry for the nonparametric component η. Specifically, a quantity based on the ratio of two Kullback-Leibler distances can be used to diagnose the feasibility of a reduced model η, the smaller the ratio is, the more feasible the reduced model is. Thus, the penalized Cox semiparametric relative risk approach provides a flexible tool for identifying relevant variables in both the parametric and nonparametric components.

4. The Modified Partial Least Squares Method for Dimension Reduction in the Cox Model

Partial least squares (PLS) [25] is a classical dimension reduction method of dealing with a large number of covariates. By constructing new variables which are linear combination of the original variables, it fully utilizes the information and a proper regression analysis can be conducted using the new variables. Different from the principal components (PCs) analysis, partial least squares utilizes the information contained in both the response variable and the predictor variables to construct new variables. This complicates its direct application to censored survival data since the response variable is subject to right censoring. Nguyen and Rocke [26] applied the standard PLS methods of Wold [25] directly to survival data and used the resulted PLS components in the Cox model for predicting survival time. Since the approach did not take into account that some of the survival time are censored and not exactly the underlying time to event, the resulting components are questionable and may induce bias. Alternatively, by reformulating the Cox model into a generalized linear model, Park et al. [27] applied the formulation of PLS of Marx [28] to derive the PLS components. Despite its validity, the introduction of many additional nuisance parameters in the reformulation makes the algorithm fail to converge when the number of covariates is large.

Li and Jiang [15] proposed a modified partial least squares method for the Cox model by constructing the components based on repeated least square fitting of residuals and Cox regression fitting. Let $w_{ij} \propto \mathrm{var}(V_{ij})$ be the weights and $\sum_{i=1}^{n} w_{ij} = 1$. First, let $X_j = (Z_{1j}, \ldots, Z_{nj})^T$ and define

$$V_{1j} = X_j - z_{\cdot j}\mathbf{1}, \tag{4.1}$$

where $z_{\cdot j} = (1/n)\sum_{i=1}^{n} Z_{ij}$, and $\mathbf{1}$ is an n-dimensional vector of all elements 1. After fitting the Cox model with one covariate at one time, we obtain the maximize partial likelihood estimate $\widehat{\beta}_{1j}$ for the predictor variable V_{1j}, $j = 1, \ldots, p$. Combining these estimates, we get the first component

$$T_1 = \sum_{j=1}^{p} w_{1j}\widehat{\beta}_{1j}V_{1j}. \tag{4.2}$$

The information in X that is not in T_1 can be written as the residuals of regressing $V_{1,j}$ on T_1

$$V_{2,j} = V_{1,j} - \frac{T_1^T V_{1j}}{T_1^T T_1}T_1. \tag{4.3}$$

By performing the Cox regression analysis with T_1 and V_{1j} (one j at a time), we obtain the maximized partial likelihood estimates $\widehat{\beta}_{2j}$ and consequently get the second component

$$T_2 = \sum_{j=1}^{p} w_{2j}\widehat{\beta}_{2j}V_{2j}. \tag{4.4}$$

This procedure extends iteratively in a natural way to give component $T_2, \ldots, T_K$, where the maximum value of K is the sample size n. Specifically, suppose that T_i has just been constructed, and to construct T_{i+1}, we first regress V_{ij} against T_i and denote the residual as $V_{(i+1),j}$, which can be written as

$$V_{(i+1),j} = V_{i,j} - \frac{T_i^T V_{ij}}{T_i^T T_i} T_i. \tag{4.5}$$

Then we fit the Cox relative risk model

$$\lambda(t \mid Z) = \lambda_0(t)\exp\left(\beta_1 T_1 + \cdots + \beta_i T_i + \beta_{(i+1),j}V_{(i+1),j}\right), \tag{4.6}$$

and obtain the maximum partial likelihood estimates $\widehat{\beta}_{(i+1),j}$ and

$$T_{i+1} = \sum_{j=1}^{p} w_{(i+1)j}\widehat{\beta}_{(i+1)j}V_{(i+1)j}. \tag{4.7}$$

With the components $T_1, \ldots, T_K$, a standard Cox regression model can be fitted and the risk score can be obtained as

$$\widehat{\beta}_1 T_1 + \cdots + \widehat{\beta}_K T_K, \tag{4.8}$$

where $\widehat{\beta}_j,\ j = 1, \ldots, K$ is the maximum partial likelihood estimate of β_j when we fit the Cox relative risk model

$$\lambda(t \mid Z) = \lambda_0(t)\exp\left(\beta_1 T_1 + \cdots + \beta_K T_K\right). \tag{4.9}$$

This can then be used for estimating the hazard function for future samples on the basis of their X values. By examining the coefficients of X values in the final model with K components, one can rank the covariate effects by the risk score. The number of K can be chosen by applying the cross-validation.

5. Comparison of Different Methods and Numerical Examples

5.1. Comparison Using Survival Prediction

In the previous sections, four different approaches have been used to identify the relevant factors which have influential effects on the survival time. In practice, it would be important

and interesting to compare different methods which can be done by using certain measure of survival prediction. To assess the performance of the methods, the data set is first divided into the training sample and the test sample randomly. For example, a v-fold cross-validation divide the sample into v parts randomly. One part is retained as test set while the rest $v-1$ folds are used as the training set. The training sample gives the estimated risk score for a given model (method) and then used in test sample for prediction. There are many measures for survival prediction. One of them mimics the random clinical trial in assigning the test sample into two groups—one "good" group and one "bad" group. Whether a subject in the test sample falls into a good group or a bad group depends on whether his/her risk score is smaller than a threshold value for the risk score. The log-rank test can then be used to test the hypothesis that there is no difference between the two groups. The smaller the P value the resulting log-rank test has, the better predictive power the estimated risk score has, which translates into the better performance of the method/model. The dataset is randomly split into training and test sample and hence for a large number of replications, the comparison of different methods can be made by looking at the summary of the P-values of the log-rank test, say, the median.

The disadvantage of the log-rank test is that the subjects are only assigned to two groups and the risk score is only utilized in comparing with a threshold value. The information contained in the risk score which is continuous is not fully utilized for survival prediction. Alternatively, we can fit a Cox regression for the test sample using the risk score estimated from the training sample as a single covariate. The predictive power of the estimated risk score can be indicated by the significance of the risk score covariate in the fitted Cox regression model for the test sample. Again, the obtained P values using different methods in a large number of replications can help us assess their performance in terms of survival prediction.

5.2. Simulation Studies

We conduct simulation studies to compare different methods. As a simple illustration, we focus on the univariate shrinkage method (US) and the penalized shrinkage method (PS) reviewed in Section 2. We set the sample size $n = 500$ and the number of covariates $p = 250, 500$, and 1000, respectively. The covariates are jointly normally distributed with equal correlation coefficient $\rho = 0.5$. The first six covariates are the only relevant variables with $\beta_1 = \beta_3 = \beta_5 = 1$ and $\beta_2 = \beta_4 = \beta_6 = -1$. The baseline hazard function in (2.1) is set to be constant 1 and the censoring time is generated from the $niform(0, \tau)$, where τ is chosen to yield the censoring proportion 30%. For the univariate shrinkage method, the top ranked variables with significance at 0.05 after Bonferroni's correction will be selected. For the penalized shrinkage method, the sure independence screening procedure preselects $n/(4 \log n) = 20$ and the penalized partial likelihood method is then applied to obtain the final model. As a third method, we directly use the EBIC to select a subset model. We report the median squared estimation error (MSE) and the squared estimation error is defined as

$$\sum_{j=1}^{p} \left| \widehat{\beta}_j - \beta_j \right|^2. \tag{5.1}$$

We also report the average number of selected variables (MMS), the average positive selection, and false discovery rates (PSR and FDR), where

$$\begin{aligned} \text{PSR} &= \frac{\sum_{j=1}^{N} \nu\left(s_j^* \bigcap s_0\right)}{N\nu(s_0)}, \\ \text{FDR} &= \frac{\sum_{j=1}^{N} \nu\left(s_j^* / s_0\right)}{\sum_{j=1}^{N} \nu\left(s_j^*\right)}, \end{aligned} \tag{5.2}$$

$N = 200$ is the number of replications, s_0 denotes the true model, and s_j^* denotes the selected model in the jth replication. The simulation results are summarized in Table 1. From Table 1, we can see that both the PS and the EBIC perform better than the US. Compared with the EBIC, the PS selects slightly more variables and has relatively larger FDRs and PSRs.

5.3. A Real Example

We analyzed microarray data by the lung cancer dataset from Beer et al. [29]. The dataset consists of gene expressions of 4966 genes for 83 patients. The patients were classified according to the progression of the disease. Sixty four patients were classified as stage I. Nineteen patients were classified as stage III. For each of the 83 patients, the survival time as well as the censoring status is available. Other covariate variables in addition to the gene expressions are age, gender, and smoking status. Our aim is to study the association of survival time with the gene expressions adjusting for the effects of the other covariates via the Cox model with parametric relative risk. The US, PS, and EBIC are used to select variables. We divide the 83 patients into two groups by randomly assigning 32 of the 64 stage I and 9 of 19 stage III patients to the training group and the remaining patients to the test group. By adjusting for the covariate (gender, age, smoking) effects, we fit the Cox model with the selected genes and construct a risk index. The 50th percentile of the risk index from the training group is employed as the threshold. We then apply the threshold to test dataset to define the low-risk and high-risk groups. To assess the predictability of the so-defined discriminant criterion, we perform a log-rank test of the difference of survivals of the two groups defined by the risk index. If the survival times of the two groups can be well separated (measured by the P-value of the log-rank test), then the method has a better predictability. We therefore use the resulting median P-value (among 1000 random splitting data into training and test sets) as the measure of prediction accuracy of different methods. The results are summarized in Table 2. It is shown that the PS and EBIC have comparable predictability which is much better than the US.

6. Further Work

We review in this paper some recently developed methods for high-dimensional regression analysis in genetic studies with censored survival outcomes. The identification of relevant variable that have the influential effects on the survival time leads to a better understanding of disease and gene/environment association for many complex diseases. Although the Cox model is widely used to examine the covariate effects through the relative risk, the

proportional hazard assumption may be violated in practice, for example, when there are long-term survivors. In some situations, other alternative models such as the additive risks model, the proportional odds model, or more generally the semiparametric transformation models may fare better. Furthermore, as we discussed before, the linearity assumption may not be tenable either. It would be interesting to develop parallel methodologies in these alternative models.

Although the Cox semiparametric relative risk relaxes the assumption to some extent, the classification of the covariates into the parametric component (with linearity assumption) and the nonparametric component (without linearity assumption) is challenging and unsolved. The problem would be more difficult when both the proportional hazards assumption and the linearity assumption are violated. In the presence of a large number of genetic and environment factors, undoubtedly we have to make necessary assumptions on the underlying structure to proceed. It is worth investigating that how the nonproportionality and the linearity assumptions alone or jointly with each other impact on the high-dimensional regression analysis. In particular, how sensitive the identification of relevant variables is to the misspecification of the model and whether there are other good structures to be postulated which have appealing properties and are most suitable for the high-dimensional regression analysis.

It is also worthy to note that for model selection, there are two different purposes. One is selection consistency such as the oracle properties. The other is the prediction accuracy. While the prediction accuracy can be well assessed by cross-validation, the selection consistency should be assessed by using the FDRs and PSRs.

Acknowledgments

The authors are very grateful to Professor Yongzhao Shao and three anonymous references for many helpful comments which improved the presentation of the paper. This work was supported by the grant from National University of Singapore (R-155-000-112-112).

References

[1] L. Breiman, "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, vol. 24, pp. 2350–2383, 1996.
[2] R. Tibshirani, "Regression shrinkage and selection via the lass," *Journal of the Royal Statistical Society B*, vol. 58, pp. 267–288, 1996.
[3] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
[4] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.
[5] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society B*, vol. 70, no. 5, pp. 849–911, 2008.
[6] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.
[7] J. Chen and Z. Chen, "Extended BIC for small-n-large-P sparse GLM," *Statistica Sinica*. In press.
[8] R. Tibshirani, "The Lasso method for variable selection in the cox model," *Statistics in Medicine*, vol. 16, pp. 385–395, 1997.
[9] J. Fan and R. Li, "Variable selection for Cox's proportional hazards model and frailty model," *The Annals of Statistics*, vol. 30, no. 1, pp. 74–99, 2002.
[10] H. H. Zhang and W. Lu, "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, 2007.

[11] H. Zou, "A note on path-based variable selection in the penalized proportional hazards model," *Biometrika*, vol. 95, no. 1, pp. 241–247, 2008.
[12] R. J. Tibshirani, "Univariate shrinkage in the Cox model for high dimensional data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, pp. 3498–3528, 2009.
[13] J. Bradic, J. Fan, and J. Jiang, "Regularization for Cox's proportional hazards model with NP-dimensionality," vol. 39, no. 6, pp. 3092–3120, 2011.
[14] P. Du, S. Ma, and H. Liang, "Penalized variable selection procedure for Cox models with semiparametric relative risk," *The Annals of Statistics*, vol. 38, no. 4, pp. 2092–2117, 2010.
[15] H. Li and G. Jiang, "Partial Cox regression analysis for high-dimensional microarray gene expression data," *Bioinformatics*, vol. 20, 1, pp. i208–i215, 2004.
[16] P. Verweij and H. van Houwelingen, "Cross-validation in survival analysis," *Statistics in Medicine*, vol. 12, pp. 2305–2314, 1993.
[17] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[18] D. Paul, E. Bair, T. Hastie, and R. Tibshirani, "'Preconditioning' for feature selection and regression in high-dimensional problems," *The Annals of Statistics*, vol. 36, no. 4, pp. 1595–1618, 2008.
[19] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of LASSO and Dantzig selector," *Annals of Statistics*, vol. 37, pp. 1705–1732, 2009.
[20] Y. Kim, H. Choi, and H.-S. Oh, "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1665–1673, 2008.
[21] J. Bradic, J. Fan, and W. Wang, "Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection," *Journal of Royal Statistical Society Series B*. In press.
[22] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.
[23] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, pp. 1–22, 2010.
[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004, With discussion, and a rejoinder by the authors.
[25] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, P. R. Krishaiaah, Ed., pp. 391–420, Academic Press, New York, NY, USA, 1966.
[26] D. Nguyen and D. M. Rocke, "Partial least squares proportional hazard regression for application to DNA microarray data," *Bioinformatics*, vol. 18, pp. 1625–1632, 2002.
[27] P. J. Park, L. Tian, and I. S. Kohane, "Linking expression data with patient survival times using partial least squares," *Bioinformatics*, vol. 18, pp. S120–S127, 2002.
[28] B. D. Marx, "Iteratively reweighted partial least squares estimation for generalized linear regression," *Technometrics*, vol. 38, pp. 374–381, 1996.
[29] D. G. Beer, S. L. Kardia, C. C. Huang et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8, pp. 816–824, 2002.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 652569, 13 pages
doi:10.1155/2012/652569

Review Article

Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies

Qiong Yang[1] and Yuanjia Wang[2]

[1] *Department of Biostatistics, Boston University School of Public Health, 810 Mass Avenue, Boston, MA 02118, USA*
[2] *Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10027, USA*

Correspondence should be addressed to Qiong Yang, qyang@bu.edu

Received 30 March 2012; Accepted 21 May 2012

Academic Editor: Yongzhao Shao

Multivariate phenotypes are frequently encountered in genetic association studies. The purpose of analyzing multivariate phenotypes usually includes discovery of novel genetic variants of pleiotropy effects, that is, affecting multiple phenotypes, and the ultimate goal of uncovering the underlying genetic mechanism. In recent years, there have been new method development and application of existing statistical methods to such phenotypes. In this paper, we provide a review of the available methods for analyzing association between a single marker and a multivariate phenotype consisting of the same type of components (e.g., all continuous or all categorical) or different types of components (e.g., some are continuous and others are categorical). We also reviewed causal inference methods designed to test whether the detected association with the multivariate phenotype is truly pleiotropy or the genetic marker exerts its effects on some phenotypes through affecting the others.

1. Introduction

Association studies, where the correlation between a genetic marker and a phenotype is assessed, are useful for mapping genes influencing complex diseases. With reduction of genotyping cost, completion of the HapMap Project [1], and more recently the 1000 Genomes Project [2], genome-wide association studies (GWAS) with several hundred thousands to tens of millions genotyped and/or imputed single nucleotide polymorphisms (SNPs) have become a common approach nowadays to search for genetic determination of complex traits.

In the study of complex diseases, several correlated phenotypes, or a multivariate (MV) phenotype with several components, may be measured to study a disorder or trait. For example, hypertension is evaluated using systolic and diastolic blood pressures;

a person's cognitive ability is usually measured by tests in domains including memory, intelligence, language, executive function, and visual-spatial function. The tests within and between domains are correlated. Most published GWAS only analyzed each individual phenotype separately, although results on related phenotypes may be reported together. Published single phenotype GWAS have successfully identified a large number of novel genetic variants predisposing to a variety of complex traits [3, 4]. However, majority of the identified genetic variants only explain a small fraction of total heritability defined as between individual phenotype variability attributable to genetic factors [4, 5]. It has been hypothesized that current GWAS may be underpowered to detect many genetic variants of moderate-to-small effects. Joint analysis of correlated phenotypes can exploit the correlation among the phenotypes, which may lead to better power to detect additional genetic variants with small effects across multiple traits or pleiotropy effects. Furthermore, joint analysis avoids multiple testing penalty incurred in analyzing each phenotype separately. Therefore, it is important to identify appropriate methods that fully utilize information in multivariate phenotypes to detect novel genetic loci in genetic association studies.

In addition to discovery of novel loci of potential pleiotropy effects, it is also important to detangle the complex relationship between phenotype components and genetic variants One of the frequently asked questions is whether a genetic variant affects multiple phenotypes simultaneously (pleiotropy) or affects one phenotype through affecting another phenotype. In this paper, we review methods for both purposes.

2. Methods for Detecting Association Using Multivariate Phenotypes

For all the methods mentioned in this section, the null hypothesis is no association between a single genetic marker and any components of a multivariate (MV) phenotype; the alternative hypothesis is the genetic maker associated with at least one phenotype component. Here we review methods for an MV phenotype consisting of all continuous, all categorical, or all time-to-event components, and methods for MV phenotypes consisting of a mixture of different types of components.

2.1. Regression Models

Regression models for clustered observations such as linear and generalized mixed effects models, generalized estimating equations, and frailty models can be used to analyze the association of a genetic marker with all continuous, categorical, or survival multivariate phenotypes.

2.1.1. Mixed Effects Models

Mixed effects models such as linear mixed effects model (LME) and generalized linear mixed effects model (GLMM) involve using fixed effects for the genetic marker effect and random effects to account for correlation among multivariate phenotypes [6, 7].

Let y_{jk} denote the kth ($k = 1, \ldots, K$) continuous component of the K-dimensional phenotype of the jth ($j = 1, \ldots, J$) individual. Let g_j be the genotype of a genetic marker of

the jth individual, and $X(g_j)$ a score of the genotype. The linear mixed effects model takes the following form:

$$y_{jk} = \beta_0 + \beta_k X(g_j) + \eta_{jk} + e_{jk}, \tag{2.1}$$

where β_0 is the intercept or other genetic or environmental fixed effects; β_k is the fixed effect size of $X(g_j)$ on the kth phenotype; $\eta_{jk}(k = 1,\ldots,K) \sim N(0,\Sigma)$ are the random effects correlated within jth person; e_{jk} is the random errors iid. $\sim N(0,\sigma_e^2)$. Between any two individuals, $\eta_{jk}, k = 1,\ldots,K$ are independent. Within a person, $\eta_{jk}, k = 1,\ldots,K$ are correlated. The null hypothesis that the genetic marker is not associated with any phenotype component corresponds to $H_0 : \beta_1 =,\ldots,\beta_K = 0$. The estimation of variance parameters and fixed effect parameters can be obtained using restricted maximum likelihood method (REML) [8, 9].

When y_{jk} is categorical, it can be modeled with generalized mixed effects model (GLMM) as follows:

$$E(y_{jk} \mid \eta_k) = \mu^{-1}(\beta_0 + \beta_k X(g_j) + \eta_{jk}), \tag{2.2}$$

where μ is a link function and μ^{-1} is its inverse. For Gaussian distributed traits, μ is the identity link, thus (2.2) is identical to the linear mixed effects model (2.1); for binary traits, μ is the logit link $\mu(x) = \ln(x/1-x)$. For links other than identity function, the likelihood for this model contains integrals without a close form solution. All existing algorithms for likelihood maximization are either based on theoretical or numerical approximation [10, 11].

The null hypothesis under the LME or GLMM can be tested using the likelihood ratio test or Wald chi-squared test. They can be implemented using SAS PROC Mixed or *R* lme4 package function *lmer()*. The Wald chi-squared test statistic takes the form $\boldsymbol{\beta}^T \mathrm{cov}(\boldsymbol{\beta})^{-1}\boldsymbol{\beta} \sim \chi_K^2$, where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)$ is estimated using (2.1) or (2.2). For example, Kraja et al. [12] have employed a model similar to (2.1) to the analyses of bivariate continuous metabolic traits. We can also fit a model assuming $\beta_1 = \cdots = \beta_K = \beta$, that is, $E(y_{jk} \mid \eta_k) = \mu^{-1}(\beta_0 + \beta X(g_j) + \eta_{jk})$, where a single degree-of-freedom (df) test $\widehat{\beta}/\mathrm{se}(\widehat{\beta})$ can be used to test the null hypothesis. This test can be more powerful than the multi-df Wald chi-squared test if the effect sizes are in the same direction and not very different. It, however, may lack power if the $\beta_1,\ldots,\beta_K$ are very different, especially have different signs and cancel each other out.

2.1.2. Frailty Models

When the phenotypes are correlated survival times, frailty models can be used to fit the association model. Suppose the survival or censoring times are t_{kj} for the kth $(k = 1,\ldots,K)$ phenotype of the jth $(j = 1,\ldots,J)$ individual. Let g_j be the genotype of a genetic marker of the jth individual, and $X(g_j)$ a score of the genotype as follows:

$$h(t_{kj}; X(g_j)) = h_0(t_{kj}) \exp(\beta_0 + \beta X(g_j) + \eta_{kj}), \tag{2.3}$$

where $\eta_{kj}(j = 1,\ldots,J)$ are subject specific random effects following $N(0,\Sigma)$, and Σ is a K-dimensional correlation matrix. This is the Gaussian frailty model. There is another class of frailty models where $\exp(\eta_{kj})$ follows a gamma distribution. A Gaussian or gamma frailty

model assuming an exchangeable correlation within a person can be fitted using *coxph()* in the survival package of *R* by including a *frailty()* term in the regressor. In addition, including a *cluster()* term in *coxph()* fits generalized estimating equations (GEE) type of model that assumes an independent working correlation matrix [13]. Frailty models with an arbitrary prespecified Σ can be fitted with the *coxme()* in *R* coxme package for Gaussian random effects model.

Fitting a mixed effects (frailty) model requires predetermining the correlation matrix Σ of random effects η_{jk} within jth person. The correlation between the phenotypes y_{jk} within a person is attributable to the random effects η_{jk} and the fixed effects of the genetic marker. However, since the fixed effects are unknown, it is impossible to directly infer the correlation among the random effects. Misspecifying the correlation among random effects may result in bias in the inference on fixed effects. But the bias seems to be small for genetic association studies [14, 15].

2.1.3. Generalized Estimating Equations

Different from mixed effects model is a class of models called marginal models. Instead of having random effects as regressors in addition to random errors to model correlation in multivariable response, marginal models collapse the random effects and random residual errors in the model. Generalized estimating equations (GEE) [16] solve the quasi-likelihood score function as follows:

$$\sum_{j=1}^{n} \left(\frac{\partial \mu_j}{\partial \boldsymbol{\beta}} \right)^t V_j^{-1} (Y_j - \mu_j) = 0, \tag{2.4}$$

where $V_j = A_j^{1/2} R(\alpha) A_j^{1/2}$, and $R(\alpha)$ is the working correlation matrix for the residual correlation. The variance and covariance of $\boldsymbol{\beta}$ is estimated with the so-called robust variance estimator [16]. Similar to the LME, single- or multi-df Wald test statistic can be usually used to test that the genetic marker is not associated with any of the phenotypes.

In our experience, GEE results are inflated with low minor frequency SNPs and not as powerful as LME in general [15, 17]. However, GEE is robust to misspecification of response distribution or association model and thus can be used when the LME shows bias or inflation due to these reasons.

2.2. Variable Reduction Method

Variable reduction approaches are in general only applicable to MV phenotype consisting of all continuous phenotypes that are approximately normal distributed. It derives a single or a few new phenotypes that are linear combinations of the original phenotypes, for example,

$$\widetilde{Y} = a_1 Y_1 + a_2 Y_2 + \cdots + a_K Y_K. \tag{2.5}$$

Existing methods include principal components analysis (PCA) where for the first component, $a_i, i = 1, \ldots, K$ are coefficients that maximize the variance of $\widetilde{Y}$; principal component of heritability (PCH) with coefficients maximizing the total heritability of $\widetilde{Y}$ [18]

and penalized PCH applicable to high-dimensional data [19, 20]; and principal components of heritability with coefficients maximizing the quantitative trait locus (QTL) heritability (PCQH) of $\widetilde{Y}$ [21–24], that is, the variance explained by the genetic marker. The PCQH approaches are designed to maximize the individual phenotype variation explained by the genetic marker and thus may be more powerful than PCA and PCH in genetic association studies.

2.2.1. PCQH Approaches

The approaches proposed by Lange et al. [21, 25] and Klei et al. [23] involve using a subset of the sample to estimate the coefficients in (2.5) that maximize the correlation between $\widetilde{Y}$ and the genetic marker. Specifically, in the estimation sample, the total phenotype variance is partitioned into QTL variance and residual variance as follows:

$$V_p = V_q + V_\varepsilon, \tag{2.6}$$

where V_p is the $K \times K$ total phenotype variance-covariance matrix, V_q the QTL variance matrix, and V_ε the residual variance matrix. Let $A = (\alpha_1, \ldots, \alpha_K)$, then the variance of $\widetilde{Y} = A^t Y$ explained by the genetic marker is

$$h_A^2 = \frac{A^t V_q A}{A^t V_p A}. \tag{2.7}$$

A that maximizes h_A^2 can be obtained by solving the following generalized eigen system [18]:

$$V_q A = \lambda V_p A. \tag{2.8}$$

$V_q = \mathrm{var}(\beta_1 X, \ldots, \beta_K X)$ can be approximated by $\Gamma 1 1^t \Gamma$, where $\Gamma = \mathrm{diag}(|\beta_1|\sigma_x, \ldots, |\beta_K|\sigma_x)$, σ_x is the sample standard deviation of the score of genotype $X(g)$ across all individuals, β_i is estimated using the least squared estimator of $Y_i = \alpha + \beta_i X(g) + \varepsilon$, and $1 = (\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_K))$.

Lange et al. [21, 25] approaches are only applicable to family-based association design. They suggest using the noninformative families or parental genotypes to estimate A because these data will not contribute directly to the family-based association tests (FBAT). Then perform FBAT of $\widetilde{Y}$ on $X(g)$. However, FBAT has low power in the absence of population stratification [26] compared to population based approaches. Klei et al's. [23] is a population-based association approach where they randomly split the sample into two subsets: one used to estimate A, the other used to test the association of $\widetilde{Y}$ with $X(g)$ via a linear regression model: $\widetilde{Y} = \alpha + \beta X(g) + \varepsilon$. This ensures valid P value in the association test.

2.2.2. Canonical Correlation Analysis

Canonical correlation analysis seeks coefficients so that the squared correlation between $\widetilde{Y}$ in (2.5), and the score of genetic marker, $X(g)$, is maximized. Here $\widehat{\rho} = \mathrm{corr}(\widetilde{Y}, X)$ is called

Table 1: Relationship between MANOVA test statistics and canonical correlation for association test of multivariate phenotype and a genetic marker.

MANOVA test	$f(\hat{\rho})$
Roy's largest root	$\hat{\rho}^2$
Hotelling-Lawley trace	$\hat{\rho}^2/1-\hat{\rho}^2$
Wilks lambda	$1-\hat{\rho}^2$
Pillai-Bartlett trace	$\hat{\rho}^2$

estimated canonical correlation. To obtain $\hat{\rho}$, the covariance matrix of Y and X is partitioned as follows:

$$\operatorname{cov}\begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}, \tag{2.9}$$

where Σ_{YY} is the $K \times K$ matrix of the variance-covariance matrix of Y, Σ_{YX} and its transpose Σ_{XY} are $K \times 1$ and $1 \times K$matrix of the covariance matrix between Y and X, Σ_{XX} is the variance of X, a scalar. All these submatrices can be estimated using the respective sample co-variance matrix. The canonical correlation, $\hat{\rho} = \Sigma_{XY}A/\left(A^t\Sigma_{YY}A\Sigma_{XX}\right)^{1/2}$, is solved as the squared root of the largest eigenvalue of $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$, and the corresponding eigenvector A contains the coefficients for constructing $\widetilde{Y}$. multivariate analysis of variance (MANOVA) tests correspond to evaluating canonical correlation. Table 1 details the relationship between $\hat{\rho}$ and commonly reported test statistics in MANOVA of a multivariate phenotype Y on $X(g)$ [27].

These tests are implemented in SAS PROC GLM and *R* function *summary.manova()*. As part of the PLINK package specifically developed for genetic analysis, Ferreira et al. [24] implemented the Wilks lambda, and its P value is obtained from F-approximation $F = (\hat{\rho}^2/K)/((1-\hat{\rho}^2)/(n-K-1))$.

Canonical correlation analysis shares similarity with PCQH [23] in that both estimate a linear combination of original phenotypes, so that the genotype score explains most of the variation (in terms of percent of total variance and squared correlation, resp.) of the new phenotype. The difference between the two approaches is that the canonical correlation analysis evaluates squared correlation using whole sample, while PCQH estimates the loadings using a subset of the sample and test the association in the rest of the sample. Extensive simulation studies performed in [28]. The author of [28] showed that MANOVA via Wilk's lambda was substantially more powerful than PCQH [23] with $K = 5$ phenotypes.

2.3. Combining Test Statistics from Univariate Analysis

An alternative way to analyze multivariate phenotypes is to perform univariate phenotype-genotype association test for each phenotype individually and then combine the test statistics from the univariate analysis. The advantage of such approach is the simplicity, that is, the methods to deal with univariate phenotypes are generally simpler than methods for MV phenotypes. It is especially useful for analyzing multivariate phenotype consisting of components of different types of distributions such as continuous, dichotomous, and survival. Regression methods for analyzing such multivariate phenotype are generally

complicated and not trivial to implement for MV phenotype with dimension > 2, see for example, [29, 30].

In recent years, researchers have generated large amount of univariate GWAS results for a variety of complex traits. Methods that combine the univariate results of multiple traits to detect genetic markers associated with multiple phenotypes are appealing.

2.3.1. Methods for Homogeneous Genetic Effects across Phenotypes

Assume that $\mathbf{T} = [T_1, T_2, \ldots, T_K]^T$ is a vector of K test statistics obtained from association analyses of each individual component phenotype against the genetic marker. Assume that $\mathbf{T}$ follows a multivariate normal distribution with mean $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_K)^T$ and a nonsingular covariance matrix $\boldsymbol{\Sigma}$. For example, $\mathbf{T}$ can be the β coefficients from least squared regression model for individual components or the t-test statistics from the regression models. The null hypothesis of no association to any phenotypes is H_0: $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_K)^T = \mathbf{0}$. O'Brien [31–33] suggested the following linear combination of $T_1, T_2, \ldots, T_K$, with weight $\mathbf{e} = (1, 1, \ldots, 1)^T$ of length K:

$$S = \mathbf{e}^T \Sigma^{-1} \mathbf{T} \tag{2.10}$$

when $\tau_1 = \tau_2 = \cdots = \tau_K \neq 0$ (2.10) is the most powerful test among a class of tests statistics that are linear combinations of $T_1, T_2, \ldots, T_K$. Under the null hypothesis, S follows the normal distribution with mean 0 and variance $\mathbf{e}^T \Sigma^{-1} \mathbf{e}$. To estimate $\boldsymbol{\Sigma}$ with GWAS results, Yang et al. [34] suggested using the sample covariance matrix of the statistics on a large number of SNPs genomewide with little or no linkage disequilibrium among them (say HapMap $r^2 < 0.1$).

The power of O'Brien's method depends on the assumption $\tau_1 = \tau_2 = \cdots = \tau_K$. When the means are very different or with opposite signs, O'Brien's method may not be efficient. Yang et al. proposed a sample splitting approach that replaces the uniform weight $\mathbf{e}^T$ by weights $\mathbf{w}$ estimated using a portion of the sample and only used the remaining sample to estimate $\mathbf{T}$ in (2.10), that is, $S = \mathbf{w}^T \Sigma^{-1} \mathbf{T}$. To overcome the variability introduced by a random sample splitting, Yang et al. also evaluated a cross-validation approach that averages the test statistics of 10 random splitting samples. The results showed that when $\tau_1, \tau_2, \ldots, \tau_K$ are of different magnitude or in opposite directions, O'Brien's method is less powerful than Yang et al., which indicates room for improvements for O'Brien's method. However, the sample splitting and cross-validation methods are less powerful than O'Brien's method with homogeneous effect sizes.

2.3.2. Methods for Heterogeneous Genetic Effects across Phenotypes

The limitation of O'Brien statistic is that it is not powerful for heterogeneous effects across multiple phenotypes, especially if some effects are of opposite directions. Another class of statistics that takes a quadratic form of the vector of the individual association statistic may overcome the limitation. For example, the following Wald chi-squared type test statistic was mentioned in Xu et al. [32].

$$S_w = \mathbf{T}^T \Sigma^{-1} \mathbf{T}. \tag{2.11}$$

The difference between (2.10) and (2.11) is that the vector $\mathbf{e} = (1, 1, \ldots, 1)^T$ is replaced by the $\mathbf{T}$ in (2.11). S follows a chi-squared distribution with degree of freedom equal to the number of the phenotypes K or rank of Σ if it is not full rank. Due to the "curse of dimensionality," power of (2.11) is diminishing with the increased number of phenotypes. Similar problem has been extensively studied and discussed in high-dimensional data analysis field and most recently in the analyses of multiple rare variants. Borrowing ideas from these fields, we propose the following test statistic that may be more powerful than (2.10) and (2.11) with heterogeneous effects.

$$S_{\text{sq}} = \mathbf{T}^T\mathbf{T} = \sum_{i=1}^{K} t_i^2. \tag{2.12}$$

The difference between (2.12) and (2.11) is that there is no variance-covariance matrix in (2.12). This statistic was first proposed by Pan [35] to analyze multiple rare or common variants against a single phenotype, where the t_i is the beta coefficient for the ith genetic variant. Different from Pan [35], here t_i is the association statistic for the ith phenotype with a single marker. Based on the groundwork of Zhang [36], Pan [35] pointed out that the distribution of (2.11) is a mixture of single degree-of-freedom chi-squared variates, $\sum_{i=1}^{K} c_i\chi_1^2$ where c_is are the eigen values of Σ, that is, the variance-covariate matrix of t_i. The distribution of (2.12) can be well approximated by $a\chi_d^2 + b$ with

$$a = \frac{\sum_{i=1}^{K} c_i^3}{\sum_{i=1}^{K} c_i^2}, \qquad b = \sum_{i=1}^{K} c_i - \frac{\left(\sum_{i=1}^{K} c_i^2\right)^2}{\sum_{i=1}^{K} c_i^3}, \qquad d = \frac{\left(\sum_{i=1}^{K} c_i^2\right)^3}{\left(\sum_{i=1}^{K} c_i^3\right)^2}. \tag{2.13}$$

The P value is calculated as $p(\chi_d^2 > (S_{\text{Sq}} - b)/a)$. The degree of freedom of the S_{sq} may be less than K with highly correlated phenotypes. In addition, (2.12) does not have the problem of instability observed for (2.10) and (2.11) when some of the components are highly correlated (in one of our applications, a correlation ~0.7 has resulted in inflated results for (2.10) and (2.11)). We have developed an R package CUMP (combining univariate results of multivariate phenotypes) that have implemented all the aforementioned combining statistics approaches. The software can be downloaded at (http://people.bu.edu/qyang/), and a short report of this software is submitted [37].

3. Identifying Pleiotropy

All the aforementioned methods can be used to detect association that is potentially due to pleiotropy. But they do not answer the question if the detected association is truly pleiotropy, that is, the marker locus affects all components of the MV phenotype directly. The detect association can affect some of the phenotypes and/or mediate through these phenotypes to affect the other phenotypes. Vansteelandt et al. [38] illustrated potential confounding mechanism between the genotype of a genetic marker and a phenotype using a causal diagram (Figure 1): the association between the genotype, denoted as G, and the response phenotype Y can occur through the paths connecting the two variables along all unbroken sequences of edges regardless of the direction of the arrows, given that there are no colliders

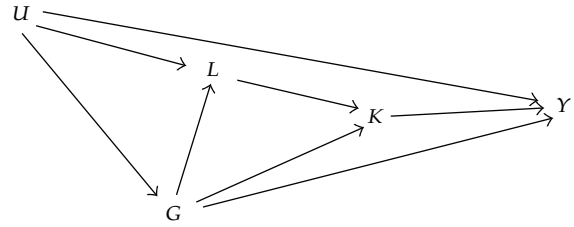


Figure 1: Causal diagram showing potential confounding mechanisms for the association between the genotype of a genetic marker G and the phenotype Y. The variable K denotes the intermediate phenotype, L the collection of known environmental and genetic risk factors, and U the unknown environmental and genetic risk factors such as population stratification and unknown genetic variants in linkage disequilibrium with G.

(i.e., variables in which two arrows converge, e.g., variables K and L in Figure 1) in the sequence [39].

The genotype G may be associated with Y due to (1) direct causal effect, that is, $G \rightarrow Y$; (2) through intermediate phenotype or risk factors, that is, $G \rightarrow K \rightarrow Y$or $G \rightarrow L \rightarrow K \rightarrow Y$; (3) because of confounding factors, that is, $G \leftarrow U \rightarrow Y$ or $G \leftarrow U \rightarrow L \rightarrow K \rightarrow Y$.

The authors showed that two commonly used approaches to detangle the complex relationship between phenotypes, genotype, and traditional risk factors are flawed. The first commonly used approach derives the residuals of Y regressing on K, say $\widetilde{Y} = Y - \beta K$, and then the association between G and $\widetilde{Y}$ is tested. The disadvantage of this approach is that not only the direct causal effect of K on Y is removed but also any indirect effect of K on Y through G (e.g., $K \leftarrow G \rightarrow Y$ and $Y \leftarrow U \leftarrow L \leftarrow G \rightarrow K$) and other factors (e.g., $K \leftarrow L \rightarrow U \rightarrow Y$). Therefore β may be biased in the presence of confounding factors which leads to biased test of G with $\widetilde{Y}$.

The second commonly used approach tests the direct effect of G on Y in a regression model including K and L as covariates. Adjustment of K removes the relationship between G and Y through $G \rightarrow K \rightarrow Y$; however, because K is a collider (Figure 1), the adjustment of K induces a spurious association [39, 40] along the path $G \rightarrow K \rightarrow L \leftarrow U \rightarrow Y$. Additionally, adjusting for L induces spurious associated through the path $G \rightarrow L \leftarrow U \rightarrow Y$.

To overcome the limitation of the two commonly adapted approaches, Vansteelandt et al. [38] proposed a least squared regression model to estimate the direct effect size of K on Y. This regression model includes the suspected intermediate phenotype, the score of the genetic marker genotype, $X(G)$, and other common risk factors between the two phenotypes as regressors:

$$E(Y_i) = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 L_i. \tag{3.1}$$

The estimated effect size of the phenotype represents the direct effect of the K on Y, that is, not confounded by the effect of X mediated through any of the covariates. Then, a new phenotype is created as the residual of the response subtract the effect of K only $\widetilde{Y}_i = Y_i - \overline{y} - \widehat{\gamma}_1(K_i - \overline{k})$. Then, whether the G only exerts its effect on Y through K can be tested using any standard association test statistic between the residual and the X. A negative result indicates that G only exerts its effect on Y through K while a positive result indicates that the G has a direct

effect on Y and/or a spurious effect through other confounders. Extensions of the method to dichotomous and time-to-event outcomes have been proposed [41, 42].

4. Discussion

In this paper, we reviewed methods available for joint analyzing correlated phenotypes in genetic association studies. Some of these methods are designed to detect potential association with multiple phenotypes (pleiotropy), while the others are designed to test whether the detected association with the MV phenotype is truly pleiotropy or the genetic marker exerts its effects on some phenotypes through affecting the others.

For methods designed to detect association, each method has its own pros and cons. Random effects model requires knowledge of residual correlation, and misspecifying the correlation may incur inflation or power loss. Generalized estimating equations are robust to misspecification of residual correlation, but it is inflated for low-frequency variants and less powerful than random effects model in our experience. Variable reduction approaches are appealing because correlated outcomes are reduced to a single or fewer number of uncorrelated outcomes. However, in the presence of missing data in the outcomes, individuals with missing data do not contribute to the analysis, which may result in power loss. The approaches combining univariate association results are more flexible than the other methods especially when MV phenotypes consist of a mixture of continuous, discrete, and/or time-to-events data. Regression approaches have been developed to deal with such phenotypes. But they are generally complicated and few available software implements these methods. Since univariate association results are used, individuals with incomplete observations still contribute to the analysis of available phenotypes. Simulations on all continuous phenotypes indicated that the power of O'Brien's method, one of the approaches combining univariate association results is similar to regression and variable reduction methods when the effects size are similar across multiple phenotypes [34].

All the approaches introduced here for population based approaches assume unrelated individuals. When there are related individuals in the data, not accounting for family structure can result in inflation or power loss. Extension of introduced methods to account for family data are possible. For example, one may add a random effect in mixed effects model to account for family structure. For approaches combining univariate association results, a model that account for family structure need be used in the univariate analyses.

In terms of computational cost, mixed effects models may be most time consuming since maximization of likelihood is required.

Finally, it has been shown that traditional causal inference is useful in distinguishing true pleiotropy from other mechanisms that also result in genetic association with multiple phenotypes. A related causal inference in recent genetic literature is Mendelian randomization test [43–45]. This approach can be used to infer whether an intermediate phenotype has a causal effect on an outcome phenotype, using genetic marker(s) in association with the intermediate phenotype. Unlike a phenotype that is subject to the influence of uncontrolled environmental factors and/or reverse causation of another phenotype, genotype(s) of genetic marker(s) is(are) free of influence of environmental factors and reverse causation. For this approach, marker genotype(s) is(are) used as an instrument variable. This test requires that there is no pleiotropy effect of the genetic marker on outcome phenotype. Association of the genotype and outcome phenotype indicates that the intermediate phenotype may causally affect the outcome phenotype.

5. URLs to Software Mentioned in This Paper

SAS: http://www.sas.com/,

R: http://www.r-project.org/,

CUMP: http://cran.r-project.org/web/packages/CUMP/index.html,

coxme: http://cran.r-project.org/web/packages/coxme/index.html,

gee: http://cran.r-project.org/web/packages/gee/index.html,

survival: http://cran.r-project.org/web/packages/survival/index.html,

lme4: http://cran.r-project.org/web/packages/lme4/index.html,

PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/.

Acknowledgments

Q. Yang's work is supported by the National Heart, Lung, and Blood Institute's Framingham Heart Study (Contract no. N01-HC-25195) and Grant no. R01HL093328 and R01HL093029. Y. Wang's work is supported by NIH Grants nos. R03AG031113-01A2 and 1R01NS073671-01 1.

References

[1] International HapMap Consortium, K. A. Frazer, D. G. Ballinger et al., "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851–861, 2007.

[2] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[3] L. A. Hindorff, H. A. Junkins, P. N. Hall, J. P. Mehta, and T. A. Manolio, "A catalog of published genome-wide association studies," *National Human Genome Research Institute*, 2011, http://www.genome.gov/gwastudies/.

[4] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[5] E. E. Eichler, J. Flint, G. Gibson et al., "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[6] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[7] G. M. Fitzmaurice and N. M. Laird, "A likelihood-based method for analysing longitudinal binary responses," *Biometrika*, vol. 80, no. 1, pp. 141–151, 1993.

[8] H. D. Patterson and R. Thompson, "Recovery of inter-block information when block sizes are unequal," *Biometrika*, vol. 58, pp. 545–554, 1971.

[9] D. A. Harville, "Maximum likelihood approaches to variance component estimation and to related problems," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 320–340, 1977.

[10] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, pp. 9–25, 1993.

[11] D. M. Bates and S. DebRoy, "Linear mixed models and penalized least squares," *Journal of Multivariate Analysis*, vol. 91, no. 1, pp. 1–17, 2004.

[12] A. T. Kraja, D. Vaidya, J. S. Pankow et al., "A bivariate genome-wide approach to metabolic syndrome: STAMPEED Consortium," *Diabetes*, vol. 60, no. 4, pp. 1329–1339, 2011.

[13] T. M. Therneau, P. M. Grambsch, and V. S. Pankratz, "Penalized survival models and frailty," *Journal of Computational and Graphical Statistics*, vol. 12, no. 1, pp. 156–175, 2003.

[14] R. M. Pfeiffer, A. Hildesheim, M. H. Gail et al., "Robustness of inference on measured covariates to misspecification of genetic random effects in family studies," *Genetic Epidemiology*, vol. 24, no. 1, pp. 14–23, 2003.

[15] M. H. Chen, X. Liu, F. Wei et al., "A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees," *Genetic Epidemiology*, vol. 35, no. 7, pp. 650–657, 2011.
[16] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.
[17] L. A. Cupples, H. T. Arruda, E. J. Benjamin et al., "The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports," *BMC Medical Genetics*, vol. 8, supplement 1, 2007.
[18] J. Ott and D. Rabinowitz, "A principal-components approach based on heritability for combining phenotype information," *Human Heredity*, vol. 49, no. 2, pp. 106–111, 1999.
[19] Y. Wang, Y. Fang, and M. Jin, "A ridge penalized principal-components approach based on heritability for high-dimensional data," *Human Heredity*, vol. 64, no. 3, pp. 182–191, 2007.
[20] Y. . Wang, Y. Fang, and S. Wang, "Clustering and principal-components approach based on heritability for mapping multiple gene expressions," *BMC Proceedings*, vol. 1, supplement 1, p. S121, 2007.
[21] C. Lange, K. van Steen, T. Andrew et al., "A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1544–6115, 2004.
[22] C. Lange, D. L. DeMeo, and N. M. Laird, "Power and design considerations for a general class of family-based association tests: quantitative traits," *American Journal of Human Genetics*, vol. 71, no. 6, pp. 1330–1341, 2002.
[23] L. Klei, D. Luca, B. Devlin, and K. Roeder, "Pleiotropy and principal components of heritability combine to increase power for association analysis," *Genetic Epidemiology*, vol. 32, no. 1, pp. 9–19, 2008.
[24] M. A. R. Ferreira and S. M. Purcell, "A multivariate test of association," *Bioinformatics*, vol. 25, no. 1, pp. 132–133, 2009.
[25] C. Lange, E. K. Silverman, X. Xu, S. T. Weiss, and N. M. Laird, "A multivariate family-based association test using generalized estimating equations: FBAT-GEE," *Biostatistics*, vol. 4, no. 2, pp. 195–206, 2003.
[26] Y. S. Aulchenko, D. J. de Koning, and C. Haley, "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis," *Genetics*, vol. 177, no. 1, pp. 577–585, 2007.
[27] K. E. Muller and B. L. Peterson, "Practical methods for computing power in testing the multivariate general linear hypothesis," *Computational Statistics and Data Analysis*, vol. 2, no. 2, pp. 143–158, 1984.
[28] H. Wu, *Methods for genetic association studies using longitudinal and multivariate phenotypes in families [Ph.D. thesis]*, Boston University, Boston, Mass, USA, 2009.
[29] G. M. Fitzmaurice and N. M. Laird, "Regression models for mixed discrete and continuous responses with potentially missing values," *Biometrics*, vol. 53, no. 1, pp. 110–122, 1997.
[30] J. Liu, Y. Pei, C. J. Papasian, and H. W. Deng, "Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations," *Genetic Epidemiology*, vol. 33, no. 3, pp. 217–227, 2009.
[31] P. C. O'Brien, "Procedures for comparing samples with multiple endpoints," *Biometrics*, vol. 40, no. 4, pp. 1079–1087, 1984.
[32] X. Xu, L. Tian, and L. J. Wei, "Combining dependent tests for linkage or association across multiple phenotypic traits," *Biostatistics*, vol. 4, no. 2, pp. 223–229, 2003.
[33] L. J. Wei and W. E. Johnson, "Combining dependent tests with incomplete repeated measurements," *Biometrika*, vol. 72, no. 2, pp. 359–364, 1985.
[34] Q. Yang, H. Wu, C. Y. Guo, and C. S. Fox, "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests," *Genetic Epidemiology*, vol. 34, no. 5, pp. 444–454, 2010.
[35] W. Pan, "Asymptotic tests of association with multiple SNPs in linkage disequilibrium," *Genetic Epidemiology*, vol. 33, no. 6, pp. 497–507, 2009.
[36] J.-T. Zhang, "Approximate and asymptotic distributions of chi-squared-type mixtures with applications," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 273–285, 2005.
[37] X. Liu and Q. Yang, "CUMP: an R package for analyzing multivariate phenotypes in genetic association studies".
[38] S. Vansteelandt, S. Goetgeluk, S. Lutz et al., "On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects," *Genetic Epidemiology*, vol. 33, no. 5, pp. 394–405, 2009.

[39] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.
[40] J. M. Robins, "Data, design, and background knowledge in etiologic inference," *Epidemiology*, vol. 12, no. 3, pp. 313–320, 2001.
[41] P. J. Lipman, K. Y. Liu, J. D. Muehlschlegel, S. Body, and C. Lange, "Inferring genetic causal effects on survival data with associated endo-phenotypes," *Genetic Epidemiology*, vol. 35, no. 2, pp. 119–124, 2011.
[42] S. Vansteelandt, "Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models," *Biometrika*, vol. 97, no. 4, pp. 921–934, 2010.
[43] G. D. Smith and S. Ebrahim, "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology*, vol. 32, no. 1, pp. 1–22, 2003.
[44] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith, "Mendelian randomization: using genes as instruments for making causal inferences in epidemiology," *Statistics in Medicine*, vol. 27, no. 8, pp. 1133–1163, 2008.
[45] P. M. McKeigue, H. Campbell, S. Wild et al., "Bayesian methods for instrumental variable analysis with genetic instruments ("Mendelian randomization"): example with urate transporter SLC2A9 as an instrumental variable for effect of urate levels on metabolic syndrome," *International Journal of Epidemiology*, vol. 39, no. 3, pp. 907–918, 2010.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 485174, 16 pages
doi:10.1155/2012/485174

Review Article

Mixed Modeling with Whole Genome Data

Jing Hua Zhao and Jian'an Luan

MRC Epidemiology Unit & Institute of Metabolic Science, Addrenbrooke's Hospital, Box 285, Hills Road, Cambridge CB2 0QQ, UK

Correspondence should be addressed to Jing Hua Zhao, jinghua.zhao@mrc-epid.cam.ac.uk

Received 2 March 2012; Accepted 20 April 2012

Academic Editor: Yongzhao Shao

Objective. We consider the need for a modeling framework for related individuals and various sources of variations. The relationships could either be among relatives in families or among unrelated individuals in a general population with cryptic relatedness; both could be refined or derived with whole genome data. As with variations they can include oliogogenes, polygenes, single nucleotide polymorphism (SNP), and covariates. *Methods*. We describe mixed models as a coherent theoretical framework to accommodate correlations for various types of outcomes in relation to many sources of variations. The framework also extends to consortium meta-analysis involving both population-based and family-based studies. *Results*. Through examples we show that the framework can be furnished with general statistical packages whose great advantage lies in simplicity and exibility to study both genetic and environmental effects. Areas which require further work are also indicated. *Conclusion*. Mixed models will play an important role in practical analysis of data on both families and unrelated individuals when whole genome information is available.

1. Introduction

Genomewide association studies (GWASs) have successfully identified many genetic variants consistently associated with human diseases or other traits. Both unrelated individuals in a population or related individuals in families have been involved in such studies. There is a variety of issues which merit further consideration.

Our concern here is on correlations among individuals, which are "the central piece of information" [1] in detection and characterization of gene-trait association. Consideration of these correlations has traditionally limited to family data whose critical role in genetic epidemiological study ranges from familial aggregation, segregation, linkage to association [2], and special attention is required in the analysis compared to unrelated individuals from a population. Correlations arise naturally among relatives but can be relevant to

population-based study as well given that relatedness can also be established among unrelated individuals based on whole-genome data in GWASs [3]. The correlations are linked to a long attempt to model influence of multiple genes on a specific phenotype. Specifically, Fisher [4] assumed that a quantitative trait results from many genes with variable small to moderate effects. Concrete evidence of multiple genetic influence has been revealed by recent waves of GWASs on height [5], blood pressure [6], lipids [7], obesity [8], and so forth, leading to the note in [9]. Gene-environment interaction and common environment can be considered similarly.

There is a relativelysmall literature in human genetics to iterate mixed models to account for heterogeneity among groups of individuals compared to the general statistics literature where genetic applications have been acknowledged [10, pages 190–192] and [11, pages 4864–4871]. This is likely due to the complexity with a generic implementation. We therefore conduct a survey of the framework with exploration of general software environments. As will be seen below, it readily applies to human genetics when correlations within these groups are explicitly modeled. The familiar form accommodate effects of major or oligogenes, polygenes, common environment, and unique environment, which collectively contribute to variance of the trait and known as "variance component models" [12, 13]. For instance, individuals' body weight (kg) divided by height^2 (m^2), referred as body mass index (BMI, kg/m^2) and commonly used as surrogate of obesity, varies with the broad heritable background of individuals (polygenes), sex, age, family membership, susceptible genes such as *FTO* [14] (which has a major effect serving an example of oliogene), where sex and age can be considered as fixed effects while variability attributable to (expected) correlation between members of family as with *FTO* are random effects [15–18]. The flexibility of such a framework may be missing in various computer programs (see http://linkage.rockefeller.edu). As for outcome of interest, it is usually quantitative or binary traits, with [19] as an exception. The implementation we consider will be *SAS* (see http://www.sas.com) [20] and *R* (see http://www.r-project.org) [21] with a Cox model counterpart [22]. A note on Bayesian counterpart is also ready [23–25], especially for linkage [1], association [26] and implementation in *Morgan*. To save space, we consequently omit reference to programs when they are available from the lists given here.

We attempt to connect various models in our survey paying special attention to their use in data analysis. We show that with generic facilities as available from *R*, we can accommodate additional outcomes such as count, survival, as well as account for information such as identity-by-descent (IBD) or common environment. We will illustrate with the family data available to genetic analysis workshops (GAWs) (see http://www.gaworkshop.org) 16 and 17. We will also discuss the implications of whole genome data availability via connection to earlier literature.

2. Models

As will soon become clear, the framework is essentially motivated from the usual general linear model (GLM) or generalized linear mixed model (GLMM) allowing for correlated random effects, including the Cox regression model. We will briefly describe the models as an analogy between GLM and GLMM but will not go into details of their estimation procedures, as both are widely available.

2.1. GLM

We start from the usual GLM disregarding familial correlations. Let the phenotypes of n individuals in a family be $(y_1, \ldots, y_n)$, its distribution is exponential

$$f(y_i, \theta_i, \varphi) = \exp\left[\frac{y_i\theta_i - b_i(\theta_i)}{\varphi} + c(y_i, \varphi)\right], \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, φ a scale or dispersion parameter. Furthermore, let $E[y_i] = \mu_i$ and let this be connected to a linear predictor using link function $g(\cdot)$ by $\eta_i = g(\mu_i) = X_i\beta$, where X_i is a vector of covariates and β the regression coefficient(s). For simplicity, only canonical link is used so that $\theta_i = \mu_i$. It can be shown [27] that the expectation $E(y_i) = \mu_i = b'(\theta_i)$ and variance $V(y_i) = \varphi b''(\theta_i)$. Some special cases as with their properties are well-recognized [28], for which models involving continuous and binary outcomes are most common.

Normal: $y_i \sim N(\mu_i, \sigma_i^2)$, we have $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $\varphi = \sigma_i^2$, $b'(\theta_i) = \theta_i$, $\varphi b''(\theta_i) = \sigma_i^2$ and an identity link.

Binomial: $y_i \sim \text{Binom}(n, \mu_i)$, $\theta(\mu_i) = \ln(\mu_i/(1-\mu_i))$, $b(\theta_i) = \ln(1+\exp(\theta_i))$, $\varphi = 1/n$, $b'(\theta_i) = \exp(\theta_i)/(1+\exp(\theta_i))$, $\varphi b''(\theta_i) = \mu_i(1-\mu_i)/n$, and a logit link $g(\mu_i) = \ln(\mu_i/(1-\mu_i))$.

Analysis of censored survival data can be molded into the framework [29]. Let t_i denote the event time, c_i the censoring time and $\delta_i = I(t_i \le c_i)$ the event indicator for unit $i, i = 1, \ldots, n$; the basic Cox model with vector of explanatory variables X_i is specified via a hazard function $\lambda_i(t) = \lambda_0(t)\exp(X_i\beta)$, where $\lambda_0(t)$ is the baseline hazard function. The partial likelihood (PL) for the standard Cox model can be expressed as follows:

$$\text{PL}(\beta) = \prod_{i=1}^{n}\left[\frac{\exp(X_i\beta)}{\sum_{j\in R(t_i)}\exp(X_j\beta)}\right]^{\delta_i}, \tag{2.2}$$

where n failure times have been ordered such that $t_1 < \cdots < t_n$ and $R(t_i)$ is the "risk set," the number of cases that are at risk of experiencing an event at time t_i.

Although GLM lays the foundation in many applications of general statistics, it largely serves a motivating role for models that are capable to account for familial correlations. As shown below, this is achieved with introduction of (correlated) random effects as in GLMM, but it is also linked with other models.

2.2. GLMM

We now consider model involving individual i, $i = 1, \ldots, N$, where N is the total number of individuals in our sample.

Polygene

Let P denote the polygene representing independent genes of small effect, which follows a multivariate normal distribution with covariance matrix

$$g(\mu_i) = X_i\beta + P_i. \tag{2.3}$$

The likelihood for all relatives is furnished with specification of the distribution of $P = (P_1, \ldots, P_N)$ with covariance

$$\Sigma_P = 2\Phi\sigma_P^2, \tag{2.4}$$

where $\Phi \equiv \{\phi_{ij}\}_{n\times n}$ and ϕ_{ij} is the kinship coefficient, defined such that, given two individuals, one with genes (g_i, g_j) and the other with genes (g_k, g_l), the quantity is $(1/4)(P(g_i \equiv g_k) + P(g_i \equiv g_l) + P(g_j \equiv g_k) + P(g_j \equiv g_l))$, where $\equiv$ represents probability that two genes sampled at random from each individual are IBD. The kinship coefficients for MZ twins, DZ twins/full-sibs, parent-offspring, half-sibs, and unrelated individuals are 0.5, 0.25, 0.25, 0.125, and 0, respectively.

The likelihood function for model (2.3) has the following form:

$$L(y_1, \ldots, y_N) = \int L(y \mid P)L(P)dP, \tag{2.5}$$

where $L(y \mid P) = \prod_{i=1}^{N} f(y_i \mid P)$ and $L(P) = \left(\sqrt{2\pi|\Sigma_P|}\right)^{-1} \exp[-P'\Sigma_P^{-1}P/2]$ only involve with random effects, noting that it is assumed that, given random effects in the model, the phenotypic values among n relatives are independent and that the parameters of interest in (2.4) are the variances involving polygene (σ_P^2). Regarding the statistical inference of random effects, since the parameter under the null hypothesis is on the boundary of the parameter space, the test for a specific $\sigma_k^2 = 0$, likelihood ratio statistic testing for the hypothesis that $H_0 : \sigma_P^2 = 0$ versus $H_A : \sigma_P^2 > 0$, is referred to a $0.5\chi_0^2 + 0.5\chi_1^2$ distribution or a score statistic as outlined in [11, 19, page 2961].

Oligogene

Suppose that a major gene M is also involved, independently and normally distributed with mean 0 and variances σ_M^2, then the covariance matrix has the form

$$\Sigma_M = \sigma_M^2\Pi, \tag{2.6}$$

where $\Pi \equiv \{\pi_{ij}\}_{N\times N}$ in which π_{ij} is the proportion of alleles shared (IBD) at the major gene between relatives i and j which can be estimated from a multipoint data, so that when it acts additively with polygene P, the likelihood is furnished with an extended covariance

$$\Sigma_{M,P} = \Sigma_M + \Sigma_P. \tag{2.7}$$

For a test of a strictly positive variance associated with a polygene versus polygene and an oligogene, the log likelihood ratio test statistic is referred to $0.5\chi_1^2 + 0.5\chi_2^2$ [30].

Multiple Random Effects

The framework in (2.3) includes the common distributions such as normal, gamma, binomial and Poisson as special cases. For simplicity, we consider a quantitative trait, whose probability density function is normal and a statistical model is as follows:

$$y = X\beta + U + \epsilon, \tag{2.8}$$

and $U \sim N(0, \Sigma)$, $\epsilon \sim N(0, \sigma^2)$, $\mathrm{Cov}(U, \epsilon) = 0$. The expression of Σ^{-1} relative to the precision $1/\sigma^2$ of ϵ as a Cholesky factorization $\Delta'\Delta$, that is, $\Sigma^{-1}/(1/\sigma^2) = \Delta'\Delta$ led to the term *relative precision factor* for Δ [31]. Note that the partition of effects as being fixed and random (H_A: genetic effect) can be compared to a sporadic model (H_0: no genetic effect) $y = X_1\beta_1 + X_2\beta_2 + e$, where both β_1 and β_2 are fixed effects, the involvement of Σ or more specifically Σ^{-1} as a "ridge factor" creates shrinkage in the random effects solutions to the normal equations, that is, "regression towards the mean."

We will see an example from the GAW17 data below that a quantitative trait Q1 is influenced by polygenic background and specific gene *VEGFC* as captured by kinship or relationship matrix and IBD matrix, respectively. This prompts the need to consider multiple random effects. We therefore pursue (2.8) further. As in [32], write $y = X\beta + Z_1 a_1 + \cdots + Z_k a_k + \epsilon$ with the usual assumption that y is $N \times 1$ vector of observations, X an $N \times p$ known matrix, not necessarily of full column rank, β a vector of fixed effects, Z_i a known $N \times r_i$ matrix of rank r_i, a_i random effects with $E(a_i) = 0$, $\mathrm{cov}(a_i) = \sigma_i^2 I_{r_i}, \mathrm{cov}(a_i, a_j) = 0, i \neq j, \mathrm{cov}(a_i, \epsilon) = 0, i, j = 1, \ldots, k$, ϵ an $N \times 1$ vector of errors with $E(\epsilon) = 0, \mathrm{cov}(\epsilon) = \sigma^2 I_N$. Then $E(y) = X\beta$ and $\mathrm{cov}(y) = \Sigma = \sigma^2 I_N + \sum_{j=1}^{k} \sigma_j^2 Z_j Z_j'$. This turns out to be critical to explore the covariance structure involving more (k) parameters $(\sigma_1^2, \ldots, \sigma_k^2)$ in the form

$$\sum\left(\sigma_1^2, \ldots, \sigma_k^2\right) = \sum_1\left(\sigma_1^2\right) + \cdots + \sum_k\left(\sigma_k^2\right), \tag{2.9}$$

where $\Sigma_i(\sigma_i^2)$ has the form of $\sigma_i^2 H_i$, $i = 1, \ldots, k$ with σ_i^2 being the unknown parameter and H_i a (known) coefficient matrix. It will also hold when different variance components such as multiple major genes of interest, gene-gene, gene-environment interactions, common shared environment are to be modeled. For significance test, Case 4 in [30] serves as a general guideline.

A closely related model is the so-called *marginal or population-average model* whereby familial relationship can be specified for e, namely, generalized estimating equations (GEEs) [12, 33]. Given $\mu_i = E(y)$, $V_i = \mathrm{Var}(y)$, it has the form

$$\sum_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(y_i - \mu_i) = 0, \tag{2.10}$$

for which only link function and variance need to be specified. Parameter estimates are consistent even when variance structure is misspecified, but the ability to use (2.9) is an apparent advantage.

We now turn to the Cox model. First, the consideration of an unobserved family specific random effect is often termed as frailty model, such that families with a larger value

of the frailty will experience the event at earlier times and most "frail" individuals will fail early [34]. Now we allow for correlated frailty and, in analogy to model (2.3) and [22], the appropriate model with random effect U_i becomes $\lambda_i(t) = \lambda_0(t)\exp(X_i\beta + U_i)$. Assuming the parameters of interest are β and σ^2 we have

$$\mathrm{PL}(\beta, U) = \prod_{i=1}^{N}\left[\frac{\exp(X_i\beta + U_i)}{\sum_{j\in R(t_i)}\exp(X_j\beta + U_i)}\right]^{\delta_i}. \tag{2.11}$$

The so-called integrated log likelihood is derived as

$$L = \int \mathrm{PL}(\beta, U)L(U)dU. \tag{2.12}$$

A more tractable solution is via a Laplace approximation for an approximate marginal log likelihood that can be maximized by a penalized partial likelihood (PPL) for parameters (β, σ^2), $\mathrm{PPL}(\beta, U) = \log(\mathrm{PL}(\beta, U)) - U^T\Sigma^{-1}U/2$, followed by a profile likelihood function involving only σ^2.

Furthermore, we can take advantage of the generic form of covariance in other types of models as well. A straightforward yet remarkably useful extension is the multivariate model. For instance, consider (2.8) with m phenotypes. Let $y = (y_{11}, \ldots, y_{1N}, \ldots, y_{mN})^T$ be a vector of m multivariate phenotypes for N individuals. Let β be a vector of dimension mp of the regression coefficients for the p covariates including a vector of 1's corresponding to the overall mean, $X = I_m \otimes X_{N,p}$, an $mN \times mp$ known matrix of covariate values. An analogy to (2.7) and (2.8) lead to the variance-covariance matrix of the m phenotypes with dimension $mN \times mN$ is

$$\Sigma = A \bigotimes \Pi + B \bigotimes R + C \bigotimes I, \tag{2.13}$$

where R is the $N \times N$ matrix of the coefficients of relationship, Π is an $N \times N$ matrix of estimated proportion of alleles IBD, and A, B, C are oligogenic, polygenic, and residual variance-covariance matrices each with dimension $m \times m$.

2.3. Meta-Analysis

One indispensable element in current GWASs is meta-analysis, typically involving findings from both unrelated individuals in a population and those from family data. While we have seen that mixed models are appropriate for a variety of traits in family-based association studies, broadly models for meta-analysis also fall into the same framework as described above. One can imagine a meta-analysis involving individual participant data (IPD). A good summary of approaches for IPD meta-analysis is available [35].

In the two-step approach, the individual participant data are first analysed in each separate study independently by using a statistical method appropriate for the type of data being analysed; for example, a linear regression model might be fitted for continuous responses such as blood pressure, or Cox regression might be applied for time to event data. (This step produces aggregate data for each study including effect estimate and its standard error). These data are then synthesised in the second step using a suitable model

for meta-analysis of aggregate data, such as one that weights studies by the inverse of the variance while assuming fixed or random effects across studies. In the one-step approach, the individual participant data from all studies are modelled simultaneously while accounting for the clustering of participants within studies. This approach again requires a model specific to the type of data being synthesised, alongside appropriate specification of the assumptions of the meta-analysis (e.g., of fixed or random effects across studies).

The two-step approach is the usual one used in various GWAS consortia while a one-step approach for all studies in our context could involve unrelated population-based samples and family data in the meta-model as long as the correlation structure is appropriately specified. The practicality of both approaches has been illustrated in the literature [36, 37] but, in view of the complexity involving in such a framework, and the practical difficulty that a researcher may not have access to individual data from all studies, we refrain ourselves from such a consideration for now but remain focusing on family data as illustrated with both simulated and real data.

2.4. Related Results and Implementations

There have been concerns in the literature regarding large number of units each with bounded size [38] and a large number of random effects [39]. In our context large number of families, each with bounded members, consistent estimate of the random effect is difficult to obtain though fixed effects and variance components will be consistent. However, Type I error rate and power have been explored before [19, 22, 26, 40], so there will be more on specific examples.

Instead of using purposely written programs, we chose to use *R*, for its wide availability and many other features [41], and in particular procedures to fit models described earlier are to a great extent available, including generic procedures from *nlme*, *lme4,* and *gee*, among others, but package designed for family data is *pedigreemm* with *lmekin* for linear mixed models available from *coxme*. We will also compare them to *SAS*, due to its ability to deal with large data, and great flexibility in model specification.

3. Examples

We consider two examples from GAWs 17 and 16, which involve simulated and real data widely available and allow for a lot of experiments to be done.

3.1. GAW17 Data

Data distributed by GAW17 were based on a collection of unrelated individuals and their genotypes were generated from the 1000 Genomes Project (see http://www.1000genomes.org/), from which a sample of 697 individuals in 8 extended families and their genotypes and phenotypes was available. A total of 202 founders in the family data set were chosen at random from the set of unrelated individuals. Replicates of the trait were generated 200 times, but the simulated genotypes remain constant over replicates. The traits made available were Q1, Q2, Q4, and AFFECTED (coded 0 = no 1 = yes) with covariates AGE and SMOKE. The variables describing family structures were ID, FA, MO, SEX (1 = men, 2 = women). Fully informative IBD information was available for 3205 genes.

We chose to examine traits Q1, Q2, and AFFECTED as representatives of quantitative and qualitative traits. According to [42], vascular endothelial growth factor (*VEGF*) pathway was enriched and here vascular endothelial growth factor C (*VEGFC* (see http://en.wikipedia.org/wiki/Vascular_endothelial_growth_factor_C)) was chosen as a causal variant associated with Q1 but not Q2. Q1 also increased with age, and the fact that AFFECTED is a function of Q1 offers the possibility to furnish a logistic regression model and explore age at onset via a Cox model. For illustration, we used age as surrogate for age onset. Being aware of the fact that this was only an approximation, whenever multipe affected individuals within a sibship are available, their average age was used. Causal variants and associate genes provide information on power of association testing statistics while the noncausal counterparts provide analogous results on Type I error rate.

The statistical significance was assessed according to log likelihood ratio tests between models using relationship only versus using both relationship and IBD information. The computation for this is relatively fast; results for all 200 replicates took 1 hour and 48 minutes on our 20-node Linux clusters each with 16 GB RAM and 4 CPUs using Sun grid engines. The nominal significance levels are shown in Table 1, which reveal that the tests are both close to the expected levels under H_0 and H_A.

Gene-based analysis was also conducted for Q1 involving all 3205 genes and the results are shown with selected candidates highlighted in Figure 1, which agree with the simulated model in which the significant regions were in VEGFC/VEGFA.

As one would be keen to see various parameter estimates in a real analysis, we also provide results associated with replicate one. Q1 as based on restricted maximum likelihood (REML) is shown in Table 2. The models with relationship only and with both relationship and IBD information have −2 Res(tricted) log likelihood being 1789.5 and 1775.2, respectively while Akaike Information Criteria (AIC) being 1793.5 and 1781.2, respectively so that using IBD information improved fit for Q1 (smaller AIC). For AFFECTED the results based on maximum pseudolikelihood are shown in Table 3 and those from Cox model in Table 4. Note that the improvement in terms of −2 log pseudolikelihood from 3434.4 to 3445.7 was also substantial. To explore the multivariate model (2.13) involving the polygenic effects for Q1, Q2, and Q4, the six parameters (σ_{11}, σ_{21}, σ_{22}, σ_{31}, σ_{32}, σ_{33}) in the variance-covariance matrix have been expressed according to (2.9). The appropriate matrices associated with all parameters are constructed a priori. These are then subject to procedures such as PROC MIXED and *lmekin*. The joint model of Q1, Q2, Q4 is shown in Table 5.

The implementations are provided in Supplementary information available online at doi: 10.1155/2012/485174. While code blocks shown there are appropriate for one instance, it is preferable to use *SAS*'s output delivery system (ODS) to save various results into databases.

3.2. The Framingham Heart Study

The Framingham Heart Study is under the direction of National Heart, Lung, and Blood Institute (NHLBI) which began in 1948 with the recruitment of adults from the town of Framingham, Massachusetts. Data available for GAW16 were 7130 individuals from the original cohort (373), the first generation cohort (2760), and the third generation cohort (3997) with sex, age, height, weight, blood pressure, lipids, smoking, and drinking. Data as outlined in [43] was used here, where 6848 had genotype data for at least one of the four specified SNPs (rs1121980, rs9939609, rs17782313, and rs17700633). Data for 96 individuals without

any phenotype data but with genotype data and an additional 227 individuals without being assigned a family ID were excluded from analyses. Additionally, four individuals had no data on weight; 86 observations were measured at <18 years of age, and therefore were excluded. The 6,520 remaining individuals were part of 962 families, among which 2073 individuals had completed four visits. Meanwhile, there were also 365 cases of diabetes with their ages of onset.

Kinship information was obtained from family structure and used for genotype-trait association. Computer program *PLINK* [44] with the *−genome* option was also used to infer correlations ($\widehat{\pi}$) using whole genome data. A total of 8485 SNPs on Affymetrix 500 K chips were derived from a panel of 45620 informative autosomal SNPs used in our consortium analysis. This led to estimates for $6520(6520-1)/2 = 21251940$ pairs of relationship. The genetic distance according to $|\pi - \widehat{\pi}|$ [45], that is, sum(abs*(EZ-PI_HAT), na.rm=TRUE)*, is 3421.724. Approximately half (10478474) had $\widehat{\pi}$ of 0.01 or more. Although there was a good agreement between kinship according to the specified family structures and $\widehat{\pi}$, 11207 pairs of individuals deemed to be unrelated had $\widehat{\pi}$ between 0.1–0.3 and 12 of which were greater than 0.3.

Both types of relationship matrices were used for the Cox model via *kinship* and *bdsmatrix.ibd* functions in *R*. The frailty and polygenic models had log likelihoods of −1788.53, −1791.93 with variance estimates 0.10^2 and 0.02^2, respectively. However, with inferred relationship the log likelihood turned out to be −1762.69 and variance estimate 0.24^2. Similar model for BMI at wave 1 was also fitted; a family specific random intercept model yielded log likelihood of −19273.26 and variance 3.44, while a correlated random intercept model gave log likelihood −19379.3 and variance 0.01^2 with comparable results from inferred relationship though with a smaller residual error. The results on diabetes might have suggested a substantial genetic effect while for BMI the use of inferred relationship performed equally well with a model using explicit family structures.

4. Discussion

The models we have considered extend counterparts for unrelated sample by taking into account correlation within and heterogeneity between families. To a large extent, we have presented an appreciation of models and implementations for related individuals using mixed models. At the meantime, we have envisaged a whole range of analyses that can be put in the framework. However, compared to [13] and especially [19], our development is more incremental and helps to gain insight into more complicated models. As a key feature of the model specification, oligogenes, polygenes, common environment, gene-environment interaction, and multivariate data are accommodated in a coherent framework via appropriate covariance structure. The generic nature has enabled a range of genetic association studies. Our interpretation of the model also naturally extends the model for quantitative traits outlined by [19, 46]. It has been recognized that for longitudinal data some commonly used covariance structures, such as compound symmetry, can be expressed as "linear covariance of dimension k" [47, page 258]. Although it could be more involved, it may be possible in our context. Data as in consortium meta-analysis analysis is also perceived in broader framework consisting of both unrelated and related individuals.

We should be aware that mixed models are quite general and may well be linked to other models. For instance, we noticed that model (2.10) is reminiscent of an approach proposed for generalized method of moments [48]. An example as with its link with

individual empirical Bayes estimates has been provided by [49, 50]. A reviewer has brought to our attention recent work on nonparametric methods for longitudinal data [51] and the utility of mixed models in controlling for bias of population stratification (e.g., [52]). This paper has limited coverage of literature on longitudinal analysis of family data, mainly owing to the fact that there is greater difficulty in implementation via general software package. However, this is expected to change. To our knowledge, little work has been done on joint analysis of individual data in the GWAS meta-analysis context. In view of the popularity of consortium data analysis, it will be appealing to have the appropriate mechanism to make it possible.

The models and their implementations are connected with whole genome data in several ways. First, the transition from the variance components models in earlier literature becomes more explicit. More specifically, the models described here are appropriate for GWAS where genetic variants coupled with a high resolution map are available. In general, the variance component associated with a major gene as in (2.7) is a function of the recombination rate (r) [12], that is, $\sigma_M^2 f(r, \pi_{ij})$, where π_{ij} represents identity-by-descent sharing between a pair of individuals i, j for the marker locus; with dense marker, we can assume that $r = 0$ which is also true with (2.9). Second, as in the Framingham data there is a further benefit with dense genetic markers such that they can be used to infer family structure [53] or (global) IBD information [54]. The availability of the deep sequencing data and a long list of established genes are likely to give greater weight on use of family data [55]. It is also desirable that cryptic relatedness in population-based sample can be appropriately taken into account in association analysis. In our own EPIC-Norolk GWAS, samples with cryptic relatedness have been excluded at the quality control stage [56]. It is interesting to note that *coxme* was developed for handling large pedigrees involving sparse matrices; the availability of whole genome data will alter the scenario slightly but nevertheless remain in the same framework. Third, more work is required to shorten computing time. In the literature, it has been proposed to absorb the relationship in the model for quantitative trait by multiplying inverse of the kinship matrix followed by a linear regression, or using residuals from a phenotype-covariate only regression as outcome in a model including SNPs as in *GenABEL*. In principle one can extend the idea to multivariate or longitudinal models where the residuals are obtained only once for GWAS or incorporating regional information before turning to SNP-specific analysis. There are also alternative approaches such as retrospective methods found in *Merlin*. With its greater requirement in computation the "measured genotype" approach here remains intuitive especially for gene-environment characterization. To this point, associate projects such as *BORDICEA* (see http://www.srl.cam.ac.uk/genepi/boadicea/boadicea_home.html) and *BayesMendel* (see http://bcb.dfci.harvard.edu/BayesMendel/) have contributed to the success of work on *R* described here.

A reviewer has expressed interest regarding the Type I error linking to results shown in Table 1. We believe that data as distributed by GAW17 as they were (200 replicates) are not ideal for assessing Type I error and possibly require a bootstrap procedure. In general, from our experience (and personal communications with Profs. Douglas Bates and Terry Therneau), this is a difficult issue and possibly problem specific. In fact, in the recent implementation of GLMM in *lme4*, the associate p values for fixed effects are not shown which nevertheless may leave users with temptation to employ normal approximation. Although we have not conducted extensive numerical experiments, results from GAW17 and the Framingham Study have indicated good performance of these models, and that of the inferred relationship based on whole genome data is impressive. Since only directly

Table 1: Nominal significance according to *VEGFC*.

	Q1	Q2	AFFECTED
Significance level	Power	Type I error	Power
.05	.989	.060	.880
.01	.907	.016	.730
.001	.665	0	.555
.0001	.412	0	.420
.00001	.225	0	.305
.000001	.104	0	.200

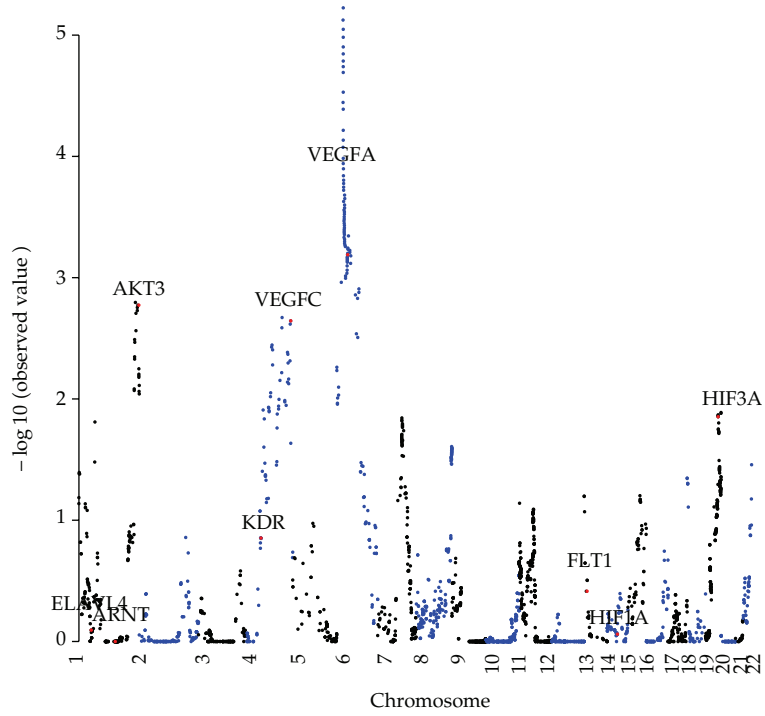


Figure 1: Manhattan plot of Q1 and IBD information where the true loci are highlighted.

genotyped Affy500K SNPs were used, the addition of imputed genotypes, say based on the HapMap, should help to improve the inference. Its use in the usual genomewide association analysis should be considered.

Our attention lies on the implementation by taking advantages of the available implementation in general statistical computing environment. The clarification of the implementation in these should facilitate practical analysis of family data. Although these models are conceptually simple, availability of their implementation vary, notably the ability to allow for both oligogenes and polygenes in a GLMM framework. For *R*, these are at least possible with *nlme*, *lme4*, and additionally *coxme*. At the moment, applications of packages in *R* are often restricted with *lmekin* in *coxme* offering outcomes only on continuous outcome but for *pedigreemm* it is unable to handle complex covariance structure. It is desirable that a function called *nlmekin* can be developed as with *pedigreemm* expanded to incorporate

Table 2: Q1 and *VEGFC* under a linear model.

Model/parameter	Estimate	SE	$z/t^{\dagger}$	−2 Res log likelihood	AIC
Kinship				1789.5	1793.5
σ_P^2	0.5488	0.08262	6.64		
SEX	−0.2379	0.04614	−5.16		
AGE	0.01014	0.001345	7.54		
SMOKE	0.36894	0.07280	5.07		
Kinship + IBD				1775.2	1781.2
σ_P^2	0.4157	0.08713	4.77		
σ_M^2	0.1076	0.03846	2.80		
SEX	−0.2488	0.04542	−5.48		
AGE	0.01044	0.001334	7.82		
SMOKE	0.3821	0.07181	5.32		

$^{\dagger}z$ is for variance components while t for fixed effects.

Table 3: AFFECTED and *VEGFC* under a logistic model.

Model/parameter	Estimate	SE	t	−2 log pseudolikelihood
Kinship				3434.4
σ_P^2	1.3170	0.4376		
SEX	−0.00822	0.2042	−0.04	
AGE	0.07181	0.006047	11.87	
SMOKE	0.9098	0.2285	3.98	
Kinship + IBD				3445.7
σ_P^2	0.6918	0.5989		
σ_M^2	0.4868	0.3698		
SEX	0.006923	0.2048	0.03	
AGE	0.07211	0.006114	11.79	
SMOKE	0.9429	0.2290	4.12	

Table 4: AFFECTED and *VEGFC* under a Cox model.

Model/parameter	Estimate	SE	z	Integrated/penalized likelihoods†
Kinship				−998.8/−980.6
σ_P^2	0.2073			
SEX	0.05267	0.1541	0.34	
SMOKE	0.5000	0.1622	3.08	
Kinship + IBD				−996.1/−967.3
σ_P^2	0.002690			
σ_M^2	0.3615			
SEX	0.07146	0.1603	0.43	
SMOKE	0.5560	0.1696	3.28	

† The log likelihood under the null is −1003.9.

Table 5: Q1, Q2, and Q4 under a multivariate polygenic model.

	Estimate	SE	Log likelihood
	Linear coefficients		−1393.867
c1	0.565	0.108	
c2	0.531	0.109	
c3	0.526	0.109	
Sex	−0.005	0.043	
Age	−0.013	0.001	
Smoke	−0.019	0.051	
	Variance coefficients		
σ_{11}	4.219	0.227	
σ_{12}	−0.103	0.166	
σ_{22}	4.542	0.244	
σ_{31}	0.601	0.178	
σ_{32}	−0.108	0.183	
σ_{33}	5.115	0.275	

additive covariance structures. *SAS*, MIXED, GLIMMIX, and NLMIXED together provide a rich source of practical modeling functionality though the Cox model counterpart is not available. The tackling of various issues has led to efficient algorithm [25]. When the interest is on correlation between multiple traits, the use of *nlme* for multivariate longitudinal data in unrelated individuals has been described [57]. In general, this could be complicated with longitudinal familial data without [58] or with [59] consideration of relationship. In study of obesity-related traits, *FTO* has been shown to be strongly associated with BMI and supported by cross-sectional data as in [14], longitudinal data as in [43] and data across life span as in [60]. Our previous attempt [43], was based on a three-level model and it would be of interest to use kinship information as well.

While the framework we have outlined is comprehensive, we feel that our "proof of concepts" here awaits for extensive testing. It is also desirable that the current implementation can be optimized in computing time. A lot of work has been done for quantitative genetics in plants and animals. Our experience indicated that the running time with *SAS* was longer time than *R*. However, in an analysis of longitudinal lung function data in the EPIC-Norfolk study, we have shown that although an individual analysis could be slow, it is possible to perform an analysis for GWAS using *SAS* and Linux clusters so that ~2.5 M SNPs would finish within 14 hours when running each chromosome on a separate node. It is likely that it benefited from *SAS* caching frequently-used instructions. Greater proportion of coding in *C/C++* should also be helpful. Given the utility of the popular environments can be shown, their take-up in genomewide association studies will be quick and it is very much in line with efforts in other disciplines where large volume of data is involved.

Acknowledgments

A lot of the insights were gained during analysis of GAWs 14, 16, 17 and in particular maintenance of the *R* counterpart of the *S-PLUS* package *kinship* (http://mayoresearch.mayo.edu/mayo/research/biostat/upload/kinship.pdf) by the first author. The authors are therefore very grateful of the pioneering work and advices given by Profs Terry Therneau,

Beth Atkinson, and Mariza de Andrade all at the Mayo Clinic and interactions with many other colleagues elsewhere. The comprehensive *R* archive network (CRAN (http://cran.r-project.org)) as with Professors Kurt Hornik and Brian Ripley has been a constant source of support. The work presented here was partly done for CompBio2011 and useR!2011. They wish to thank Drs. Qihua Tan, Fuzhong Xue, Wendi Qian, and Luigi Palla for their participation and comments during the GAWs 16 & 17 analysis which led to this work, Dr Wendi Qian's comments on *SAS* PROC GLIMMIX, and Dr Antonis Antoniou's suggestion of using average age within a sibship to approximate age at onset. The example regarding twins was due to a query from Dr. Marcel de van Hoed. They are also grateful of the Editor for communications which led to the work on the paper and three anonymous reviewers for their insightful comments which led to its improvement. The work reported here also allows us for making minor changes to the syntax shown in [36]. Professor Peter McCullagh from University of Chicago and Dr. David Clifford from CSIRO have kindly provided advices regarding the use of *regress*.

References

[1] D. C. Thomas and W. J. Gauderman, "Gibbs sampling methods in genetics," in *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richard, and D. J. Spiegelhalter, Eds., pp. 419–440, Chapman & Hall/CRC, London, UK, 1996.

[2] D. C. Thomas, *Statistical Methods in Genetic Epidemiology*, University Press, Oxford, UK, 2004.

[3] J. Yang, B. Benyamin, B. P. McEvoy et al., "Common SNPs explain a large proportion of the heritability for human height," *Nature Genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[4] R. A. Fisher, "The correlation between relatives on the supposition of mendelian inheritance," *Transactions of the Royal Society of Edinburgh*, vol. 52, pp. 399–433, 1918.

[5] H. L. Allen, K. Estrada, G. Lettre et al., "Hundreds of variants clustered in genomic loci and biological pathways affect human height," *Nature*, vol. 467, no. 7317, pp. 832–838, 2010.

[6] G. B. Ehret, P. B. Munroe, K. M. Rice et al., "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, vol. 478, no. 7367, pp. 103–109, 2011.

[7] T. M. Teslovich, K. Musunuru, A. V. Smith et al., "Biological, clinical and population relevance of 95 loci for blood lipids," *Nature*, vol. 466, no. 7307, pp. 707–713, 2010.

[8] E. K. Speliotes, C. J. Willer, S. I. Berndt et al., "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index," *Nature Genetics*, vol. 42, no. 11, pp. 937–948, 2010.

[9] R. Plomin, C. M. A. Haworth, and O. S. P. Davis, "Common disorders are quantitative traits," *Nature Reviews Genetics*, vol. 10, no. 12, pp. 872–878, 2009.

[10] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models*, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, NY, USA, 2001.

[11] SAS Institute, *SAS/STAT 9.3 User's Guide*, SAS Publishing, Cary, NC, USA, 2011.

[12] C. I. Amos, "Robust variance-components approach for assessing genetic linkage in pedigrees," *American Journal of Human Genetics*, vol. 54, no. 3, pp. 535–543, 1994.

[13] J. Blangero, J. T. Williams, and L. Almasy, "Variance component methods for detecting complex trait loci," *Advances in Genetics*, vol. 42, pp. 151–181, 2001.

[14] T. M. Frayling, N. J. Timpson, M. N. Weedon et al., "A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity," *Science*, vol. 316, no. 5826, pp. 889–894, 2007.

[15] N. E. Morton and C. J. MacLean, "Analysis of family resem blance. III. complex segregation of quantitative traits," *American Journal of Human Genetics*, vol. 26, pp. 489–503, 1974.

[16] J. L. Hopper and J. D. Mathews, "Extensions to multivariate normal models for pedigree analysis," *Annals of Human Genetics*, vol. 46, no. 4, pp. 373–383, 1982.

[17] K. Lange and M. Boehnke, "Extensions to pedigree analysis. IV. Covariance components models for multivariate traits," *American Journal of Medical Genetics*, vol. 14, no. 3, pp. 513–524, 1983.

[18] S. J. Hasstedt, "A mixed-model likelihood approximation on large pedigrees," *Computers and Biomedical Research*, vol. 15, no. 3, pp. 295–307, 1982.

[19] M. P. Epstein, J. E. Hunter, E. G. Allen, S. L. Sherman, X. Lin, and M. Boehnke, "A variance-component framework for pedigree analysis of continuous and categorical outcomes," *Statistics in BioSciences*, vol. 1, no. 2, pp. 181–198, 2009.
[20] A. M. Saxton, Ed., *Genetic Analysis of Complex Traits Using SAS*, SAS Publishing, 2004.
[21] A. I. Vazquez, D. M. Bates, G. J. M. Rosa, D. Gianola, and K. A. Weigel, "Technical note: an R package for fitting generalized linear mixed models in animal breeding," *Journal of Animal Science*, vol. 88, no. 2, pp. 497–504, 2010.
[22] V. S. Pankratz, M. de Andrade, and T. M. Therneau, "Random-effects cox proportional hazards model: general variance components methods for time-to-event data," *Genetic Epidemiology*, vol. 28, no. 2, pp. 97–109, 2005.
[23] V. Ducrocq and G. Casella, "A bayesian analysis of mixed survival models," *Genetics Selection Evolution*, vol. 28, no. 6, pp. 505–529, 1996.
[24] D. Sorensen and D. Gianola, *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, Springer, New York, NY, USA, 2002.
[25] P. Waldmann, "Easy and flexible Bayesian inference of quantitative genetic parameters," *Evolution*, vol. 63, no. 6, pp. 1640–1643, 2009.
[26] P. R. Burton, K. J. Scurrah, M. D. Tobin, and L. J. Palmer, "Covariance components models for longitudinal family data," *International Journal of Epidemiology*, vol. 34, no. 5, pp. 1063–1079, 2005.
[27] J. M. Lachin, *Biostatistical Methods: The Assessment of Relative Risks*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2011.
[28] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2004.
[29] J. Whitehead, "Fitting Cox's regression model to survival data using GLIM," *Journal of the Royal Statistical Society*, vol. 29, no. 3, pp. 268–275, 1980.
[30] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*, Springer, New York, NY, USA, 2000.
[31] J. Pinheiro and D. M. Bates, *Mixed Effects Models in S and S-PLUS*, Springer, 2000.
[32] R. B. Bapat, *Linear Algebra and Linear Models*, Universitext, Springer, London, UK, 3rd edition, 2012.
[33] P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*, vol. 25, Oxford University Press, Oxford, UK, 2nd edition, 2002.
[34] J. P. Klein and M. L. Moeschberger, *Survival Analysis-Techniques for Censored and Truncated Data*, Springer, 2nd edition, 2003.
[35] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: rationale, conduct, and reporting," *British Medical Journal*, vol. 340, p. c221, 2010.
[36] J. H. Zhao, J. Luan, R. J. F. Loos, and N. Wareham, "On genotype-phenotype association using SAS," in *Proceedings of the 2nd International Conference on Computational Bioscience*, pp. 428–433, Cambridge, Mass, USA, 2011.
[37] T. D. Pigott, *Advances in Meta-Analysis*, Springer, 2012.
[38] J. Neyman and E. L. Scott, "Consistent estimates based on partially consistent observations," *Econometrica*, vol. 16, pp. 1–32, 1948.
[39] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society*, vol. 67, no. 3, pp. 427–444, 2005.
[40] N. J. Schork, "Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations," *American Journal of Human Genetics*, vol. 53, no. 6, pp. 1306–1319, 1993.
[41] J. H. Zhao and Q. Tan, "Integrated analysis of genetic data with R," *Human Genomics*, vol. 2, no. 4, pp. 258–265, 2006.
[42] L. Almasy, T. D. Dyer, J. M. Peralta et al., "Genetic Analysis Workshop 17 mini-exome simulation," *BMC Proceedings*, vol. 5, article S2, supplement 9, Article ID S2, 2011.
[43] J. Luan, B. Kerner, J. H. Zhao et al., "A multilevel linear mixed model of the association between candidate genes and weight and body mass index using the framingham longitudinal family data," *BMC Proceedings*, vol. 3, article S115, supplement 7, 2009.
[44] S. Purcell, B. Neale, K. Todd-Brown et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.
[45] A. Sanchez, J. Ocaña, and F. Utzet, "Sampling theory, estimation, and significance testing for Prevosti's estimate of genetic distance," *Biometrics*, vol. 51, no. 4, pp. 1216–1235, 1995.

[46] M. de Andrade, E. Atkinson, E. Lunde, C. I. Amos, and J. Chen, "Estimating genetic components of variance for quantitative traits in family studies using the multic," Tech. Rep., Mayo Clinic, 2006.
[47] E. F. Vonesh and V. M. Chinchilli, *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, vol. 154 of *Statistics: Textbooks and Monographs*, Marcel Dekker, NewYork, NY, USA, 1997.
[48] G. Yin, "Bayesian generalized method of moments," *Bayesian Analysis*, vol. 4, no. 2, pp. 191–208, 2009.
[49] T. Moger, O. O. Aalen, K. Heimdal, and H. K. Gjessing, "Analysis of testicular cancer data using a frailty model with familial dependence," *Statistics in Medicine*, vol. 23, no. 4, pp. 617–632, 2004.
[50] O. O. Aalen, O. Borgan, and H. K. Gjessing, *Survival and Event History Analysis: A Process Point of View*, Statistics for Biology and Health, Springer, NewYork, NY, USA, 2008.
[51] Y. Wang, C. Huang, Y. Fang, Q. Yang, and R. Li, "Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing," *Journal of the Royal Statistical Society*, vol. 61, no. 1, pp. 1–24, 2012.
[52] J. Yu, G. Pressoir, W. H. Briggs et al., "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nature Genetics*, vol. 38, no. 2, pp. 203–208, 2006.
[53] A. G. Day-Williams, J. Blangero, T. D. Dyer, K. Lange, and E. M. Sobel, "Linkage analysis without defined pedigrees," *Genetic Epidemiology*, vol. 35, no. 5, pp. 360–370, 2011.
[54] L. Han and M. Abney, "Identity by descent estimation with dense genome-wide genotype data," *Genetic Epidemiology*, vol. 35, no. 6, pp. 557–567, 2011.
[55] J. R. Lupski, J. G. Reid, C. Gonzaga-Jauregui et al., "Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy," *The New England Journal of Medicine*, vol. 362, no. 13, pp. 1181–1191, 2010.
[56] R. J. F. Loos, C. M. Lindgren, S. Li et al., "Common variants near MC_4R are associated with fat mass, weight and risk of obesity," *Nature Genetics*, vol. 40, no. 6, pp. 768–775, 2008.
[57] S. Bandyopadhyay, B. Ganguli, and A. Chatterjee, "A review of multivariate longitudinal data analysis," *Statistical Methods in Medical Research*, vol. 20, no. 4, pp. 299–330, 2011.
[58] B. C. Sutradhar, *Dynamic Mixed Models for Familial Longitudinal Data*, Springer Series in Statistics, Springer, NewYork, NY, USA, 2011.
[59] J. M. Soler and J. Blangero, "Longitudinal familial analysis of blood pressure involving parametric (co)variance functions," *BMC Genetics*, vol. 4, article S87, supplement 1, 2003.
[60] R. Hardy, A. K. Wills, A. Wong et al., "Life course variations in the associations between *FTO* and MC_4R gene variants and body size," *Human Molecular Genetics*, vol. 19, no. 3, pp. 545–552, 2010.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 830575, 18 pages
doi:10.1155/2012/830575

Research Article

Finding Transcription Factor Binding Motifs for Coregulated Genes by Combining Sequence Overrepresentation with Cross-Species Conservation

Hui Jia[1] and Jinming Li[1,2]

[1] *School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551*
[2] *Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou Dadao Bei1838, Guangzhou 510515, China*

Correspondence should be addressed to Jinming Li, jmli@smu.edu.cn

Received 1 March 2012; Accepted 29 April 2012

Academic Editor: Xiaohua Douglas Zhang

Novel computational methods for finding transcription factor binding motifs have long been sought due to tedious work of experimentally identifying them. However, the current prevailing methods yield a large number of false positive predictions due to the short, variable nature of transcriptional factor binding sites (TFBSs). We proposed here a method that combines sequence overrepresentation and cross-species sequence conservation to detect TFBSs in upstream regions of a given set of coregulated genes. We applied the method to 35 *S. cerevisiae* transcriptional factors with known DNA binding motifs (with the support of orthologous sequences from genomes of *S. mikatae*, *S. bayanus*, and *S. paradoxus*), and the proposed method outperformed the single-genome-based motif finding methods *MEME* and *AlignACE* as well as the multiple-genome-based methods *PHYME* and *Footprinter* for the majority of these transcriptional factors. Compared with the prevailing motif finding software, our method has some advantages in finding transcriptional factor binding motifs for potential coregulated genes if the gene upstream sequences of multiple closely related species are available. Although we used yeast genomes to assess our method in this study, it might also be applied to other organisms if suitable related species are available and the upstream sequences of coregulated genes can be obtained for the multiple closely related species.

1. Introduction

To understand the mechanisms that regulate the gene expression in eukaryotes is a major challenge in modern molecular biology. Gene regulation is accomplished by a number of regulatory proteins called transcriptional factors (TFs), which bind to specific DNA motifs in the promoter region of the target gene. TFs and their binding motifs interact with each other

and help cells to respond to diverse stimuli. Identifying TFBSs in the upstream region of coregulated genes (genes regulated by a common TF) is crucial for inferring gene regulatory networks, since these motifs might be the building blocks of the regulatory network structures [1]. Most DNA binding motifs contain 6–25 bps and have a range of variability. Regulatory systems can take advantage of the variability in the binding sites to better control transcription [2]. Classical computational motif finding methods can be classified into two major categories: (1) enumerative methods, which explore all possible motifs up to a certain length; (2) local search algorithms using statistic approaches such as EM [3–6] and Gibbs sampling [7–9]. Under the second category, *MEME* [10–12] and *AlignACE* [13, 14] are two computer programs used frequently in finding motifs in a group of related DNA sequences. Recently, comparative genomics approaches such as phylogenetic footprinting have been developed for identifying TFBSs based on the premise that selective pressure causes functional elements to evolve at a slower rate than that of nonfunctional sequences [15]. Phylogenetic footprinting is mostly applied to finding well-conserved regions in a set of orthologous sequences from multiple species [15–18]. Although substantial progresses have been made in developing computational methods for predicting transcription factor binding motifs, currently available motif finding tools still yield many false positives due to relatively short length and high variability of DNA binding motifs. These motif finding tools with standard parameter settings usually report one putative TFBS out of 500 to 5000 bps, whereas only 0.1% of the predictions is likely to be functional [19]. Recently, gene expression profile analysis using microarray data and statistical clustering has resulted in numerous sets of potential co-regulated genes. Furthermore, the complete sequencing of more and more eukaryotic genomes makes it easier to obtain the upstream sequences of these co-regulated genes. Hence the development of a novel method with improved specificity in predicting transcription factor binding motifs for co-regulated genes becomes necessary and feasible.

We proposed here a method of finding TF binding motifs by considering both sequence overrepresentation in promoter regions and their conservation across closely related species. DNA binding motifs are believed to appear more frequently in the upstream regions of the genes being regulated, and these motifs are usually conserved across multiple closely related species. We use the degree of sequence conservation among multiple species as an additional constraint to reduce the false positive predictions. For a given set of co-regulated genes from a certain organism, we collect orthologous sequences from multiple closely related species and align them using multiple alignment programs such as *ClustalW* [20]. The statistically overrepresented sequences will be firstly selected as initial motif candidates, and then we evaluate their conservation in the alignments of orthologous upstream sequences of coexpressed genes. A statistic procedure based on the above principles was designed to scan for potential motifs, and a Perl script was written to conduct the procedure. To evaluate the proposed method, we collected 35 yeast TFs with known DNA binding motifs, and for genes co-regulated by each of these TFs, we searched the upstream regions for potential binding motifs. We compared our method with single-genome-based motif finding methods such as *MEME* and *AlignACE*, as well as with the multiple-genome based methods such as *PHYME* and *Footprinter*; the results suggested that the rank of the known binding motifs among the predictions of our method are generally higher than that using other methods.

2. Results

We used 35 well-studied yeast transcription factors (see Table 1) to evaluate the proposed method. The criteria for selecting the TFs are (1) their true DNA binding motifs are known;

(2) the orthologous genes are available in all the four yeast species, and the upstream sequences of these genes are also available. For each TF, we built two sets of genes, namely, the positive set (*PS*) and the negative set (*NS*). The *PS* consisted of all the genes that are known to be co-regulated by the TF (see Table 1), whereas the *NS* consisted of randomly selected genes from the *S. cerevisiae* genome. The *NS* was used to introduce the background information and serve as a control in our motif finding process. For each gene in both *PS* and *NS,* we extracted its promoter region sequences from the genomes of four yeast species, namely, *S. cerevisiae*, *S. mikatae*, *S. bayanus,* and *S. paradoxus*. We took *S. cerevisiae* as the principal species in our study.

The method was implemented using a PERL script to find potential binding motifs in the upstream sequences of the genes co-regulated by a given TF (see Table 1 for the 35 TFs considered in this study). We found the known binding motifs for 25 out of the 35 TFs. In Table 2, we listed the known DNA binding motif and the motif found using our method for each TF.

We compared our method with the single-genome-based motif finding methods *MEME* and *AlignACE*, as well as with the multiple-genome-based methods *PHYME* and *Footprinter* for the majority of these transcription factors. We used the upstream sequences of *S. cerevisiae* genes in the *PS* of each TF as the input of *MEME* and *AlignACE*. All the parameters were set to default when we used *AlignACE* to find motifs. To apply *MEME* to the motif finding, we set the minimum length of the potential motif to 6, and we set the number of motifs expected to be found to the same as the number of motifs predicted by our method. The results are listed in Tables 3 and 4, respectively. Since our method takes into account the conservation of candidate motif sequences among multiple species, the number of predicted motifs found for each TF is in general less than that found by *AlignACE* (Table 3) or *MEME* (Table 4). Tables 3 and 4 showed that our method is more efficient in finding the true motifs than *AlignACE* or *MEME*, in the sense that it returned less predicted motifs, and the ranks of the known motifs are also generally higher than those in the output of *AlignACE* or *MEME*. For example, there were 11 potential motifs found by *AlignACE* for *STE12*, and the known motif of *STE12* ranked second in the output; however, using our method only one motif was found, and it was the known motif. The results for other TFs showed the same tendency. *AlignACE* and *MEME* could only find the known binding motifs for 14 and 12 TFs, respectively, out of the 25 TFs whose known binding motifs were found using our method. Our method could not find the known binding motifs for 10 TFs among the 35 (Table 5) with any of the three parameter threshold settings. Out of these 10 TFs, using *AlignACE* and *MEME* we can find known binding motifs for 5 and 3 TFs, respectively.

Unlike single-genome-based motif finding methods such as *AlignACE* and *MEME*, our method uses the sequence information from multigenomes, so it is more reasonable to compare it with *PHYME* and *Footprinter,* which are two popular multiple-genome-based methods. For a given TF, we found that *Footprinter* usually yields overwhelming number of predictions, and this makes it difficult to do a comparison. To apply *PHYME* to the motif finding, we set the motif length limit to 17, which is the maximum length of all known binding motifs of the 35 TFs. For each regulon, the number of motifs predicted was set to 10 and the motifs were searched on both strands. The results were listed in Table 6. Using *PHYME* we found known motifs for 23 TFs, among them there were 6 TFs whose known binding motifs were not found using our method. From Table 6, we can see that our method and *PHYME* nearly have the same power in motif finding; however, the ranks of the known motifs found using our method are generally higher than those found by *PHYME*. Table 7 gives a list of the TFs whose known binding motifs could be found using our method but could not be found by *MEME*, *AlignACE,* and *PHYME*. Our method could not find the known binding motifs for

Table 1: Transcription factors and the genes being regulated.

TF	Number of co-regulated genes	Genes regulated by the TF
Ste12	9	YBR083, YCL055W, YFL027C, YJL170C, YLR452C, YML047C, YMR232W, YNL279W, YPL156C
Gal4	10	GAL2, GAL3, GAL1/10, GAL7, MTH1, FUR4, PCL10, GAL80, PGM2, GCY1
MET31	8	YEL015W, YEL016C, STR3, MET16, NUT2, SSN8, YJL060W, YEL072W
Mbp1	18	YEL018W, MMS21, YCK2, MCD1, MCM2, RPS9A, MOT1, OPY2, CLB5, YER071C, VTC1, YJL045W, MSH6, YNR009W, HXT10, YER087C-A, TOF1, YNL274C
Leu3	10	YDR279W, LEU1, OAC1, YOR271C, YDL228C, YHR209W, YHR207C, BAT1, ILV2, RRP6
Cbf1	16	YAL026C, YBR089C-A, YBR225W, YDR438W, YIL074C, YIL126W, YIL127C, YJL167W, YJL168C, YJL209W, YJR010W, YKL191W, YKL192C, YNL094W, YNL095C, YNL282W
Ace2	1	YLR286C
Gcn4	6	YBL103C, YDL170W, YKL015W, YLR451W, YML099C, YNL103W
Abf1	15	YAL038W, YBR248C, YCR012W, YFL038C, YFR031C, YGL234W, YGR059W, YHR174W, YIL160C, YJL166W, YKL112W, YLR203C, YLR204W, YOR116C, YPR110C
Hap1	4	YEL039C, YJR048W, YML054C, YOR065W
Ino4	6	YDR050C, YER026C, YGR157W, YHR123W, YMR084W, YNR016C
Mcb	6	YDL102W, YDL164C, YJL194W, YMR199W, YNL102W, YOR074C
Mse	1	YGR059W
Nbf	1	YJL153C
Pdr3	2	YBL005W, YGR281W
Pho4	2	YDR481C, YGR233C
Put3	1	YHR037W
Rap1	8	YFL014W, YFR031C, YGL123W, YKL062W, YLR399C, YNL216W, YOL082W, YPR102C
Swi5	2	YDL227C, YNL327W
Uasino	1	YJL153C
Uasrad	2	YCR066W, YGL058W
Adr1	2	YDR256C, YMR303C
Mig1	7	YBR019C, YBR020W, YDR009W, YDR146C, YIL162W, YKL109W, YPL248C
T4c	2	YJL106W, YJL153C
Uasphr	14	YBR114W, YDL200C, YDR217C, YEL037C, YER095W, YER142C, YGL058W, YIL066C, YJL026W, YJR035W, YJR052W, YML032C, YNL250W, YPL022W
Ap-1	1	YGR209C
Bas2	2	YCL030C, YGL234W
Csre	2	YER065C, YNL117W
Mac1	11	YDR058C, YDR075W, YER145C, YER146W, YGR136W, YJR049C, YJR050W, YNL250W, YNL251C, YPR110C, YPR111W
Gcr1	2	YAL038W, YGR215W
Mcm1	17	YAL040C, YBR160W, YBR202W, YDR146C, YDR403W, YER111C, YFL026W, YGL008C, YGR108W, YJL159W, YJL194W, YKL178C, YKL209C, YKR066C, YNL277W, YPR113W, YPR119W

Table 1: Continued.

TF	Number of co-regulated genes	Genes regulated by the TF
Reb1	12	YCR012W, YDL164C, YDR007W, YDR050C, YDR146C, YER086W, YFL039C, YGL026C, YNL216W, YOL004W, YOL006C, YPL231W
Rox1	2	YDR044W, YPR065W
Scb	2	YDL227C, YMR199W
Sff	3	YDR146C, YGR108W, YPR119W-

10 TFs out of the 35 (Table 5). For these 10 TFs, *AlignACE* and *MEME* can find known binding motifs for 5 and 3 TFs, respectively. With *PHYME*, we can find known binding motifs for 6 TFs out of these 10 (Table 6).

3. Discussion

Transcription factors and their DNA binding sites are two of the most important functional elements in eukaryotic genomes. A thorough study of the interactions of the two is important for mapping the regulatory pathways and understanding the potential function of the genes regulated by the TFs [21]. In the past decade, clustering of gene expression profiles obtained from large-scale DNA microarray experiments has been successfully used in identifying coexpressed genes [22, 23], and we believed that these coexpressed genes may share common regulators that bind to their upstream regions. Finding the TF binding motifs of these potentially co-regulated genes becomes critical for understanding the interaction of the genes and their regulators [24–27]. So far the binding specificities have been well characterized only for a small number of TFs [19, 21]. TFBSs are usually quite short (around 6–25 bp) and degenerate, which leads to the difficulties in finding them reliably using current motif finding tools. Even though the *ab initio* motif finding tools have been used successfully in many cases, their performances are far from satisfying. The major drawback of these tools is that they produce many false positive predictions. Under default parameter settings, they yield usually tens or hundreds of putative motifs, and it is difficult to judge which candidate motifs out of them are functional [19]. Phylogenetic footprinting methods have been proposed recently [15–18], by which the interspecies comparative sequence information is used for helping to signal the presence of TF binding sites that might not have been predicted using sequences from a single genome. For example, binding sites found in human sequences that are also found in orthologous mouse or other mammalian sequences are far more likely to be functional than those found only in human [28]. We refer to these short orthologous sequences that are conserved over 6 bp or more as phylogenetic footprints.

Our method proposed here considers both overrepresentation and cross-species conservation of potential binding motifs. We used binomial test to determine the statistically overrepresented candidate sequences, and the average relative entropy of the aligned sequence block was used to measure the cross-species conservation of these candidates. The relative entropy is a popular measure of the degree of conservation at a site in a DNA or protein sequence alignment [29]. In our method, the input data are the upstream sequences of two groups of genes, namely, the co-regulated genes of a TF (*PS*) and the control genes (*NS*) selected randomly from the genome of the principal species under study, as well as the

Table 2: Comparison between the known motifs and the motifs found using our method. Different parameter threshold settings are used in our motif finding. (a) P-value in a magnitude of 10^{-6} (after Bonferroni adjustment), $\text{ARE}_p = 1.0$, and Z-value = 2.0; (b) P-value = 0.01 (without Bonferroni adjustment), $\text{ARE}_p = 1.0$, and Z-value = 2.0; (c) P-value = 0.01 (without Bonferroni adjustment), $\text{ARE}_p = 0.8$, and Z-value = 2.0.

TF	Genes in PS	Known motif	Motif found	P-value	Z-value	ARE_p
Ste12[(a)]	9	TGAAACA	TGAAACA	$5.6e{-}12$	3.68	1.00
Gal4[(a)]	10	CGGNNNNNNNNNNCCG	CGGNNNNNNNNNNCCG	$3.7e{-}12$	2.43	1.22
Mbp1[(a)]	18	ACGCGTNA	ACGCGT	$3.0e-7$	3.37	1.35
Leu3[(a)]	10	CCGGNNCCGG	CGGNNNCGG	$7.0e{-}13$	3.15	1.27
Cbf1[(a)]	16	RTCACRTG	CACGTG	$7.7e{-}13$	2.98	1.19
MET31[(a)]	8	CTGTGGC	TGTGGC	$6.7e-7$	3.39	1.06
Abf1[(b)]	15	TCRNNNNNNACG	TCANNNNNNACG	$1.3e-3$	3.73	1.26
Ace2[(b)]	1	GCTGGT	TGCTGGT	$1.4e-3$	6.07	1.55
Gcn4[(b)]	6	TGANTN	ATGACT	$8.7e-4$	4.45	1.10
Hap1[(b)]	4	CGGNNNTANCGG	TGCCGNNNNNNNCGG	$2.3e-4$	6.09	1.64
Ino4[(b)]	6	CATGTGAAAT	CATGTT	$2.9e-4$	5.60	1.31
Mcb[(b)]	6	WCGCGW	CGCNTCG	$4.1e-4$	4.66	1.36
Mse[(b)]	1	CRCAAAW	GACNCAA	$8.3e-3$	4.05	1.19
Nbf[(b)]	1	ATGYGRAWW	CATGTG	$5.9e-3$	5.85	1.36
Pdr3[(b)]	2	TCCGYGGA	TCCNNGGA	$4.3e-4$	2.88	1.03
Pho4[(b)]	2	CACGTK	GCGCGT	$1.8e-3$	3.55	1.20
Put3[(b)]	1	CGGNNNNNNNNNNCCG	TCGNNNNNNNNNNNCG	$2.6e-4$	4.65	1.51
Rap1[(b)]	8	RMACCCA	GTCNNNNNCCCAT	$8.8e-3$	3.16	1.01
Swi5[(b)]	2	KGCTGR	TGCTGG	$6.5e-4$	4.45	1.19
Uasino[(b)]	1	ATCTGAAWW	CATGTG	$5.9e-3$	5.83	1.36
Uasrad[(b)]	2	WTTTCCCGS	TCCNGCT	$1.1e-3$	4.42	1.24
Adr1[(c)]	2	TCTCC	CTCCNNNNNTCC	$1.6e-3$	2.18	0.88
Mig1[(c)]	7	CCCCRNNWWWWW	ACCCCA	$7.2e-3$	2.18	0.82
Uasphr[(c)]	14	CTTCCT	TCTNNNNNNNNNNTCCT	$2.2e-3$	2.38	0.93
T4c[(c)]	2	TTTTCTYCG	TTTTCNNNNNTCC	$1.2e-3$	2.69	0.96

orthologous sequences from other species, which are closely related to the principal species. Usually the co-regulated genes are collected through wet lab experiments or predicted through gene expression profile analysis using microarray data. The upstream sequences of genes in *PS* and *NS* could be extracted from the genome of the principal species, and the corresponding upstream sequences from other species could be obtained by doing BLAST [30] or by downloading from the publicly available databases.

Three parameters are considered in our method: (1) P value, which is used to evaluate the overrepresentation of a candidate sequence, (2) average relative entropy ARE_P of S_{OP}, which gives the degree of conservation of a candidate motif, (3) Z-value, which is used to assess the statistical significance of the conservation. In order to have a balanced consideration of the sensitivity and the specificity and to cope with different situations, we applied three different parameter threshold settings to scan for candidate motifs, and they are (a) P-value in a magnitude of 10^{-6} (after Bonferroni correction),

Table 3: Comparison to *AlignACE*. For each TF, we listed the rank of the known motif in the predictions. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	*AlignACE*		Our method	
	The number of motifs found	The rank of the known motif	The number of motifs found	The rank of the known motif
Ste12[(a)]	11	2	1	1
Gal4[(a)]	9	2	1	1
Leu3[(a)]	22	3	1	1
Mbp1[(a)]	20	7	2	1
Cbf1[(a)]	29	1	4	2
Met31[(a)]	20	15	1	1
Abf1[(b)]	10	4	5	3
Ace2[(b)]	7	Not found	4	2
Gcn4[(b)]	11	4	39	6
Hap1[(b)]	6	Not found	23	1
Ino4[(b)]	13	Not found	22	4
Mcb[(b)]	18	2	11	1
Mse[(b)]	6	Not found	5	5
Nbf[(b)]	6	1	16	15
Pdr3[(b)]	12	Not found	10	2
Pho4[(b)]	8	Not found	9	1
Put3[(b)]	2	Not found	4	1
Rap1[(b)]	13	6	14	11
Swi5[(b)]	9	Not found	3	1
Uasino[(b)]	4	Not found	14	13
Uasrad[(b)]	4	Not found	20	3
Adr1[(c)]	4	2	14	2
Mig1[(c)]	30	2	32	30
Uasphr[(c)]	14	Not found	47	13
T4c[(c)]	9	1	28	6

ARE_P = 1.0, and Z-value = 2.0; (b) P-value = 0.01 (without Bonferroni correction), ARE_P = 1.0, and Z-value = 2.0; (c) P-value = 0.01 (without Bonferroni correction), ARE_P = 0.8, and Z-value = 2.0. Theoretically, we can find most of the known motifs as long as we make the criteria for overrepresentation and conservation loose enough, but the less strict criteria may result in numerous putative motifs that are actually false positives. Considering the high cost of verifying a predicted motif through lab experiment, we used firstly a strict criterion for candidate motif screening, so parameter setting (a) was set as default in our method. Using this strict parameter threshold setting we may miss some true TF binding motifs (see Tables 3 and 4), especially those without very high-level statistical significance of overrepresentation, and the method may not be able to return any predictions. We loosen the criteria by using setting (b) or setting (c) in actual motif finding process, if using the default threshold setting, we can find no hit at all. Setting (b) has a moderate criterion for overrepresentation, so it allows more candidate motif to pass the screening. With setting (c), we loosen the criterion

Table 4: Comparison to *MEME*. *MEME* requests a predetermined number of predicted motifs as its input, and we let it be the number of motifs predicted using our method. For each TF, we listed the rank of the known motif in the predictions. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	*Meme*		Our Method	
	The number of motifs found	The rank of the known motif	The number of motifs found	The rank of the known motif
Ste12[(a)]	1	Not found	1	1
Gal4[(a)]	1	1	1	1
Leu3[(a)]	1	1	1	1
Mbp1[(a)]	2	1	2	1
Cbf1[(a)]	4	1	4	2
Met31[(a)]	1	Not found	1	1
Abf1[(b)]	5	Not found	5	3
Ace2[(b)]	4	Not found	4	2
Gcn4[(b)]	39	Not found	39	6
Hap1[(b)]	9	Not found	23	1
Ino4[(b)]	22	10	22	4
Mcb[(b)]	11	1	11	1
Mse[(b)]	5	Not found	5	5
Nbf[(b)]	16	3	16	15
Pdr3[(b)]	10	1	10	1
Pho4[(b)]	9	4	9	1
Put3[(b)]	4	Not found	4	1
Rap1[(b)]	14	Not found	14	11
Swi5[(b)]	3	Not found	3	1
Uasino[(b)]	14	Not found	14	13
Uasrad[(b)]	20	1	20	3
Adr1[(c)]	10	Not found	14	2
Mig1[(c)]	32	2	32	30
Uasphr[(c)]	47	Not found	47	13
T4c[(c)]	28	23	28	6

of the degree of conservation, since there do exist some known TF binding motifs with ARE_P less than 1.0 (see Table 2).

The method proposed here is, nevertheless, not a replacement of the prevailing motif tools such as *MEME* and *AlignACE*. The major limitation of our method is its strong prerequisite. Multiple closely related species and the upstream sequences of each co-regulated gene for all species under study are requested, and in many cases these prerequisites may not be satisfied, so the method is, therefore, not generally applicable. Another problem is how to choose the appropriate species to evaluate the cross-species conservation. In principle, the species selected in the study should be close enough so that the conservation of motif sequences could be detected in a multiple alignment, in the meanwhile their evolutionary distances should not be too close, so that the signals could be distinguished from the noises [31]. The number of species used in the method is also a factor that may need

Table 5: The TFs whose known binding sites cannot be found using our method. The expected number of motifs predicted by *MEME* was set at 10.

TF	Known motif	Using *AlignACE*	Rank of the known motif/total predictions	Using *MEME*	Rank of the known motif/total predictions
Ap-1	TTANTAA	Not found		TTAGTAA	3/10
Bas2	TAATRA,TAANTAA	Not found		Not found	
Csre	YCGGAYRRAWGG	Not found		GTCCGGAC	8/10
Mac1	GAGCAAA	GGAAGCAAA	17/33	Not found	
Gcr1	CWTCC	ATTGTTTTCC	5/5	Not found	
Mcm1	CCNNNWWRGG	TTACCNNNTAGGAAA	2/11	TTTCCTAATTAGGAAA	1/10
Reb1	YYACCCG	TTACCCGCACGGC	3/8	Not found	
Rox1	YYNATTGTTY	Not found		Not found	
Scb	CNCGAAA	AAGCCACGAAAA	1/13	Not found	
Sff	GTMAACAA	Not found		Not found	

to be considered. We recommended three or four, since using too many species may bring up strong noise and reduce the detection power of the method.

After comparing with the motif finding software such as *MEME*, *AlignACE*, and *PHYME*, we can reach the following conclusions: (1) Our method screens for candidate motifs in terms of both overrepresentation and conservation, therefore, it gives relatively less predicted motifs for a group of co-regulated genes (Tables 3 and 4), hence it is helpful for reducing false positive predictions; (2) The rank of known motif in the output of our method is in general higher (Tables 3 and 4), and this is of practical importance, since we usually focus only on putative binding motifs with high ranks despite the large number of predicted motifs; (3) unlike the most common motif finding tools, our method requests no prior inputs such as the length of the motifs or the number of predictions. Although we used yeast genomes to assess our method, it could also be applied to other organisms if suitable related species are available and the upstream sequences of co-regulated genes could be obtained for the multiple species.

4. Materials and Methods

4.1. Materials

In this study, we considered gene promoter regions of four yeast species, namely, *S. cerevisiae*, *S. mikatae*, *S. bayanus*, and *S. paradoxus*. All these four are members of the *Saccharomyces sensu stricto* group. The last three are believed to be separated from *S. cerevisiae* by an estimated 5–20 million years of evolution and are found to have sufficient sequence similarity to *S. cerevisiae* such that orthologous regions can be aligned reliably [32].

We obtained the information about gene regulation network of *S. cerevisiae* from the database SCPD (The Promoter Database of Saccharomyces cerevisiae) [33], which

Table 6: Comparison to *PHYME*. For each TF, we listed the number of predicted motifs and the rank of the known motif in the predictions. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	Known motif	Found by our method	Rank	Found by *PHYME*	Rank
Ste12[(a)]	TGAAACA	TGAAACA	1	TGAAACA	3
Gal4[(a)]	CGGNNNNNNNNNNCCG	CGGNNNNNNNNNNCCG	1	CCGAATAGTCTGCCCCG	8
Mbp1[(a)]	ACGCGTNA	ACGCGT	1	ACGCGTCA	3
Leu3[(a)]	CCGGNNCCGG	CGGNNNCGG	1	CCGGTACCGG	3
Cbf1[(a)]	RTCACRTG	CACGTG	2	GTCACGTG	2
MET31[(a)]	CTGTGGC	TGTGGC	1	Not found	
Abf1[(b)]	TCRNNNNNNACG	TCANNNNNNACG	3	Not found	
Ace2[(b)]	GCTGGT	TGCTGGT	2	Not found	
Gcn4[(b)]	TGANTN	ATGACT	6	TGAGTC	6
Hap1[(b)]	CGGNNNTANCGG	TGCCGNNNNNNNCGG	1	Not found	
Ino4[(b)]	CATGTGAAAT	CATGTT	4	Not found	
Mcb[(b)]	WCGCGW	CGCNTCG	1	ACGCGT	1
Mse[(b)]	CRCAAAW	GACNCAA	5	CACAAAA	3
Nbf[(b)]	ATGYGRAWW	CATGTG	15	ATGTGAAAT	1
Pdr3[(b)]	TCCGYGGA	TCCNNGGA	1	TCCGCGGA	2
Pho4[(b)]	CACGTK	GCGCGT	1	Not found	
Put3[(b)]	CGGNNNNNNNNNNCCG	TCGNNNNNNNNNNNCG	1	Not found	
Rap1[(b)]	RMACCCA	GTCNNNNNCCCAT	11	AAACCGA	4
Swi5[(b)]	KGCTGR	TGCTGG	1	TGCTGAATG	1
Uasino[(b)]	ATCTGAAWW	CATGTG	13	Not found	
Uasrad[(b)]	WTTTCCCGS	TCCNGCT	3	TTTCCCAC	4
Adr1[(c)]	TCTCC	CTCCNNNNNTCC	2	ACTCC	4
Mig1[(c)]	CCCCRNNWWWWW	ACCCCA	30	CCCCGCCCC	4
Uasphr[(c)]	CTTCCT	TCTNNNNNNNNNNTCCT	13	GCTTTCTTT	8
T4c[(c)]	TTTTCTYCG	TTTTCNNNNNTCC	6	TTTTTCTTTT	1
Ap-1	TTANTAA	Not found		Not found	
Bas2	TAATRA,TAANTAA	Not found		TAATAG	8
Csre	YCGGAYRRAWGG	Not found		Not found	
Mac1	GAGCAAA	Not found		GAGAAAA	3
Gcr1	CWTCC	Not found		Not found	
Mcm1	CCNNNWWRGG	Not found		CCGTTTTGGG	5
Reb1	YYACCCG	Not found		CTACCCG	5
Rox1	YYNATTGTTY	Not found		Not found	
Scb	CNCGAAA	Not found		CACGAAA	1
Sff	GTMAACAA	Not found		GTTAACAA	6

Table 7: The TFs whose known binding motifs cannot be found by *MEME*/*AlignACE*/*PHYME*, but can be found using our method. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	Genes in *PS*	Our method	Known motif
Ace2[(b)]	1	GCTGGT	TGCTGGT
Hap1[(b)]	4	TGCCGNNNNNNNCGG	CGGNNNTANCGG
Mse[(b)]	1	GNCACAA	CRCAAAW
Put3[(b)]	1	TCGNNNNNNNNNNNCG	CGGNNNNNNNNNNCCG
Swi5[(b)]	2	TGCTGG	KGCTGR
Uasino[(b)]	1	CATGTG	ATCTGAAWW
Uasphr[(c)]	14	TCTNNNNNNNNNNTCCT	CTTCCT

contained TFs and genes co-regulated by them. The upstream region sequences of the co-regulated genes of each TF for all the four yeast species were downloaded from http://www.broad.mit.edu/.

The genes known to be co-regulated by specific TFs such as *STE12* and *GAL4* were used to evaluate the method. We let *PS* (positive set) denote the collection of *S. cerevisiae* genes co-regulated by a common TF, and we built an *NS* (negative set) by randomly selected *S. cerevisiae* genes. For each gene in both *PS* and *NS,* we extracted the promoter region sequences for all the four species and aligned them using multiple sequence alignment program *ClustalW*.

4.2. Methods

The method proposed here requests promoter region sequences from multiple closely related ortholog species. Usually we are interested in motif finding for only one of the species, namely, the principal species, whereas the sequences from other species are helpful for the reduction of false positives. For a given TF, we need two sets of genes, namely, positive set (*PS*) and negative set (*NS*). *PS* consists of the genes co-regulated by the TF, whereas *NS* consists of genes randomly collected from the genome of the principal species.

4.3. Finding Overrepresented Sequences

We only consider the principal species for finding overrepresented sequences. We first search the promoter regions of the genes in *NS* for each possible sequence pattern of length M ($6 \leq M \leq 17$) that satisfying the following constraints: the first three nucleotides in the left flank and the last three nucleotides in the right flank are the core elements and fixed, between the two core elements there might be M_0 nucleotides ($M_0 = 0, 1, 2, \ldots, 11$) and each of them could be any of the nucleotides A, C, T, and G. So within the L-bp upstream region of a gene, there are $L - M + 1$ possible locations that can be occupied by a sequence pattern of length M. We call the fraction as the background probability of the given sequence pattern

$$p = \frac{c_n}{c}, \tag{4.1}$$

where c_n is the number of total occurrences of the pattern in the promoter regions of genes in NS, and c is the total number of possible locations for an M-bp sequence in promoter regions of the genes in NS. In the same way, we can obtain the number of the pattern occurrences K and the total number (N) of possible M-bp locations in the promoter regions of the genes in PS. Using binomial distribution, we can calculate the probability of the pattern occurring more than K times as following [27, 34]:

$$P = \sum_{k=K}^{N} \frac{N!}{(N-k)!k!} p^k (1-p)^{N-k}, \tag{4.2}$$

where p is the background probability. We choose the sequence patterns with P less than a threshold p^* (usually in magnitude of 10^{-6} after Bonferroni adjustment) for further analysis. If the overlap of two sequences is longer than 80% of one of the two, we eliminate the sequence with larger P-value from the collection of overrepresented sequences. Both DNA strands are considered when we calculate the number of occurrences for a given sequence in the upstream region. If the sequence is a palindrome, we just use the count in one strand as the total occurrence.

4.4. Bonferroni Adjustment

We used the Bonferroni adjustment to the multiple statistical tests for determining overrepresented sequences, so that it was more "difficult" for any single test to be significant. The adjustment was accomplished by setting the P-value threshold at the common significant level (usually 0.05 or 0.01) divided by the number of tests being performed. In our case, the p^* was set as 0.05 divided by the number of all possible sequences in the form of $NNNnn\ldots nnNNN$, where NNN stands for three fixed nucleotides and $nn\ldots nn$ stands for unfixed number (from 0 to 11) of nucleotides.

4.4.1. Relative Entropy and Conservation Criteria

Let α be the background nucleotide distribution and β the nucleotide distribution at a given position in a multiple sequence alignment. For the two probability distributions α and β, the relative entropy (also known as Kullback-Leibler "distance") is defined by [29, 35]

$$H(\beta||\alpha) = \sum_{i=1}^{4} \beta_i \log \frac{\beta_i}{\alpha_i}. \tag{4.3}$$

We can prove that relative entropy is always a nonnegative value, and it reflects the extent of the deviation of actual nucleotide distribution from background distribution at a given site in the alignment. The larger the value, the greater the deviation between the actual distribution and the background distribution at that site [29].

Given an overrepresented sequence O, we search for its occurrences in the alignment of upstream sequences from the four species for each gene in PS and NS, respectively. If we find an occurrence in the alignment of a gene in PS, we extract the corresponding sequence block from the alignment and put the four segments that form this block to a sequence set S_{OP}. Similarly, we also build a sequence set S_{ON} for genes in NS.

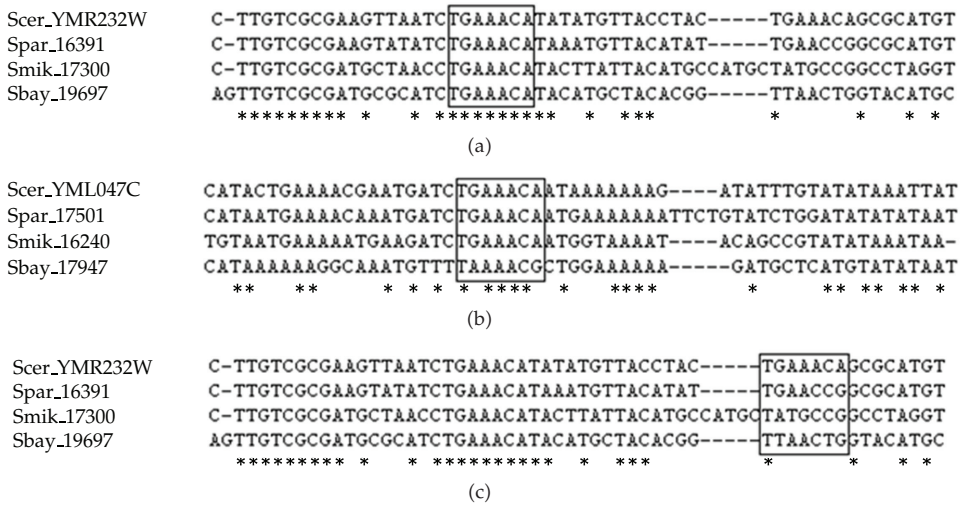


Figure 1: DNA binding motif *TGAAACA* of transcription factor *STE12*. The known DNA binding motif *TGAAACA* of transcription factor *STE12* for *S. cerevisiae* genes YML047C (a), YLR452C (b), and YCL055W (c) is conserved in the alignment of orthologous gene promoter regions of closely related yeast species, namely, *S. cerevisiae*, *S. mikatae*, *S. bayanus*, and *S. paradoxus*.

We further align all the sequences in S_{OP} and S_{ON}, respectively. These two alignments are used to evaluate the degree of conservation of O across closely related species. We define the average relative entropy (ARE_P) of S_{OP} as

$$\mathrm{ARE}_P = \frac{\sum_{i=1}^{M} EP_i}{M}, \tag{4.4}$$

where EP_i is the relative entropy at the position i of the alignment of the sequences in S_{OP}, and M is the length of O. If O is not found in the alignment of upstream sequences for any gene in NS, then we deposit O to the collectionof candidate motifs for further consideration. Otherwise, we could also calculate the average relative entropy ARE_N for the sequences in nonempty set S_{ON}. We define a Z-score as

$$Z = \frac{\mathrm{ARE}_P - \mathrm{ARE}_N}{\sqrt{s_N^2/M}}, \tag{4.5}$$

where s_N is the standard deviation of the relative entropies at different positions of the multiple upstream sequence (across multiple species) alignments of genes in NS.

Binding motifs tend to be conserved in the orthologous species (see Figures 1, 2, and 3), so we remove the sequences that are overrepresented but not conserved from our collection of candidate sequences. We set the Z-score threshold as 2, such that the sequences with $Z > 2$ are kept as the candidate sequences for further consideration.

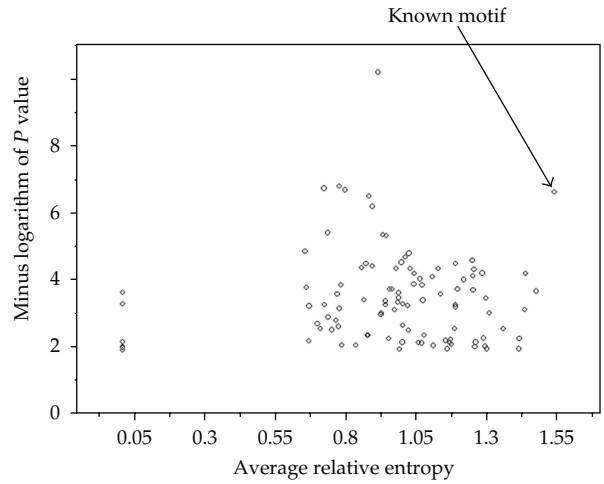


Figure 2: Average relative entropy of the motifs. We choose the 100 sequences found in the upstream of Ace2-regulated *S. cerevisiae* genes with the smallest overrepresentation *P*-values and compute their average relative entropies in the multiple alignments of orthologous upstream sequences of the four related species. The results were displayed with a scatter plot of *P*-value versus average relative entropy. The arrow points to the known binding motif *GCTGGT* of Ace2 in *S. cerevisiae*. The average relative entropy of the known binding motif is greater than that of most other sequences.

4.5. Building a Profile for a Candidate Sequence

Each candidate sequence will be searched for in the alignment of upstream sequences (from the multiple species) of each gene in *PS*. If an instance is found in any of the species, we extract the corresponding alignment block for further consideration. We use e_B to denote the average of the relative entropies at M different positions of an alignment block of length M. For each block, we set $h_P = \mu_P + 2(\sigma_P/\sqrt{M})$ as our cutoff value for block selection, where μ_P and σ_P are the mean and the standard deviation of the relative entropies, respectively, at different positions in the alignments of upstream sequences (from multiple species) of the genes in *PS*. For a given candidate sequence, we use all the blocks with e_B greater than h_P to build a profile to represent the candidate motif. For example, we search for a candidate sequence *GTTTCA* in the alignments of upstream sequences of genes in *PS*. If we can find it in any species in the alignments, we extract the corresponding alignment block, calculate e_B, and compare it with h_P to decide whether we keep this block for profile building. Using all the blocks selected, we calculate the base frequencies at each position and create thereafter the profile to represent the initial candidate motif. Both strands are considered when we build the profile.

4.6. Species-Specific PSSM Building

The profile obtained above represents the initial candidate motif derived from all the ortholog species. Usually we are only interested in the motif finding for one species, which is named as principal species in our analysis, and it is necessary to build a species-specific *PSSM* (Position Specific Score Matrix) for the candidate motif [36]. For the genes in *PS,* which are from the principal species, we search for the candidate motif in their upstream sequences in terms

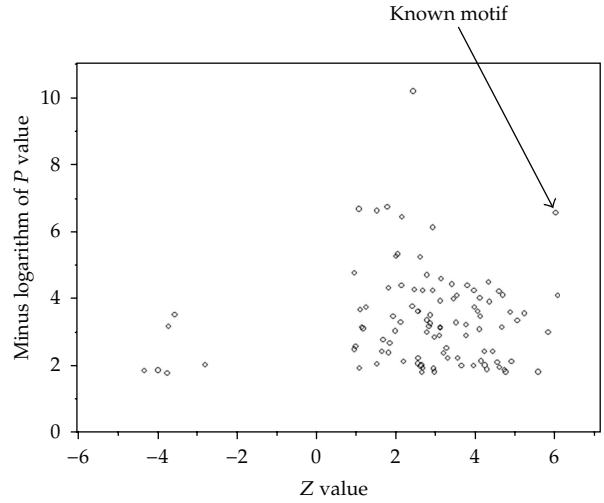


Figure 3: Z-value of the motifs. We choose the 100 sequences found in the upstream of *Ace2*-regulated *S. cerevisiae* genes with the smallest overrepresentation P-values and compute their Z-values in the multiple alignments of orthologous upstream sequences of the four related yeast species. The arrow points to the known motif *GCTGGT* of Ace2 in *S. cerevisiae*. The Z-value of the known binding motif is greater than that of the most other sequences.

of the initial motif profile, and all the significant hits found are used in building the final species-specific *PSSM*. The profile search is performed as follows. For each M-bp segment of upstream sequences of the genes in PS, we calculate a score

$$Sc = \prod_{i=1}^{M} q_i, \tag{4.6}$$

where q_i is the probability of observing the ith nucleotide of the segment, which is defined by the position-specific nucleotide distribution in the initial profile of the candidate motif. To determine the significance criterion, we calculate Scs for all the possible M-bp segments of the upstream sequences (for principal species only) of genes in NS and rank these scores in the descending order. We use the 0.001-quantile of these ranked scores, denoted as Sc^*, as the threshold value to determine whether a match is significant in the profile search. For example, if there are 1000 genes in the NS and the length of each promoter region is L-bp, then there are totally $1000^*(L-M+1)$ possible segments, so we have $1000^*(L-M+1)$ scores. We sort the scores in the descending order and set the nth value as the cutoff score Sc^* with $n = L-M+1$. We calculate Sc for each possible segment in the upstream sequences (principal species only) of the genes in PS. If $Sc \geq Sc^*$, we deposit the segment into I, which is the set of the incidences of the candidate motif.

4.7. Optimal Motif Length

Let k be the number of sequence segments in I. In order to determine the optimal length of the potential motif, we extend 0 to 5 bp in both flanks of each M-bp segment in I according to its mother sequence in the gene upstream region. So we have totally 36 possible combinations

(left flank extended by $M_L = 0, 1, 2, 3, 4,$ or 5 bp; right flank extended by M_R = 0, 1, 2, 3, 4, or 5 bp). For each possible combination (M_L, M_R), we put the newly added flanks into a block with k rows and $M_L + M_R$ columns. We calculate the average relative entropies of all 36 blocks and choose the combination (M_{L^*}, M_{R^*}) that delivers the maximum average relative entropy e_{B^*} for further consideration. In the meanwhile, we randomly generate 1000 sequence blocks, each with k rows and $M_{L^*} + M_{R^*}$ columns, in terms of the background nucleotide distribution α. We calculate the mean e_{rand} and the standard deviation s_{rand} of the average relative entropies of these 1000 blocks. If e_{B^*} is greater than $e_{\text{rand}} + 2s_{\text{rand}}$, then we accept the extension (M_{L^*}, M_{R^*}) and set the final motif length at $M + M_{L^*} + M_{R^*}$; otherwise, we still keep the original motif length M. The extended sequences (M_{L^*} bp in left flank and M_{R^*} bp in right flank) of the segments in I form a new sequence set I_e, which is the set of the incidences of the extended motif. Using all the sequences in I_e, we build the *PSSM* for a general representation of the final motif.

4.8. Implementation

We used a PERL script to implement the method. The script and the example of input data are available upon request.

Abbreviations

TFBS: Transcription factor binding sites
TF: Transcription factor
bp: Base pair
EM: Expectation maximization.

Conflict of Interests

The authors have declared that no competing interests exist.

Acknowledgments

This work is financially supported by the Nanyang Technological University Research Grant RG64/06 (to JMLI) and a start-up grant from Southern Medical University and Guangdong Province (to JMLI).

References

[1] T. I. Lee, N. J. Rinaldi, F. Robert et al., "Transcriptional regulatory networks in *Saccharomyces cerevisiae*," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.
[2] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.
[3] Elkan TLBaC, *Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers*, AAAI Press, Menlo Park, Calif, USA, 1994.
[4] C. E. Lawrence and A. A. Reilly, "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins*, vol. 7, no. 1, pp. 41–51, 1990.

[5] L. R. Cardon and G. D. Stormo, "Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments," *Journal of Molecular Biology*, vol. 223, no. 1, pp. 159–170, 1992.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] G. Thijs, K. Marchal, M. Lescot et al., "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *Journal of Computational Biology*, vol. 9, no. 2, pp. 447–464, 2002.

[8] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats," *Protein Science*, vol. 4, no. 8, pp. 1618–1632, 1995.

[9] S. Sinha and M. Tompa, "Discovery of novel transcription factor binding sites by statistical overrepresentation," *Nucleic Acids Research*, vol. 30, no. 24, pp. 5549–5560, 2002.

[10] T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 3, pp. 21–29, 1995.

[11] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, 1994.

[12] T. L. Bailey and M. Gribskov, "Combining evidence using *P*-values: application to sequence homology searches," *Bioinformatics*, vol. 14, no. 1, pp. 48–54, 1998.

[13] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnology*, vol. 16, no. 10, pp. 939–945, 1998.

[14] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*," *Journal of Molecular Biology*, vol. 296, no. 5, pp. 1205–1214, 2000.

[15] M. Blanchette and M. Tompa, "Discovery of regulatory elements by a computational method for phylogenetic footprinting," *Genome Research*, vol. 12, no. 5, pp. 739–748, 2002.

[16] L. A. McCue, W. Thompson, C. S. Carmack et al., "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes," *Nucleic Acids Research*, vol. 29, no. 3, pp. 774–782, 2001.

[17] M. Blanchette, B. Schwikowski, and M. Tompa, "Algorithms for phylogenetic footprinting," *Journal of Computational Biology*, vol. 9, no. 2, pp. 211–223, 2002.

[18] S. Sinha, M. Blanchette, and M. Tompa, "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences," *BMC Bioinformatics*, vol. 5, article 170, 2004.

[19] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.

[20] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[21] P. Qiu, "Recent advances in computational promoter analysis in understanding the transcriptional regulatory network," *Biochemical and Biophysical Research Communications*, vol. 309, no. 3, pp. 495–501, 2003.

[22] L. Ma, J. Li, L. Qu et al., "Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways," *Plant Cell*, vol. 13, no. 12, pp. 2589–2607, 2001.

[23] G. B. Fogel, D. G. Weekes, G. Varga et al., "Discovery of sequence motifs related to coexpression of genes using evolutionary computation," *Nucleic Acids Research*, vol. 32, no. 13, pp. 3826–3835, 2004.

[24] M. Caselle, F. Di Cunto, and P. Provero, "Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes," *BMC Bioinformatics*, vol. 3, article 7, 2002.

[25] L. Mao, C. Mackenzie, J. H. Roh, J. M. Eraso, S. Kaplan, and H. Resat, "Combining microarray and genomic data to predict DNA binding motifs," *Microbiology*, vol. 151, no. 10, pp. 3197–3213, 2005.

[26] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 6, pp. 3339–3344, 2003.

[27] P. M. Haverty, U. Hansen, and Z. Weng, "Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification," *Nucleic Acids Research*, vol. 32, no. 1, pp. 179–188, 2004.

[28] A. Prakash and M. Tompa, "Discovery of regulatory elements in vertebrates through comparative genomics," *Nature Biotechnology*, vol. 23, no. 10, pp. 1249–1256, 2005.
[29] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1998.
[30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
[31] S. W. Doniger, J. Huh, and J. C. Fay, "Identification of functional transcription factor binding sites using closely related Saccharomyces species," *Genome Research*, vol. 15, no. 5, pp. 701–709, 2005.
[32] E. Herrero, "Evolutionary relationships between *Saccharomyces cerevisiae* and other fungal species as determined from genome comparisons," *Revista Iberoamericana de Micologia*, vol. 22, no. 4, pp. 217–222, 2005.
[33] J. Zhu and M. Q. Zhang, "SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*," *Bioinformatics*, vol. 15, no. 7-8, pp. 607–611, 1999.
[34] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor, "Toucan: deciphering the cis-regulatory logic of coregulated genes," *Nucleic Acids Research*, vol. 31, no. 6, pp. 1753–1764, 2003.
[35] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, New York, NY, USA, 1959.
[36] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2004.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 320425, 19 pages
doi:10.1155/2012/320425

Research Article

Control of the False Discovery Proportion for Independently Tested Null Hypotheses

Yongchao Ge[1] and Xiaochun Li[2]

[1] *Department of Neurology, Mount Sinai School of Medicine, One Gustave L. Levy Place, P.O. Box 1137, New York, NY 10029, USA*
[2] *Division of Biostatistics, School of Medicine, New York University, 650 First Avenue, 5th Floor, New York, NY 10016, USA*

Correspondence should be addressed to Yongchao Ge, yongchao.ge@mssm.edu

Received 14 December 2011; Accepted 8 February 2012

Academic Editor: Yongzhao Shao

Consider the multiple testing problem of testing m null hypotheses $H_1, \ldots, H_m$, among which m_0 hypotheses are truly null. Given the P-values for each hypothesis, the question of interest is how to combine the P-values to find out which hypotheses are false nulls and possibly to make a statistical inference on m_0. Benjamini and Hochberg proposed a classical procedure that can control the false discovery rate (FDR). The FDR control is a little bit unsatisfactory in that it only concerns the expectation of the false discovery proportion (FDP). The control of the actual random variable FDP has recently drawn much attention. For any level $1-\alpha$, this paper proposes a procedure to construct an upper prediction bound (UPB) for the FDP for a fixed rejection region. When $1 - \alpha = 50\%$, our procedure is very close to the classical Benjamini and Hochberg procedure. Simultaneous UPBs for all rejection regions' FDPs and the upper confidence bound for the unknown m_0 are presented consequently. This new proposed procedure works for finite samples and hence avoids the slow convergence problem of the asymptotic theory.

1. Introduction

In this paper, we consider the problem of testing m null hypotheses $H_1, \ldots, H_m$, among which m_0 hypotheses are truly null. We shall assume that P-values are available for individual hypotheses. In a seminal paper, Benjamini and Hochberg [1] proposed the false discovery rate (FDR) as an alternative to the classically defined family-wise error rate (FWER). The proposed FDR achieves a good balance between the P-value itself and the FWER correction [2]; the former may give too many false positives, and the latter may give too many false negatives. However, the control of the FDR is a little bit unsatisfactory in that it only concerns the expectation of the false discovery proportion (FDP). In practice, researchers may be

interested in more detailed statistical inference on the actual random variable FDP, not just its expectation. The goal of this paper is to provide a simple procedure to control the FDP.

Let us first introduce some notation. Given m hypotheses $H_1, \ldots, H_m$, let the complete index set be $M = \{1, \ldots, m\}$, M_0 the unknown subset of M for which the null hypotheses are true, and $M_1 = M \setminus M_0$ the subset for which null hypotheses are false. Denote that $m_0 = |M_0|, m_1 = |M_1|$, where $|\cdot|$ denotes the cardinality of a set. The P-values for testing the m hypotheses are $P_1, \ldots, P_m$. A *fixed rejection region* for the P-values can conveniently be taken as $[0, t]$ $(0 < t < 1)$. The value of t could be 0.05, for example. Define the number R_t of all rejected hypotheses and the number V_t of falsely rejected hypotheses, respectively,

$$R_t = \sum_{i=1}^{m} I(P_i \le t), \qquad V_t = \sum_{i \in M_0} I(P_i \le t). \tag{1.1}$$

Following the notation of Korn et al. [3], and Genovese and Wasserman [4], Lehmann and Romano [5], the *false discovery proportion* is defined to be the proportion of falsely rejected null hypotheses among the rejected ones,

$$Q_t = \frac{V_t}{R_t}, \tag{1.2}$$

where the ratio is defined to be zero when the denominator is zero. For a given fixed rejection region $[0, t]$, R_t, V_t, and Q_t are random variables. R, V, and Q will be shorthand for R_t, V_t, and Q_t respectively, when the rejection region $[0, t]$ is clear from the context. The *false discovery rate* of Benjamini and Hochberg [1] is

$$\mathrm{FDR} = E(Q). \tag{1.3}$$

A good understanding of Q will lead investigators to pick an appropriate rejection region $[0, t]$ of P-values. As Q is an unobservable random variable depending on the observed P-values and the rejection region $[0, t]$, the quantity FDR just describes the expectation of this random variable Q. One way to have a more detailed statistical inference on the random variable Q is to derive its distribution, which is very difficult unless a strong assumption can be imposed on the P-values from the false null hypotheses. A conservative approach is to compute an upper prediction bound for Q so that we can safeguard against excessive type I errors. In Section 2, for a fixed rejection region $[0, t]$, we can compute an *upper prediction bound* (UPB) $\overline{Q}_{1-\alpha}(t)$ for Q_t such that

$$\mathrm{pr}\left(Q_t \le \overline{Q}_{1-\alpha}(t)\right) \ge 1 - \alpha. \tag{1.4}$$

If $\overline{Q}_{1-\alpha}(t)$ had been a nonrandom variable, then it should be always no less than the $1 - \alpha$ quantile of the random variable Q_t. When $1 - \alpha = 50\%$, our procedure is very close to the classical BH procedure of Benjamini and Hochberg [1]. In other words, the BH procedure gives us an approximate 50% upper prediction bound (UPB) for Q. With different degrees of being conservative, one should take $1 - \alpha$ at 0.9, 0.95, and 0.99 to ensure high coverage of the false discovery proportion. We also describe how to compute an upper confidence bound

(UCB) for m_0, the number of true null hypotheses. The UCB for m_0 can be used to improve the estimate $\overline{Q}_{1-\alpha}(t)$. In practice, the rejection region $[0,t]$ needs to be adapted to the actual dataset. In Section 3, we give a procedure to construct an upper prediction band for Q_t for all $t \in (0,1)$, and this upper prediction band can be used to pick a data-defined rejection region $[0,\tau]$ of P-values such that the false discovery proportion Q can be controlled at target level γ with prediction accuracy $1-\alpha$, that is,

$$\mathrm{pr}\left(Q_\tau \le \gamma\right) \ge 1-\alpha. \tag{1.5}$$

Thus with probability at least $1-\alpha$, the value of Q is γ or less. For the independent true null P-values, Genovese and Wasserman [4], Meinshausen and Rice [6] also worked on the control of the FDP in the sense of the above equation. However, their results are based on asymptotic theory, while our focus is on the finite-sample results and avoids the slow convergence problem of the asymptotic theory. Other works such as Lehmann and Romano [5], Romano and Shaikh [7, 8], and van der Laan et al. [9] proposed procedures that allow dependence in the P-values but have potentially lost statistical power as the dependence information is not exploited. Section 4 presents a focused statistical inference by restricting the rejection regions onto $\{[0,t] : t \in [t_0, t_0']\}$, which unifies the results of Sections 2 and 3. Section 5 generalizes the results from independent data to less-independent situations, and Section 6 gives our discussion.

2. Finding a $1-\alpha$ UPB for the False Discovery Proportion for a Fixed Rejection Region

For the sake of simplicity, we will first assume that the P-values from the true null hypotheses are following mutually independently uniform distribution $U[0,1]$. We have no further assumptions on the P-values from false null hypotheses. This assumption is the same as in Benjamini and Hochberg [1]. In Section 5 we will generalize the result to less independent situations. For a fixed rejection region $[0,t]$ of the P-values, we would like to find the $1-\alpha$ upper prediction bound (UPB) for the false discovery proportion Q_t. As we mentioned in Section 1, the distribution of Q_t is unknown. However, for any given experimental data, the total number of rejections, R_t, can be easily obtained by (1.1). Under the assumption that true null P-values are independently distributed as $U[0,1]$, V_t has a binomial distribution $\mathrm{Bin}(m_0,t)$. Let U_i, $i = 1,\ldots,m$ be random variables mutually independently distributed as $U[0,1]$, and $N_{m_0,t} = \sum_{i=1}^{m_0} I(U_i \le t)$ distributed as $\mathrm{Bin}(m_0,t)$, hence,

$$V_t \overset{d}{=} N_{m_0,t}. \tag{2.1}$$

The $1-\alpha$ quantile for V_t is the $1-\alpha$ quantile $C_{1-\alpha}(m_0,t)$ of the distribution $\mathrm{Bin}(m_0,t)$. Here $C_{1-\alpha}(m_0,t)$ is defined as

$$C_{1-\alpha}(m_0,t) = \min\left\{k : \mathrm{pr}(N_{m_0,t} \le k) \ge 1-\alpha\right\}. \tag{2.2}$$

As R_t can be computed from the observed data, a $1-\alpha$ UPB for Q_t can be estimated by

$$\overline{Q}_{1-\alpha}(m_0,t)=\frac{C_{1-\alpha}(m_0,t)}{R_t}. \tag{2.3}$$

Lemma 2.1. *For any given* $0\le m_1\le m_2\le m$,

(a) $C_{1-\alpha}(m_1,t)\le C_{1-\alpha}(m_2,t)$,

(b) $m_1-C_{1-\alpha}(m_1,t)\le m_2-C_{1-\alpha}(m_2,t)$, *and*

(c) *let* $g(k)=C_{1-\alpha}(k,t)$ *and* $h(k)=k-C_{1-\alpha}(k,t)$. *The values that* $g(k+1)-g(k)$ *and* $h(k+1)-h(k)$ *take can only be zero or one.*

Proof. By noting that $N_{m_1,t}\le N_{m_2,t}$ when $m_1\le m_2$, we have $\mathrm{pr}(N_{m_1,t}\le k)\ge \mathrm{pr}(N_{m_2,t}\le k)$. Applying the definition of $C_{1-\alpha}$, we obtain the result for part (a).

Note that

$$\begin{aligned}C_{1-\alpha}(m_1,t)&=\min\{k:\mathrm{pr}(N_{m_1,t}\le k)\ge 1-\alpha\}\\&=\min\{k:\mathrm{pr}(m_1-N_{m_1,t}\ge m_1-k)\ge 1-\alpha\}\\&=m_1-\max\{k':\mathrm{pr}(m_1-N_{m_1,t}\ge k')\ge 1-\alpha\}\quad(\text{use } k' \text{ to replace } m_1-k).\end{aligned} \tag{2.4}$$

Therefore,

$$m_1-C_{1-\alpha}(m_1,t)=\max\{k':\mathrm{pr}(m_1-N_{m_1,t}\ge k')\ge 1-\alpha\}. \tag{2.5}$$

Similarly, we can obtain that

$$m_2-C_{1-\alpha}(m_2,t)=\max\{k':\mathrm{pr}(m_2-N_{m_2,t}\ge k')\ge 1-\alpha\}. \tag{2.6}$$

By noting that

$$m_1-N_{m_1,t}=\sum_{i=1}^{m_1}I(U_i>t)\le\sum_{i=1}^{m_2}I(U_i>t)=m_2-N_{m_2,t}, \tag{2.7}$$

we have

$$\mathrm{pr}(m_1-N_{m_1,t}\ge k')\le\mathrm{pr}(m_2-N_{m_2,t}\ge k'). \tag{2.8}$$

Combining this with (2.5) and (2.6) leads to the result for part (b).

Parts (a) and (b), respectively, say that $g(k)$ and $h(k)$ are both increasing functions of k. Simple algebra can establish that $\{g(k+1)-g(k)\}+\{(h(k+1)-h(k)\}=1$. Both $\{g(k+1)-g(k)\}$ and $\{h(k+1)-h(k)\}$ are nonnegative due to the increasing property of functions $g(k)$ and $h(k)$, and hence $0\le g(k+1)-g(k)\le 1$ and $0\le h(k+1)-h(k)\le 1$. The only values that $\{g(k+1)-g(k)\}$ and $\{h(k+1)-h(k)\}$ take can only be zero and one as functions $g(k)$ and $h(k)$ only take integer values. Thus, we complete the proof for part (c). □

Lemma 2.2. *For any given t, $0 < t < 1$, $\overline{Q}_{1-\alpha}(m_0, t)$ of (2.3) is a $1-\alpha$ UPB for the false discovery proportion Q_t, that is,*

$$\mathrm{pr}\Big(Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\Big) \ge 1-\alpha. \tag{2.9}$$

The proof is straightforward by using the fact that $V_t \overset{d}{=} N_{m_0,t}$. We have

$$\begin{aligned}
\mathrm{pr}\Big(Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\Big) &= \mathrm{pr}\left(\frac{V_t}{R_t} \le \frac{C_{1-\alpha}(m_0, t)}{R_t},\ R_t > 0\right) + \mathrm{pr}(R_t = 0) \\
&= \mathrm{pr}(V_t \le C_{1-\alpha}(m_0, t),\ R_t > 0) + \mathrm{pr}(R_t = 0) \\
&= \mathrm{pr}(V_t \le C_{1-\alpha}(m_0, t)) \\
&\ge 1-\alpha.
\end{aligned} \tag{2.10}$$

In the third line, we have used the fact that $\{V_t \le C_{1-\alpha}(m_0, t)$ and $R_t = 0\}$ is the same as the set $\{R_t = 0\}$, which is obtained by noting that V_t must be zero when $R_t = 0$. Following this proof, we can easily see that

$$\left\{Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\right\} = \{V_t \le C_{1-\alpha}(m_0, t)\}. \tag{2.11}$$

The basic construction of $\overline{Q}_{1-\alpha}(m_0, t)$ in (2.3) is the idea central to formulating prediction inference for Q_t. In practice, m_0 is an unknown parameter. The most conservative approach is to replace m_0 with m, in which case we obtain a *conservative* $1-\alpha$ UPB for Q_t. The independence assumption among true null P-values can be used to give a confidence inference for m_0; thus, we can find a better estimate of the UPB for Q_t. For any given $0 < \lambda < 1$, a $1-\alpha$ UCB for m_0 is given by

$$\overline{m}_{0,1-\alpha}(\lambda) = \begin{cases} \max\limits_{k=0,\ldots,m-1}\{k : h(k) = m - R_\lambda\} & \text{if } h(m) < m - R_\lambda, \\ m & \text{otherwise,} \end{cases} \tag{2.12}$$

where $h(k) = k - C_{1-\alpha}(k, \lambda)$ as defined in Lemma 2.1(c). Since $h(0) = 0$ and $h(k+1) - h(k)$ takes value of only zero and one, there exists at least one k, $k \in \{0, \ldots, m-1\}$ such that $h(k) = m - R_\lambda$ when $h(m) < m - R_\lambda$. Therefore, $\overline{m}_{0,1-\alpha}(\lambda)$ in (2.12) is well defined. The parameter λ in (2.12) is used to construct a UCB for m_0; more discussion about it can be seen in Remark 2.6 of the following theorem.

Theorem 2.3. (a) *$\overline{m}_{0,1-\alpha}(\lambda)$ is a conservative $1-\alpha$ UCB for m_0, that is,*

$$\mathrm{pr}(m_0 \le \overline{m}_{0,1-\alpha}(\lambda)) \ge 1-\alpha. \tag{2.13}$$

(b) *Especially, if λ takes the same value as t in the P-value rejection region, then*

$$\mathrm{pr}\Big(Q_t \le \overline{Q}_{1-\alpha}(\overline{m}_{0,1-\alpha}(t), t),\ m_0 \le \overline{m}_{0,1-\alpha}(t)\Big) \ge 1-\alpha. \tag{2.14}$$

Proof. Use $\overline{m}_0$ as a shorthand of $\overline{m}_{0,1-\alpha}(\lambda)$ in this proof. We want to establish that

$$\{m_0 \le \overline{m}_0\} = \{h(m_0) \le h(\overline{m}_0)\}. \tag{2.15}$$

The fact that function $h(k)$ is increasing in k leads to $\{m_0 \le \overline{m}_0\} \subseteq \{h(m_0) \le h(\overline{m}_0)\}$. On the other hand, if $m_0 > \overline{m}_0$, then $\overline{m}_0$ is strictly less than m, and we must have $h(\overline{m}_0) = m - R_\lambda$ according to (2.12). $\overline{m}_0$ is the maximum of k such that $h(k) = m - R_\lambda$, and hence $h(m_0) \neq m - R_\lambda = h(\overline{m}_0)$ as $m_0 > \overline{m}_0$. The increasing property of $h(k)$ leads to $h(m_0) \ge h(\overline{m}_0)$. Combining this with $h(m_0) \neq h(\overline{m}_0)$, we obtain that $h(m_0) > h(\overline{m}_0)$; therefore, we conclude

$$\{m_0 > \overline{m}_0\} \subseteq \{h(m_0) > h(\overline{m}_0)\} \tag{2.16}$$

and complete the proof of (2.15).

Note that

$$\begin{aligned}
\{h(m_0) \le h(\overline{m}_0)\} &= \{h(m_0) \le m - R_\lambda, \overline{m}_0 < m\} \cup \{h(m_0) \le h(m),\ \overline{m}_0 = m\} \\
&= \{h(m_0) \le m - R_\lambda\ , \overline{m}_0 < m\} \cup \{\overline{m}_0 = m\} \quad \big(h(m_0) \le h(m) \text{ always holds}\big) \\
&\supseteq \{h(m_0) \le m - R_\lambda\},
\end{aligned} \tag{2.17}$$

we have

$$\begin{aligned}
\operatorname{pr}(m_0 \le \overline{m}_0) &= \operatorname{pr}\big(h(m_0) \le h(\overline{m}_0) \ge \operatorname{pr}(m_0 - C_{1-\alpha}(m_0, \lambda) \le m - R_\lambda)\big) \\
&\ge \operatorname{pr}(m_0 - C_{1-\alpha}(m_0, \lambda) \le m_0 - V_\lambda) \quad (\text{using } m - R_\lambda \ge m_0 - V_\lambda) \\
&= \operatorname{pr}(V_\lambda \le C_{1-\alpha}(m_0, \lambda)) \\
&\ge 1 - \alpha \quad \Big(\text{Note that } V_\lambda \overset{d}{=} N_{m_0,\lambda}\Big).
\end{aligned} \tag{2.18}$$

Hence, we have the proof of part (a). When λ is set to be t, we have that the set $\{m_0 \le \overline{m}_{0,1-\alpha}(t)\}$ contains $\{V_t \le C_{1-\alpha}(m_0, t)\}$ from the above derivation, and that $\{Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\} = \{V_t \le C_{1-\alpha}(m_0, t)\}$ from the derivation of Lemma 2.2. Therefore,

$$\begin{aligned}
\left\{Q_t \le \overline{Q}_{1-\alpha}(\overline{m}_{0,1-\alpha}(t), t),\ m_0 \le \overline{m}_{0,1-\alpha}(t)\right\} &\supseteq \left\{Q_t \le \overline{Q}_{1-\alpha}(m_0, t),\ m_0 \le \overline{m}_{0,1-\alpha}(t)\right\} \\
&\supseteq \{V_t \le C_{1-\alpha}(m_0, t)\}.
\end{aligned} \tag{2.19}$$

Thus, we have the proof of part (b) of this theorem. □

Remark 2.4. Theorem 2.3 gives researchers a good sense of the total number m_0 of true null hypotheses. Other papers, for example, Storey et al. [10], Benjamini and Hochberg [11], and Langaas et al. [12], gave only point estimates of m_0 or $\pi_0 = m_0/m$. Part (a) gives a confidence inference for m_0, and part (b) gives a simultaneous statement for the Q_t and m_0, which is more interesting. Meinshausen [13] gives a confidence for m_0 by using resampling methods, while ours exploited the independence information so that it works for finite samples.

Remark 2.5. Theorem 2.3 of Göb [14] implies that for a binomial distribution, the difference between the median and mean is less than 1, that is, $|C_{0.5}(m_0, t) - m_0 t| < 1$. From (2.3), we know the 50% UPB for the Q_t can be estimated by $C_{0.5}(m_0, t)/R_t$. This UPB for Q_t is very close to $m_0 t/R_t$ with a difference smaller than $1/R_t$. Replacing m_0 by m in $m_0 t/R_t$ is equivalent to the classical BH procedure. For a very large R_t, the term $1/R_t$ can be ignored, and the BH procedure offers an approximate estimate of the 50% UPB for Q_t.

Remark 2.6. When k is large, the distribution $\mathrm{Bin}(k, \lambda)$ can be closely approximated by $N(k\lambda, k\lambda(1-\lambda))$. Let $z_{1-\alpha}$ be the $1-\alpha$ quantile of a standard normal distribution. After some algebraic manipulations, we obtain a $1-\alpha$ UCB for m_0

$$\overline{m}_{0,1-\alpha}(\lambda) \approx \left\{ \frac{1}{2(1-\lambda)} \left(z_{1-\alpha}\sqrt{\lambda(1-\lambda)} + \sqrt{z_{1-\alpha}^2 \lambda(1-\lambda) + 4(m - R_\lambda)(1-\lambda)} \right) \right\}^2. \tag{2.20}$$

Taking $1-\alpha = 0.5$, we have $(m - R_\lambda)/(1-\lambda)$, which is equivalent to (2.3) of Storey et al. [10]. For most practical applications, one can set the value of $\lambda = 0.5$. Fine tuning of the parameter λ will be discussed in Section 3.3.

When the rejection region $[0, t]$ is small, the UCB for m_0 obtained from part (b) of Theorem 2.3 may be too conservative. It may be advantageous to have separate values for λ and t. Part (a) of Theorem 2.3 implies the following.

Corollary 2.7. *Replacing m_0 by its the upper confidence bound $\overline{m}_{0,1-\alpha_2}(\lambda)$ and α by α_1 in (2.3), we define*

$$\overline{Q}_{1-\alpha_1,1-\alpha_2}(t, \lambda) = \frac{C_{1-\alpha_1}(\overline{m}_{0,1-\alpha_2}(\lambda), t)}{R_t}. \tag{2.21}$$

Then, $\overline{Q}_{1-\alpha_1,1-\alpha_2}(t, \lambda)$ is a conservative $1-\alpha$ $(\alpha = \alpha_1 + \alpha_2)$ UPB for the false discovery proportion Q_t.

3. Upper Prediction Bounds and Simultaneous Inferences

3.1. The Setup

In Section 2, the UPBs for Q are only valid for a *fixed* rejection region $[0, t]$ of P-values. In practice, researchers will not fix the rejected region $[0, t]$ but adapt it to the actual data. The logic is the same as with single hypothesis testing. In single hypothesis testing with nested rejection regions $\{\Gamma_\alpha : 0 < \alpha < 1\}$, for an observed statistic T, one will find the rejection region that contains the observed statistic with the smallest type I error α, that is,

$$P\text{-value}(T) = \min\{\alpha : T \in \Gamma_\alpha\}. \tag{3.1}$$

The same logic can be applied to our false discovery proportion. In this case, we will try to find the largest rejection region $[0, t]$ such that the false discovery proportion Q is not more than γ, say 10%, with probability $1-\alpha$. Define

$$\tau = \max\left\{ t : \overline{Q}_{1-\alpha_1,1-\alpha_2}(t, \lambda) \le \gamma \right\}. \tag{3.2}$$

We then reject any hypothesis whose P-value is no greater than τ. If τ is independent of Q and $\overline{Q}$, then we can expect that

$$\mathrm{pr}(Q_\tau \le \gamma) \ge \mathrm{pr}\left(Q_\tau \le \overline{Q}_{1-\alpha_1,1-\alpha_2}(\tau,\lambda)\right) \ge 1-\alpha. \tag{3.3}$$

Asymptotically, τ and $(Q,\overline{Q})$ may be independent: this question is open for future research. To overcome the independence assumption of τ and $(Q,\overline{Q})$, we seek an alternative approach: to find simultaneous UPBs for all rejection regions $[0,t]$, $t \in (0,1)$, that is, to find an *upper prediction band* $\overline{Q}_t$ such that

$$\mathrm{pr}\left(Q_t \le \overline{Q}_t \text{ for } t \in (0,1)\right) \ge 1-\alpha. \tag{3.4}$$

Hence we have the simultaneous inferences on Q_t for each rejection region $[0,t]$, $t \in (0,1)$. Following the definition of $C_{1-\alpha}(n,t)$ in (2.2) to construct the UPB for Q_t, we want to define the simultaneous critical values of $N_{n,t}$. Using the distribution of $\max_{t\in(0,1)} N_{n,t}$ is unwise as $\max_{t\in(0,1)} N_{n,t} = N_{n,1}$, which takes value n with probability one. A better approach is to center $N_{n,t}$, that is,

$$\sup_{t\in(0,1)} (N_{n,t} - nt). \tag{3.5}$$

This leads to a test statistic related to the Kolmogorov-Smirnov test statistic, which gives an upper confidence band for a cumulative distribution function $F(x)$. It turns out that this method leads to very high UPBs when t is close to zero or one. Therefore, we normalize $N_{n,t}$, that is,

$$\widetilde{Z}_n = \frac{\sup_{t\in(0,1)}(N_{n,t} - nt)}{nt(1-t)}. \tag{3.6}$$

Note that $\widetilde{Z}_n$ is continuously distributed even though each $N_{n,t}$ is discretely distributed. Let $\widetilde{z}_{1-\alpha}(n)$ be the $1-\alpha$ quantile of $\widetilde{Z}_n$, that is,

$$\mathrm{pr}\left(\widetilde{Z}_n \le \widetilde{z}_{1-\alpha}(n)\right) = 1-\alpha. \tag{3.7}$$

We can then redefine $\overline{Q}$ as

$$\widetilde{Q}_{1-\alpha}(m_0,t) = \frac{m_0 t + \widetilde{z}_{1-\alpha}(m_0)\sqrt{m_0 t(1-t)}}{R_t}. \tag{3.8}$$

Corresponding to Lemma 2.2 and Corollary 2.7, we have similar results below.

Corollary 3.1. *For any given* $0 < t < 1$, $\widetilde{Q}_{1-\alpha}(m_0, t)$ *of* (3.8) *is an exactly* $1-\alpha$ *upper prediction band for the false discovery proportion* Q_t, *that is,*

$$\mathrm{pr}\Big(Q_t \le \widetilde{Q}_{1-\alpha}(m_0, t)\ \forall t \in (0,1)\Big) = 1-\alpha. \tag{3.9}$$

Corollary 3.2. *Denote that* $\overline{m}_0 = \overline{m}_{0,1-\alpha_2}(\lambda)$. *Define*

$$\widetilde{Q}_{1-\alpha_1,1-\alpha_2}(t,\lambda) = \frac{\overline{m}_0 t + \widetilde{z}_{1-\alpha_1}(\overline{m}_0)\sqrt{\overline{m}_0 t(1-t)}}{R_t}. \tag{3.10}$$

Let $\alpha = \alpha_1 + \alpha_2$. *Then* $\widetilde{Q}_{1-\alpha_1,1-\alpha_2}(t,\lambda)$ *is a conservative* $1-\alpha$ *upper prediction band for the false discovery proportion* Q_t, *that is,*

$$\mathrm{pr}\Big(Q_t \le \widetilde{Q}_{1-\alpha_1,1-\alpha_2}(t,\lambda)\ \forall t \in (0,1)\Big) \ge 1-\alpha. \tag{3.11}$$

Remark 3.3. Using the same idea as in the proof of Lemma 2.2, the proof of the above corollaries is straightforward after converting the comparison between Q and $\widetilde{Q}$ to the comparison between V_t and $m_0 t + \widetilde{z}_{1-\alpha}(m_0)\sqrt{m_0 t(1-t)}$. This conversion provides a powerful tool for understanding the false discovery proportion.

Remark 3.4. The formulation of $\widetilde{Q}_{1-\alpha}(m_0, t)$ in (3.8) is motivated by the normal approximation of $N_{n,t}$. But our definition of $\widetilde{Q}_{1-\alpha}(m_0, t)$ gives *exact* UPBs simultaneously for all $t \in (0,1)$ due to the exactness of the quantile $\widetilde{z}_{1-\alpha}$.

Remark 3.5. Meinshausen and Rice [6] and Donoho and Jin [15] also utilize the empirical process $\widetilde{Z}_n$. However, they focus on the asymptotic theory for $\widetilde{Z}_n$, which may face the slow convergence problem described in the next section. Our focus is on the finite sample control.

Remark 3.6. Let $f(t) = (k-nt)/\sqrt{nt(1-t)}$. After some simplifications, we have $f'(t) \cdot \left(\sqrt{nt(1-t)}\right)^3 = -n^2 t(1-t) - 1/2 \cdot (k-nt)[n(1-2t)] = -n/2 \cdot [k(1-t) + (n-k)t]$, and then $f(t)$ is a decreasing function in t, $0 < t < 1$ for $0 \le k \le n$. Equation (3.6) can be simplified to be

$$\max_{k=1,\ldots,n} \sup_{t \in [U_{(k)}, U_{(k+1)})} \frac{k-nt}{\sqrt{nt(1-t)}} = \max_{k=1,\ldots,n} \frac{k - nU_{(k)}}{\sqrt{nU_{(k)}\left(1-U_{(k)}\right)}}, \tag{3.12}$$

where $U_{(k)}$ is the kth smallest ordered one among the n samples of $U[0,1]$ distribution. This formulation facilitates the computation of the distribution of $\widetilde{Z}_n$ by Monte Carlo methods. The standard error associated with the Monte Carlo simulations in computing the probability in (3.7) is no greater than $\sqrt{\alpha(1-\alpha)/B}$, where B is the number of simulations.

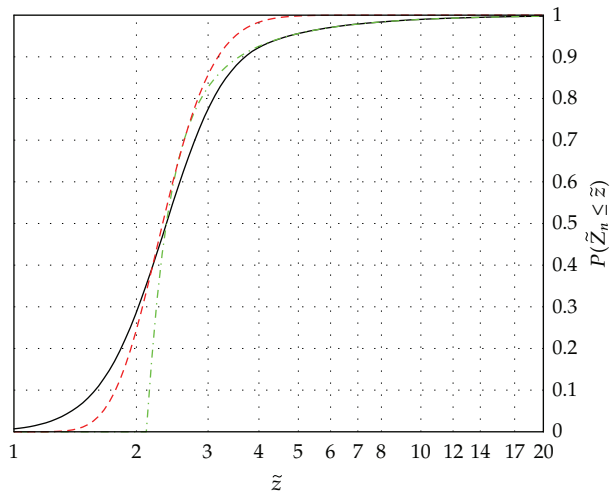


Figure 1: Plot of the probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$ for $1 \leq \tilde{z} \leq 20$ with $n = 10^5$. The x-axis is plotted on a log scale. The black solid curve is computed from 10^6 Monte Carlo simulations. The red-dashed curve is based on (3.13). The green dot-dashed curve is computed from (3.14).

3.2. Computing the Distribution of $\tilde{Z}_n$

In order to make simultaneous inferences, we need to know the distribution of $\tilde{Z}_n$ defined in (3.6). Example 1 of Jaeschke [16] showed that asymptotically, for any x,

$$\lim_{n\to\infty} \text{pr}\left(\tilde{Z}_n \leq \frac{x + 2\ln\ln n + (1/2)\ln\ln\ln n - (1/2)\ln \pi}{\sqrt{2\ln\ln n}}\right) = \exp\left[-\exp(-x)\right]. \tag{3.13}$$

This implies that $\tilde{Z}_n / \sqrt{2\ln\ln n}$ converges to 1 in probability as n goes to ∞. Jaeschke [16] claimed that this probability convergence is of almost no practical use. This is where we need to be cautious using asymptotic results. Figure 1 shows the poor approximation of the asymptotic result, even for a very large $n = 10^5$. Noe and Vandewiele [17] gave an iterative algorithm to compute the exact probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$. Their algorithm is only good for very small n due to the computational time and propagation of precision errors in representing real numbers in computer. Equation (24) of their paper gives an approximate formula for $n = 1, \ldots, 100$,

$$\begin{aligned}\text{pr}\left(\tilde{Z}_n \leq \tilde{z}\right) \approx{}& 1 - (\tilde{z})^{-2} - \left(2 - 3n^{-1}\right)(\tilde{z})^{-4} - \left(10 - 57n^{-1} + 48n^{-2}\right)(\tilde{z})^{-6} \\ &- \left(74 - 1021n^{-1} + 2743n^{-2} - 1797n^{-3}\right)(\tilde{z})^{-8} \\ &- \left(706 - 19123n^{-1} + 111905n^{-2} - 213619n^{-3} + 120132n^{-4}\right)(\tilde{z})^{-10}.\end{aligned} \tag{3.14}$$

This approximation is very good for $\tilde{z} \geq 4$ but is away from the true probability when $\tilde{z} < 4$. For our applications, the 50% quantile (median) of $\tilde{Z}$ is very useful, but the approximation of (3.14) is poor there.

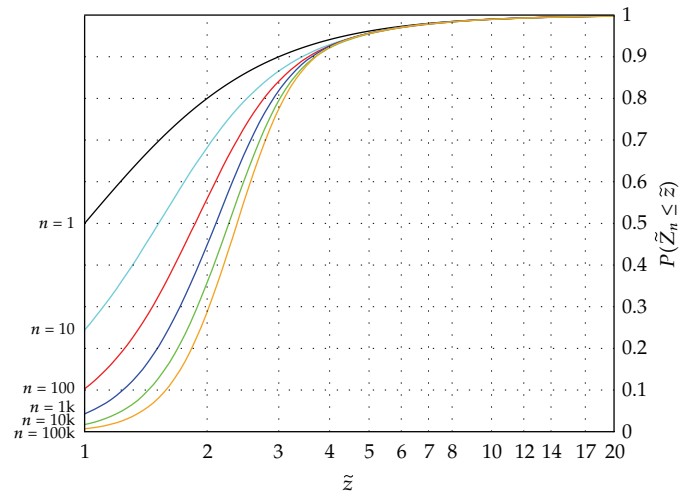


Figure 2: Plot of the probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$ for $1 \leq \tilde{z} \leq 20$. The probability is computed with 10^6 Monte Carlo simulations. The curves from the top to the bottom correspond to $n = 1, 10, 10^2, 10^3, 10^4$, and 10^5.

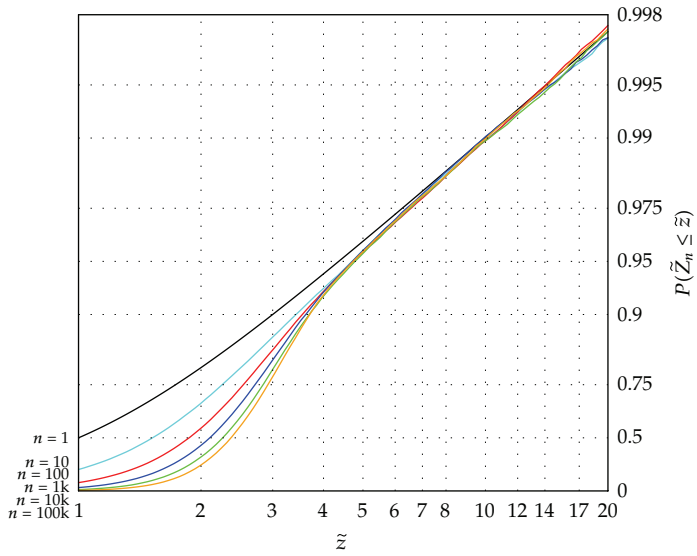


Figure 3: A blowup of Figure 2. This shows the part of the probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$ that is close to 1 as $1-P$ is drawn on a log scale for the y-axis.

In order to overcome the above poor approximation, we propose to use the Monte Carlo method to obtain the probability $\text{pr}(\tilde{Z} \leq \tilde{z})$. Figures 2 and 3 give the probability for $\tilde{z} \in [1, 20]$ with 10^6 simulations for $n = 1, 10, 10^2, 10^3, 10^4$, and 10^5. The Monto Carlo method generated quantiles $\tilde{z}$ for $n = 1, \ldots, 100$ are almost the same as those quantiles that were able to be computed by the exact algorithm in Noe and Vandewiele [17]. Our two figures show that the distribution of $\tilde{Z}_n$ does not change dramatically from $n = 1, \ldots, 10^5$. This property is beneficial for a multiple testing problem with large number of hypotheses, as it will not be overpenalized. Table 1 gives the quantiles of $\tilde{Z}_n$ with $n = 10^5$.

Table 1: The quantiles of $\widetilde{Z}_n$ of Figure 3, where $n = 10^5$. This table is estimated by 10^6 Monto Carlo simulations. The column $1-\alpha$ gives the probabilities, and the column $\tilde{z}$ gives the quantiles of $\widetilde{Z}_n$.

$1-\alpha$	$\tilde{z}$	$1-\alpha$	$\tilde{z}$	$1-\alpha$	$\tilde{z}$	$1-\alpha$	$\tilde{z}$
0.5	2.37	0.69	2.76	0.88	3.52	0.9675	5.74
0.51	2.39	0.7	2.78	0.89	3.61	0.97	5.95
0.52	2.41	0.71	2.81	0.9	3.7	0.9725	6.19
0.53	2.42	0.72	2.84	0.905	3.75	0.975	6.47
0.54	2.44	0.73	2.86	0.91	3.81	0.9775	6.78
0.55	2.46	0.74	2.89	0.915	3.88	0.98	7.23
0.56	2.48	0.75	2.92	0.92	3.96	0.9825	7.73
0.57	2.5	0.76	2.95	0.925	4.05	0.985	8.31
0.58	2.52	0.77	2.99	0.93	4.15	0.9875	9.04
0.59	2.54	0.78	3.02	0.935	4.26	0.99	10.04
0.6	2.56	0.79	3.06	0.94	4.4	0.9925	11.56
0.61	2.58	0.8	3.09	0.945	4.55	0.995	14.24
0.62	2.6	0.81	3.13	0.95	4.73	0.9955	14.87
0.63	2.62	0.82	3.18	0.9525	4.84	0.996	15.75
0.64	2.64	0.83	3.22	0.955	4.95	0.9965	16.6
0.65	2.66	0.84	3.27	0.9575	5.08	0.997	17.97
0.66	2.69	0.85	3.32	0.96	5.22	0.9975	19.69
0.67	2.71	0.86	3.39	0.9625	5.36		
0.68	2.73	0.87	3.45	0.965	5.55		

3.3. More about the Upper Confidence Bound for m_0

In computing the UCB for m_0 and consequently the UPB for Q_t, we rely on the unspecified parameter λ. A conventional choice of λ is 0.5. It is tempting to use $\min_{\lambda\in(0,1)}\overline{m}_{0,1-\alpha}(\lambda)$ as the best UCB for m_0. This approach should be avoided as it may lead to an overoptimistic UCB. We can use the same idea in computing the simultaneous upper prediction bounds for Q_t to find an UCB for m_0. Equation (2.20) motivates to the following theorem.

Theorem 3.7. *Define* $\widetilde{m}_{0,1-\alpha}(\lambda)$ *as*

$$\left\{\frac{1}{2(1-\lambda)}\left(\overline{\overline{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)}+\sqrt{\left(\overline{\overline{z}}_{1-\alpha}(m)\right)^2\lambda(1-\lambda)+4(m-R_\lambda)(1-\lambda)}\right)\right\}^2, \tag{3.15}$$

where

$$\overline{\overline{z}}_{1-\alpha}(m)=\max_{n=1,\ldots,m}\widetilde{z}_{1-\alpha}(n). \tag{3.16}$$

Let

$$\widetilde{m}_{0,1-\alpha}=\min_{\lambda\in(0,1)}\widetilde{m}_{0,1-\alpha}(\lambda). \tag{3.17}$$

Using $\tilde{m}_{0,1-\alpha}$ to replace m_0 in (3.8) *results in $\tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$. We have*

$$\mathrm{pr}\Big(\mathrm{m}_0 \le \tilde{\mathrm{m}}_{0,1-\mathrm{ff}},\ \mathrm{Q_t} \le \tilde{\mathrm{Q}}_{1-\mathrm{ff}}(\tilde{\mathrm{m}}_{0,1-\mathrm{ff}}, \mathrm{t})\ \forall \mathrm{t} \in (0,1)\Big) \ge 1 - \mathrm{ff}. \tag{3.18}$$

Thus simultaneously $\tilde{m}_{0,1-\alpha}$ is a $1-\alpha$ UCB for m_0 and $\tilde{Q}$ is a $1-\alpha$ upper prediction band.

Proof. Note that when $x > 0$,

$$x(1-\lambda) - \sqrt{x}\bar{\tilde{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)} - (m - R_\lambda) \le 0 \tag{3.19}$$

if and only if

$$x \le \left\{ \frac{1}{2(1-\lambda)} \left(\bar{\tilde{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)} + \sqrt{\left(\bar{\tilde{z}}_{1-\alpha}(m)\right)^2 \lambda(1-\lambda) + 4(m - R_\lambda)(1-\lambda)} \right) \right\}^2. \tag{3.20}$$

Therefore,

$$\begin{aligned}
\{m_0 \le \tilde{m}_{0,1-\alpha}\} &= \{m_0 \le \tilde{m}_{0,1-\alpha}(\lambda)\ \forall \lambda \in (0,1)\} \\
&= \left\{ m_0(1-\lambda) - \sqrt{m_0}\bar{\tilde{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)} - (m - R_\lambda) \le 0\ \forall \lambda \in (0,1) \right\} \\
&= \left\{ \max_{\lambda \in (0,1)} \frac{m_0(1-\lambda) - (m - R_\lambda)}{\sqrt{m_0 \lambda(1-\lambda)}} \le \bar{\tilde{z}}_{1-\alpha}(m) \right\} \\
&\supseteq \left\{ \max_{\lambda \in (0,1)} \frac{V_\lambda - m_0\lambda}{\sqrt{m_0\lambda(1-\lambda)}} \le \tilde{z}_{1-\alpha}(m_0) \right\}.
\end{aligned} \tag{3.21}$$

The last step follows: (i) $\bar{\tilde{z}}_{1-\alpha}(m)$ is no less than $\tilde{z}_{1-\alpha}(n)$ for any $n \le m$, and (ii) $m - R_\lambda \ge m_0 - V_\lambda$. The fact (ii) gives

$$m_0(1-\lambda) - (m - R_\lambda) \le m_0(1-\lambda) - (m_0 - V_\lambda) = V_\lambda - m_0\lambda. \tag{3.22}$$

Following the same idea as in the proof of Theorem 2.3 part (b), we can show that the set $\{m_0 \le \tilde{m}_{0,1-\alpha}$ and $Q_t \le \tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$ for all $t \in (0,1)\}$ is a superset of $\{\max_{t\in(0,1)}(V_t - m_0 t)/\sqrt{m_0 t(1-t)} \le \tilde{z}_{1-\alpha}(m_0)\}$. Therefore,

$$\begin{aligned}
&\mathrm{pr}\Big(m_0 \le \tilde{m}_{0,1-\alpha},\ Q_t \le \tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)\ \forall t \in (0,1)\Big) \\
&\quad \ge \mathrm{pr}\left(\max_{t\in(0,1)} \frac{V_t - m_0 t}{\sqrt{m_0 t(1-t)}} \le \tilde{z}_{1-\alpha}(m_0) \right) \\
&\quad = 1 - \alpha \quad \Big(\text{Note that } V_t \overset{d}{=} N_{m_0,t}\Big).
\end{aligned} \tag{3.23}$$

For any given α and γ,
(1) Compute $\tilde{m}_{0,1-\alpha}(\lambda)$ of (3.15) for some pre-specified λ_i's, say $\lambda_i = i/1000$, for $i = 1, \ldots, 999$.
(2) Compute $\tilde{m}_{0,1-\alpha} = \min_i \tilde{m}_{0,1-\alpha}(\lambda_i)$. This $\tilde{m}_{0,1-\alpha}$ is the $1-\alpha$ UCB for m_0.
If $\tilde{m}_{0,1-\alpha}$ exceeds m, replace it by m.
(3) Sort the observed P-values such that $P_{(1)} \leq \cdots \leq P_{(m)}$, and use (3.8) to compute the $1-\alpha$ simultaneous UPBs for the false discovery proportion Q, that is, for $i = 1, \ldots, m$,
$\tilde{Q}_{1-\alpha}(P_{(i)}) = (1/i)\ (\tilde{m}_{0,1-\alpha} P_{(i)} + \tilde{z}_{1-\alpha}(\tilde{m}_{0,1-\alpha})\sqrt{\tilde{m}_{0,1-\alpha} P_{(i)}(1 - P_{(i)})})$.
If $\tilde{Q}_{1-\alpha}(P_{(i)})$ exceeds 1, replace it by 1.
(4) Compute $\tau = \max\{P_{(i)} : \tilde{Q}_{1-\alpha}(P_{(i)}) \leq \gamma\}$,
reject the hypotheses whose P-values are no greater than τ,
which ensures that the false discovery proportion Q is not exceeding γ with probability $1-\alpha$.

Algorithm 1: Compute the simultaneous UPBs for the false discovery proportion and the UCB for m_0.

Readers should note that the maximum quantile $\bar{\tilde{z}}_{1-\alpha}(m)$ defined in (3.16) is only used to construct $\tilde{m}_{0,1-\alpha}$, The construction of $\tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$ itself does not use the maximum but the quantile $\tilde{z}_{1-\alpha}(m)$, while $\tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$ still depends on the maximum quantiles indirectly through $\tilde{m}_{0,1-\alpha}$. □

3.4. The Algorithm

Putting all these pieces together, we describe the procedure to compute the upper prediction band for Q_t and the UCB for m_0 in Algorithm 1. Note that we have to compute the quantile $\bar{\tilde{z}}_{1-\alpha}(m)$ and $\tilde{z}_{1-\alpha}(\tilde{m}_{0,1-\alpha})$. This is very time consuming for large m, which is typically from thousands to tens of thousands. The computationally time can be reduced by the following strategies.

(1) After careful study of the two equations (3.8) and (3.15), we find that if we replace all $\tilde{z}_{1-\alpha}(n)$ and $\bar{\tilde{z}}_{1-\alpha}(n)$ by $\bar{\tilde{z}}_{1-\alpha}(N)$, where $N \geq n$, the conclusions of Corollaries 3.1 and 3.2 and Theorem 3.7 still hold.

(2) The quantile $\tilde{z}_{1-\alpha}(n)$ is an increasing function of n, as shown by the Monte Carlo simulations in Figures 2 and 3. The rigorous mathematical proof of this finding is open to future research. For practical applications, we can first use Monte Carlo simulations to verify this property for the range of n that is related to the project and then replace all $\bar{\tilde{z}}_{1-\alpha}(n)$ by $\tilde{z}_{1-\alpha}(n)$.

(3) Figure 2 shows that $\tilde{z}_{1-\alpha}(n)$ is very close to $\tilde{z}_{1-\alpha}(N)$ if n is close to a large N, say more than 100. Therefore, in practical computations, we can first compute and store a representative sequences of the quantiles $\tilde{z}_{1-\alpha}(n)$ for $n = n_1, \ldots, n_I$, and consequently we can get an upper bound for $\tilde{z}_{1-\alpha}(n)$ for $n = 1, \ldots, m$. In computing the quantiles $\tilde{z}_{1-\alpha}(n)$, we recommend to have at least 10^4 Monte Carlo simulations in order to get an accurate quantile computation for tail part. Even with 10^6 simulations, we still see a small amount of random noise in the tail part in Figure 3.

4. A Focused Inference on Q and m_0: A Unified Approach

In many applications, it may be unnecessary to compute the simultaneous UPBs for Q_t for all t in $(0,1)$ and using $\tilde{m}_{1-\alpha}(\lambda)$ for all λ in $(0,1)$ to derive a $1-\alpha$ UCB for m_0. In most applications, it may be reasonable to restrict the rejections onto $\{[0,t] : t \in [t_0, t_0']\}$. The t_0 can take value of $0.01/m$ based on Bonferroni FWER control at level 0.01. It is rare to consider a smaller rejection region than this. The t_0' can take value of 0.05 as it is rare to consider a larger rejection region than $[0, 0.05]$ even in a single hypothesis testing problem. For the same reason, we can also restrict λ onto $[\lambda_0, \lambda_0']$ in (3.17). The interval $[\lambda_0, \lambda_0']$ can be taken as a region close to one as the minimum of $\tilde{m}_{0,1-\alpha}(\lambda)$ is reached when λ is close to 1 [18], but if λ is too close to 1, $\tilde{m}_{0,1-\alpha}(\lambda)$ is not stable. One good choice of λ_0 can be 0.8, and λ_0' can be 0.95.

The above scenario is a focused inference on Q and m_0. The $\tilde{z}$ in Section 3.2 will be a little bit more conservative for us. We can redefine $\widetilde{Z}_n^*$ as in the following:

$$\widetilde{Z}_n^* = \max\left(\frac{\sup_{t\in[t_0,t_0']}(N_{n,t}-nt)}{\sqrt{nt(1-t)}}, \frac{\sup_{\lambda\in[\lambda_0,\lambda_0']}(N_{n,\lambda}-n\lambda)}{\sqrt{n\lambda(1-\lambda)}}\right). \tag{4.1}$$

From this $\widetilde{Z}_n^*$ we can define the $1-\alpha$ quantile $\tilde{z}_{1-\alpha}^*(n)$, and derive results similar to Theorem 3.7. Figure 4 shows quantiles $\tilde{z}_{1-\alpha}^*(n)$ for $n = 1, \ldots, 10^5$ for $[t_0, t_0'] = [0.01/n, 0.05]$ and $[\lambda_0, \lambda_0'] = [0.8, 0.95]$. Table 2 gives the numerical values of $\tilde{z}_{1-\alpha}^*(n)$ for $n = 10^5$. It clearly shows that $\tilde{z}_{1-\alpha}^*(n)$ is around 10% smaller than the unrestricted quantiles $\tilde{z}_{1-\alpha}^*(n)$. For small values of α, say that $\alpha \le 0.01$, the former is at least 25% smaller than the latter.

Corollary 4.1. *Define* $\tilde{m}_{0,1-\alpha}^*(\lambda)$ *as*

$$\left\{\frac{1}{2(1-\lambda)}\left(\overline{\tilde{z}}_{1-\alpha}^*(m)\sqrt{\lambda(1-\lambda)} + \sqrt{\left(\overline{\tilde{z}}_{1-\alpha}^*(m)\right)^2\lambda(1-\lambda) + 4(m-R_\lambda)(1-\lambda)}\right)\right\}^2, \tag{4.2}$$

where $\overline{\tilde{z}}_{1-\alpha}^*(m) = \max_{n=1}^m \tilde{z}_{1-\alpha}^*(n)$. *Let*

$$\tilde{m}_{0,1-\alpha}^* = \min\left(\min_{\lambda\in[\lambda_0,\lambda_0']}\tilde{m}_{0,1-\alpha}^*(\lambda), \min_{\lambda\in[t_0,t_0']}\tilde{m}_{0,1-\alpha}^*(\lambda)\right). \tag{4.3}$$

Define $\widetilde{Q}^*$

$$\widetilde{Q}_{1-\alpha}^*(m_0,t) = \frac{m_0 t + \tilde{z}_{1-\alpha}^*(m_0)\sqrt{m_0 t(1-t)}}{R_t}. \tag{4.4}$$

Replacing m_0 *by* $\tilde{m}_{0,1-\alpha}^*$ *results in* $\widetilde{Q}_{1-\alpha}^*(\tilde{m}_{0,1-\alpha}^*, t)$. *We have that* $\tilde{m}_{0,1-\alpha}^*$ *is a* $1-\alpha$ *UCB for* m_0, *and* $\widetilde{Q}^*$ *is a* $1-\alpha$ *upper prediction band for* Q *for* $t \in [t_0 \cdot t_0']$, *that is,*

$$\mathrm{pr}\left(m_0 \le \tilde{m}_{0,1-\alpha}^*,\ Q_t \le \widetilde{Q}_{1-\alpha}^*\left(\tilde{m}_{0,1-\alpha}^*, t\right)\ \forall t \in [t_0, t_0']\right) \ge 1-\alpha. \tag{4.5}$$

Table 2: The quantiles of $\widetilde{Z}_n^*$ of Figure 4, where $n = 10^5$. The column $1-\alpha$ gives the probabilities, and the column $\widetilde{z}^*$ gives the quantiles of $\widetilde{Z}_n^*$.

$1-\alpha$	$\widetilde{z}^*$	$1-\alpha$	$\widetilde{z}^*$	$1-\alpha$	$\widetilde{z}^*$	$1-\alpha$	$\widetilde{z}^*$
0.5	2.1	0.69	2.55	0.88	3.38	0.9675	5.15
0.51	2.12	0.7	2.58	0.89	3.47	0.97	5.28
0.52	2.14	0.71	2.61	0.9	3.57	0.9725	5.43
0.53	2.16	0.72	2.64	0.905	3.62	0.975	5.58
0.54	2.18	0.73	2.67	0.91	3.68	0.9775	5.78
0.55	2.21	0.74	2.7	0.915	3.74	0.98	5.99
0.56	2.23	0.75	2.73	0.92	3.82	0.9825	6.2
0.57	2.25	0.76	2.77	0.925	3.89	0.985	6.48
0.58	2.27	0.77	2.8	0.93	3.97	0.9875	6.82
0.59	2.3	0.78	2.84	0.935	4.06	0.99	7.23
0.6	2.32	0.79	2.88	0.94	4.18	0.9925	7.73
0.61	2.35	0.8	2.92	0.945	4.3	0.995	8.37
0.62	2.37	0.81	2.96	0.95	4.43	0.9955	8.52
0.63	2.39	0.82	3.01	0.9525	4.52	0.996	8.63
0.64	2.42	0.83	3.06	0.955	4.61	0.9965	8.77
0.65	2.44	0.84	3.11	0.9575	4.7	0.997	8.94
0.66	2.47	0.85	3.17	0.96	4.79	0.9975	9.12
0.67	2.5	0.86	3.23	0.9625	4.89		
0.68	2.52	0.87	3.3	0.965	5.02		

Note that the $1-\alpha$ UCB for m_0 takes not only the minimum of $\tilde{m}^*_{0,1-\alpha}(\lambda)$ for $\lambda \in [\lambda_0, \lambda_0']$, but also the minimum of $\tilde{m}^*_{0,1-\alpha}(t)$ for $t \in [t_0, t_0']$. This advantage is due to the construction of the $\widetilde{Z}_n^*$, which takes maximum over these two intervals. The details of the calculation are summarized in Algorithm 2. The proof of this corollary is the same as that in Theorem 3.7.

By setting $\lambda_0 = \lambda_0' = t$ and $t_0 = t_0' = t$, this corollary is equivalent to Theorem 2.3 through some algebra manipulations, while Theorem 2.3 uses the exact confidence bound from the binomial distribution without relying on the quantiles of $\widetilde{Z}_n^*$. Furthermore, Theorem 3.7 is exact a special case of Corollary 4.1 by setting $\lambda_0 = 0$, $\lambda_0' = 1$, and $t_0 = 0$, $t_0' = 1$ and by considering open intervals rather close intervals. The focused inference thus unifies both the fixed rejection approach and simultaneous approach. We should be cautious of selecting $[t_0, t_0']$ and $[\lambda_0, \lambda_0']$ based on the observed data, which may result in overoptimistic false discovery proportions. These settings have to be decided before the data are generated. A careful study of choosing appropriate values for $[t_0, t_0']$ and $[\lambda_0, \lambda_0']$ is open for future research.

5. Generalizing the Results to Less-Independent Situations

The results of Sections 2, 3, and 4 are based on the assumption that the true null P-values are independently distributed as $U[0,1]$. Given this, we need no further assumptions concerning the false null P-values. This independence assumption can be weakened as in the following:

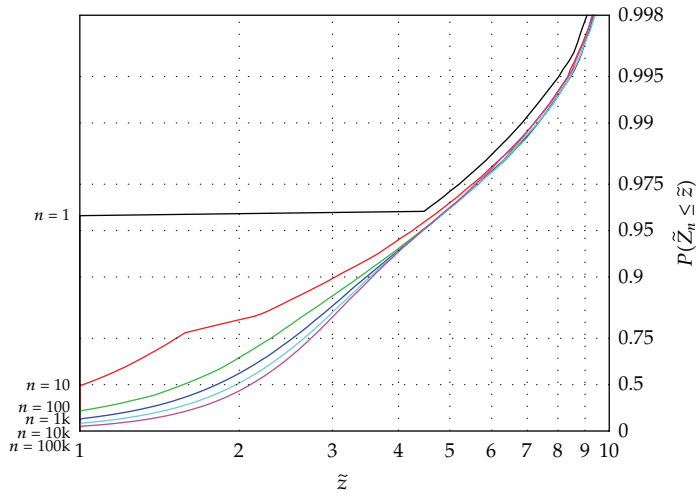


Figure 4: The probability distribution of $\tilde{Z}_n^*$ is computed with 10^6 Monte Carlo simulations. The interval $[t_0, t_0']$ takes value of $[0.01/n, 0.05]$, and interval $[\lambda_0, \lambda_0']$ takes value of $[0.80, 0.95]$. The curves from the top to the bottom correspond to $n = 1, 10, 10^2, 10^3, 10^4$, and 10^5.

For any given α and γ, choose $t_0 = 0.01/m$, $t_0' = 0.05$, $\lambda_0 = 0.8$, $\lambda_0' = 0.95$.
(1) Compute $\tilde{m}^*_{0,1-\alpha}(\lambda)$ of (4.2) for some pre-specified λ_i's in the region (λ_0, λ_0') and pre-specified t_i's in the region (t_0, t_0'), say $\lambda_i = \lambda_0 + (\lambda_0' - \lambda_0)i/1000$, $t_i = t_0 + (t_0' - t_0)i/1000$ for $i = 0, \ldots, 1000$.
(2) Compute $\tilde{m}^*_{0,1-\alpha} = \min(\min_i \tilde{m}^*_{0,1-\alpha}(\lambda_i), \min_i \tilde{m}^*_{0,1-\alpha}(t_i))$. This $\tilde{m}^*_{0,1-\alpha}$ is the $1-\alpha$ UCB for m_0. If $\tilde{m}^*_{0,1-\alpha}$ exceeds m, replace it by m.
(3) Sort the observed P-values such that $P_{(1)} \leq \cdots \leq P_{(m)}$, and use (4.4) to compute the $1-\alpha$ UPB for the false discovery proportion Q, that is, for $P_{(i)} \in [t_0, t_0']$
$\tilde{Q}^*_{1-\alpha}(P_{(i)}) = (1/i)(\tilde{m}^*_{0,1-\alpha} P_{(i)} + \tilde{z}^*_{1-\alpha}(\tilde{m}^*_{0,1-\alpha})\sqrt{\tilde{m}^*_{0,1-\alpha} P_{(i)}(1 - P_{(i)})})$.
If $\tilde{Q}^*_{1-\alpha}(P_{(i)})$ exceeds 1, replace it by 1.
(4) Compute $\tau = \max\{P_{(i)} \in [t_0, t_0'] : \tilde{Q}^*_{1-\alpha}(P_{(i)}) \leq \gamma\}$, reject the hypotheses whose P-values are no greater than τ, which ensures that the false discovery proportion Q is not exceeding γ with probability $1 - \alpha$.

Algorithm 2: Focused simultaneous inferences on the UPBs for the false discovery proportion and the UCB for m_0.

Binomial Dominant Condition: One has $V_t \overset{d}{\leq} N_{m_0,t}$ for $0 < t < 1$.

The notation $X \overset{d}{\leq} Y$ means that random variable X is stochastically no greater than random variable Y, that is, $\text{pr}(X \leq x) \geq \text{pr}(Y \leq x)$ for any x. Replacing the independence assumption by the binomial dominant condition, the results corresponding to Lemma 2.2, Theorem 2.3, and Corollary 2.7 in Section 2 still hold for a fixed rejection region. For the simultaneous UPBs in Sections 3 and 4, we need a stronger assumption than the binomial dominant condition as the joint distribution of $\{V_t, t \in (0,1)\}$ needs to be specified. We can replace the binomial dominant condition by the following.

Joint Binomial Dominant Condition: $(V_{t_1}, \ldots, V_{t_k}) \overset{d}{\leq} (N_{m_0,t_1}, \ldots, N_{m_0,t_k})$ for any $k = 1, 2, \ldots$, and $t_1, \ldots, t_k \in (0,1)$. Here $N_{m_0,t} = \sum_{i=1}^{m_0} I(U_i \leq t)$, and U_i, $i = 1, \ldots, m_0$ are mutually independently distributed as distribution $U[0,1]$. The notation $(X_1, \ldots, X_k) \overset{d}{\leq} (Y_1, \ldots, Y_k)$ means for any $x_1, \ldots, x_k$, $\text{pr}(X_1 \leq x_1, \ldots, X_k \leq x_k) \geq \text{pr}(Y_1 \leq x_1, \ldots, Y_k \leq x_k)$.

Replacing the independence assumption by this joint binomial dominant condition, the results in Sections 3 and 4 are still valid for the upper prediction band for Q and the UCB for m_0. A special case for this joint binomial dominant condition is that when the true null P-values are independent with distribution stochastically no smaller than $U[0,1]$. This happens when the null hypothesis is composite or the statistic to test the null hypothesis is not a continuous random variable.

More generally, we would like the construction of upper prediction band for Q not to rely on the independence assumption or any kind of weak dependence assumption (the binomial dominant condition or the joint binomial dominant condition). The method of Romano and Shaikh [7, 8] can be applied without any assumptions on the dependence, but may potentially have lost power due to that the correlation structure of the data has not been exploited. A resampling procedure [13, 19] has been proposed to address this limitation.

6. Discussion

The method of this paper applies to data where true null P-values are independent, or to slightly dependent data where the joint binomial dominant condition is satisfied. This assumption does not rely on any specification for the false null P-values. In this paper we used the idea of considering a fixed rejection region to construct a UPB for Q_t and a UCB for m_0. By utilizing the normalized empirical process $\widetilde{Z}_n = \sup_{t\in(0,1)}(N_{n,t} - nt)/\sqrt{nt(1-t)}$, we find simultaneous UPBs for Q_t for all $t \in (0,1)$ and can further modify the construction of the UCB for m_0. The result of Theorem 3.7 gives the joint statement about the UCB for m_0 and the simultaneous UPBs for the false discovery proportions Q. A focused approach in Corollary 4.1 unifies the result of the fixed rejection region method and the simultaneous approach.

The method in this paper is based on finite samples and avoids the slow convergence problem of the asymptotic theory for the empirical process $\widetilde{Z}_n$. The Monte Carlo simulations give very accurate estimates of the quantiles for $\widetilde{Z}_n$. The standard error associated with the Monte Carlo simulations in computing the probability in (3.7) is no greater than $\sqrt{\alpha(1-\alpha)/B}$, where B is the number of simulations.

In the dataset where the test statistics are not independent or do not satisfy joint binomial dominant condition, the method in this paper may not be guaranteed to work. One can alternatively use the methods proposed in Romano and Shaikh [7, 8], Meinshausen [13], Ge et al. [19]. The method proposed in this paper can be potentially extended to dependent data by using resamplings, and this work is open for future research.

Acknowledgments

The authors thank Terry Speed, Stuart Sealfon, Carol Bodian, Sylvan Wallenstein, John Mandeli, Samprit Chatterjee, and Jim Godbold for their discussions of this work. They thank the editor and referees for helpful comments that have led to an improved paper. This

work was partly supported by National Institute of Allergy and Infectious Diseases with the contract HHSN266200500021C.

References

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.
[2] C. Genovese and L. Wasserman, "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society B*, vol. 64, no. 3, pp. 499–517, 2002.
[3] E. L. Korn, J. F. Troendle, L. M. McShane, and R. Simon, "Controlling the number of false discoveries: application to high-dimensional genomic data," *Journal of Statistical Planning and Inference*, vol. 124, no. 2, pp. 379–398, 2004.
[4] C. Genovese and L. Wasserman, "A stochastic process approach to false discovery control," *The Annals of Statistics*, vol. 32, no. 3, pp. 1035–1061, 2004.
[5] E. L. Lehmann and J. P. Romano, "Generalizations of the familywise error rate," *The Annals of Statistics*, vol. 33, no. 3, pp. 1138–1154, 2005.
[6] N. Meinshausen and J. Rice, "Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses," *The Annals of Statistics*, vol. 34, no. 1, pp. 373–393, 2006.
[7] J. P. Romano and A. M. Shaikh, "On stepdown control of the false discovery proportion," in *Lehmann Symposium—Optimality*, vol. 49 of *IMS Lecture Notes—Monograph Series*, pp. 33–50, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2nd edition, 2006.
[8] J. P. Romano and A. M. Shaikh, "Stepup procedures for control of generalizations of the familywise error rate," *The Annals of Statistics*, vol. 34, no. 4, pp. 1850–1873, 2006.
[9] M. J. van der Laan, S. Dudoit, and K. S. Pollard, "Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 15, 2004.
[10] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society B*, vol. 66, no. 1, pp. 187–205, 2004.
[11] Y. Benjamini and Y. Hochberg, "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 1, pp. 60–83, 2000.
[12] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society B*, vol. 67, no. 4, pp. 555–572, 2005.
[13] N. Meinshausen, "False discovery control for multiple tests of association under general dependence," *Scandinavian Journal of Statistics*, vol. 33, no. 2, pp. 227–237, 2006.
[14] R. Göb, "Bounds for median and 50 percentage point of binomial and negative binomial distribution," *Metrika*, vol. 41, no. 1, pp. 43–54, 1994.
[15] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *The Annals of Statistics*, vol. 32, no. 3, pp. 962–994, 2004.
[16] D. Jaeschke, "The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals," *The Annals of Statistics*, vol. 7, no. 1, pp. 108–115, 1979.
[17] M. Noe and G. Vandewiele, "The calculation of distributions of Kolmogorov-Smirnov type statistics including a table of significance points for a particular case," *Annals of Mathematical Statistics*, vol. 39, pp. 233–241, 1968.
[18] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.
[19] Y. Ge, S. C. Sealfon, and T. P. Speed, "Multiple testing and its applications to microarrays," *Statistical Methods in Medical Research*, vol. 18, no. 6, pp. 543–563, 2009.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 419832, 12 pages
doi:10.1155/2012/419832

Research Article

A Multinomial Ordinal Probit Model with Singular Value Decomposition Method for a Multinomial Trait

Soonil Kwon,[1] Mark O. Goodarzi,[1] Kent D. Taylor,[1] Jinrui Cui,[1] Y.-D. Ida Chen,[1] Jerome I. Rotter,[1] Willa Hsueh,[2] and Xiuqing Guo[1]

[1] *Medical Genetics Institute, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Paciffc Theatres Building, 4th Floor, Los Angeles, CA 90048, USA*
[2] *The Methodist Hospital Research Institute, The Methodist Hospital, 6670 Bertner Street R8-103, Houston, TX 77030, USA*

Correspondence should be addressed to Xiuqing Guo, xiuqing.guo@cshs.org

Received 12 March 2012; Accepted 31 March 2012

Academic Editor: Yongzhao Shao

We developed a multinomial ordinal probit model with singular value decomposition for testing a large number of single nucleotide polymorphisms (SNPs) simultaneously for association with multidisease status when sample size is much smaller than the number of SNPs. The validity and performance of the method was evaluated via simulation. We applied the method to our real study sample recruited through the Mexican-American Coronary Artery Disease study. We found 3 genes (SORCS1, AMPD1, and PPARα) to be associated with the development of both IGT and IFG, while 5 genes (AMPD2, PRKAA2, C5, TCF7L2, and ITR) with the IGT mechanism only and 6 genes (CAPN10, IL4, NOS3, CD14, GCG, and SORT1) with the IFG mechanism only. These data suggest that IGT and IFG may indicate different physiological mechanism to prediabetes, via different genetic determinants.

1. Introduction

Genome-wide association studies (GWASs) examine genetic variants across the entire genome to improve the understanding of genetic components underlying complex human disease. With whole-genome genotyping techniques that allow GWAS to involve hundreds of thousands of single nucleotide polymorphisms (SNPs), many studies have successfully identified novel genetic components for many diseases or related quantitative traits. However, the sample size is often limited due to the difficulty of recruiting patients and/or

the cost of research. This leads to the situation that the number of SNPs (m) is much larger than the samples (n) available, that is, $m \gg n$, and makes the traditional statistical methods unsuitable for analyzing multiple SNPs simultaneously. Most of current GWASs deal with this shortcoming by performing single SNP association test, which analyzes one SNP at a time and results in a huge multiple testing problem. These motivated us to develop methods that can avoid the multiple testing problem, in other words, methods that can evaluate multiple SNPs simultaneously when $m \gg n$. We first introduced the iterative Bayesian variable selection (IBVS) method [1], which analyzes all SNPs simultaneously when $m \gg n$ and uses the Bayesian variable selection [2] iteratively to find SNPs that are associated with disease. The method was successfully applied to the simulated rheumatoid arthritis data provided by the Genetic Analysis Workshop 15 (GAW15). We later introduced the Bayesian classification with singular value decomposition (BCSVD) method [3]. The method applies the singular value decomposition (SVD) to the covariate matrix, which is usually the genotype data in GWAS and reduces the dimension of parameters to be estimated to the number of samples. This makes the method feasible to handle multiple SNPs simultaneously when $m \gg n$. The validation of the method was demonstrated by applying to the simulated data provided by GAW16. We now extend our method from binary disease to the analysis of polytomous ordinal response variables. We propose here a multinomial ordinal probit model with singular value decomposition method. We show the validity of the newly developed method by applying it to simulated data sets as well as to a real study sample to identify genes contributing to two different mechanisms for prediabetes, namely, impaired glucose tolerance (IGT) and impaired fasting glucose (IFG). With the simulated data sets, we demonstrate that this new method is superior to single SNP analysis method and, with the real data, identify different genes for each mechanism.

2. Method and Materials

2.1. Multinomial Probit Model with Singular Value Decomposition

Logit and probit models are statistical models that are widely used for the analysis of categorical (ordinal/nominal) data. The difference between these two models is the choice of the link function relating the linear predictor to the expected value; the probit model uses the inverse normal cumulative distribution, and the logit model uses the logit transformation. As discussed by Greene [4], in most cases, the choice of the link function is largely a matter of taste. We utilized the probit model here to analyze data with polytomous ordinal response variables. In general, the multinomial ordinal probit model can be expressed by latent (unobserved) continuous variables associated with categorical responses. Let us assume that responses $y_1, y_2, \ldots, y_n$ are observed, where y_i takes one of the J-ordered categories and $\theta_1, \ldots, \theta_J$ are real numbers of bin boundaries, which satisfy that $-\infty = \theta_1 \leq \cdots \leq \theta_J = \infty$. As discussed by Albert and Chib [5], we denote that $z_1, z_2, \ldots, z_n$ are latent continuous random variables. We assume that the latent variable (z_i) associated with a categorical outcome (y_i) can be explained in terms of an underlying linear model, and that the observed response y_i has category j if and only if z_i falls between θ_{j-1} and θ_j. The multinomial ordinal probit model is equivalent to the following model:

$$\begin{aligned} z_i &= x_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad i = 1, \ldots, n, \\ y_i = j &\iff \theta_{j-1} < z_i \leq \theta_j, \quad j = 1, \ldots, J, \end{aligned} \tag{2.1}$$

where x_i is a $1 \times m$ vector of the explanatory variables for the ith sample, and β is a $m \times 1$ vector of parameters to be estimated. In vector-matrix notation, we can have the multinomial ordinal probit model

$$z = X\beta + \epsilon, \tag{2.2}$$

where z is the $n \times 1$ vector of latent variables, X is the $n \times m$ matrix of the explanatory variables, β is the $m \times 1$ vector of unknown regression coefficients, and ϵ follows an independent standard multivariate normal distribution, $\epsilon \sim N(0, I_n)$. By applying SVD to the matrix X in (2.2), when $\text{rank}(X) = n$, the matrix can be expressed as $X' = ADF'$, where A is the $m \times n$ singular value factor loading matrix with orthonormal columns so that $A'A = I_n$, F is the $n \times n$ SVD orthogonal factor matrix with $F'F = FF' = I_n$, and $D = \text{diag}(d_1, \ldots, d_n)$, the diagonal matrix of positive singular values, ordered as $d_1 \geq \cdots \geq d_n > 0$. When $\text{rank}(X) = r < n$, the smallest $n - r$ singular values in D are replaced with 0, that is, $d_1 \geq \cdots \geq d_r > d_{r+1} = \cdots = d_n = 0$. Therefore, in the product $X' = ADF'$, the last $n - r$ columns of both A and F for which $d_{r+1} = \cdots = d_n = 0$ are ignored since they interact with the block of zeros in D. Hence, this leads to another form of SVD, $X' = A_r D_r F_r'$, that is, the product of the first r columns of A, the upper $r \times r$ block of D, and the first r columns of F. Since the difference between the both scenario is only in dimension of matrices in SVD, we assume that $\text{rank}(X) = n$ in the rest part of the paper for convenience. Thus, the model in (2.2) with the SVD of X can be written as follows:

$$z = X\beta + \epsilon = \left(ADF'\right)'\beta + \epsilon = FDA'\beta + \epsilon = L\gamma + \epsilon, \tag{2.3}$$

where $L = FD$ and $\gamma_{n\times 1} = A'_{n\times m}\beta_{m\times 1}$. Therefore, z, the $n \times 1$ vector of latent variables in (2.3), has a multivariate normal distribution, that is, $z \sim N(L\gamma, I_n)$. As shown in (2.3), γ is expressed by a linear combination of the original parameters (β). Hence, we call γ as the vector of superfactors. The model in (2.3) represents a massive dimension reduction from m to n parameters. The regression model with m parameters reduced to that with n parameters derived from the SVD of the covariate matrix X. Therefore, the statistical inference on the original parameter turns into the superfactors. Let $p_i = (p_{i1}, \ldots, p_{i_J})$ denote the vector of probabilities associated with the assignment of the ith sample into categories $1, \ldots, J$, where p_{ij} denote the probability that a sample falls into category j. From (2.1) and (2.3), it follows that

$$p_{ij} = \int_{\theta_{j-1}}^{\theta_j} \phi(z - l_i\gamma)dz = \Pr(\theta_{j-1} < Z_i < \theta_j) = \Phi(\theta_j - l_i\gamma) - \Phi(\theta_{j-1} - l_i\gamma), \tag{2.4}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative density function of the standard normal distribution, respectively, and l_i is the ith row of matrix $L = FD$. Let $y = (y_1, \ldots, y_n)$ denote the vector of responses observed for all samples. Then, the probability of observing data y is given as follow:

$$\Pr(y \mid p_i) = \prod_{i=1}^{n}\prod_{j=1}^{J} p_{ij}^{I\{y_i=j\}}. \tag{2.5}$$

From (2.4) and (2.5), the log likelihood function for (γ, θ) can be written as

$$\ln L(\gamma, \theta) = \sum_{i=1}^{n}\sum_{j=1}^{J} I\{y_i = j\} \ln\left[\Phi(\theta_j - l_i\gamma) - \Phi(\theta_{j-1} - l_i\gamma)\right]. \tag{2.6}$$

2.2. Model Fitting with Maximum Likelihood Estimation

The maximum likelihood estimates (MLEs) of the superfactors, γ, in (2.3) can be obtained by the iteratively reweighted least squares (IRLSs) procedure [6] using the log likelihood function for (γ, θ) in (2.6). The procedure can be briefly described as follows. Let η denote the vector of all model parameters, that is, $\eta = (\theta_2, \ldots, \theta_{J-1}, \gamma_1, \ldots, \gamma_{n-J+2})$. Note that θ_1 and θ_J are not included in this vector because their values are assumed to be 0 and ∞, respectively, for the purpose of model identifiability. Also note that the $(J-2)$ smallest singular values together with their corresponding factors are dropped from the parameters since the number of parameters must not exceed the number of samples. Assuming that $J = 4$, define that

$$\mathcal{C}_i = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \qquad \mathcal{L}_i = \begin{bmatrix} 1 & 0 & 0 & -l_i \\ 0 & 1 & 0 & -l_i \\ 0 & 0 & 1 & -l_i \end{bmatrix}, \tag{2.7}$$

and $\mathcal{H}_i = \operatorname{diag}(f_{i1}, \ldots, f_{iJ-1})$, where f_{ij} denotes the derivative of the standard normal cumulative distribution function at $\theta_j - l_i\gamma$. Take $W_i = \operatorname{diag}(p_i)$, where p_i is the $J \times 1$ vector of probabilities that the ith individual falls in each category, that is, $p_i = (p_{i1}, \ldots, p_{iJ})'$, and let $\mathcal{N}_i$ be a $J \times 1$ vector of observation, that is, $\mathcal{N}_i = (I\{y_i = 1\}, \ldots, I\{y_i = J\})'$. After initialization of all elements, the iteration $s+1$ $(s = 1, 2, \ldots)$ can be written as

$$\eta^{s+1} = \eta^s + \left(\sum_{i=1}^{n} \mathcal{L}_i' \mathcal{H}_i^{(s)} \mathcal{C}_i' W_i^{-1(s)} \mathcal{C}_i \mathcal{H}_i^{(s)} \mathcal{L}_i\right)^{-1} \sum_{i=1}^{n} \mathcal{L}_i' \mathcal{H}_i^{(s)} \mathcal{C}_i' W_i^{-1(s)} \left(\mathcal{N}_i - p_i^{(s)}\right). \tag{2.8}$$

The MLE of η can be found by performing the process recursively until the change between η^{s+1} and η^s is negligible.

2.3. General Solution for the Original Parameters

We have discussed how to estimate the superfactor (γ) in (2.3) thus far. Since the primary interest is to find SNPs that are significantly associated with a disease, it is necessary to transform the superfactor (γ) to the original parameters (β) in (2.1). The equation $\gamma = A'\beta$ in (2.3) can be utilized for the transformation even though A is $m \times n$ nonsquare matrix. As discussed by Graybill [7], the unique solution for β can be achieved by taking the generalized inverse matrix of A' as A since $A'A = I_n$. Therefore, the unique solution for SNP effect (β) can be calculated by $\beta = A\gamma$.

2.4. Selection of Significant SNPs

Finding significant SNPs is the same as testing if each SNP effect (β_i, $i = 1, \ldots, m$) is statistically significant, that is, testing the hypothesis: H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$, $i = 1, \ldots, m$. The simple method is to use Wald's test statistic, which forms $(\widehat{\beta} - \beta)/\mathrm{se}(\widehat{\beta})$ and assumes a normal distribution. However, when $m \gg n$, it is hard to calculate $\mathrm{se}(\widehat{\beta})$ directly from the data. We therefore utilized permutation test to select significant SNPs. The rationale behind the test is that, under the null hypothesis, the estimate of β obtained from the raw (unpermuted) data is similar to the estimate of β obtained from the permuted data. That is, the difference between two estimates is closed to zero under H_0. With this idea, we can construct Wald's test statistic as follows. Let $\widehat{\beta}_i$ $(i = 1, \ldots, m)$ be the estimate of the ith SNP effect from the raw data and $\widehat{\beta}_i^k$ $(k = 1, \ldots, K)$ be the estimate of the ith SNP effect from the kth-permuted data. Let us define $\beta_i^{d_k}$ as the difference between $\widehat{\beta}_i$ and $\widehat{\beta}_i^k$, that is, $\beta_i^{d_k} = \widehat{\beta}_i - \widehat{\beta}_i^k$. Then, Wald's test statistic can be as follows:

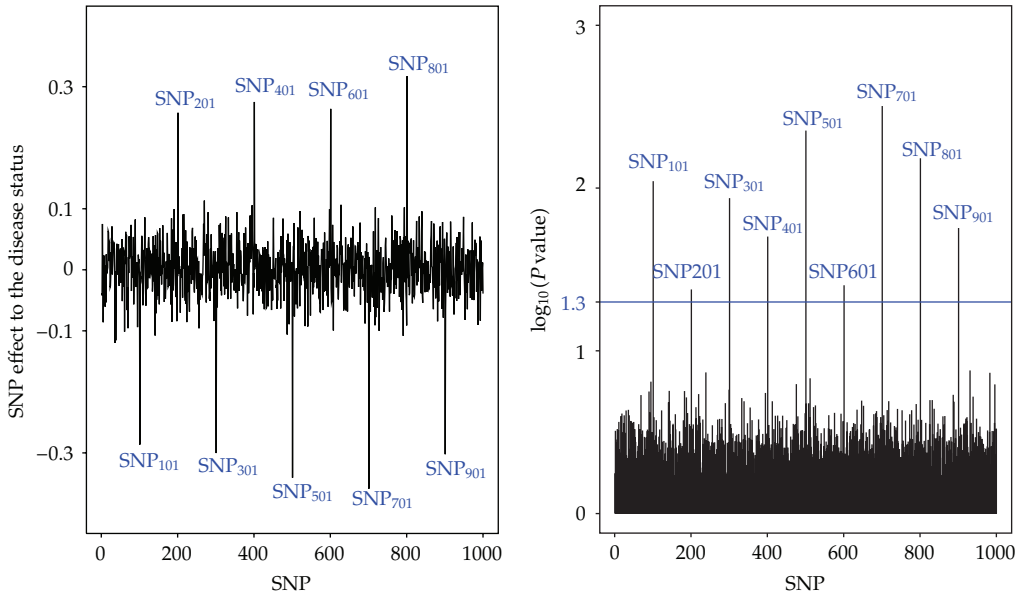
$$\Lambda_i = \frac{\overline{\beta}_i^d}{\mathrm{se}\left(\overline{\beta}_i^d\right)}, \quad i = 1, \ldots, m, \tag{2.9}$$

where $\overline{\beta}_i^d$ is the sample mean of $\beta_i^{d_k}$'s, which is $\overline{\beta}_i^d = (1/K) \sum_{k=1}^{K} \beta_i^{d_k}$, and $\mathrm{se}(\overline{\beta}_i^d)$ is the standard error of $\overline{\beta}_i^d$. Under the null hypothesis, the statistic Λ_i defined in (2.9) follows approximately standard normal distribution when k is large. P value for rejecting the null hypothesis at a significance level $\alpha = 0.05$ can be utilized to identify significant SNPs.

2.5. Application of the Multinomial Probit Model with SVD

2.5.1. Simulated Multinomial Ordinal Data

The validity of the proposed method was evaluated using simulated data sets. The procedure of data generation was composed of three steps: generating genotype data with certain genetic model, generating the latent variable, and defining the disease status variable by applying the predefined bin boundaries. The brief scheme of each step is as follows: we first generated 10 sets of the simulated genotype data under an additive genetic model, each set consists of 100 samples and 1000 SNPs. From (2.1), we can notice that the latent variable (z_i) consists of two parts: the expected value $(x_i\beta)$ and the random error (ϵ_i). In order to generate the expected value, we assumed that, for each sample, 9 out of the 1000 SNPs (every 101th SNP, except the last one) contribute to disease status $x_i\beta = \beta_1 \cdot \mathrm{SNP}_{101} + \beta_2 \cdot \mathrm{SNP}_{201} + \cdots + \beta_1 \cdot \mathrm{SNP}_{801} + \beta_2 \cdot \mathrm{SNP}_{901}$, where β_1 and β_2 are set as -1 and 1, respectively. Hence, the latent variable can be obtained from the sum of the expected values $(x_i\beta)$ and the random error generated from standard normal distribution. We then generated disease status variable (y_i) assuming 3 disease development stages. Therefore, when applying the proposed method to the simulated data sets, we would expect 9 strong signals corresponding to each of the 9 disease-associated SNPs. We also compared results obtained from the proposed method with that from single SNP analysis with multinomial ordinal probit model.



(a) Estimates of SNP effects by the proposed method (b) P values of testing SNP effects ($-\log_{10}(P \text{ value})$)

Figure 1: Analysis of the simulated data sets with multinomial ordinal probit model with SVD.

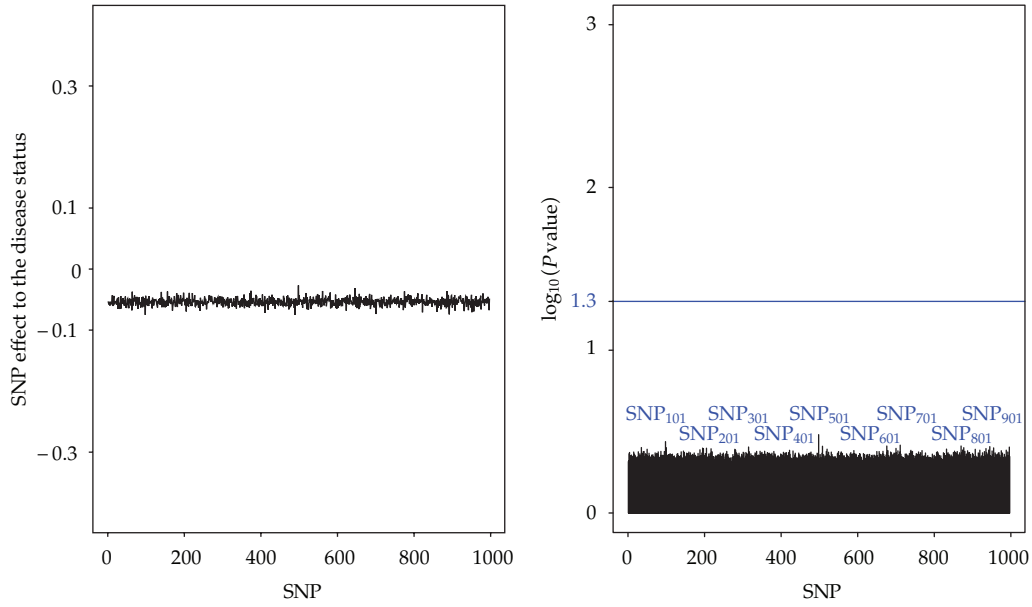
2.6. Mexican-American Coronary Artery Disease (MACAD) Study

We also applied the proposed method to study sample recruited through the Mexican-American Coronary Artery Disease (MACAD) study [8, 9]. The study population consists of probands who are Mexican American aged between 45 and 75 with coronary artery disease: spouses of probands, adult offspring (≥18), and their spouses. For the offspring generation, we performed oral glucose tolerance test and genotyped 132 SNPs in 32 genes selected based on a prior relationship to insulin physiology. The goal of the study herein was to identify genes involved in the development of IGT and/or IFG, where IGT was defined as a 2 hr glucose level between 140 and 199 mg/dL and IFG defined as a fasting glucose level between 100 and 125 mg/dL. In order to identify and compare genes affecting the development of IGT and/or IFG, we generated two study samples, for which each sample has 3 disease stages (D1) both 2 hr and fasting glucoses normal $(N/N)(n_1 = 60)$, IGT only (IGT/N) $(n_2 = 31)$ and IGT and IFG (IGT/IFG) $(n_3 = 15)$ (D2) both 2 hr and fasting glucoses normal $(N/N)(n_1 = 60)$, IFG only (N/IFG) $(n_2' = 34)$ and IGT and IFG (IGT/IFG) $(n_3 = 15)$.

3. Results and Discussion

3.1. Simulated Multinomial Data

Figure 1 summarizes the results of association analyses when applying the multinomial ordinal probit model with SVD to the simulated data sets. All numbers shown in the figures are the average of the estimates obtained from the 10 simulated data sets. As mentioned previously, we expected 9 strong signals corresponding to the 9 SNPs designed to be associated



(a) Estimates of SNP effects by single SNP analysis (b) P values of testing SNP effects ($-\log_{10}(P \text{ value})$)

Figure 2: Analysis of the simulated data sets with single SNP analysis.

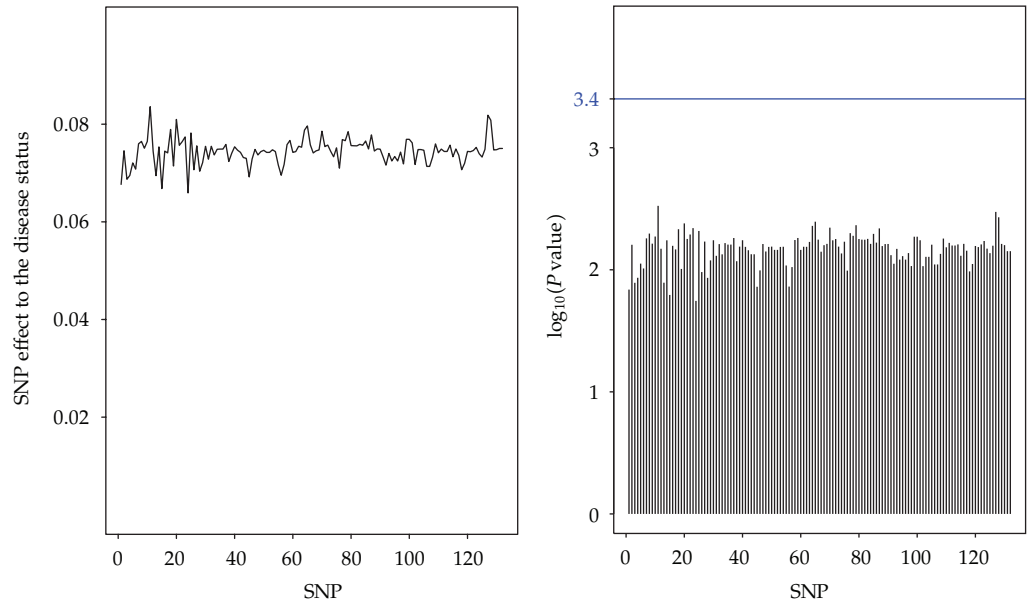
with disease development when generating the simulated data sets, and 9 were observed in our analysis. Similar results from the single SNP analysis were shown in Figure 2.

Figure 1(a) summarizes MLEs of SNP effects calculated with the multinomial ordinal probit model with SVD. The figure shows that almost all MLEs except 9 were between −0.1 and 0.1, while there were 9 large MLEs (4 around 0.3, 5 around −0.3) corresponding to the 9 SNPs contributed to disease status. Figure 1(b) gives P values in $-\log_{10}$ scale for testing SNP effects. The line in Figure 1(b) corresponds to significance level $\alpha = 0.05$. 9 SNPs were clearly separated from the rest and had $-\log_{10}(P \text{ value}) > 1.3$.

Figure 2(a) summarizes MLEs of SNP effects obtained by the single SNP analysis and shows that no signal was strong enough to be distinguished from all other signals. The P values are given in Figure 2(b) in $-\log_{10}$ scale. SNP_{501} in the middle of the figure had a relatively strong signal compared to all others. However, the $-\log_{10}(P \text{ value})$ was much less than 1.3, which corresponds to significance level $\alpha = 0.05$. Thus, no SNPs were identified as statistically significant from the single SNP analysis method. In contrast to the fact that no SNP was identified as statistically significant by the single SNP analysis, the multinomial ordinal probit model with SVD method was able to identify all 9 SNPs contributing to disease status as statistically significant at significance level $\alpha = 0.05$. These results indicated that the proposed method should be reliable for the analysis of large-scale genome-wide association data that have polytomous ordinal responses when $m \gg n$.

3.2. *Mexican-American Coronary Artery Disease (MACAD) Study*

We analyzed the data sets D1 and D2 (see methods) generated from a subsample of subjects recruited through a coronary artery disease proband in the Mexican-American Coronary



(a) Estimates of parameters by single-SNP analysis (b) P values of testing SNP effects ($-\log_{10}(P \text{ value})$)

Figure 3: Analysis of genes for IGT/IFG through IGT pathway (Data Set D1) with single-SNP analysis.

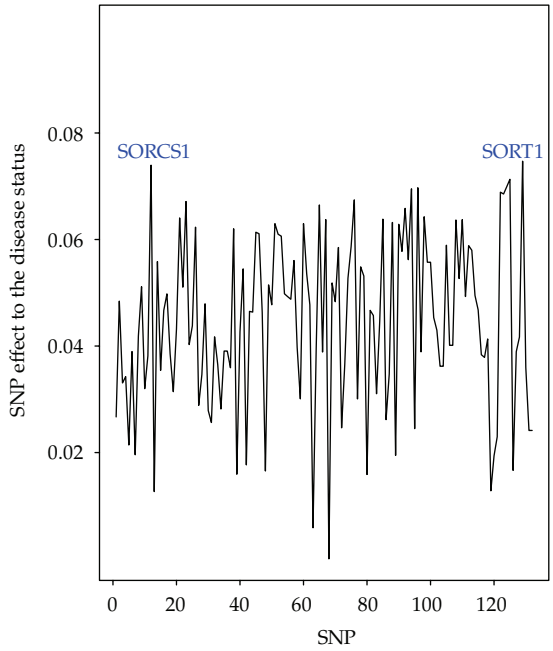
Artery Disease Project as described in the method section, using both the multinomial ordinal probit model with SVD method and the single SNP analysis method.

Figure 3 summarizes the analysis results with the data D1 (N/N-IGT/N-IGT/IFG) using the single SNP analysis. Figure 3(a) gives MLEs of SNP effects. Figure 3(b) plots P values of association analysis in $-\log_{10}$ scale. With Sidak correction, which is often used to correct multiple testing problem, the adjusted significance level should be $1 - (1 - \alpha)^{1/m}$, where α is significance level, and m represents the number of tests. Thus, the corrected $-\log_{10}(P \text{ value})$ threshold for significance level $\alpha = 0.05$ is 3.4, which corresponds to the line in Figure 3(b). We applied the adjusted significance level to the P values in Figure 3(b) since the P values are before correcting multiple testing problem. No SNP was identified as statistically significant (Figure 3(b)).

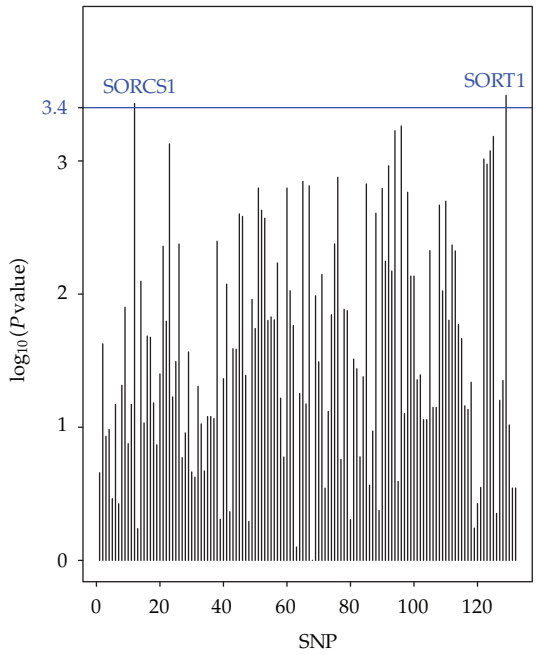
The data set D2 (N/N-N/IFG-IGT/IFG) was analyzed with the same method, and analysis results are given in Figure 4. Figure 4(a) plots MLEs of the SNP effects. Since P values in Figure 4(b) are before the the multiple testing correction, we used 3.4 as the $-\log_{10}(P \text{ value})$ threshold corresponding to 0.05 significance level after the multiple testing correction. Two SNPs corresponding to SORC1 and SORT1 were found significant.

We then also analyzed D1 and D2 with the multinomial ordinal probit model with SVD method. Figure 5 summarizes the analysis results for data D1. Figure 5(a) plots MLEs of SNP effects. Figure 5(b) plots P values in $-\log_{10}$ scale for testing SNP effects. The multiple testing correction does not need to be applied now since the method tests all SNPs simultaneously. With the 1.3 P value threshold, which corresponds to 0.05 significance level, we identified that 8 out of the 32 candidate genes (SORCS1, AMPD1, PPARα, AMPD2, PRKAA2, C5, TCF7L2, and ITR) were associated with the disease path defined in D1.

The multinomial ordinal probit model with SVD method was applied to data set D2 as well. The results are shown in Figure 6. In Figure 6(a), MLEs of the SNP effects were

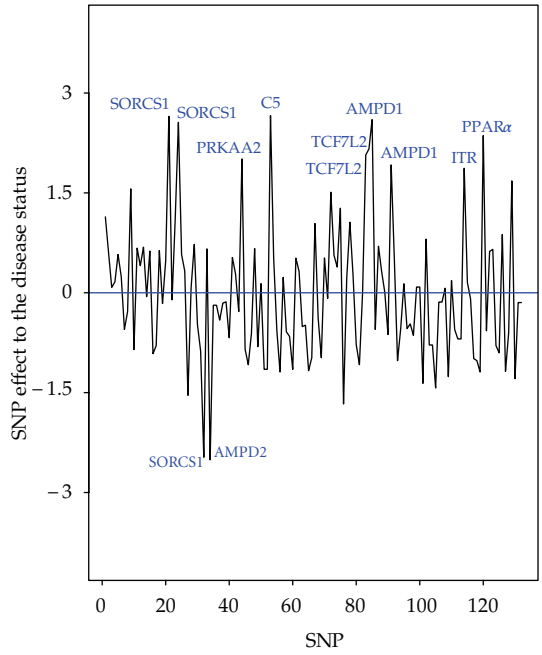


(a) Estimates of parameters by single SNP-analysis

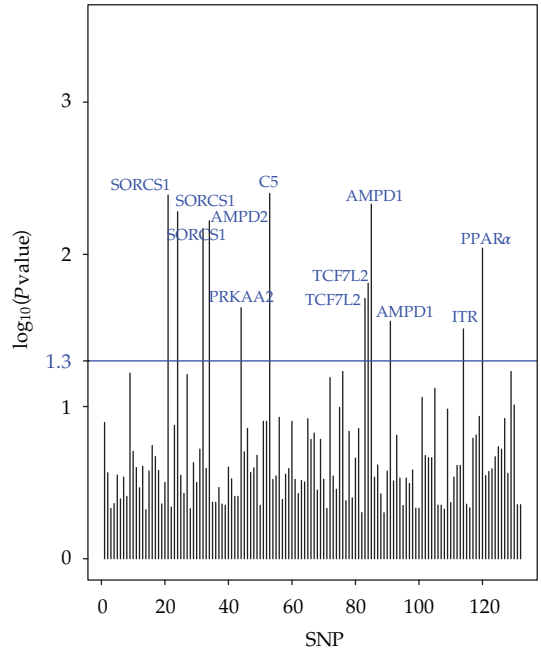


(b) P values of testing SNP effects ($-\log_{10}(P$ value))

Figure 4: Analysis of genes for IGT/IFG through IFG pathway (Data Set D2) with single SNP-analysis.

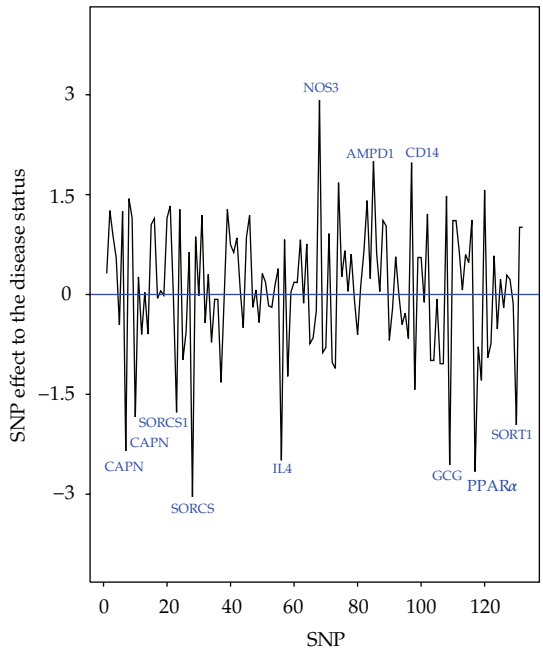


(a) Estimates of parameters by the proposed method

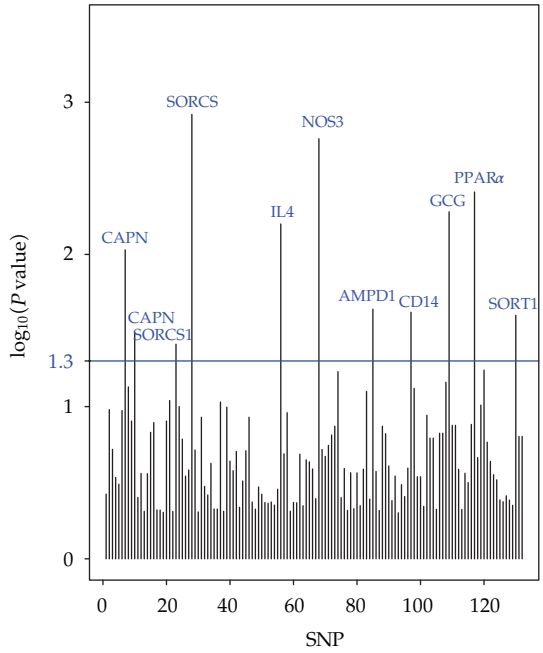


(b) P values of testing SNP effects ($-\log_{10}(P$ value$)$)

Figure 5: Analysis of Genes for IGT/IFG through IGT pathway (Data Set D1) with multinomial ordinal probit model with SVD.



(a) Estimates of parameters by the proposed method



(b) P values of testing SNP effects ($-\log_{10}(P \text{ value})$)

Figure 6: Analysis of genes for IGT/IFG through IFG pathway (Data Set D2) with multinomial ordinal probit model with SVD.

summarized. Figure 6(b) plots the P values in $-\log_{10}$ scale for testing the SNP effects. It showed that 11 SNPs corresponding to 9 out of 32 candidate genes (SORCS1, AMPD1, PPARα, CAPN10, IL4, NOS3, CD14, GCG, and SORT1) have $-\log_{10}(P$ value) greater than the 1.3 P value threshold. From the analyses of D1 and D2, we found that SNPs in 3 genes (SORCS1, AMPD, and PPARα) were associated with both IGT and IFG; SNPs in 5 genes (AMPD2, PRKAA2, C5, TCF7L2, and ITR) were associated with IGT only; SNPs in 6 genes (CAPN, IL4, NOS3, CD14, GCG, and SORT1) were associated with IFG only. These results suggest that IGT and IFG may indicate different pathways to diabetes, with different genetic determinants.

Thus, using both simulated data and a real study sample, we demonstrated that multinomial ordinal probit model with SVD method can be utilized to identify associated markers involved in disease development when multidisease stages are considered. For relatively small size of data set used in the paper, which is 100 samples and 1000 SNPs for the simulation study, the computation took about less than 10 minutes to complete. However, the computation time might be a concern when applying this method to large data set, such as GAWS with millions of SNPs and thousands of samples.

Acknowledgments

This research was partly supported by Mexican American Coronary Artery Disease (MACAD) study Grant NIH-NHLBI HL 088457, Diabetes Endocrinology Research Center (DERC) Grant NIH-NIDDK DK063491, and UCLA Clinical and Translational Science Institute (CTSI) Grant NIH-NCATS UL1TR000124.

References

[1] S. Kwon, D. Wang, and X. Guo, "Application of an iterative Bayesian variable selection method in a genome-wide association study of rheumatoid arthritis," *BMC Proceedings*, vol. 1, supplement 1, article S109, 2007.

[2] E. I. George and R. E. McCulloch, "Approaches for bayesian variable selection," *Statistica Sinica*, vol. 7, no. 2, pp. 339–373, 1997.

[3] S. Kwon, J. Cui, K. D. Taylor, R. Azziz, M. O. Goodarzi, and X. Guo, "Application of Bayesian classification with singular value decomposition method in Genome-wide association study of rheumatoid arthritis," *BMC Proceedings*, vol. 3, supplement 7, article S9, 2009.

[4] W. H. Greene, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1997.

[5] J. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[6] J. Jansen, "Fitting regression models to ordinal data," *Biometrical Journal*, vol. 33, no. 7, pp. 807–815, 1991.

[7] F. A. Graybill, *Theory and Application of the Linear Model*, Duxbury Press, Belmont, Calif, USA, 1976.

[8] M. O. Goodarzi, X. Guo, K. D. Taylor et al., "Determination and use of haplotypes: ethnic comparison and association of the lipoprotein lipase gene and coronary arter disease in Mexican-Americans," *Genetics in Medicine*, vol. 5, no. 4, pp. 322–327, 2003.

[9] M. O. Goodarzi, X. Guo, K. D. Taylor et al., "Lipoprotein lipase is a gene for insulin resistance in Mexican Americans," *Diabetes*, vol. 53, no. 1, pp. 214–220, 2004.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 375935, 19 pages
doi:10.1155/2012/375935

Research Article

Predicting Disease Onset from Mutation Status Using Proband and Relative Data with Applications to Huntington's Disease

Tianle Chen,[1] Yuanjia Wang,[1] Yanyuan Ma,[2] Karen Marder,[3] and Douglas R. Langbehn[4]

[1] *Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA*
[2] *Department of Statistics, Texas A&M University, College Station, TX 77843, USA*
[3] *Departments of Neurology and Psychiatry and Sergievsky Center and the Taub Institute, Columbia University Medical Center, New York, NY 10032, USA*
[4] *Department of Psychiatry and Biostatistics (Secondary), University of Iowa, Iowa City, IA 52242, USA*

Correspondence should be addressed to Yuanjia Wang, yw2016@columbia.edu

Received 15 December 2011; Accepted 22 February 2012

Academic Editor: Yongzhao Shao

Huntington's disease (HD) is a progressive neurodegenerative disorder caused by an expansion of CAG repeats in the IT15 gene. The age-at-onset (AAO) of HD is inversely related to the CAG repeat length and the minimum length thought to cause HD is 36. Accurate estimation of the AAO distribution based on CAG repeat length is important for genetic counseling and the design of clinical trials. In the Cooperative Huntington's Observational Research Trial (COHORT) study, the CAG repeat length is known for the proband participants. However, whether a family member shares the huntingtin gene status (CAG expanded or not) with the proband is unknown. In this work, we use the expectation-maximization (EM) algorithm to handle the missing huntingtin gene information in first-degree family members in COHORT, assuming that a family member has the same CAG length as the proband if the family member carries a huntingtin gene mutation. We perform simulation studies to examine performance of the proposed method and apply the methods to analyze COHORT proband and family combined data. Our analyses reveal that the estimated cumulative risk of HD symptom onset obtained from the combined data is slightly lower than the risk estimated from the proband data alone.

1. Introduction

Huntington's disease (HD) is a severe, autosomal dominantly inherited neurodegenerative disorder that affects motor, cognitive, and psychiatric function and is uniformly fatal. HD is caused by the expansion of CAG trinucleotide repeats at the huntingtin gene (IT15)

[1, 2]. Affected individuals typically begin to show motor signs around 30–50 years of age and typically die 15–20 years after the disease onset [3]. Despite identification of the causative gene, there is currently no treatment that modifies disease progression.

One large genetic epidemiological study of HD, the Cooperative Huntington's Observational Research Trial (COHORT), including 42 Huntington study group research centers in North America and Australia, was initiated in 2005 and concluded in 2011 [4–6]. Participants in COHORT (probands) underwent a clinical evaluation and DNA from whole blood was genotyped for the length of the CAG-repeat huntingtin mutation. Since 2005, COHORT probands from sites with IRB approval have participated in family history interviews and have provided information on HD affection status in their family members. While CAG repeat length is ascertained in probands, the high cost of conducting in-person interviews of family members prevents the collection of all family members' blood samples. However, family members' age-at-onset (AAO) of HD and vital status are obtained through systematic interviews of the probands or the family members themselves. Although a relative's HD genotype is unavailable, the corresponding distribution of the HD gene can be estimated based on the relative's relationship with the proband, the proband's mutation status, and assumptions regarding within-family similarity of CAG length [7, 8].

In a genetic counseling setting, subjects with CAG repeats of 36 or greater are defined as carrying the HD mutation (carrier; [9]), and CAG less than 36 is defined as screened negative, or noncarrier [9]. It is known that there is an inverse association between the CAG repeat length and AAO of HD, that is, the longer the repeat length, the earlier the motor onset [10]. Modeling such a relationship as well as the conditional distribution of HD onset given CAG repeat length accurately and precisely is important for genetic counseling and the design of clinical trials for HD. The AAO of HD onset is subject to right censoring by constraints of the observation periods. Carriers who have not been diagnosed with HD are right-censored for AAO. Several formulae were proposed in the literature to estimate the survival function of age at HD diagnosis given CAG repeat length (e.g., [9–11]). Langbehn et al. [10] have shown that the standard semiparametric survival models, such as the Cox proportional hazards model, do not fit the HD data and proposed a new logistic-exponential parametric model. Specifically, the conditional distribution of HD onset given the CAG repeat length is modeled as a logistic function, with a location and a scale parameter both depending on CAG through nonlinear relationships. Using a large clinical data set, they observed that separate exponential relationships with CAG length gave excellent empirical goodness of fit to both the mean AAO and its variance. Other parametric models, such as Gamma distribution, have also been proposed in the literature [12, 13]. Langbehn et al. [14] examine several AAO models in the literature and show the superior performance of Langbehn et al. [10] in terms of predicting the two-year probability of new HD diagnosis with independent prospective data.

None of the aforementioned existing methods can be directly used to analyze COHORT family data because family members are not always genotyped and their HD mutation status is unknown. The inclusion of family data contributes additional information; however, the unobserved HD mutation sharing status in family members (CAG-elongated or not) complicates the analysis. To see this, note that the affected parent carrying huntingtin mutation has a 50% chance of transmitting the mutation to an offspring. An added complexity is that the likelihood of the offspring having a higher CAG repeat than the parent is higher if the parent is the father. Since the offspring is not genotyped, whether he or she carries expanded CAG repeats is unknown. In this work, we treat the unknown huntingtin gene sharing status in first-degree family members (CAG-elongated or not) as missing data and

use the EM algorithm to carry out the maximum likelihood estimation of the proband and family data jointly. Conditionally on the transmission status in family members, we use the logistic-exponential model in Langbehn et al. [14] to model the AAO as a function of CAG repeat length. We perform simulation studies to examine finite sample performances of the proposed methods. Finally, we apply these methods to analyze the COHORT proband and family combined data. Our results show a slightly lower estimated cumulative risk of HD symptom onset using the combined data compared to using proband data alone.

2. Methods

We start by introducing some notations. For the ith subject, let T_i denote the age-at-onset of HD, let δ_i be the event indicator, let C_i denote the censoring time, and let $X_i = \min(T_i, C_i)$. Let A_i denote the CAG repeat length. Langbehn et al. [10] model distribution of T_i given A_i by a logistic function. The cumulative distribution function (CDF) given A_i is

$$F(t \mid A_i) = \Pr(T_i \le t \mid A_i) = \frac{1}{1 + e^{-[t-\mu(A_i)]/s(A_i)}}, \tag{2.1}$$

and the density function is

$$f(t \mid A_i) = \frac{e^{-[t-\mu(A_i)]/s(A_i)}}{s(A_i)\left\{1 + e^{-[t-\mu(A_i)]/s(A_i)}\right\}^2}. \tag{2.2}$$

Here $\mu(A_i)$ is a location parameter depending on the covariate A_i and $s(A_i)$ is a scale parameter depending on A_i. Let $S(t \mid A_i) = 1 - F(t \mid A_i)$ denote the survival function of HD onset. The location and scale parameters have the following relationship with the mean and variance of T_i given A_i:

$$E(T_i A_i) = \mu(A_i), \qquad \operatorname{var}(T_i A_i) = \pi^2 3 s^2(A_i). \tag{2.3}$$

Various parametric functions for the location and scale parameters were compared in Langbehn et al. [10, 14], and the exponential function provides the best fit. Therefore, we use the same model where

$$\begin{aligned} \mu(A_i) &= \mu_1 + \exp(\mu_2 - \mu_3 A_i), \\ \operatorname{var}(A_i) &= \sigma_1 + \exp(\sigma_2 - \sigma_3 A_i). \end{aligned} \tag{2.4}$$

Substitute these into $F(t \mid A_i)$ and $f(t \mid A_i)$ to obtain a parametric model for the distribution of AAO of HD with six parameters, $\beta = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)^T$. Langbehn et al. [10] fitted estimates of $\beta = (21.54, 9.56, 0.146, 35.55, 17.72, 0.327)^T$.

2.1. Proband-Only Analysis

First, consider probands' data where all A_i's are observed. Since a subject's AAO of HD is subject to the right censoring, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} f^{\delta_i}(X_i \mid A_i; \beta) S^{1-\delta_i}(X_i \mid A_i; \beta), \tag{2.5}$$

and the log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} \left\{ -\delta_i \log[s(A_i)] - \frac{X_i - \mu(A_i)}{s(A_i)} - (1+\delta_i) \log\left[1 + e^{-(X_i - \mu(A_i))/s(A_i)}\right] \right\}. \tag{2.6}$$

The maximum likelihood estimate (MLE) of the parameters, $\widehat{\beta}$, can be obtained via a general-purpose optimization algorithm such as Newton-Raphson or Nelder-Mead implemented in the R program version 2.13.1. The variance-covariance matrix of $\widehat{\beta}$ is estimated by the inverse of the estimated Hessian matrix

$$\widehat{\text{cov}}\left(\widehat{\beta}\right) = \left[H\left(\widehat{\beta}\right)\right]^{-1}. \tag{2.7}$$

The standard error of the estimated survival function, $\widehat{S}(t \mid A_i)$, is then estimated by the Delta method, that is,

$$\widehat{\text{var}}\left[\widehat{S}(t \mid A_i)\right] = G^T\left(\widehat{\beta}\right) \widehat{\text{var}}\left(\widehat{\beta}\right) G\left(\widehat{\beta}\right), \tag{2.8}$$

where the gradient vector

$$G\left(\widehat{\beta}\right) = \left.\frac{\partial S(t \mid A_i)}{\partial \beta}\right|_{\beta=\widehat{\beta}}. \tag{2.9}$$

Since the parameters are estimated by maximum likelihood, it is straightforward to carry out likelihood ratio tests (LRTs) to compare the model fit from the COHORT data with the one obtained by applying parameters from other studies such as Langbehn et al. [10] to the COHORT data. Here, twice the difference in the log-likelihood follows an asymptotic chi-square distribution with 6 degrees of freedom.

2.2. Incorporating Family Members

Next, we consider incorporating family members' AAO data. We do not directly observe whether a family member shares the huntingtin mutation with the proband, but we do have data regarding family members' age-at-onset of the first symptoms, as well as the family members' current ages. When we incorporate the additional family data, the likelihood for the survival takes a mixture form. Let p_i denote the probability of the ith subject sharing

a deleterious allele with a proband and therefore becoming a carrier. Such probabilities are calculated based on Mendelian transmission and a family member's relationship to the proband [8]. For example, offspring and siblings of a carrier proband have a probability of 50% of receiving the huntingtin allele that contains the CAG expansion (Homozygotes for HD are extremely rare since prevalence of HD in general population is rare). We assume that, conditioning on a family member receiving the expanded huntingtin allele, the CAG repeat length is the same as observed in the proband, although this is a simplification [7]. For subjects who receive a wild-type allele (CAG < 36), their probability of developing HD is zero, thus $f(t \mid A_i < 36) = 0$, and $S(tA_i < 36) = 1$, for all t. For the family members, the likelihood is

$$L(\beta) = \prod_{i=1}^{n} \left[p_i f^{\delta_i}(X_i \mid A_i; \beta) S^{1-\delta_i}(X_i \mid A_i; \beta) + (1 - p_i)(1 - \delta_i) \right], \tag{2.10}$$

where the above second term follows from the assumption that noncarriers do not develop HD. Note that for all carrier probands we observe $p_i = 1$, thus the likelihood reduces to (2.5).

The above likelihood can be maximized by a combination of EM and Newton-Raphson algorithms. Let G_i denote the unobserved carrier status indicator for the ith family member (i.e., $G_i = 1$ indicates a family member receives a mutation and $G_i = 0$ indicates otherwise). Then the complete data log-likelihood is

$$\sum_{i=1}^{n} I(G_i = 1) \{ \delta_i \log[f(X_i \mid A_i; \beta)] + (1 - \delta_i) \log[S(X_i \mid A_i; \beta)] \}. \tag{2.11}$$

At the $(k+1)$th iteration of the E-step, we compute the conditional expectation of the complete data log-likelihood, given the observed data. Essentially, we compute

$$\begin{aligned} w_i^{(k+1)} &= E\left[I(G_i = 1) \mid X_i, \delta_i, \beta^{(k)} \right] \\ &= \frac{p_i f^{\delta_i}(X_i \mid A_i; \beta^{(k)}) S^{1-\delta_i}(X_i \mid A_i; \beta^{(k)})}{p_i f^{\delta_i}(X_i \mid A_i; \beta^{(k)}) S^{1-\delta_i}(X_i \mid A_i; \beta^{(k)}) + (1 - p_i)(1 - \delta_i)}. \end{aligned} \tag{2.12}$$

In the M-step, we update $\beta^{(k+1)}$ by maximizing the weighted log-likelihood

$$\sum_{i=1}^{n} w_i^{(k+1)} \{ \delta_i \log[f(X_i \mid A_i; \beta)] + (1 - \delta_i) \log[S(X_i \mid A_i; \beta)] \} \tag{2.13}$$

using the Newton-Raphson algorithm developed for the proband data.

Since for the combined analysis, the parameters are estimated by maximizing the likelihood through an EM algorithm, the standard asymptotic theory applies and the standard errors of parameters can be estimated by inverting the expected or observed information matrix based on the log-likelihood of the observed data. When there is missing data and an EM algorithm is used to obtain the MLE, the information matrix based on the observed data likelihood can be difficult to compute analytically or computationally. In such situations, Louis [15] proposed to compute the observed information matrix in terms of the conditional

moments of the first and second derivatives of the complete data log likelihood which can be obtained easily under the EM algorithm framework. In some cases, these moments are easier to compute than the corresponding derivatives of the incomplete, observed data log-likelihood.

However, in our application, the derivatives of the observed data log likelihood are easy to compute. Thus, we computed the gradient and Hessian matrix of the observed data log-likelihood directly and estimated the standard errors of $\widehat{\beta}$ by the inverse of the Hessian matrix and estimated the standard errors of $\widehat{F}(t)$ by the Delta method similar to the proband-only analysis. Simulation studies in the next section show satisfactory performance of this direct and relatively simpler approach.

3. Simulation Studies

We conducted two simulation studies closely related to the observed COHORT data to illustrate the performance of the Newton-Raphson optimization and the EM algorithm [16]. In all our optimization procedures, we centered both A_i and X_i. Since the direct optimization and EM algorithm need reasonable initial values, we fitted two nonlinear least square (NLS) to the observed sample mean and variance of the AAO on subjects with $\delta_i = 1$. To be specific, we fit

$$m_1(a_i) = \mu_1 + \exp(\mu_2 - \mu_3 a_i), \qquad s_1^2(a_i) = \sigma_1 + \exp(\sigma_2 - \sigma_3 a_i), \tag{3.1}$$

where $m_1(a_i)$ and $s_1^2(a_i)$ are the sample mean and variance for all subjects with $A_i = a_i$, respectively. The six NLS estimators were used as the initial values for further optimization. We denoted the estimated β from the centered data as $\widehat{\beta}_c$. For each simulation, the uncentered $\widehat{\beta}$ were then calculated based on $\widehat{\beta}_c$ and the sample mean of A_i and X_i.

We restricted simulations to CAG repeat lengths between 41 and 56 to guard against sensitivity to the extremely high or low CAG repeats to be consistent with Langbehn et al. [10]. For the analysis of proband data, we generated a sample of 2000 subjects, each with a CAG length ranging from 41 to 56 that follows a multinomial distribution in which the probability $\mathrm{pr}(A_i = a)$ equals to the observed proportion of $A_i = a$ in the COHORT proband data set. The failure times T_i were simulated from the distribution (2.1), where the parameters β were fixed at the values fitted from the COHORT proband data (see next section for their values). The censoring times, C_i, were generated from a rescaled Beta distribution with a scale and shape parameter of four. The parameters for the Beta distribution were chosen so that the proportion of censored subjects is the same in the simulated data and the observed COHORT proband data.

For the analysis of the combined proband and family data, we generated a sample of 4000 subjects. We assume the same proportion of the probands and relatives as observed in the combined COHORT data. For the family members, the probabilities p_i were generated by resampling the observed p_i's in the COHORT data. With a given p_i for each subject, we simulated his or her huntingtin carrier status from a Bernoulli distribution with success probability p_i. For family members simulated to receive an expanded CAG repeat (carriers), their CAG repeats A_i were set to be the same as the probands and their failure times were simulated from (2.5) with β fixed at estimates from the COHORT combined data. For

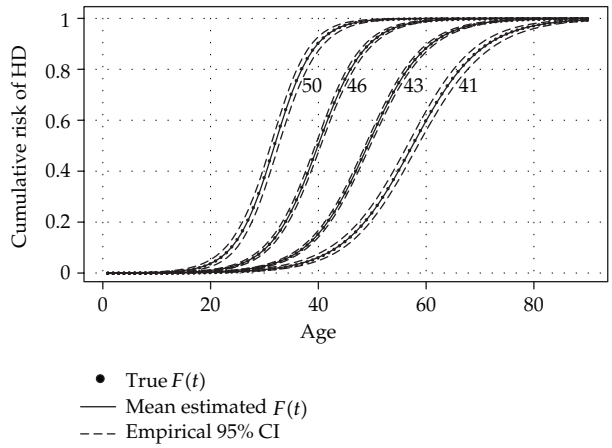


Figure 1: Estimated CDF of HD onset for $A_i = 41, 43, 46$, and 50 with simulated proband data.

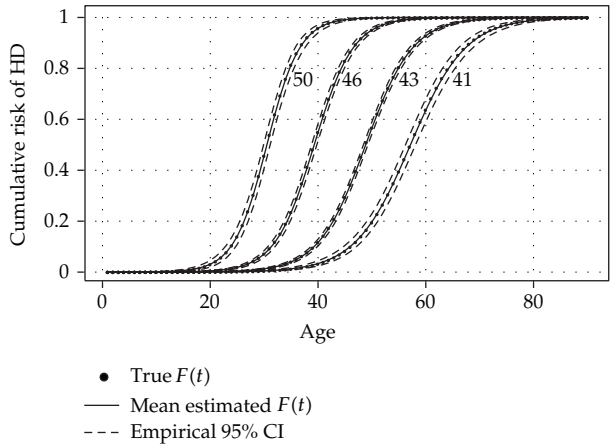


Figure 2: Estimated CDF of HD onset for $A_i = 41, 43, 46$, and 50 with simulated combined proband and relative data.

noncarrier family members, their failure times were set to be infinity and their $X_i = C_i$. We used the same censoring distribution for generating C_i as in the first simulation study.

We provide simulation results of the proband only and combined analyses in Tables 1 and 2. We present mean $\widehat{F}(t \mid A_i)$, empirical standard deviation of $\widehat{F}(t \mid A_i)$, and the mean estimated standard error of $\widehat{F}(t \mid A_i)$ at various ages in. We see from these tables that mean $\widehat{F}(t \mid A_i)$ is very close to true $F(t \mid A_i)$ in both studies. The mean estimated standard errors of $\widehat{F}(t \mid A_i)$ are close to the empirical standard deviations, indicating that the estimation of variability is appropriate. Figures 1 and 2 present three curves of $\widehat{F}(t \mid A_i)$ at $A_i = 41$, 46, 50 and their 95% empirical confidence intervals for the proband data and combined data, respectively. We see that $\widehat{F}(t \mid A_i)$ coincide with the circles representing true $F(t \mid A_i)$ at various ages.

Table 1: Simulation 1 (proband data). Estimated CDF and standard errors from the direct optimization of proband-only analysis, $n = 2000$, 1000 replications.

	CAG = 41				CAG = 46				CAG = 50			
Age	$F(t \mid A_i)$	Mean $\widehat{F}(t \mid A_i)$	Empi: sd	Mean $\widehat{\text{sd}}$	$F(t \mid A_i)$	Mean $\widehat{F}(t \mid A_i)$	Empi: sd	Mean $\widehat{\text{sd}}$	$F(t \mid A_i)$	Mean $\widehat{F}(t \mid A_i)$	Empi: sd	Mean $\widehat{\text{sd}}$
10	0.0001	0.0001	0.0000	0.0000	0.0003	0.0003	0.0001	0.0001	0.0011	0.0012	0.0005	0.0004
20	0.0006	0.0007	0.0002	0.0002	0.0049	0.0048	0.0008	0.0009	0.0301	0.0309	0.0066	0.0060
30	0.0046	0.0049	0.0011	0.0012	0.0717	0.0709	0.0066	0.0068	0.4560	0.4578	0.0253	0.0248
40	0.0322	0.0335	0.0051	0.0054	0.5492	0.5487	0.0171	0.0162	0.9577	0.9572	0.0084	0.0077
50	0.1944	0.1972	0.0162	0.0160	0.9505	0.9509	0.0052	0.0056	0.9984	0.9983	0.0007	0.0006
60	0.6368	0.6358	0.0227	0.0219	0.9967	0.9967	0.0006	0.0007	0.9999	0.9999	0.0000	0.0000
70	0.9272	0.9252	0.0102	0.0108	0.9998	0.9998	0.0001	0.0001	1.0000	1.0000	0.0000	0.0000
80	0.9893	0.9887	0.0025	0.0026	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
90	0.9985	0.9984	0.0005	0.0005	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000

Table 2: Simulation 2 (combined proband and relative data). Estimated CDF and standard errors from the EM algorithm with combined proband and family analysis, $n = 4000$, 1000 replications.

	CAG = 41				CAG = 46				CAG = 50			
Age	$F(t \mid A_i)$	Mean $\widehat{F}(t \mid A_i)$	Empi: sd	Mean $\widehat{\text{sd}}$	$F(t \mid A_i)$	Mean $\widehat{F}(t \mid A_i)$	Empi: sd	Mean $\widehat{\text{sd}}$	$F(t \mid A_i)$	Mean $\widehat{F}(t \mid A_i)$	Empi: sd	Mean $\widehat{\text{sd}}$
10	0.0006	0.0006	0.0002	0.0002	0.0010	0.0010	0.0002	0.0002	0.0025	0.0026	0.0008	0.0008
20	0.0028	0.0029	0.0007	0.0007	0.0102	0.0102	0.0014	0.0014	0.0373	0.0374	0.0069	0.0068
30	0.0134	0.0137	0.0023	0.0023	0.0928	0.0928	0.0069	0.0070	0.3754	0.3751	0.0241	0.0238
40	0.0609	0.0616	0.0069	0.0069	0.5041	0.5042	0.0148	0.0143	0.9031	0.9030	0.0139	0.0132
50	0.2373	0.2378	0.0149	0.0146	0.9099	0.9100	0.0076	0.0074	0.9931	0.9930	0.0020	0.0019
60	0.5987	0.5979	0.0200	0.0188	0.9901	0.9901	0.0015	0.0014	0.9996	0.9995	0.0002	0.0002
70	0.8773	0.8761	0.0133	0.0125	0.9990	0.9990	0.0002	0.0002	1.0000	1.0000	0.0000	0.0000
80	0.9717	0.9710	0.0050	0.0047	0.9999	0.9999	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
90	0.9940	0.9937	0.0015	0.0014	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000

4. COHORT Data Analysis Results

COHORT is a multicenter observational study of individuals in the HD community. COHORT recruitment is open to subjects who have HD symptoms and signs (manifest HD), subjects who have an expanded CAG repeat but have not yet developed symptoms of HD (presymptomatic), subjects who have an HD affected parent but have not been tested and do not have symptoms (at risk), subjects who have an affected grandparent (secondary risk), and control subjects who are not at risk for HD. Information available on participating probands include genetic status (whether or not they carry HD mutation, and the number of CAG repeats), clinical diagnosis of HD, and the timing of symptom onset and timing of diagnosis. In our analyses, only probands with expanded CAG (CAG $\geq$ 36) and their family members were included. Details of the cohort are cited in a publication in press [6].

We first describe the proband and family data in the COHORT study. Information on CAG repeat length and age was available for 1357 probands with CAG repeats varying from 36 to 100 (Table 3). There were 3409 first-degree relatives available from 675 probands. We do not have information on whether some of the probands are from the same family. We show the descriptive statistics for the relatives stratified by relationship type in Table 4. Each proband potentially has three versions of age-at-the-first-symptom (rater's report, subject's self-report, and a family member's report). We gave the rater reported AAO of symptom the highest priority. If the rater reported version is not available, we then used subject report. If neither rater nor subject's self-report is available, we then used the family member's report. Twenty-one subjects whose self-reported and rater-reported AAO of symptom differed by greater than 15 years were removed. Our proband data set has 1151 subjects with CAG length between 41 and 56 and was used for the proband-only analysis. Similar to Langbehn et al. [10], we restricted the analysis to CAG repeat lengths between 41 and 56 to guard against sensitivity to the extremely high or low CAG repeats and against bias due to likely under ascertainment (relative to the population) of subjects with CAG length between 36 and 40.

Information on CAG repeat length, age at time of evaluation and the probability of being a carrier (receiving huntingtin mutation from the proband) was available for 2851 family members of 1151 probands. In the proband data set, both individuals with manifest HD and presymptomatic carriers (24%) are included. Their age-at-diagnosis and age-at-first-motor sign were recorded. Among 1151 probands, 876 (76%) subjects had experienced HD onset and the average AAO of the HD diagnosis was 44 years of age (standard deviation: 10.7). There were 54% females and 94% Caucasians. Our combined proband and family data set has 4002 subjects. In this combined data set, 51% were females and 35% subjects had experienced HD onset. Among the 4002 subjects, 467 are singletons (probands with no family member included). The other 3535 subjects belong to 623 pedigrees with an average size of 5.674 (sd = 2.609) members. In the combined data, there are two different probabilities of being a carrier: $p_i = 1$ (1199 subjects with known CAG expansions or known HD onset) or $p_i = 0.5$ (2803 subjects). Among the 2851 family members, 966 are parents of the probands, 1095 are siblings of the probands, and 790 are children of the probands.

When using the age-at-diagnosis in our proband data as T_i, the estimated cumulative risk of HD is

$$\widehat{F}(t \mid A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}} \frac{[t - 16.284 - \exp(8.325 - 0.111A_i)]}{\sqrt{22.379 + \exp(15.657 - 0.284A_i)}}\right\}\right)^{-1}. \tag{4.1}$$

Table 3: Descriptive statistics of the COHORT proband data.

		Numbers and ages for a CAG repeat length																						
		36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57+	Total
At risk	Number	2	5	15	21	43	55	68	57	31	28	18	9	5	2	1	0	2	0	0	0	0	0	362
	Ave age	61	64	48	55	50	45	42	39	37	31	34	34	27	23	30		35						42
	Min age	60	61	26	37	25	21	21	18	19	19	20	21	20	21	30		23						18
	Max age	62	69	66	70	88	67	71	62	51	44	51	53	40	25	30		47						88
	sd	1	3	11	9	14	11	11	10	9	7	9	9	9	3	.		17						13
	%	0.6	1.4	4.1	5.8	11.9	15.2	18.8	15.7	8.6	7.7	5.0	2.5	1.4	0.6	0.3	0.0	0.6	0.0	0.0	0.0	0.0	0.0	
Affected	Number	2	1	6	7	67	128	148	144	143	93	83	47	34	21	18	9	7	10	6	3	3	15	995
	Ave age	54	68	55	53	60	55	51	48	44	41	38	36	33	31	31	30	28	23	26	26	23	20	45
	Min age	49	68	46	25	37	28	17	19	21	16	25	21	20	19	22	23	22	11	18	25	17	12	11
	Max age	59	68	67	77	82	76	76	67	67	58	53	48	46	44	39	35	35	29	31	29	28	27	82
	sd	7	.	7	19	10	9	9	8	8	8	6	6	6	6	5	4	5	6	5	2	6	4	12
	%	0.2	0.1	0.6	0.7	6.7	12.9	14.9	14.5	14.4	9.3	8.3	4.7	3.4	2.1	1.8	0.9	0.7	1.0	0.6	0.3	0.3	1.5	
Total	Number	4	6	21	28	110	183	216	201	174	121	101	56	39	23	19	9	9	10	6	3	3	15	1357

Table 4: Descriptive statistics of the first-degree relatives of COHORT proband subjects stratified by relationship.

		Relationship			
		Parents	Siblings	Children	Total
Not affected	Number	739	1110	931	2780
	Ave age	70	50	26	42
	Min age	27	0	0	18
	Max age	111	93	62	88
	sd	13	15	14	13
	%	26.6	39.9	33.5	
Affected	Number	379	237	13	629
	Ave age	45	42	36	45
	Min age	18	7	23	11
	Max age	82	70	44	82
	sd	11	11	7	12
	%	60.3	37.7	2.1	
Total	Number	1118	1347	944	3409

Table 5: Mean and standard deviation of the AAO estimated from the model (2.1) for four analyses.

	Langbehn data		COHORT data					
			Probands diagnosis*		Probands symptom**		Combined symptom†	
CAG	Mean	SD	Mean	SD	Mean	SD	Mean	SD
41	57.06	10.50	59.84	8.78	57.74	9.13	59.33	11.68
43	48.06	8.62	51.17	7.31	49.32	7.90	50.63	9.60
46	38.66	7.08	41.29	5.97	39.66	6.57	41.20	7.59
48	34.32	6.57	36.31	5.47	34.75	5.95	36.69	6.79
50	31.08	6.28	32.32	5.16	30.80	5.50	33.21	6.28

* : using proband age-at-diagnosis data;
**: using proband age-at-first-symptom data;
†: using proband and relative combined age-at-first-symptom data.

The estimated parameters for the CDF from the proband-only analysis are slightly different from the ones obtained from Langbehn et al. [10]. Our estimated mean and standard deviation of the AAO of HD is about 1 to 3 years later than the ones obtained in Langbehn et al. [10], and the standard deviation (SD) is slightly smaller (Table 5). In addition, the estimated CDF is smaller for most A_i values using COHORT data. We ran a joint likelihood ratio test on the goodness-of-fit of parameters obtained in Langbehn et al. [10] and the P value was less than 0.001 (test statistic = 66.0). When analyzing the age-at-first-symptom in our proband data, the estimated cumulative risk of HD is

$$\widehat{F}(t \mid A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}} \frac{[t - 14.266 - \exp(7.987 - 0.104A_i)]}{\sqrt{28.933 + \exp(17.130 - 0.312A_i)}}\right\}\right)^{-1}. \tag{4.2}$$

We present $\widehat{F}(t \mid A_i)$ curves for age-at-diagnosis and age-at-symptom at various CAG lengths and their 95% confidence intervals for the proband data in Figure 3. It can be seen that with a given A_i, the estimated probability of having the first symptoms of HD is higher than

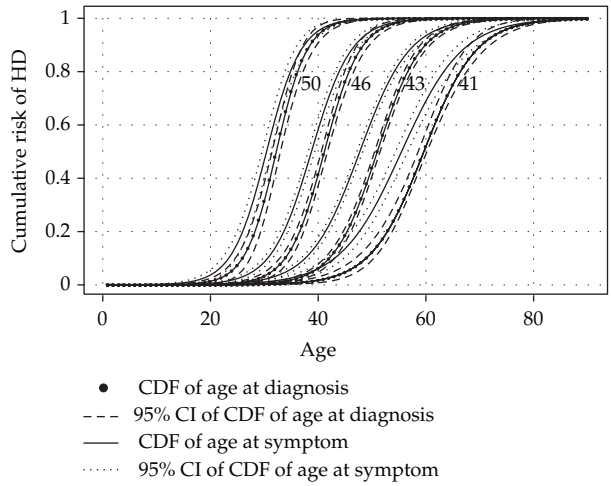


Figure 3: Estimated CDFs of age-at-diagnosis and age-at-first-symptom of HD for $A_i = 41, 43, 46$, and 50 with COHORT proband data.

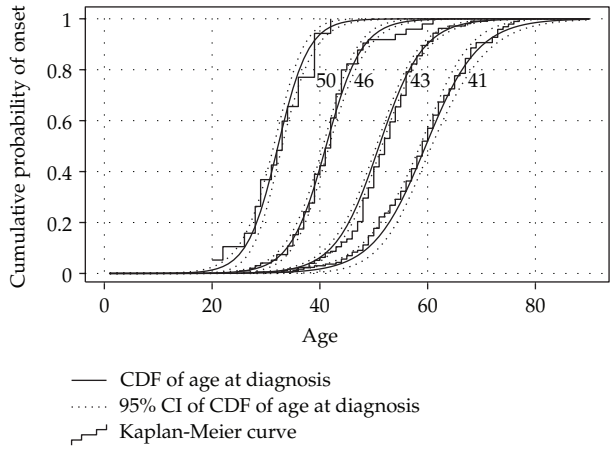


Figure 4: Kaplan-Meier curve and estimated CDF of age-at-diagnosis of HD for $A_i = 41, 43, 46$, and 50 with COHORT proband data.

the probability of a diagnosis of HD at the same age. This is consistent with the intuition that symptoms of HD will be observed before a diagnosis. The mean AAO of first symptom is estimated to be about 2 years earlier than AAO of diagnosis (Table 5) and the standard deviation of the former is slightly larger, indicating that reported age-at-first-symptom is more variable. It is unclear to what extent this difference represents true physical variability in illness development versus possibly lower reliability in the retrospective reporting of symptom onset [17].

As a sensitivity analysis, we compared the estimated CDF based on the parametric model with a nonparametric Kaplan-Meier estimator for subjects with a given A_i. Figure 4 presents this comparison using probands' age-at-diagnosis data. We show in the figure that the parametric model fit is consistent with the Kaplan-Meier fit. However, as expected, the confidence interval for the parametric model estimate at a given age is narrower than

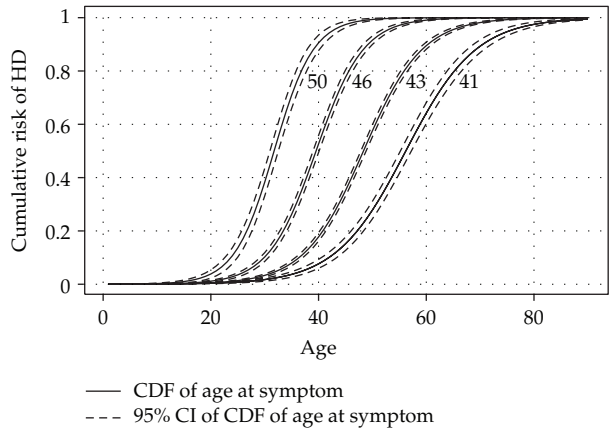


Figure 5: Estimated CDF of age-at-first-symptom of HD for $A_i = 41, 43, 46$, and 50 with COHORT combined proband and relative data.

the Kaplan-Meier estimate (results not shown). The figure comparing age-at-symptom models is similar and therefore omitted.

We reanalyzed only the AAO of the first symptom using the combined proband and family data, since the age-at-diagnosis was not available for family members who are not seen in person. The estimated cumulative risk of HD at age t is

$$\widehat{F}(t \mid A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}} \frac{[t - 18.832 - \exp(8.461 - 0.118A_i)]}{\sqrt{32.365 + \exp(14.823 - 0.248A_i)}}\right\}\right)^{-1}. \tag{4.3}$$

The corresponding $\widehat{F}(t \mid A_i)$ curves at various CAG lengths and their 95% confidence intervals are shown in Figure 5. In Table 5, we compare the estimated mean and SD of the AAO from the proband and combined data. We can see that the estimated mean AAOs for several CAGs are similar regardless of whether family members are included. The SD estimated from the model is larger for the combined data. This is a reflection of the observed data in that there is a wider range of AAO in the combined data than in the proband data. For example, the SD for CAG = 41 of the former is 11 years, whereas it is 10 years in the probands, and the SD for CAG = 42 is 10 in the combined and 8 in the probands.

One of the utilities of the estimated curves is to estimate the conditional probability of having an HD onset (or staying HD free) in the next five or ten years, given a subject has not had an onset by a given age. Similar to Langbehn et al. [10], in Table 6, we present such conditional probabilities in five-year intervals for a subject without HD at age 40 and with given CAG repeats. For example, a 40-year presymptomatic subject with a CAG of 42 has a probability of 34% (CI: 32%, 36%) for developing HD in the next 10 years (by age 50), while for a subject with a CAG of 50 this probability increases to 0.93 (CI: 0.91, 0.95).

5. Discussion

We propose methods to predict disease risk from a known mutation (or to estimate the penetrance function). For most complex diseases, predicting the AAO of a disease

Table 6: Conditional survival probabilities estimated from the COHORT combined data.

CAG	45 years	50 years	55 years	60 years	65 years	70 years
36	0.01 (0.00, 0.02)	0.02 (0.00, 0.04)	0.04 (0.00, 0.08)	0.07 (0.01, 0.13)	0.11 (0.20, 0.20)	0.17 (0.07, 0.28)
37	0.01 (0.00, 0.02)	0.03 (0.01, 0.06)	0.06 (0.02, 0.11)	0.11 (0.05, 0.18)	0.18 (0.27, 0.27)	0.28 (0.17, 0.39)
38	0.02 (0.01, 0.03)	0.05 (0.02, 0.08)	0.10 (0.06, 0.15)	0.18 (0.12, 0.25)	0.29 (0.38, 0.38)	0.43 (0.33, 0.53)
39	0.03 (0.02, 0.04)	0.08 (0.05, 0.11)	0.17 (0.12, 0.21)	0.29 (0.23, 0.35)	0.44 (0.52, 0.52)	0.60 (0.52, 0.69)
40	0.05 (0.04, 0.06)	0.14 (0.11, 0.16)	0.27 (0.23, 0.31)	0.44 (0.39, 0.50)	0.62 (0.68, 0.68)	0.77 (0.72, 0.82)
41	0.08 (0.07, 0.09)	0.22 (0.19, 0.24)	0.41 (0.37, 0.44)	0.61 (0.57, 0.65)	0.78 (0.81, 0.81)	0.88 (0.86, 0.91)
42	0.13 (0.12, 0.14)	0.34 (0.32, 0.36)	0.57 (0.54, 0.60)	0.77 (0.74, 0.79)	0.89 (0.90, 0.90)	0.95 (0.94, 0.96)
43	0.21 (0.20, 0.22)	0.48 (0.46, 0.51)	0.72 (0.70, 0.75)	0.87 (0.86, 0.89)	0.95 (0.95, 0.95)	0.98 (0.97, 0.98)
44	0.31 (0.29, 0.33)	0.63 (0.60, 0.65)	0.83 (0.81, 0.85)	0.93 (0.92, 0.95)	0.97 (0.98, 0.98)	0.99 (0.99, 0.99)
45	0.43 (0.40, 0.45)	0.74 (0.72, 0.77)	0.90 (0.88, 0.92)	0.96 (0.96, 0.97)	0.99 (0.99, 0.99)	>0.99 (0.99, >0.99)
46	0.53 (0.50, 0.56)	0.82 (0.80, 0.85)	0.94 (0.93, 0.95)	0.98 (0.97, 0.99)	0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
47	0.61 (0.57, 0.64)	0.87 (0.85, 0.89)	0.96 (0.95, 0.97)	0.99 (0.98, 0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
48	0.66 (0.63, 0.70)	0.90 (0.88, 0.92)	0.97 (0.96, 0.98)	0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
49	0.70 (0.66, 0.74)	0.92 (0.90, 0.94)	0.98 (0.97, 0.99)	0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
50	0.73 (0.68, 0.77)	0.93 (0.91, 0.95)	0.98 (0.97, 0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
51	0.74 (0.69, 0.80)	0.94 (0.91, 0.96)	0.98 (0.98, 0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
52	0.76 (0.70, 0.82)	0.94 (0.91, 0.97)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
53	0.77 (0.70, 0.83)	0.95 (0.92, 0.98)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
54	0.77 (0.70, 0.85)	0.95 (0.92, 0.98)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
55	0.78 (0.70, 0.86)	0.95 (0.92, 0.99)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)
56	0.78 (0.70, 0.87)	0.95 (0.92, 0.99)	0.99 (0.98, >0.99)	>0.99 (0.99, >0.99)	>0.99 (>0.99, >0.99)	>0.99 (>0.99, >0.99)

from genetic markers such as single-nucleotide polymorphisms (SNPs) continue to be a challenging issue [18]. Even with diseases like HD where the gene is identified, the predictive model can be complicated: a special feature of HD is that the mutation severity is quantifiable and varies significantly among the affected population. This contrasts with the typical categorical approach needed, for example, in genome-wide association studies. The proposed methods are also applicable to other expanded trinucleotide repeat diseases similar to HD.

One of the contributions of this work is to use the family data as well as the proband data to maximize available information in building a model. Our results reveal that the estimated risk obtained from the combined proband and family data is slightly lower than the risk estimated from the proband data alone. It is possible that the proband data consists of a biased clinical sample of gene positive or HD-affected subjects (e.g., subjects with more severe disease or with earlier onset may be more likely to participate; presymptomatic subjects might be undersampled) and is therefore not a fair representative sample of the entire HD population, especially underrepresenting subjects at risk. The plausibility of such underascertainment is so strong for CAG lengths of 40 or less [7] that we did exclude observations within that range from analysis. The family data may be a better representative of the population since the family members are included in the analysis only through the inclusion of the probands. Although proband may participate the study because they had HD or they had more severe symptoms of HD, the relatives were not included based on their CAG repeat lengths or affection status. Of course, some of the family members will not share an expanded CAG repeat huntingtin with the probands and therefore are noncarriers who will never develop HD.

Note that our estimated cumulative risk of onset of a positive HD diagnosis in the proband data is also slightly lower than Langbehn et al. [10] which also examined age-at-HD diagnosis. We estimated later mean AAO for each CAG repeat length shorter than 54 than did Langbehn et al. [10]. For example, the mean AAO of HD diagnosis for probands with a CAG of 42 in the former data was 3 years later and, for a CAG of 43, it was 4 years later (Table 3). On average, for all subjects with a CAG between 41 and 50, the mean AAO in Langbehn data was 2 years earlier than in the COHORT data. More detailed comparisons are presented in Table 5. There are several possible reasons for these differences. The model end point, AAO, should probably be considered to be slightly different in the two models. The outcome in Langbehn et al. [10] was earliest age at which a clinician documented an irreversible objective sign of the illness. This may occur earlier than the point at which an actual diagnosis of manifest HD is given. (Many clinicians wait until there are several such signs.) This may also occur, however, at a point that is later than the proband's or family's first report of subjective symptoms or their first perception of disease signs. In the CAG range of 41–49, the Langbehn et al. means are very close to the symptom onset means in the current data. For longer CAG lengths, the Langbehn et al. estimates more closely resemble the current models for disease diagnosis. Possible systematic variability between the clinicians in the two studies may also account for the differences in the estimates.

Other potential differences between the data sources include potential research-center-specific heterogeneity in diagnostic and rating conventions and slight variations in the methods used to determine CAG repeat length. In the Langbehn study, these were measured by a variety of laboratories while in the COHORT they were all measured in the same laboratory.

We do note that the differences between the fitted models here and those in Langbehn et al. are substantially smaller than differences among other formulae in the literature [14]. AAO probabilities, conditioned on current age, are especially similar. In HD research and

genetic counseling, these conditional probabilities are perhaps the most commonly used statistic deriving from these formulae. Finally, the logistic-exponential form of the parametric model proposed in Langbehn et al. [10] does indeed fit the empirical AAO distributions quite well in the COHORT data. This validates use of this relatively complicated survival model for HD AAO research and may encourage considerations of quantitative biological mechanisms that would generate exponential relationships between CAG and both AAO and its variance.

There has often been ambiguity in the modeling literature concerning the exact meaning of HD "onset." The first onset of observable signs or reportable symptoms of HD generally occurs before the actual diagnosis of clinically manifest HD is given. Much of the earlier modeling literature, reviewed in Langbehn et al. [14], does not clearly address this distinction, although the resultant formulas have often been used for subsequent prediction of HD diagnosis [14]. The event modeled in Langbehn et al. [10] was "the first time that neurological signs representing a permanent change from the normal state was identified in a patient." This might be considered to the concept of "subject's first noted symptom" rather than age of diagnosis. Nonetheless, this model has been used frequently as a predictor of future diagnosis in HD [14]. In the current study, we do distinguish between first symptom onset and diagnosis.

Here, we assumed Mendelian transmission of huntingtin without interference so that the CAG length does not change from parents to offspring. There are several possible violations of these assumptions. CAG lengths do, in reality, vary somewhat among family members, and those inheriting the gene from their father have, on average, a slightly longer CAG repeat length than their father. The probability of this occurring is much lower if inheritance is from the mother [19]. An explanation is that there are many more biological opportunities for the CAG length to change in the father's process of sperm formation than in the mother's process of egg formation. These processes and their dynamics have been studied extensively in vitro [7, 20], but we know of no well-verified in vivo dynamic population genetics models. Assuming the CAG length does not change from father to offspring may lead to a slightly lower estimated risk for affected fathers of probands.

Consistent with Langbehn et al. [10] and other studies [20, 21], we estimated reduced penetrance for lower CAG repeat lengths ($\leq$40). We point out that the parameter estimates from the current model do not include subjects with CAG less than 41; therefore, the risk estimates for these subjects are extrapolations. However, it is conceivable that as long as the inverse relationship between AAO and CAG still holds for the lower CAGs, the life time disease risk for these subjects will be less than 100%, since the life time risk for a CAG of 41 is about 100%.

In the literature, no proportional odds model has been fitted to model the age-at-onset of HD. Proportional odds model, or along a similar line, transformation model, belongs to the semiparametric model framework and is beyond the scope of this paper. We are currently investigating semiparametric models other than the Cox proportional hazards model.

Finally, we stress that our current model does not include other observed covariates, such as additional genetic polymorphisms. In addition, we assumed conditional independence of family members' age-at-onset (AAO) of HD given their CAG repeats. This assumption implies that we do not account for residual correlation among family members' AAO caused by factors other than the CAG repeats, such as life style factors. When there exists such residual correlation, point estimates from our current approach are still consistent hence still valid, although the standard error estimates are no longer correct. A practical limitation of using family members' AAO data is that they may be less reliable than the data directly collected from the probands. This limitation applies to all other diseases, especially those

with late onset. This limitation can be more pronounce when there is incomplete penetrance and variability of phenotype. Future work would consider incorporating such measurement error in the analysis. Lastly, the proposed methods do not include possible unobserved effects that may be site or clinician-specific and perhaps related to the interpretation of the point of "onset." Future research will focus on incorporating observed covariates and adding family-specific random effects to account for residual familial aggregation.

Acknowledgments

Y. Wang's research is supported by NIH Grants R03AG031113-01A2 and R01NS073671-01. Samples and/or data from the COHORT study, which receives support from HP Therapeutics, Inc., were used in this study. The authors thank the Huntington Study Group COHORT investigators and coordinators who collected data and/or samples used in this study, as well as participants and their families, who made this work possible.

References

[1] C. A. Ross, "When more is less: pathogenesis of glutamine repeat neurodegenerative diseases," *Neuron*, vol. 15, no. 3, pp. 493–496, 1995.

[2] C. A. Ross and S. J. Tabrizi, "Huntington's disease: from molecular pathogenesis to clinical treatment," *The Lancet Neurology*, vol. 10, pp. 83–98, 2010.

[3] T. Foroud, J. Gray, J. Ivashina, and P. M. Conneally, "Differences in duration of Huntington's disease based on age at onset," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 66, no. 1, pp. 52–56, 1999.

[4] K. Kieburtz and Huntington Study Group, "The unified Huntington's disease rating scale: reliability and consistency," *Movement Disorder*, vol. 11, pp. 136–142, 1996.

[5] E. R. Dorsey, C. A. Beck, M. Adams et al., "TREND-HD communicating clinical trial results to research participants," *Archives of Neurology*, vol. 65, no. 12, pp. 1590–1595, 2008.

[6] E. R. Dorsey and Huntington Study Group COHORT Investigators, "Characterization of a large group of individuals with Huntington disease and their relatives enrolled in the COHORT study," *PLoS ONE*, vol. 7, no. 2, Article ID e29522, 2012.

[7] D. Falush, E. W. Almqvist, R. R. Brinkmann, Y. Iwasa, and M. R. Hayden, "Measurement of mutational flow implies both a high new-mutation rate for huntington disease and substantial under ascertainment of late-onset cases," *The American Journal of Human Genetics*, vol. 68, pp. 373–385, 2000.

[8] Y. Wang, L. N. Clark, E. D. Louis et al., "Risk of Parkinson disease in carriers of Parkin mutations: estimation using the kin-cohort method," *Archives of Neurology*, vol. 65, no. 4, pp. 467–474, 2008.

[9] D. C. Rubinsztein, J. Leggo, R. Coles et al., "Phenotypic characterization of individuals with 30–40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36–39 repeats," *American Journal of Human Genetics*, vol. 59, no. 1, pp. 16–22, 1996.

[10] D. R. Langbehn, R. R. Brinkman, D. Falush, J. S. Paulsen, and M. R. Hayden, "A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length," *Clinical Genetics*, vol. 65, no. 4, pp. 267–277, 2004.

[11] O. C. Stine, N. Pleasant, M. L. Franz, M. H. Abbott, S. E. Folstein, and C. A. Ross, "Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15," *Human Molecular Genetics*, vol. 2, no. 10, pp. 1547–1549, 1993.

[12] C. Gutierrez and A. MacDonald, *Huntington Disease and Insurance. I: A Model of Huntington Disease*, Genetics and Insurance Research Centre (GIRC), Edinburgh, UK, 2002.

[13] C. Gutierrez and A. MacDonald, "Huntington disease, critical illness insurance and life insurance," *Scandinavian Actuarial Journal*, vol. 4, pp. 279–313, 2004.

[14] D. R. Langbehn, M. R. Hayden, and J. S. Paulsen, "CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches," *American Journal of Medical Genetics*, vol. 153, no. 2, pp. 397–408, 2010.

[15] T. Louis, "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 44, pp. 226–233, 1982.

[16] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[17] K. Marder, G. Levy, E. D. Louis et al., "Accuracy of family history data on Parkinson's disease," *Neurology*, vol. 61, no. 1, pp. 18–23, 2003.

[18] J. Kang, J. Cho, and H. Zhao, "Practical issues in building risk-predicting models for complex diseases," *Journal of Biopharmaceutical Statistics*, vol. 20, no. 2, pp. 415–440, 2010.

[19] B. Kremer, E. Almqvist, J. Theilmann et al., "Sex-dependent mechanisms for expansions and contractions of the CAG repeat on affected Huntington disease chromosomes," *American Journal of Human Genetics*, vol. 57, no. 2, pp. 343–350, 1995.

[20] C. T. McMurray, "Mechanisms of trinucleotide repeat instability during human development," *Nature Reviews Genetics*, vol. 11, no. 11, pp. 786–799, 2010.

[21] R. R. Brinkman, M. M. Mezei, J. Theilmann, E. Almqvist, and M. R. Hayden, "The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size," *The American Journal of Human Genetics*, vol. 60, no. 5, pp. 1202–1210, 1997.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 537474, 15 pages
doi:10.1155/2012/537474

Research Article

Testing Homogeneity in a Semiparametric Two-Sample Problem

Yukun Liu,[1] Pengfei Li,[2] and Yuejiao Fu[3]

[1] *Department of Statistics and Actuarial Science, School of Finance and Statistics, East China Normal University, Shanghai 200241, China*
[2] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[3] *Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3*

Correspondence should be addressed to Yuejiao Fu, yuejiao@mathstat.yorku.ca

Received 18 November 2011; Accepted 24 January 2012

Academic Editor: Yongzhao Shao

We study a two-sample homogeneity testing problem, in which one sample comes from a population with density $f(x)$ and the other is from a mixture population with mixture density $(1-\lambda)f(x)+\lambda g(x)$. This problem arises naturally from many statistical applications such as test for partial differential gene expression in microarray study or genetic studies for gene mutation. Under the semiparametric assumption $g(x) = f(x)e^{\alpha+\beta x}$, a penalized empirical likelihood ratio test could be constructed, but its implementation is hindered by the fact that there is neither feasible algorithm for computing the test statistic nor available research results on its theoretical properties. To circumvent these difficulties, we propose an EM test based on the penalized empirical likelihood. We prove that the EM test has a simple chi-square limiting distribution, and we also demonstrate its competitive testing performances by simulations. A real-data example is used to illustrate the proposed methodology.

1. Introduction

Let $x_1, \ldots, x_{n_0}$ be a random sample from a population with distribution function F, and let $y_1, \ldots, y_{n_1}$ be a random sample from a population with distribution function H. Testing whether the two populations have the same distribution, that is, $H_0 : F = H$ versus $H_1 : F \neq H$, with both F and H completely unspecified, will require a nonparametric test. Since $H_1 : F \neq H$ is a very broad hypothesis, many times one may want to consider some more specified alternative, for example, the two populations only differ in location. In the present paper, we will consider a specified alternative in which one of the two samples has a mixture structure. More specifically, we have

$$x_1, \ldots, x_{n_0} \overset{\text{i.i.d.}}{\sim} f(x), \qquad y_1, \ldots, y_{n_1} \overset{\text{i.i.d.}}{\sim} h(y) = (1-\lambda)f(y) + \lambda g(y), \tag{1.1}$$

where $f(x) = dF(x)/dx$, $g(y) = dG(y)/dy$, $h(y) = dH(y)/dy$, and $\lambda \in (0,1)$ is an unknown parameter sometimes called contamination proportion. The problem of interest is to test H_0 : $f = h$ or equivalently $\lambda = 0$. This particular two-sample problem arises naturally in a variety of statistical applications such as test for partial differential gene expression in microarray study, genetic studies for gene mutation, case-control studies with contaminated controls, or the test of a treatment effect in the presence of nonresponders in biological experiments (see Qin and Liang [1] for details).

If no auxiliary information is available, this is merely the usual two-sample goodness-of-fit problem. There has been extensive literature on it; see Zhang [2] and references therein. However, these tests are not suitable for the specific alternative with a mixture structure as they might be inferior comparing with methods that are designed for the specific alternative. In this paper, we will propose an empirical likelihood-based testing procedure for this specific mixture alternative under Anderson's semiparametric assumption [3]. Motivated by the logistic regression model, the semiparametric assumption proposed by Anderson [3] links the two distribution functions F and G through the following equation:

$$\log \frac{g(x)}{f(x)} = \alpha + \beta x, \tag{1.2}$$

where α and β are both unknown parameters. There are many examples where the logarithm of the density ratio is linear in the observations.

Example 1.1. Let F and G be the distribution functions of Binomial (m, p_1) and Binomial (m, p_2), respectively. We refer the densities f and g to the probability mass functions corresponding to F and G, respectively. Then,

$$\log \frac{g(x)}{f(x)} = m \log\left\{\frac{1-p_2}{1-p_1}\right\} + \log\left\{\frac{p_2(1-p_1)}{p_1(1-p_2)}\right\} x. \tag{1.3}$$

Example 1.2. Let F be the distribution function of $N(\mu_1, \sigma^2)$ and G the distribution function of $N(\mu_2, \sigma^2)$. Then,

$$\log \frac{g(x)}{f(x)} = \frac{1}{2\sigma^2}\left(\mu_1^2 - \mu_2^2\right) + \frac{1}{\sigma^2}(\mu_2 - \mu_1)x. \tag{1.4}$$

In practice, one may need to apply some sort of transformation to the data (e.g., logarithm transformation) in order to justify the use of the semiparametric model assumption (1.2).

Example 1.3. Let F and G be the distribution functions of $\log N(\mu_1, \sigma^2)$ and $\log N(\mu_2, \sigma^2)$, respectively. It is clear that the density ratio is a linear function of the log-transformed data:

$$\log \frac{g(x)}{f(x)} = \frac{1}{2\sigma^2}\left(\mu_1^2 - \mu_2^2\right) + \frac{1}{\sigma^2}(\mu_2 - \mu_1)\log x. \tag{1.5}$$

Example 1.4. Let F and G be the distribution functions of Gamma (m_1, θ) and Gamma (m_2, θ), respectively. In this case,

$$\log \frac{g(x)}{f(x)} = \log\left\{\frac{\Gamma(m_1)}{\Gamma(m_2)}\right\} + (m_1 - m_2)\log\theta + (m_2 - m_1)\log x. \tag{1.6}$$

The semiparametric modeling assumption (1.2) is very flexible and has the advantage of not putting any specific restrictions on the functional form of f. Under this assumption, various approaches have been proposed to test homogeneity in the two-sample problem (see [1, 4, 5] and references therein). This paper adds to this literature by introducing a new type of test statistics which are based on the empirical likelihood [6, 7].

The empirical likelihood (EL) is a nonparametric likelihood method which has many nice properties paralleling to the likelihood methods, for example, it is range-preserving, transform-respect, Bartlett correctable, and a systematic approach to incorporating auxiliary information [8–11]. In general, if the parameters are identifiable, the empirical likelihood ratio (ELR) test has a chi-square limiting distribution under null hypothesis. However, for the aforementioned testing problem, the parameters under H_0 are not identifiable, which results in an intractable null limiting distribution for the ELR test. To circumvent this problem, we would add a penalty to the log EL to penalize λ being too close to zero. Working like a soft threshold, the penalty makes the parameters roughly identifiable. Intuitively, the penalized (or modified) ELR test should restore the usual chi-square limiting distribution. Unfortunately two things hinder the direct use of the penalized ELR test. One is that, to the best of our knowledge, there is no feasible algorithm to compute the penalized ELR test statistic. The other one is that there has been no research on the asymptotic properties of the penalized ELR test. Therefore, one cannot obtain critical values for the penalized ELR test regardless through simulations or an asymptotic reference distribution. We find that the EM test [12, 13] based on the penalized EL is a nice solution to the testing problem.

The remainder of this paper is organized as follows. In Section 2, we introduce the ELR and the penalized ELR. The penalized EL-based EM test is given in Section 3. A key computational issue of the EM test is discussed in Section 4. Sections 5 and 6 contain a simulation study and a real-data application, respectively. For clarity, all proofs are postponed to the appendix.

2. Empirical Likelihood

Let $\{t_1, \ldots, t_{n_0}, t_{n_0+1}, \ldots, t_n\} = \{x_1, \ldots, x_{n_0}, y_1, \ldots, y_{n_1}\}$ denote the combined two-sample data, where $n = n_0 + n_1$. Under Anderson's semiparametric assumption (1.2), the likelihood of two-sample data (1.1) is

$$L = \prod_{i=1}^{n_0} dF(t_i) \prod_{j=n_0+1}^{n} \left[1 - \lambda + \lambda e^{\alpha+\beta t_j}\right] dF(t_j). \tag{2.1}$$

Let $p_h = dF(t_h)$, $h = 1, \ldots, n$. The EL is just the likelihood L with constraints $p_h \geq 0$, $\sum_{h=1}^{n} p_h = 1$ and $\sum_{h=1}^{n} p_h(e^{\alpha+\beta t_h} - 1) = 0$. The corresponding log-EL is

$$l = \sum_{h=1}^{n} \log p_h + \sum_{j=1}^{n_1} \log\left[1 - \lambda + \lambda e^{\alpha+\beta y_j}\right]. \tag{2.2}$$

We are interested in testing

$$H_0 : \lambda = 0 \quad \text{or} \quad (\alpha, \beta) = (0, 0). \tag{2.3}$$

Under the null hypothesis, the constraint $\sum_{h=1}^{n} p_h(e^{\alpha+\beta t_h} - 1) = 0$ will always hold and $\sup_{H_0} l = -n \log n$. Under alternative hypothesis, for any fixed (λ, α, β), maximizing l with respect to p_h's leads to the log-EL function of (λ, α, β):

$$l(\lambda, \alpha, \beta) = -\sum_{h=1}^{n} \log\left[1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)\right] - n \log n + \sum_{j=1}^{n_1} \log\left[1 - \lambda + \lambda e^{\alpha+\beta y_j}\right], \tag{2.4}$$

where ξ is the solution to the following equation:

$$\sum_{h=1}^{n} \frac{e^{\alpha+\beta t_h} - 1}{1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)} = 0. \tag{2.5}$$

Hence, the EL ratio function $R(\lambda, \alpha, \beta) = 2\{l(\lambda, \alpha, \beta) + n \log n\}$ and the ELR is denoted as $R = \sup R(\lambda, \alpha, \beta)$.

The null hypothesis H_0 holds for $\lambda = 0$ regardless of (α, β), or $(\alpha, \beta) = (0, 0)$ regardless of λ. This implies that the parameter (λ, α, β) is not identifiable under H_0, resulting in rather complicated asymptotic properties of the ELR. One may consider the modified or penalized likelihood method [14] and define the penalized log-EL function $pl(\lambda, \alpha, \beta) = l(\lambda, \alpha, \beta) + \log(\lambda)$. Accordingly the penalized EL ratio function is

$$\begin{aligned} pR(\lambda, \alpha, \beta) &= 2\{pl(\lambda, \alpha, \beta) - pl(1, 0, 0)\} \\ &= -2\sum_{h=1}^{n}\left[1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)\right] \\ &\quad + 2\sum_{j=1}^{n_1} \log\left(1 - \lambda + \lambda e^{\alpha+\beta y_j}\right) + 2\log(\lambda), \end{aligned} \tag{2.6}$$

where ξ is the solution to (2.5). The penalty function $\log(\lambda)$ goes to $-\infty$ as λ approaches 0. Therefore, λ is bounded away from 0, and the null hypothesis in (2.3) then reduces to $(\alpha, \beta) = (0, 0)$. That is, the parameters in the penalized log-EL function is asymptotically identifiable. However, the asymptotic behavior of the penalized ELR test is still complicated. Meanwhile, the computation of the penalized ELR test statistic is another obstacle of the implementation of the penalized ELR method. No feasible and stable algorithm has been found for this purpose. An EL-based EM test proposed in this paper provides an efficient way to solve the problem.

3. EL-Based EM Test

Motivated by Chen and Li [12] and Li et al. [13], we propose an EM test based on the penalized EL to test the hypothesis (2.3). The EM test statistics are derived iteratively. We first

choose a finite set of $\Lambda = \{\lambda_1, \ldots, \lambda_L\} \subset (0,\ 1]$, for instance, $\Lambda = \{0.1, 0.2, \ldots, 0.9, 1.0\}$, and a positive integer K (2 or 3 in general). For each $l = 1, \ldots, L$, we proceed the following steps.

Step 1. Let $k = 1$ and $\lambda_l^{(k)} = \lambda_l$. Calculate $(\alpha_l^{(k)}, \beta_l^{(k)}) = \operatorname{argmax}_{\alpha,\beta}\ p\ R(\lambda_l^{(k)}, \alpha, \beta)$.

Step 2. Update (λ, α, β) by using the following algorithm for $K - 1$ times.

Substep 2.1. Calculate the posterior distribution,

$$w_{jl}^{(k)} = \frac{\lambda_l^{(k)}\ \exp\left(\alpha_l^{(k)} + \beta_l^{(k)} y_j\right)}{1 - \lambda_l^{(k)} + \lambda_l^{(k)} \exp\left(\alpha_l^{(k)} + \beta_l^{(k)} y_j\right)}, \quad j = 1, \ldots, n_1, \tag{3.1}$$

and update λ by

$$\lambda_l^{(k+1)} = \arg\max_{\lambda} \left\{ \sum_{j=1}^{n_1} \left(1 - w_{jl}^{(k)}\right) \log(1 - \lambda) + \sum_{j=1}^{n_1} w_{jl}^{(k)} \log(\lambda) + \log(\lambda) \right\}. \tag{3.2}$$

Substep 2.2. Update (α, β) by $(\alpha_l^{(k+1)}, \beta_l^{(k+1)}) = \operatorname{argmax}_{\alpha,\beta} pR(\lambda_l^{(k+1)}, \alpha, \beta)$.

Substep 2.3. Let $k = k + 1$ and continue.

Step 3. Define the test statistics $M_n^{(K)}(\lambda_l) = pR(\lambda_l^{(K)}, \alpha_l^{(K)}, \beta_l^{(K)})$.

The EM test statistic is defined as

$$\mathrm{EM}_n^{(K)} = \max\left\{ M_n^{(K)}(\lambda_l), l = 1, \ldots, L \right\}. \tag{3.3}$$

We reject the null hypothesis H_0 when the EM test statistic is greater than some critical value determined by the following limiting distribution.

Theorem 3.1. *Suppose $\rho = n_1/n \in (0,1)$ is a constant. Assume the null hypothesis H_0 holds and $E(t_h) = 0$ and $\mathrm{Var}(t_h) = \sigma^2 \in (0,\infty)$ for $h = 1, \ldots, n$. For $l = 1, \ldots, L$ and any fixed k, it holds that*

$$\lambda_l^{(k)} - \lambda_l = o_p(1), \qquad \alpha_l^{(k)} = O_p\left(n^{-1}\right), \qquad \beta_l^{(k)} = \frac{\overline{y} - \overline{x}}{\lambda_l \sigma^2} + o_p\left(n^{-1/2}\right), \tag{3.4}$$

where $\overline{x} = (1/n_0) \sum_{i=1}^{n_0} x_i$ and $\overline{y} = (1/n_1) \sum_{j=1}^{n_1} y_j$.

Remark 3.2. The assumption $Et_h = 0$ is only for convenience purpose and unnecessary. Otherwise, we can replace t_h and α with $t_h - E(t_h)$ *and* $\alpha + \beta E(t_h)$.

Theorem 3.3. *Assume the conditions of Theorem 3.1 hold and* $1 \in \Lambda$. *Under the null hypothesis* (2.3), $\mathrm{EM}_n^{(K)} \to \mathcal{O}_1^2$ *in distribution, as* $n \to \infty$.

We finish this section with an additional remark.

Remark 3.4. We point out that the idea of the EM-test can also be generalized to more general models such as $\log(g(x)/f(x)) = \alpha + \beta_1 x + \cdots + \beta_k x^k$ for some integer k or $\log(g(x)/f(x)) = \alpha + \beta_1 t_1(x) + \cdots + \beta_k t_k(x)$ with $t_i(\cdot)$'s being known functions.

4. Computation of the EM Test

A key step of the EM test procedure is to maximize $pR(\lambda, \alpha, \beta)$ with respect to (α, β) for fixed λ. In this section, we propose a computation strategy which provides stable solution to this optimization problem. Throughout this section, λ is suppressed to be fixed.

The objective function is $pR(\lambda, \alpha, \beta) = G(\xi_*, \alpha, \beta)$ where

$$G(\xi, \alpha, \beta) = -2\sum_{h=1}^{n} \log\left[1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)\right] + 2\sum_{j=1}^{n_1} \log\left[1 - \lambda + \lambda e^{\alpha+\beta y_j}\right] + 2\log(\lambda) \tag{4.1}$$

and $\xi_* = \xi_*(\alpha, \beta)$ is the solution to

$$\frac{\partial G}{\partial \xi} = -2\sum_{h=1}^{n} \frac{e^{\alpha+\beta t_h} - 1}{1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)} = 0. \tag{4.2}$$

If (α, β) is the maximum point of $pR(\lambda, \alpha, \beta)$, it should generally satisfy

$$\frac{\partial G}{\partial \alpha} = -2\sum_{h=1}^{n} \frac{\xi e^{\alpha+\beta t_h}}{1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)} + 2\sum_{j=1}^{n_1} \frac{\lambda e^{\alpha+\beta y_j}}{1 - \lambda + \lambda e^{\alpha+\beta y_j}} = 0. \tag{4.3}$$

Combining (4.2) and (4.3) leads to

$$\xi = \frac{1}{n}\sum_{j=1}^{n_1} \frac{\lambda e^{\alpha+\beta y_j}}{1 - \lambda + \lambda e^{\alpha+\beta y_j}}. \tag{4.4}$$

Putting this expression of ξ back into (4.1), we have a new function

$$\begin{aligned} H(\alpha, \beta) = &-2\sum_{h=1}^{n} \log\left\{1 + \left(e^{\alpha+\beta t_h} - 1\right)\frac{1}{n}\sum_{j=1}^{n_1} \frac{\lambda e^{\alpha+\beta y_j}}{1 - \lambda + \lambda e^{\alpha+\beta y_j}}\right\} \\ &+ 2\sum_{j=1}^{n_1} \log\left(1 - \lambda + \lambda e^{\alpha+\beta y_j}\right). \end{aligned} \tag{4.5}$$

It can be verified that $H(\alpha, \beta)$ is almost surely concave in a neighborhood of $(0, 0)$ given λ, which means that maximizing $H(\alpha, \beta)$ with respect to (α, β) gives the maximum of $pR(\lambda, \alpha, \beta)$ for fixed λ. The stability of the method is illustrated by the following simulation study.

5. Simulation Study

We consider two models in Examples 1.3 and 1.4 with $\mu_1 = 0$, $\mu_2 = \mu$, and $\sigma^2 = 1$ for Example 1.3, and $m_1 = 1$, $m_2 = m$, and $\theta = 1$ for Example 1.4. Nominal levels of 0.01, 0.05, and 0.10 are considered. The logarithm transformation is applied to the original data before using the EM test. The initial set $\Lambda = \{0.1, 0.2, \ldots, 1\}$ and iteration number $K = 3$ are used to calculate the EM test statistic.

One competitive method for testing homogeneity under the semiparametric two-sample model is the score test proposed by Qin and Liang [1]. This method is based on

$$S(\alpha, \beta) = \left.\frac{\partial l(\lambda, \alpha, \beta)}{\partial \lambda}\right|_{\lambda=0} = \sum_{j=1}^{n_1} \left(e^{\alpha+\beta y_j} - 1\right), \tag{5.1}$$

where $l(\lambda, \alpha, \beta)$ is the log empirical likelihood function given in (2.4). Let $(\widehat{\alpha}_1, \widehat{\beta}_1) = \operatorname{argmax}_{\alpha,\beta} l(1, \alpha, \beta)$. The score test statistic was defined as $T_1 = S(\widehat{\alpha}_1, \widehat{\beta}_1)/(1+n_1/n_0)$, which has a χ_1^2 limiting distribution under the null hypothesis.

We compare the EM test and the score test in terms of type I error and power. We calculate the type I errors of each method under the null hypothesis based on 20,000 repetitions and the power under the alternative models based on 2,000 repetitions. For fair comparison, simulated critical values are used to calculate the power. We consider two sample sizes: 50 and 200 and $K = 1, 2, 3$. Tables 1 and 2 contain the simulation results for the log-normal models and Tables 3 and 4 for the gamma models.

The results show that the EM test and the score test have similar type I errors. For both methods, the type I errors are somehow larger than the nominal levels when the sample size is $n = 50$; they are close to the nominal levels when the sample size is increased to $n = 200$. For the log-normal models, two methods have almost the same power when the alternatives are close to each other such as $\mu = 1$; the EM test becomes much more powerful when the alternatives are distant and the sample size increases. In the case of $n = 50$, $\lambda = 0.2$, $\mu = 3$, and nominal level 0.01, the EM test has a 10% gain in power compared with the score test; the gain rushes up to almost 30% when $\lambda = 0.1$, $\mu = 3$, and the sample size increases to $n = 200$. For the gamma models, the advantage of the EM test is more obvious. For both sample sizes $n = 50$ and 200, the EM test is more powerful than the score test.

6. Real Example

We apply our EM test procedure to the drug abuse data [15] in a study of addiction to morphine in rats. In this study, rats got morphine by pressing a lever and the frequency of lever presses (self-injection rates) after six-day treatment with morphine was recorded as response variable. The data consist of the number of lever presses for five groups of rats: four treatment groups with different dose levels and one saline group (control group).

We analyzed the response variables (the number of lever presses by rats) of the treatment group at the first dose level and the control group. The data is tabulated in Table 3 of Fu et al. [5]. Following Boos and Browine [16] and Fu et al. [5], we analyze the transformed data, $\log_{10}(R+1)$ with R being the number of lever presses by rats. Instead of using the parametric models as Boos and Browine [16] and Fu et al. [5], we adopt Anderson's semiparametric approach. That is, we assume that the response variables in control group comes from $f(x)$,

Table 1: Type I error and power comparisons (%) of the EM test and the score test (SC test) for log-normal model: $n_0 = n_1 = 50$.

λ	μ	Level	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	SC test
0		10	11.9	12.2	12.2	11.5
0		5	6.3	6.5	6.5	6.4
0		1	1.6	1.6	1.6	1.9
0.1	1	10	14.8	14.5	14.5	14.6
0.1	1	5	8.5	8.6	8.6	8.6
0.1	1	1	2.5	2.5	2.5	2.4
0.1	2	10	27.2	28.1	28	25.6
0.1	2	5	17.8	18.4	18.4	16.7
0.1	2	1	6.6	6.7	6.7	6.2
0.1	3	10	47.1	48.3	48.3	41.4
0.1	3	5	34.2	35.6	35.4	30.8
0.1	3	1	15.4	15.6	15.6	14.9
0.2	1	10	25.5	25.9	26	25.6
0.2	1	5	16.4	16.4	16.4	17
0.2	1	1	5.4	5.3	5.3	5.7
0.2	2	10	62.2	62.7	62.7	56.7
0.2	2	5	50.6	51.3	51.2	45.9
0.2	2	1	28.4	28.5	28.5	24.7
0.2	3	10	88.3	88.8	88.8	81
0.2	3	5	81	82.3	82.3	73.4
0.2	3	1	61.7	61.9	61.9	51.5
0.3	1	10	43.3	42.9	42.8	42.8
0.3	1	5	31.3	31.1	31.1	31.6
0.3	1	1	14.2	14.2	14.2	13.9
0.3	2	10	88.1	88.5	88.5	84.2
0.3	2	5	80.8	80.8	80.8	76.8
0.3	2	1	61.5	61.5	61.5	55.3
0.3	3	10	99.3	99.3	99.3	97
0.3	3	5	98	98.2	98.2	94.8
0.3	3	1	93	93.2	93.2	85.2

while the response variables in treatment group comes from $h(x) = (1 - \lambda)f(x) + \lambda g(x)$ with $g(x)/f(x) = \exp(\alpha + \beta x)$. The EM test statistics for testing homogeneity under the

Table 2: Type I error and power comparisons (%) of the EM test and the score test (SC test) for log-normal model: $n_0 = n_1 = 200$.

λ	μ	Level	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	SC test
0		10	10.4	10.5	10.6	10.2
0		5	5.5	5.6	5.6	5.4
0		1	1.2	1.2	1.2	1.2
0.1	1	10	26.5	26.7	26.5	26.2
0.1	1	5	17.2	17.2	17.2	16.4
0.1	1	1	5.8	5.9	6	5.6
0.1	2	10	68.3	69	69.2	58.4
0.1	2	5	58.5	58.8	58.9	47.4
0.1	2	1	37	37.1	37.4	25.1
0.1	3	10	96.4	96.8	97	84.4
0.1	3	5	94.6	94.8	95.2	77.6
0.1	3	1	86.2	87.2	87.4	58.6
0.2	1	10	63	62.9	62.8	62.1
0.2	1	5	50.2	50	50	49.4
0.2	1	1	27.8	27.6	27.5	26.2
0.2	2	10	99.2	99.3	99.4	97.5
0.2	2	5	98.6	98.6	98.6	95
0.2	2	1	95.1	95.2	95.2	85.5
0.2	3	10	100	100	100	100
0.2	3	5	100	100	100	99.9
0.2	3	1	100	100	100	99.2
0.3	1	10	89.5	89.5	89.6	89
0.3	1	5	84	83.9	83.9	82.6
0.3	1	1	65.1	64.9	64.6	63
0.3	2	10	100	100	100	100
0.3	2	5	100	100	100	99.9
0.3	2	1	100	100	100	99.7
0.3	3	10	100	100	100	100
0.3	3	5	100	100	100	100
0.3	3	1	100	100	100	100

semiparametric two-sample model are found to be $EM_n^{(1)} = 14.090$, $EM_n^{(2)} = 14.150$, and $EM_n^{(3)} = 14.167$. Calibrated by the χ_1^2 limiting distribution, the P values are all around 0.02%.

Table 3: Type I error and power comparisons (%) of the EM test and the score test (SC test) for gamma model: $n_0 = n_1 = 50$.

λ	m	Level	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	SC test
0		10	12.2	12.5	12.5	12.1
0		5	6.4	6.6	6.6	6.7
0		1	1.4	1.4	1.4	2.3
0.1	2	10	14.9	15.1	15.2	12
0.1	2	5	8.8	8.9	8.9	6.4
0.1	2	1	2.8	2.8	2.8	0.6
0.1	3	10	19.6	19.9	19.9	14.1
0.1	3	5	13.2	13.2	13.2	7.7
0.1	3	1	4.3	4.4	4.4	1
0.1	4	10	25.5	26.4	26.5	17
0.1	4	5	17.5	17.9	17.9	9.2
0.1	4	1	6.3	6.4	6.4	1.1
0.2	2	10	22.9	22.7	22.8	17.6
0.2	2	5	14.4	14.3	14.3	9.2
0.2	2	1	4.5	4.7	4.7	1.2
0.2	3	10	39.6	39.9	40	27.4
0.2	3	5	29.1	29.5	29.5	16.7
0.2	3	1	14.3	14.4	14.4	4
0.2	4	10	61.1	61.7	61.7	37
0.2	4	5	49.2	49.6	49.6	24.1
0.2	4	1	28.4	28.6	28.6	6.6
0.3	2	10	36.3	36.4	36.4	28.6
0.3	2	5	26.1	25.9	25.9	16.9
0.3	2	1	11.9	11.9	11.9	3.1
0.3	3	10	67.2	67.2	67.2	48.9
0.3	3	5	55.8	55.8	55.8	35.1
0.3	3	1	34	34	34.1	11.3
0.3	4	10	87.9	88.1	88.2	67.5
0.3	4	5	81.8	82.2	82.2	53.4
0.3	4	1	63.1	63.3	63.4	21.4

We also applied the score test of Qin and Liang [1]. The score test statistic is 9.417 with the P value equal to 0.2% calibrated by the χ_1^2 limiting distribution. We also used the permutation

Table 4: Type I error and power comparisons (%) of the EM test and the score test (SC test) for gamma model: $n_0 = n_1 = 200$.

λ	m	Level	$EM_n^{(1)}$	$EM_n^{(2)}$	$EM_n^{(3)}$	SC test
0		10	11.2	11.3	11.3	11.1
0		5	5.9	5.9	6	5.7
0		1	1.2	1.2	1.2	1.4
0.1	2	10	23.1	22.7	22.7	19.7
0.1	2	5	14.2	14.2	14.2	11.7
0.1	2	1	5.1	5.1	5.2	3.1
0.1	3	10	39.6	39.8	39.9	29.5
0.1	3	5	29	29.4	29.6	19
0.1	3	1	13.2	13.4	13.5	4.8
0.1	4	10	62.3	62.5	62.7	37.5
0.1	4	5	52.2	52.5	52.8	26.2
0.1	4	1	32.5	33.2	33.7	8.5
0.2	2	10	49	48.9	48.9	43.8
0.2	2	5	36.6	36.7	36.5	30.6
0.2	2	1	19.4	19.4	19.4	11.4
0.2	3	10	88.2	88.2	88.4	73
0.2	3	5	81.5	81.6	81.6	61.2
0.2	3	1	64.6	64.6	64.8	34.6
0.2	4	10	98.9	98.9	98.9	87.1
0.2	4	5	98	98.1	98.1	79.7
0.2	4	1	94.3	94.2	94.2	54.5
0.3	2	10	78.5	78.5	78.6	73
0.3	2	5	70.1	70	70	62.5
0.3	2	1	48.7	48.8	48.8	34.9
0.3	3	10	99.2	99.2	99.2	96.1
0.3	3	5	98.8	98.8	98.8	93
0.3	3	1	96.5	96.5	96.5	78.8
0.3	4	10	100	100	100	99.4
0.3	4	5	100	100	100	98.7
0.3	4	1	100	100	100	92.5

methods to get the P values of the two types of tests. Based on 50,000 permutations, the P values of the three EM test statistics are all around 0.03%, and the P value of the score test is

around 0.5%. In accordance with Fu et al. [5], both methods suggest a significant treatment effect, while the proposed EM test has much stronger evidence than the score test.

Appendix

Proofs

The proofs of Theorems 3.1 and 3.3 are based on the three lemmas given below. Lemma A.1 assesses the order of the maximum empirical likelihood estimators of α and β with λ bounded away from 0 under the null hypothesis. Lemma A.2 shows that the EM iteration updates the value of λ by the amount of order $o_p(1)$. Theorem 3.1 is then proved by iteratively using Lemmas A.1 and A.2. Lemma A.3 gives an approximation of the penalized ELR for any λ bounded away from 0, based on which we prove Theorem 3.3.

Lemma A.1. *Assume the conditions of Theorem 3.1. Let* $\overline{\lambda} \in [\epsilon, 1]$ *for some constant* $\epsilon > 0$ *and* $(\overline{\alpha}, \overline{\beta}) = \operatorname{argmax}_{\alpha,\beta} pR(\overline{\lambda}, \alpha, \beta)$. *Then, we have*

$$\overline{\alpha} = O_p\left(n^{-1}\right), \quad \overline{\beta} = \frac{\overline{y} - \overline{x}}{\overline{\lambda}\sigma^2} + o_p\left(n^{-1/2}\right) \tag{A.1}$$

with $\overline{x} = 1/n_0 \sum_{i=1}^{n_0} x_i$ *and* $\overline{y} = 1/n_1 \sum_{j=1}^{n_1} y_j$.

Proof. Since $\overline{\lambda} \geq \epsilon > 0$, the parameters (α, β) in the empirical likelihood ratio are identifiable. Therefore, $(\overline{\alpha}, \overline{\beta})$ are $\sqrt{n}$-consistent to the true value $(0, 0)$, that is, $\overline{\alpha} = O_p(n^{-1/2})$ and $\overline{\beta} = O_p(n^{-1/2})$ [10].

Following the arguments in Section 4, the maximum empirical likelihood estimate $(\overline{\alpha}, \overline{\beta})$ should satisfy (here λ is suppressed to $\overline{\lambda}$)

$$\frac{\partial G\left(\overline{\xi}, \overline{\alpha}, \overline{\beta}\right)}{\partial \alpha} = -2\sum_{h=1}^{n} \frac{\overline{\xi} e^{\overline{\alpha}+\overline{\beta}t_h}}{1 + \overline{\xi}\left(e^{\overline{\alpha}+\overline{\beta}t_h} - 1\right)} + 2\sum_{j=1}^{n_1} \frac{\overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}}{1 - \overline{\lambda} + \overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}} = 0, \tag{A.2}$$

$$\frac{\partial G\left(\overline{\xi}, \overline{\alpha}, \overline{\beta}\right)}{\partial \beta} = -2\sum_{h=1}^{n} \frac{\overline{\xi} e^{\overline{\alpha}+\overline{\beta}t_h} t_h}{1 + \overline{\xi}\left(e^{\overline{\alpha}+\overline{\beta}t_h} - 1\right)} + 2\sum_{j=1}^{n_1} \frac{\overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j} y_j}{1 - \overline{\lambda} + \overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}} = 0 \tag{A.3}$$

with

$$\overline{\xi} = \frac{1}{n}\sum_{j=1}^{n_1} \frac{\overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}}{1 - \overline{\lambda} + \overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}}. \tag{A.4}$$

Applying Taylor expansion on the right-hand side of (A.4), we get

$$\overline{\xi} = \frac{n_1}{n}\overline{\lambda} + o_p(1). \tag{A.5}$$

Further applying first-order Taylor expansion to (A.2) and using (A.4), we get

$$n\overline{\xi}\left(1-\overline{\xi}\right)\overline{\alpha}+\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n}t_h\overline{\beta}-n_1\overline{\lambda}\left(1-\overline{\lambda}\right)\overline{\alpha}-\overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1}y_j\overline{\beta}=O_p(n)\left(\overline{\alpha}^2+\overline{\beta}^2\right). \tag{A.6}$$

Note that both $\overline{\alpha}$ and $\overline{\beta}$ are of order $O_p(n^{-1/2})$ and that both $\sum_{h=1}^{n}t_h$ and $\sum_{j=1}^{n_1}y_j$ have order $O_p(n^{1/2})$. Combining (A.5) and (A.6) yields $n_1\overline{\lambda}(\overline{\lambda}-n_1/n\overline{\lambda})\overline{\alpha}=O_p(1)$. Therefore, $\overline{\alpha}=O_p(n^{-1})$.

Similarly, first-order Taylor expansion of (A.3) results in

$$\begin{aligned}0=&-\overline{\xi}\sum_{h=1}^{n}t_h-\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n}t_h\overline{\alpha}-\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n}t_h^2\overline{\beta}\\&+\overline{\lambda}\sum_{j=1}^{n_1}y_j+\overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1}y_j\overline{\alpha}+\ \overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1}y_j^2\overline{\beta}+O_p(n)\left(\overline{\alpha}^2+\overline{\beta}^2\right).\end{aligned} \tag{A.7}$$

With the same reasoning as for $\overline{\alpha}$, it follows from (A.7) that

$$\left\{n_1\overline{\lambda}\left(1-\frac{n_1}{n\overline{\lambda}}\right)\sigma^2-n_1\overline{\lambda}\left(1-\overline{\lambda}\right)\sigma^2\right\}\overline{\beta}=\overline{\lambda}\sum_{j=1}^{n_1}y_j-\frac{n_1}{n}\overline{\lambda}\sum_{h=1}^{n}t_h+o_p\left(n^{1/2}\right). \tag{A.8}$$

After some algebra, we have $\overline{\beta}=(\overline{y}-\overline{x})/(\overline{\lambda}\sigma^2)+o_p(n^{-1/2})$, which completes the proof. □

Suppose that $\overline{\lambda}$, $\overline{\alpha}$, and $\overline{\beta}$ have the properties given in Lemma A.1. For $j=1,\ldots,n_1$, let $\overline{w}_j=\overline{\lambda}\exp(\overline{\alpha}+\overline{\beta}y_j)/(1-\overline{\lambda}+\overline{\lambda}\exp(\overline{\alpha}+\overline{\beta}y_j))$. The updated value of λ is

$$\overline{\lambda}^{*}=\arg\max_{\lambda}\left\{\sum_{j=1}^{n_1}(1-\overline{w}_j)\log(1-\lambda)+\sum_{j=1}^{n_1}\overline{w}_j\log(\lambda)+\log(\lambda)\right\}. \tag{A.9}$$

It can be verified that the close form of $\overline{\lambda}^{*}$ is given by $\overline{\lambda}^{*}=(1/(n_1+1))(\sum_{j=1}^{n_1}\overline{w}_j+1)$. We now show that the above iteration only changes the value of λ by an $o_p(1)$ term.

Lemma A.2. *Assume the conditions of Lemma A.1 hold. Then,* $\overline{\lambda}^{*}=\overline{\lambda}+o_p(1)$.

Proof. Let $\widehat{\lambda}=\sum_{j=1}^{n_1}\overline{w}_j/n_1$. According to Lemma A.1, $\overline{\alpha}=o_p(1)$ and $\overline{\beta}=o_p(1)$. Applying the first-order Taylor expansion, we have

$$\widehat{\lambda}=\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\overline{\lambda}\exp\left(\overline{\alpha}+\overline{\beta}y_j\right)}{1-\overline{\lambda}+\overline{\lambda}\exp\left(\overline{\alpha}+\overline{\beta}y_j\right)}=\overline{\lambda}+O_p(1)\ \left(\overline{\alpha}+\overline{\beta}\right)=\overline{\lambda}+o_p(1). \tag{A.10}$$

Some simple algebra work shows that

$$\overline{\lambda}^{*} - \widehat{\lambda} = \frac{1-\overline{\lambda}}{n_1+1} = o_p(1). \tag{A.11}$$

Therefore, $\overline{\lambda}^{*} = \overline{\lambda} + o_p(1)$, and this finishes the proof. □

Proof of Theorem 3.1. With the above two technical lemmas, the proof is the same as that of Theorem 1 in Li et al. [13] and therefore is omitted. □

The next lemma is a technical preparation for proving Theorem 3.3. It investigates the asymptotic approximation of the penalized ELR for any λ bounded away from 0.

Lemma A.3. *Assume the conditions of Theorem 3.1 and $\overline{\lambda} \in [\epsilon, 1]$ for some $\epsilon > 0$. Then,*

$$pR\left(\overline{\lambda}, \overline{\alpha}, \overline{\beta}\right) = n\rho(1-\rho)\sigma^{-2}\left(\overline{y}-\overline{x}\right)^2 + 2\log\left(\overline{\lambda}\right) + o_p(1). \tag{A.12}$$

Proof. With Lemma A.1, we have $\overline{\alpha} = O_p(n^{-1})$ and $\overline{\beta} = O_p(n^{-1/2})$. Applying second-order Taylor expansion on $pR(\overline{\lambda}, \overline{\alpha}, \overline{\beta})$ and noting that $\partial pR/\partial\boldsymbol{\alpha}|_{(\alpha,\beta)=(0,0)} = 0$, we have

$$\begin{aligned} pR\left(\overline{\lambda}, \overline{\alpha}, \overline{\beta}\right) = {} & 2\left(-\overline{\xi}\sum_{h=1}^{n} t_h + \overline{\lambda}\sum_{j=1}^{n_1} y_j\right)\overline{\beta} - \left\{\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n} t_h^2 - \overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1} y_j^2\right\}\overline{\beta}^2 \\ & + 2\log\left(\overline{\lambda}\right) + o_p(1). \end{aligned} \tag{A.13}$$

Using (A.5) and the facts that both $\sum_{h=1}^{n} t_h^2/n$ and $\sum_{j=1}^{n_1} y_j^2/n_1$ converge to σ^2 in probability, the above expression can be simplified to

$$pR\left(\overline{\lambda}, \overline{\alpha}, \overline{\beta}\right) = 2\frac{n_1 n_0}{n}\overline{\lambda}\left(\overline{y}-\overline{x}\right)\overline{\beta} - \frac{n_1 n_0}{n}\overline{\lambda}^2\sigma^2\overline{\beta}^2 + 2\log\left(\overline{\lambda}\right) + o_p(1). \tag{A.14}$$

Plugging in the approximation $\overline{\beta} = (\overline{y}-\overline{x})/(\overline{\lambda}\sigma^2) + o_p(n^{-1/2})$, we get

$$\begin{aligned} pR\left(\overline{\lambda}, \overline{\alpha}, \overline{\beta}\right) &= \frac{n_1 n_0}{n}\frac{\left(\overline{y}-\overline{x}\right)^2}{\sigma^2} + 2\log\left(\overline{\lambda}\right) + o_p(1) \\ &= n\rho(1-\rho)\sigma^{-2}\left(\overline{y}-\overline{x}\right)^2 + 2\log\left(\overline{\lambda}\right) + o_p(1). \end{aligned} \tag{A.15}$$

This completes the proof. □

Proof of Theorem 3.3. Without loss of generality, we assume $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_L = 1$. According to Theorem 3.1 and Lemma A.3, for $l = 1, \ldots, L$, we have

$$pR\left(\lambda_l^{(K)}, \alpha_l^{(K)}, \beta_l^{(K)}\right) = n\rho(1-\rho)\sigma^{-2}\left(\overline{y}-\overline{x}\right)^2 + 2\log(\lambda_l) + o_p(1). \tag{A.16}$$

This leads to

$$\mathrm{EM}_n^{(K)} = \max_{1\le l\le L} pR\left(\lambda_l^{(K)}, \alpha_l^{(K)}, \beta_l^{(K)}\right) = n\rho(1-\rho)\sigma^{-2}\left(\overline{y}-\overline{x}\right)^2 + o_p(1), \tag{A.17}$$

where the remainder is still $o_p(1)$ since the maximum is taken over a finite set.

Note that when n tends to infinity, $\sqrt{n}(\overline{y}-\overline{x}) \longrightarrow N(0,\ \sigma^2/[\rho(1-\rho)])$ in distribution. Therefore,

$$\mathrm{EM}_n^{(K)} \longrightarrow \chi_1^2 \tag{A.18}$$

in distribution as n goes to infinity. This completes the proof. □

References

[1] J. Qin and K. Y. Liang, "Hypothesis testing in a mixture case-control model," *Biometrics*, vol. 67, pp. 182–193, 2011.

[2] J. Zhang, "Powerful two-sample tests based on the likelihood ratio," *Technometrics*, vol. 48, no. 1, pp. 95–103, 2006.

[3] J. A. Anderson, "Multivariate logistic compounds," *Biometrika*, vol. 66, no. 1, pp. 17–26, 1979.

[4] T. Lancaster and G. Imbens, "Case-control studies with contaminated controls," *Journal of Econometrics*, vol. 71, no. 1-2, pp. 145–160, 1996.

[5] Y. Fu, J. Chen, and J. D. Kalbfleisch, "Modified likelihood ratio test for homogeneity in a two-sample problem," *Statistica Sinica*, vol. 19, no. 4, pp. 1603–1619, 2009.

[6] A. B. Owen, "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, vol. 75, no. 2, pp. 237–249, 1988.

[7] A. B. Owen, "Empirical likelihood ratio confidence regions," *The Annals of Statistics*, vol. 18, no. 1, pp. 90–120, 1990.

[8] P. Hall and B. La Scala, "Methodology and algorithms of empirical likelihood," *International Statistical Review*, vol. 58, pp. 109–127, 1990.

[9] T. DiCiccio, P. Hall, and J. Romano, "Empirical likelihood is Bartlett-correctable," *The Annals of Statistics*, vol. 19, no. 2, pp. 1053–1061, 1991.

[10] J. Qin and J. Lawless, "Empirical likelihood and general estimating equations," *The Annals of Statistics*, vol. 22, no. 1, pp. 300–325, 1994.

[11] S. E. Ahmed, A. Hussein, and S. Nkurunziza, "Robust inference strategy in the presence of measurement error," *Statistics & Probability Letters*, vol. 80, no. 7-8, pp. 726–732, 2010.

[12] J. Chen and P. Li, "Hypothesis test for normal mixture models: the EM approach," *The Annals of Statistics*, vol. 37, no. 5, pp. 2523–2542, 2009.

[13] P. Li, J. Chen, and P. Marriott, "Non-finite Fisher information and homogeneity: an EM approach," *Biometrika*, vol. 96, no. 2, pp. 411–426, 2009.

[14] J. Chen, "Penalized likelihood-ratio test for finite mixture models with multinomial observations," *The Canadian Journal of Statistics*, vol. 26, no. 4, pp. 583–599, 1998.

[15] J. R. Weeks and R. J. Collins, "Primary addiction to morphine in rats," *Federation Proceedings*, vol. 30, p. 277, 1971.

[16] D. D. Boos and C. Brownie, "Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment," *Biometrics*, vol. 47, pp. 1489–1504, 1991.

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 642403, 15 pages
doi:10.1155/2012/642403

Research Article

A Two-Stage Penalized Logistic Regression Approach to Case-Control Genome-Wide Association Studies

Jingyuan Zhao[1] and Zehua Chen[2]

[1] *Human Genetics, Genome Institute of Singapore, 60 Biopolis, Genome No. 02-01, Singapore 138672*
[2] *Department of Statistics and Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117546*

Correspondence should be addressed to Zehua Chen, stachenz@nus.edu.sg

Received 20 September 2011; Accepted 28 October 2011

Academic Editor: Yongzhao Shao

We propose a two-stage penalized logistic regression approach to case-control genome-wide association studies. This approach consists of a screening stage and a selection stage. In the screening stage, main-effect and interaction-effect features are screened by using L_1-penalized logistic like-lihoods. In the selection stage, the retained features are ranked by the logistic likelihood with the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and Jeffrey's Prior penalty (Firth, 1993), a sequence of nested candidate models are formed, and the models are assessed by a family of extended Bayesian information criteria (J. Chen and Z. Chen, 2008). The proposed approach is applied to the analysis of the prostate cancer data of the Cancer Genetic Markers of Susceptibility (CGEMS) project in the National Cancer Institute, USA. Simulation studies are carried out to compare the approach with the pair-wise multiple testing approach (Marchini et al. 2005) and the LASSO-patternsearch algorithm (Shi et al. 2007).

1. Introduction

The case-control genome-wide association study (GWAS) with single-nucleotide polymorphism (SNP) data is a powerful approach to the research on common human diseases. There are two goals of GWAS: (1) to identify suitable SNPs for the construction of classification rules and (2) to discover SNPs which are etiologically important. The emphasis is on the prediction capacity of the SNPs for the first goal and on the etiological effect of the SNPs for the second goal. The phrase "an etiological SNP" is used in the sense that either the SNP itself is etiological or it is in high-linkage disequilibrium with an etiological locus. Well-developed classification methods in the literature can be used for the first goal. These methods include classification and regression trees [1], random forest [2], support vector machine [3], and logic regression [4]. In this article, we focus on statistical methods for the second goal.

The approach of multiple testing based on single or paired SNP models is commonly used for the detection of etiological SNPs. Either the Bonferroni correction is applied for the control of the overall Type I error rate, see, for example, Marchini et al. [5] or some methods are used to control the false discovery rate (FDR), see, Banjamini and Hochberg [6], Efron and Tibshirani [7], and Storey and Tibshirani [8]. Other variants of multiple testing have also been advocated, see Hoh and Ott [9]. The multiple test approach considers either a single SNP or a pair of SNPs at a time. It does not adjust for the effects of other markers. If there are many loci having high sample correlations with a true genetic variant, which is common in GWAS, it is prone to result in spurious etiological loci.

It is natural to seek alternative methods that overcome the drawback of multiple testing. Such methods must have the nature of considering many loci simultaneously and assessing the significance of the loci by their synergistic effect. When the synergistic effect is of concern, adding loci spuriously correlated to an etiological locus does not contribute to the synergistic effect while the etiological locus has already been considered. Thus the drawback of multiple testing can be avoided. In this paper, we propose a method of the abovementioned nature: a two-stage penalized logistic regression approach. In the first stage of this approach, L_1-penalized logistic regression models are used together with a tournament procedure [10] to screen out apparently unimportant features (by features we refer to the covariates representing SNPs or their products). In the second stage, logistic models with the SCAD penalty [11] plus the Jeffrey's prior penalty [12] are used to rank the retained features and form a sequence of nested candidate models. The extended Bayesian information criteria (EBIC, [13, 14]) are used for the final model selection. In both stages of the approach, the features are assessed by their synergistic effects.

The two-stage strategy has been considered by other authors. For example, J. Fan and Y. Fan [15] adopted this strategy for high-dimensional classification, and Shi et al. [16] developed a two-stage procedure called LASSO-patternsearch. Sure independence screening (SIS, [17]) and its ramifications such as correlation screening and t-tests are commonly used in the screening stage. Compared with SIS approaches, the tournament screening with L_1-penalized likelihood produces less spuriously correlated features while enjoying the sure screening property possessed by the SIS approaches, see Z. Chen and J. Chen [10] and the comprehensive simulation studies by Wu [18], which has an impact on the accuracy of feature selection in the second stage, see Koh [19]. The L_1-penalized likelihood is easier to compute than that with the SCAD penalty. However, the SCAD penalty has an edge over the L_1-penalty in ranking the features so that the ranks are more consistent with their actual effects. This has been observed in simulation studies, see Zhao [20]. It is possibly due to the fact that the L_1 penalty over-penalizes those features with large effects compared with SCAD penalty that does not penalize large effects at all. Jeffrey's prior penalty is added to handle the difficulty caused by separation of data that usually presents in logistic regression models with factor covariates, see Albert and Anderson [21]. If, within any of the categories determined by the levels of the factors, the responses are all 1 or 0, it is said that there is a complete data separation. When the responses within any of the categories are almost all 1 or 0, it is referred to as a quasicomplete data separation. When there is separation (complete or quasi-complete), the maximum likelihood estimate of the corresponding coefficients becomes infinite. Jeffrey's prior penalty plays the role to shrink the parameters toward zero in the case of separation.

Logistic regression models with various penalties have been considered for GWAS by a number of authors. Park and Hastie [22] considered logistic models with a L_2-penalty. Wu et al. [23] considered logistic models with an L_1-penalty. The LASSO-patternsearch

developed by Shi et al. [16] is also based on logistic regression models. However, the accuracy for identifying etiological SNPs was not fully addressed. Park and Hastie [22] introduced the L_2-penalty mainly for computational reasons. Their method is essentially a classical stepwise procedure with AIC/BIC as model selection criteria. The method considered by Wu et al. [23] is in fact only a screening procedure. The numbers of main-effect and interaction features to be retained are predetermined and left as a subjective matter. The LASSO-patternsearch is closer to our approach. The procedure first screens the features by correlation screening based on single-feature (main-effect/interaction) models. Then a LASSO model is fitted to the retained features with its penalty parameter chosen by cross-validation. The features selected by LASSO are then refitted to a nonpenalized logistic regression model, and the coefficients of the features are subjected to hypothesis testing with varied level α. The α is again determined by cross-validation. By using cross-validation, this procedure addresses the prediction error of the selected model instead of the accuracy of the selected features. Our method is compared with the LASSO-patternsearch and the multiple test approach by simulation studies.

The two-stage penalized logistic regression approach is described in detail in Section 2. The approach is applied to a publically accessible CGEMS prostate cancer data inSection 3. Simulation studies are presented in Section 4. The paper is ended by some remarks. A supplementary document which contains some details omitted in the paper is provided at the website: http://www.stat.nus.edu.sg/~stachenz/, available online at doi: 10.1155/2012/642403.

2. The Two-Stage Penalized Logistic Regression Approach

We first give a brief account on the elements required in the approach: the logistic model for case-control study, the penalized likelihood, and the EBIC.

2.1. Logistic Model for Case-Control GWAS

Let y_i denote the disease status of individual i, 1 for case and 0 for control. Denote by x_{ij}, $j = 1, \ldots, P$, the genotypes of individual i at the SNPs under study. The x_{ij} takes the value 0, 1, or 2, corresponding to the number of a particular allele in the genotype. Here, the additive genetic mode is assumed for all SNPs. The logistic model is as follows:

$$\begin{aligned} y_i &\sim \text{Binomial}(1, \pi_i), \\ \log \frac{\pi_i}{1-\pi_i} &= \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} + \sum_{j<k} \xi_{jk} x_{ij} x_{ik}, \quad i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where x_{ij} and $x_{ij}x_{ik}$ are referred to as main-effect and interaction features, respectively, hereafter. The validity of the logistic model for case-control experiments has been argued by Armitage [24] and Breslow and Day [25]. There are two fundamental facts about the above model for GWAS: (a) the number of features is much larger than the sample size n, since P is usually huge in GWAS, this situation is referred to as small-n-large-p; (b) since there are only a few etiological SNPs, only a few of the coefficients in the model are nonzero, this phenomenon is referred to as sparsity.

2.2. Penalized Likelihood

Penalized likelihood makes the fitting of a logistic model with small-n-large-p computationally feasible. It also provides a mechanism for feature selection. Let s be the index set of a subset of the features. Let $L(\boldsymbol{\theta}(s) \mid s)$ denote the likelihood function of the logistic model consisting of features with indices in s, where $\boldsymbol{\theta}(s)$ consists of those β and ξ with their indices in s. The penalized log likelihood is defined as

$$l_p(\boldsymbol{\theta}(s) \mid \lambda) = -2 \log L(\boldsymbol{\theta}(s) \mid s) + \sum_{j \in s} p_\lambda(\theta_j), \tag{2.2}$$

where $p_\lambda(\cdot)$ is a penalty function and λ is called the penalty parameter. The following penalty functions are used in our approach:

$$\begin{aligned} L_1\text{-penalty}: &\quad p_\lambda(\theta_j) = \lambda |\theta_j|, \\ \text{SCAD penalty}: &\quad p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \le \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\}, \end{aligned} \tag{2.3}$$

where a is a fixed constant bigger than 2. The penalized log likelihood with L_1-penalty is used together with the tournament procedure [10] in the screening stage. At each application of the penalized likelihood, the parameter λ is tuned such that the minimization of the penalized likelihood yields a predetermined number of nonzero coefficients. The R package glmpath developed by Park and Hastie [26] is used for the computation, the tuning on λ is equivalent to setting the maximum steps in the glmpath function to the predetermined number of nonzero coefficients. The SCAD penalty is used in the second stage for ranking the features.

2.3. The Extended BIC

In small-n-large-p problems, the AIC and BIC are not selection consistent. To tackle the issue of feature selection in small-n-large-p problems, J. Chen and Z. Chen [13] developed a family of extended Bayes information criteria (EBIC). In the context of the logistic model described above, the EBIC is given as

$$\mathrm{EBIC}(\gamma) = -2 \log L\left(\widehat{\boldsymbol{\theta}}(s) \mid s\right) + (\nu_1 + \nu_2) \log(n) + 2\gamma \log \binom{P}{\nu_1} \binom{P(P-1)/2}{\nu_2}, \quad \gamma \ge 0, \tag{2.4}$$

where ν_1 and ν_2 are, respectively, the numbers of main-effect and interaction features and $\widehat{\boldsymbol{\theta}}(s)$ is the maximum likelihood estimate of the parameter vector in the model. It has been shown that, under certain conditions, the EBIC is selection consistent when γ is larger than $1 - \ln n/(2 \ln p)$, see J. Chen and Z. Chen [13, 14]. The original BIC, which corresponds to the EBIC with $\gamma = 0$, fails to be selection consistent when p has a $\sqrt{n}$ order.

We now describe the two-stage penalized logistic regression (TPLR) approach as follows.

2.4. Screening Stage

Let n_M and n_I be two predetermined numbers, respectively, for main-effect and interaction features to be retained. The screening stage consists of two steps: a main-effect screening step and an interaction screening step.

In the main-effect screening step, only the main-effect features are considered. Let S_M denote the index set of the main features, that is, $S_M = \{1, \ldots, p\}$. If $|S_M|$, the number of members in S_M, is not too large, minimize

$$-2 \log L(\boldsymbol{\beta} \mid S_M) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.5}$$

by tuning the value of λ to retain n_M features. If $|S_M|$ is very large. The following tournament procedure proposed in Z. Chen and J. Chen [10] is applied. Partition S_M into $S_M = \cup_k s_k$ with $|s_k|$ equal to an appropriate group size n_G, where n_G is chosen such that the minimization of the penalized likelihood with n_G features can be efficiently carried out. For each k, minimize

$$-2 \log L(\boldsymbol{\beta}(s_k) \mid s_k) + \lambda \sum_{j \in s_k} |\beta_j| \tag{2.6}$$

by tuning the value of λ to retain $n_k (\approx n_M)$ features. If $\sum_j n_j > n_G$, repeat the above process with all retained features; otherwise, apply the L_1-penalized logistic model to the retained features to reduce the number to n_M. Let s_M denote the indices of these n_M features.

The interaction screening is similar to the main-effect screening step. However, the main-effect features retained in the main-effect screening step are built in the models for interaction screening. Let S_I denote the set of pairs of the indices for all the interaction features, that is, $S_I = \{(i, j) : i < j, i, j = 1, \ldots, p\}$. Since $|S_I|$ is large in general, the tournament procedure is applied for interaction screening. Let S_I be partitioned as $S_I = \cup_k T_k$ with $|T_k| \approx n_G$. For each k, minimize

$$-2 \log L(\boldsymbol{\beta}(s_M), \boldsymbol{\xi}(T_k) \mid s_M, T_k) + \lambda \sum_{(i,j) \in T_k} |\xi_{ij}| \tag{2.7}$$

by tuning the value of λ to retain $n_k (\approx n_I)$ interaction features. Note that, in the above penalized likelihood, both the main-effect features in s_M and the interaction features in T_k are involved in the likelihood part. However, only the parameters associated with the interaction features are penalized. Since no penalty is put on the parameters associated with the main-effect features, the main-effect features are always retained in the process of interaction screening. If $\sum_k n_k > n_G$, repeat the above process with the set of all retained features; otherwise, reduce the retained features to n_I of them by one run of the minimization using an L_1-penalized likelihood.

2.5. Selection Stage

The selection stage consists of a ranking step and a model selection step. In the ranking step, the retained features (main-effect and interaction) are ranked together by a penalized

likelihood with SCAD penalty plus an additional Jeffrey's prior penalty. In the model selection step, a sequence of nested models are formed and evaluated by the EBIC.

For convenience, let the retained interaction features be referred to by a single index. Let S^* be the index set of all the main-effect and interaction features retained in the screening stage. Let $K = |S^*|$. Denote by $\boldsymbol{\theta}(S^*)$ the vector of coefficients corresponding to these features (the components of $\boldsymbol{\theta}(S^*)$ are the β's and ξ's corresponding to the retained main-effects and interactions). Jeffrey's prior penalty is the log determinant of the Fisher information matrix. Thus the penalized likelihood in the selection stage is given by

$$l_p(\boldsymbol{\theta}(S^*) \mid \lambda) = -2\log L(\boldsymbol{\theta}(S^*) \mid S^*) - \log|I(\boldsymbol{\theta}(S^*))| + \sum_{j\in S^*} p_\lambda(|\theta_j|), \tag{2.8}$$

where p_λ is the SCAD penalty and $I(\boldsymbol{\theta}(S^*))$ is the Fisher information matrix. The ranking step is done as follows. The parameter λ is tuned to a value λ_1 such that it is the smallest to make at least one component of $\boldsymbol{\beta}(S^*)$ zero by minimizing $l_p(\boldsymbol{\beta}(S^*) \mid \lambda_1)$. Let $j_K \in S^*$ be the index corresponding to the zero component. Update S^* to $S^*/\{j_K\}$, that is, the feature with index j_K is eliminated from further consideration. With the updated S^*, the above process is repeated, and another feature is eliminated. Continuing this way, eventually, we obtain an ordered sequence of the indices in S^*: $j_1, j_2, \ldots, j_K$. From the ordered sequence above, a sequence of nested models is formed as $S_k = \{j_1, \ldots, j_k\}, k = 1, \ldots, K$. For each S_k, the un-penalized likelihood $\log L(\boldsymbol{\theta}(S_k) \mid S_k)$ is maximized. The EBIC with γ values in a range $[1 - \ln n/\ln p, \gamma_{\max}]$ is computed for all these models. For each γ, the model with the smallest EBIC(γ) is identified. The upper bound of the range, $\gamma_{\max}$, is taken as a value such that no feature can be selected by the EBIC with that value. Only a few models can be identified when γ varies in the range. Each identified model corresponds to a subinterval of $[1-\ln n/\ln p, \gamma_{\max}]$. The identified models together with their corresponding subintervals are then reported.

The choice of γ in the EBIC affects the positive discovery rate (PDR) and the false discovery rate (FDR). In the context of GWAS, the PDR is the proportion of correctly identified SNPs among all etiological SNPs, and the FDR is the proportion of incorrectly identified SNPs among all identified SNPs. A larger γ results in a smaller FDR and also a lower PDR. A smaller γ results in a higher PDR and also a higher FDR. A balance must be stricken between PDR and FDR according to the purpose of the study. If the purpose is to confirm the etiological effect of certain well-studied loci or regions, one should emphasize more on a desirably low FDR rather than a high PDR. If the purpose is to discover candidate loci or regions for further study, one should emphasize more on a high PDR with only a reasonable FDR. The FDR is related to the statistical significance of the features. Measures on the statistical significance can be obtained from the final identified models and their corresponding subintervals. The upper bound of the subinterval determines the largest threshold which the effects of the features in the model must exceed. Likelihood ratio test (LRT) statistics can be used to assess the significance of the feature effects. For example, suppose a model consisting of ν_1 main-effect features and ν_2 interaction features is selected with γ in a sub-interval $(\underline{\gamma}, \overline{\gamma}]$. The LRT statistic for the significance of the feature with the lowest rank in the model must exceed the threshold $\log n + 2\overline{\gamma}\log((P - \nu_1 + 1)/\nu_1)$, if the feature is a main-effect one, and $\log n + 2\overline{\gamma}\log((P(P-1)/2 - \nu_2 + 1)/\nu_2)$, if the feature is an interaction one. The probability for the LRT to exceed the threshold is at most $Pr(\chi_1^2 > \log n + 2\overline{\gamma}\log((P-\nu_1+1)/\nu_1))$ or $Pr(\chi_1^2 > \log n + 2\overline{\gamma}\log((P(P-1)/2-\nu_2+1)/\nu_2))$ for a main-effect

or interaction feature if the feature does not actually have any effect. These probabilities, like the P-values in classical hypothesis testing, provide statistical basis for the user to determine which model should be taken as the selected model.

A final issue on the two-stage logistic regression procedure is how to determine n_M and n_I. If they are large enough, usually several times of the actual numbers, their choice will not affect the final model selection. Since the actual numbers are unknown, a strategy is to consider several different n_M and n_I. First, run the procedure with some educated guess on n_M and n_I. Then, run the procedure again using larger n_M and n_I. If the identified models by using these n_M and n_I are almost the same, the choice of n_M and n_I is appropriate. Otherwise, further values of n_M and n_I should be considered, until eventually different n_M, n_I result in the same results.

3. Analysis of CGEMS Prostate Cancer Data

The CGEMS data portal of National Cancer Institute, USA, provides public access to the summary results of approximately 550,000 SNPs genotyped in the CGEMS prostate cancer whole genome scan, see http://cgems.cancer.gov. We applied the two-stage penalized regression approach to the prostate cancer Phase 1A data in the prostate, lung, colon, and ovarian (PLCO) cancer screening trial. The dataset contains 294,179 autosomal SNPs which passed the quality controls on 1,111 controls and 1,148 cases (673 cases are aggressive, Gleason ≥ 7 or stage $\geq$ III; 475 cases are nonaggressive, Gleason < 7 and stage $<$ III). In our analysis, we put all the cases together without distinguishing aggressive and non-aggressive ones. We assumed additive genetic mode for all the SNPs.

The application of the screening stage to all the 294,179 SNPs directly is not only time consuming but also unnecessary. Therefore, we did a preliminary screening by using single-SNP logistic models. For each SNP, a logistic model is fitted and the P-value of the significance test of the SNP effect is obtained. Those SNPs with a P-value bigger than 0.05 are discarded. There are 17,387 SNPs which have a P-value less than 0.05 and are retained.

Because of the sheer huge number of features, 17,387 main features and $17,387 \times (1,7387-1)/2$ interaction features, the tournament procedure is applied in the screening stage. At the main-effect feature screening step, the main-effect features are randomly partitioned into groups of size 1,000, except one group of size 1,387, and 100 features are selected from each group. A second round of screening is applied to the selected 1,700 features out of which 100 features are retained. The interaction feature screening is applied to $17,387 \times (1,7387-1)/2$ interaction features. Each round, the retained features are partitioned into groups of size 1,000, and 50 features are selected from each group. The procedure continues until 300 interaction features are finally selected. Eventually, the 100 main-effect features and 300 interaction features are put together and screened to retain a total of 100 features (main-effect or interaction). The eventual 100 features are then subjected to the selection procedure.

The features selected by EBIC with γ in the subintervals $(0.70, 0.73]$, $(0.73, 0.77]$, and $(0.77, 0.8]$ are given in Table 1. With $\gamma = 0.8$, the largest value at which at least one feature can be selected, the following three interaction features are selected: rs1885693-rs12537363, rs7837688-rs2256142 and rs1721525-rs2243988. The effects of these features have a significance level at least $1.8250e$-11. The next largest γ value, 0.77, selects 7 additional interaction features which have a significance level at least $9.6435e$-11. The third largest γ value, 0.73, selects still 2 additional interaction features which have a significance level at least $2.7407e$-10. The chromosomal region 8q24 is the one where many previous prostate cancer studies are concentrated. It has been reported in a number of studies that rs1447295, one of the 4 tightly

Table 1: Features associated with prostate cancer from the analysis of CGEMS data ("rsXXX" denotes SNP reference).

Chromosome	Feature	Maximum γ	Significance Level
6, 7	rs1885693-rs12537363	0.80	1.824985e-11
8, 13	**rs7837688**-rs2256142	0.80	1.824985e-11
1, 21	rs1721525-rs2243988	0.80	1.824985e-11
10, 16	rs11595532-rs8055313	0.77	9.64352e-11
12, 12	rs10842794-rs10848967	0.77	9.64352e-11
9, 12	rs3802357-rs10880221	0.77	9.64352e-11
1, 2	rs3900628-rs642501	0.77	9.64352e-11
1, 16	rs10518441-rs2663158	0.77	9.64352e-11
3, 13	rs1880589-rs1999494	0.77	9.64352e-11
5, 18	rs6883810-rs11874224	0.77	9.64352e-11
13, 19	rs4274307-rs3745180	0.73	2.740672e-10
5, 19	rs672413-rs3915790	0.73	2.740672e-10

linked SNPs in the "locus 1" region of 8q24, is associated with prostate cancer, and it has been established as a benchmark for prostate cancer association studies. In the current data set, we found that rs7837688 is highly correlated with rs1447295 ($r^2 = 0.9$) and is more significant than rs1447295 based on single-SNP models. These two SNPs, which are in the same recombination block, are also physically close.

An older and slightly different version of the CGEMS prostate data has been analyzed by Yeager et al. [27] using single-SNP multiple testing approach. In their analysis, they distinguished between aggressive and non-aggressive status and assumed no structure on genetic modes. For each SNP, they considered four tests: a χ^2-test with 4 degrees of freedom based on a 3×3 contingency table, a score test with 4 degrees of freedom based on a polytomous logistic regression model adjusted for age group, region of recruitment, and whether a case is diagnosed within one year of entry to the trial, as well as the other two which are the same as the χ^2 and score tests but take into account incidence-density sampling. They identified two physically close but genetically independent regions (in a distance 0.65 centi-Morgans) within 8q24. One of the regions is where the benchmark SNP rs1447295 is located. They reported three SNPs: rs1447295 (P-value: 9.75e-05), rs7837688 (P-value: 6.52e-06) and rs6983267 (P-value: 2.43e-05), where rs7837688 is in the same region as rs1447295 and rs6983267 is in the other region. The P-values are computed from the score statistic based on incidence-density sampling polytomous logistic regression model adjusted for other covariates.

In our analysis, we identified rs7837688 but not rs1447295. This is because the penalized likelihood tends to select only one feature among several highly correlated features, which is a contrast to the multiple testing that selects all the correlated features if any of them is associated with the disease status. We failed to identify rs6983267. The possible reason could be that its effect is masked by other more significant features which are identified in our analysis. We also carried out the selection procedure with only the 100 main-effect features retained from the screening stage. It is found that rs6983267 is among the top 20 selected main-effect features with a significance level 2.3278e-05. It is interesting to notice that the two SNPs rs7837688 and rs1721525 appearing in the top three interaction features are also among the top four features selected with a maximum γ value 0.7185 when only

main-effect features are considered. Since no SNP on chromosomes other than 8q24 has been reported in other studies, we wonder whether statistically significant SNPs on other chromosomes can be ignored due to biological reasons: if not, our analysis strongly suggests that rs1721525 located on chromosome 1 could represent another region in the genome which is associated with prostate cancer, if it holds, biologically, chromosome 1 cannot be excluded in the consideration of genetic variants for prostate cancer.

4. Simulation Studies

We present results of two simulation studies in this section. In the first study, we compare the two-stage penalized logistic regression (TPLR) approach with the paired-SNP multiple testing (PMT) approach of Marchini et al. [5] under simulation settings considered by them. In the second study, we compare the TPLR approach with LASSO-patternsearch using a data structure mimicking the CGEMS prostate cancer data.

4.1. Simulation Study 1

The comparison of TPLR and PMT is based on four models. Each model involves two etiological SNPs. In the first model, the effects of the two SNPs are multiplicative both within and between loci; in the second model, the effects of the two SNPs are multiplicative within but not between loci; in the third model, the two SNPs have threshold interaction effects; in the fourth model, the two SNPs have an interaction effect but no marginal effects. The first three models are taken from Marchini et al. [5]. The details of these models are provided in the supplementary document.

Marchini et al. [5] considered two strategies of PMT. In the first strategy, a logistic model with 9 parameters is fitted for each pair of SNPs, and the Bonferroni corrected significance level $\alpha/\binom{P}{2}$ is used to declare the significant pairs. In the second strategy, the SNPs that are significant in single-SNP tests at a liberal level α_1 are identified, then the significances of all the pairs formed by these SNPs are tested using the Bonferroni corrected level $\alpha/\binom{P\alpha_1}{2}$.

In the first three models, the marginal effects of both loci are nonnegligible and can be picked up by the single-SNP tests at the relaxed significance level. In this situation, the second strategy has an advantage over the first strategy in terms of detection power and false discovery rate. In this study, we compare our approach with the second strategy of PMT under the first three models. In the fourth model, since there are no marginal effects at both loci, the second strategy of PMT cannot be applied since it will fail to pick up any loci at the first step. Hence, we compare our approach with the first strategy of PMT. However, the first strategy involves a stupendous amount of computation which exceeds our computing capacity. To circumvent this dilemma, we consider an artificial version of the first strategy; that is, we only consider the pairs which involve at least one of the etiological SNPs. This artificial version has the same detection power but lower false discovery rate than the full version. The artificial version cannot be implemented with real data since it requires the knowledge of the etiological SNPs. However, it can be implemented with simulated data and serves the purpose of comparison.

Each simulated dataset contains $n = 800$ individuals (400 cases and 400 controls) with genotypes of P SNPs. Two values of P, 1000 and 5000, are considered. The genotypes of disease loci, which are not among the P SNPs, and the disease status of the individuals are generated first. Then, the genotypes of the SNPs which are in linkage disequilibrium with the

disease loci are generated using a square correlation coefficient $r^2 = 0.5$. The genotypes of the remaining SNPs are generated independently assuming Hardy-Weinberg equilibrium. For the first three models, the effects of the disease loci are specified by the prevalence, disease allele frequencies, denoted by q, and marginal effect parameters, denoted by λ_1 and λ_2. The prevalence is set at 0.01 throughout. The two marginal effects are set equal, that is, $\lambda_1 = \lambda_2 = \lambda$. For the fourth model, the effect is specified through the coefficient in the logistic model. The coefficients are determined by first specifying ξ_{12} and then determining β_1 and β_2 through the constraints of the model while β_0 is set to -5. The definition of these parameters and the details of the data generation are given in the supplementary document.

The α_1 and α in the PMT approach are taken to be 0.1 and 0.05, respectively, the same as in Marchini et al. [5]. The γ in EBIC is fixed as 1 since it is infeasible to incorporate the consideration on the choice of γ into the simulation study. The average PDR and FDR over 200 simulation replicates under Model 1–4 are given in Tables 2–5, respectively. In Table 5, the entries of the FDR for the PMT approach are lower bounds rather than the actual FDRs, since, as mentioned earlier, only the pairs of SNPs involving at least one etiological SNP are considered in the artificial version of the first strategy of PMT, which results in less false discoveries than the full version while retaining the same positive detections.

The results presented in Tables 2–5 are summarized as follows. Under Model 1, TPLR has much lower FDR and comparable PDR compared with PMT. Under Models 2–4, the PDR of TPLR is significantly higher than PMT in all cases except Model 2 when $\lambda = 0.7, q = 0.2, P = 1000$ (0.95 versus 1) and Model 3 when $\lambda = 1, q = 0.1, P = 1000$ (0.81 versus 0.84). The overall averaged FDRs of TPLR is 0.0487 while that of PMT is 0.7604. It is seen that the FDR of TPLR is always kept at reasonably low levels but that of PMT is intolerably high, and at the same time TPLR is still more powerful than PMT for detecting etiological SNPs. From the simulation results, we can also see the impact of P on PDR and FDR. In general, the increase of P reduces PDR and increases FDR of both approaches. However, the impact on TPLR is less than that on PMT.

4.2. Simulation Study 2

The data for this simulation study is generated mimicking the structure of the CGEMS prostate cancer data. The cases and controls are generated using a logistic model with the following linear predictor:

$$\eta = \beta_0 + \sum_{j=1}^{5} \beta_j x_j + \xi_1 x_6 x_7 + \xi_2 x_8 x_9 + \xi_3 x_{10} x_{11} + \xi_4 x_2 x_{12} + \xi_5 x_{13} x_{14}, \tag{4.1}$$

where x_j's are feature values of 14 SNPs. The parameter values are taken as

$$\begin{aligned} \boldsymbol{\beta} &= (-8.65, 0.89, 1.1, 0.74, 1.18, 1.25), \\ \boldsymbol{\xi} &= (1.95, 1.62, 1.9, 1.8, 1.1). \end{aligned} \tag{4.2}$$

Table 2: The simulated average PDR and FDR under Model 1: multiplicative effects both within and between loci.

$(n,\ P)$	λ	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	0.8	0.1	0.610	0.780	0.358	0.996
	0.9	0.1	0.850	0.900	0.320	0.998
	1.0	0.1	0.960	1.000	0.219	0.999
(800,5000)	0.8	0.1	0.470	0.660	0.405	0.999
	0.9	0.1	0.750	0.870	0.380	0.999
	1.0	0.1	0.890	0.930	0.233	0.999

Table 3: The simulated average PDR and FDR under Model 2: multiplicative effects within loci but not between loci.

$(n,\ P)$	λ	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	0.5	0.1	0.265	0.175	0.086	0.352
	0.5	0.2	0.650	0.550	0.071	0.763
	0.7	0.1	0.790	0.710	0.048	0.758
	0.7	0.2	0.950	1.000	0.050	0.954
(800,5000)	0.5	0.1	0.175	0.085	0.079	0.595
	0.5	0.2	0.610	0.405	0.077	0.928
	0.7	0.1	0.720	0.480	0.062	0.776
	0.7	0.2	0.940	0.930	0.051	0.980

The SNPs in the above model mimic the 14 SNPs involved in the top 5 main-effect features and top 5 interaction features of the CGEMS prostate cancer data. The minor allele frequencies (MAF) of the SNPs, which are estimated from the prostate cancer data, are given as follows:

$$\text{MAF} = (0.31, 0.12, 0.29, 0.12, 0.13, 0.13, 0.47, 0.18, 0.29, 0.16, 0.04, 0.12, 0.36, 0.40). \tag{4.3}$$

The genotypes of these 14 SNPs are generated by using the MAF, assuming Hardy- Weinburg Equilibrium. In addition to these 14 SNPs, 20,000 noncausal SNPs are randomly selected (without replacement) from the 294,179 SNPs of the prostate cancer data in each simulation replicate. For each simulation replicate, 1,000 cases and 1,000 controls are generated. They are matched by randomly selected (without replacement) individuals from the prostate cancer data. Their genotypes at the 20,000 noncausal SNPs are taken the same as those in the prostate cancer data.

In the TPLR approach, 50 main effect features and 50 interaction features are selected in the screening stage using the tournament screening strategy. In the selection stage, EBIC(γ) values are calculated for the nested models with γ in the range 0(0.1)2, that is, from 0 to 2 in space of 0.1.

In the LASSO-patternsearch approach, at the screening stage, 0.05 and 0.002 are used as thresholds for the P-values of the main-effect features and interaction features, respectively. At the LASSO selection step, a 5-fold cross-validation is used for the choice of penalty parameter. At the hypothesis testing step, 9 α levels are considered, that is,

Table 4: The simulated average PDR and FDR under Model 3: two-locus threshold interaction effects.

$(n,\ P)$	λ	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	0.8	0.1	0.530	0.455	0.086	0.884
	0.9	0.1	0.730	0.695	0.052	0.965
	1.0	0.1	0.810	0.840	0.047	0.970
(800,5000)	0.8	0.1	0.350	0.270	0.028	0.800
	0.9	0.1	0.620	0.490	0.101	0.999
	1.0	0.1	0.712	0.657	0.060	0.982

Table 5: The simulated average PDR and FDR under Model 4: significant interaction effect but zero marginal effects.

$(n\ , P)$	ξ_{12}	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	1.9	0.1	0.828	0.702	0.012	≥ 0.550
	2.0	0.1	0.945	0.860	0.026	≥ 0.641
	2.1	0.1	0.965	0.915	0.015	≥ 0.915
(800,5000)	1.9	0.1	0.555	0.460	0.009	≥ 0.406
	2.0	0.1	0.730	0.710	0.014	≥ 0.427
	2.1	0.1	0.885	0.795	0.006	≥ 0.562

$\alpha = 10^{-k}, k = 0,1,\ldots,8$. The case $\alpha = 1$ amounts to stopping the procedure at the LASSO selection step.

Since in the TPLR approach there is not a definite choice of γ, to facilitate the comparison, we calculate PDR and FDR for each fixed γ value in the TPLR approach, and for each fixed α level in LASSO-patternsearch. The PDR and FDR are calculated separately for the detection of true main-effect and interaction features. They are also calculated for the detection of causal SNPs. A causal SNP is considered positively discovered if it is selected either as a main-effect feature or a constituent in an interaction feature. The simulated FDR and PDR over 100 replicates of TPLR with $n_M = n_I = 50$ and $\gamma = 0(0.2)2$ and those of LASSO-patternsearch with $\alpha = 10^{-k}, k = 0,\ 1,\ldots,8$ are reported in Table 6. It is actually the γ values in the higher end and α levels in the lower end that will be involved in the final selection. The comparison of the results with those values is more relevant. As shown by the bold digits in Table 6, TPLR has higher PDR and lower FDR than LASSO-patternsearch across-the-board. For the main-effect features, the lowest FDR of TPLR is 0.006 while it achieves PDR around 0.65, but the lowest FDR of LASSO-patternsearch is around 0.2 while it only achieves PDR around 0.6. The FDR and PDR on interaction features and causal SNPs have the same pattern. When the two approaches have about the same PDR, the LASSO-patternsearch has a much larger undesirable FDR than TPLR. For example, on the SNPs, when the PDR is 0.608 for TPLR and 0.609 for LASSO-patternsearch, the FDRs are, respectively, 0.041 and 0.654; on the main-effect features, when the PDR is 0.646 for both TPLR and LASSO-patternsearch, the FDRs are, respectively, 0.006 and 0.220. The ROC curves of the two approaches in identifying etiological SNPs are plotted in Figure 1. Figure 1 shows clearly that the PDR of TPLR is much higher than the PDR of LASSO-patternsearch when FDR is the same, which is true uniformly over FDR.

Table 6: Comparison of TPLR approach and LASSO-patternsearch (the PDR and FDR with subscript M, I, and S indicate the rates calculated for main-effect features, interaction features, and SNPs resp.)

TPLR Approach						
γ	PDR_M	FDR_M	PDR_I	FDR_I	PDR_S	FDR_S
0.0	0.964	0.902	0.884	0.907	0.947	0.855
0.2	0.768	0.926	0.884	0.900	0.947	0.852
0.4	0.714	0.505	0.748	0.677	0.803	0.494
0.6	0.692	0.199	0.680	0.413	0.730	0.200
0.8	0.684	0.112	0.642	0.331	**0.691**	**0.126**
1.0	**0.672**	**0.023**	**0.610**	**0.267**	**0.654**	**0.073**
1.2	**0.668**	**0.021**	**0.594**	**0.237**	**0.638**	**0.057**
1.4	**0.658**	**0.009**	**0.566**	**0.201**	**0.608**	**0.041**
1.6	**0.654**	**0.006**	**0.536**	**0.165**	**0.578**	**0.026**
1.8	**0.646**	**0.006**	**0.524**	**0.144**	**0.563**	**0.021**
2.0	**0.630**	**0.006**	**0.492**	**0.109**	**0.526**	**0.015**
LASSO-patternsearch						
α	PDR_M	FDR_M	PDR_I	FDR_I	PDR_S	FDR_S
10^{-0}	0.882	0.445	0.710	0.967	0.847	0.940
10^{-1}	0.816	0.332	0.696	0.957	0.827	0.926
10^{-2}	0.786	0.283	0.664	0.929	0.774	0.885
10^{-3}	0.718	0.241	**0.618**	**0.869**	**0.694**	**0.802**
10^{-4}	**0.646**	**0.220**	**0.556**	**0.752**	**0.609**	**0.654**
10^{-5}	**0.578**	**0.193**	**0.486**	**0.563**	**0.531**	**0.444**
10^{-6}	**0.504**	**0.184**	**0.414**	**0.355**	**0.453**	**0.254**
10^{-7}	**0.400**	**0.190**	**0.360**	**0.196**	**0.383**	**0.130**
10^{-8}	**0.332**	**0.202**	**0.292**	**0.076**	**0.316**	**0.047**

To investigate the effect of the choice of n_M and n_I, we considered $n_M = n_I = 15, 25$, and 50 which are 3, 5, and 10 times of the actual number of causal features, respectively. The simulation results show that, though there is a slight difference between the choice of 15 and the other two choices, there is no substantial difference between the choice of 25 and 50. This justifies the strategy given at the end of Section 2. The detailed results on the comparison of the choices are given in the supplementary document.

We also investigated whether the ranking step in the TPLR approach really reflects the actual importance of the features. The average ranks of the ten causal features over the 100 simulation replicates are given in Table 7.

On the average, the causal features are all among the top ten ranks. This gives a justification for the ranking step in the selection stage of the TPLR approach.

5. Some Remarks

It is a common understanding that individual SNPs are unlikely to play an important role in the development of complex diseases, and, instead, it is the interactions of many SNPs that are behind disease developments, see Garte [28]. The finding that only interaction features are selected (since they are more significant than main-effect features) in our analysis provides some evidence to this understanding. Perhaps, even higher-order interactions should be

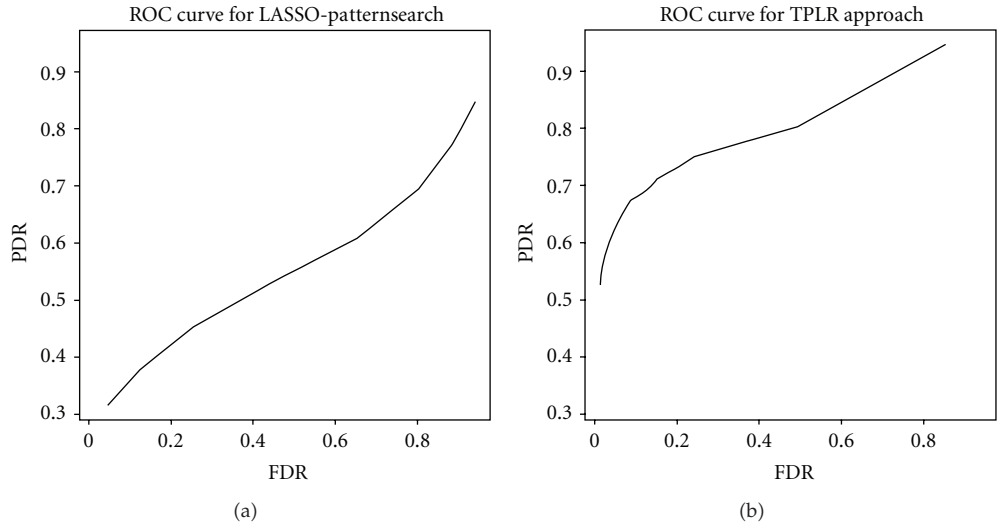


Figure 1: The ROC curves of the LASSO-patternsearch and the TPLR approach for identifying etiological SNPs.

Table 7

Features	1	2	3	4	5	(6,7)	(8,9)	(10,11)	(2,12)	(13,14)
Avg. ranks	4.7	2.0	7.2	6.1	5.4	7.6	6.8	9.2	3.0	1.1

investigated. This makes methods such as the penalized logistic regression which can deal with interactions even more desirable.

The analysis of the CGEMS prostate cancer data can be refined by replacing the binary logistic model with a polytomous logistic regression model taking into account that the genetic mechanisms behind aggressive and nonaggressive prostate cancers might be different. Accordingly, the penalty in the penalized likelihood can be replaced by some variants of the group LASSO penalty considered by Huang et al. [29]. A polytomous logistic regression model with an appropriate penalty function is of general interest in feature selection with multinomial responses, which will be pursued elsewhere.

Acknowledgments

The authors would like to thank the National Cancer Institute of USA for granting the access to the CGEMS prostate cancer data. The research of the authors is supported by Research Grant R-155-000-065-112 of the National University of Singapore, and the research of the first author was done when she was a Ph.D. student at the National University of Singapore.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, Calif, USA, 1984.
[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[3] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[4] H. Schwender and K. Ickstadt, "Identification of SNP interactions using logic regression," *Biostatistics*, vol. 9, no. 1, pp. 187–198, 2008.
[5] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, vol. 37, no. 4, pp. 413–417, 2005.
[6] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.
[7] B. Efron and R. Tibshirani, "Empirical Bayes methods and false discovery rates for microarrays," *Genetic Epidemiology*, vol. 23, no. 1, pp. 70–86, 2002.
[8] J. D. Storey and R. Tibshirani, "Statistical Methods for Identifying Differentially Expressed Genes in DNA Microarrays," *Functional Genomics: Methods in Molecular Biology*, vol. 224, pp. 149–157, 1993.
[9] J. Hoh and J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes," *Nature Reviews Genetics*, vol. 4, no. 9, pp. 701–709, 2003.
[10] Z. Chen and J. Chen, "Tournament screening cum EBIC for feature selection with high-dimensional feature spaces," *Science in China. Series A*, vol. 52, no. 6, pp. 1327–1341, 2009.
[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
[12] D. Firth, "Bias reduction of maximum likelihood estimates," *Biometrika*, vol. 80, no. 1, pp. 27–38, 1993.
[13] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.
[14] J. Chen and Z. Chen, "Extended BIC for small-n-large-P sparse GLM," *Statistica Sinica*. In press.
[15] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," *The Annals of Statistics*, vol. 36, no. 6, pp. 2605–2637, 2008.
[16] W. Shi, K. E. Lee, and G. Wahba, "Detecting disease-causing genes by LASSO-Patternsearch algorithm," *BMC Proceedings*, vol. 1, supplement 1, p. S60, 2007.
[17] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society. Series B*, vol. 70, no. 5, pp. 849–911, 2008.
[18] K. K. Wu, *Comparison of sure independence screening and tournament screening for feature selection with ultra-high dimensional feature space*, Honor's thesis, Department of Statistics & Applied Probability, National University of Singapore, 2010.
[19] W. L. H. Koh, *The comparison of two-stage feature selection methods in small-n-large-p problems*, Honor's thesis, Department of Statistics & Applied Probability, National University of Singapore, 2011.
[20] J. Zhao, *Model selection methods and their applications in genome-wide association studies*, Ph.D. thesis, Department of Statistics and Applied Probability, National University of Singapore, 2008.
[21] A. Albert and J. A. Anderson, "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, vol. 71, no. 1, pp. 1–10, 1984.
[22] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2008.
[23] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.
[24] P. Armitage, *Statistical Methods in Medical Research*, Blackwell, Oxford, UK, 1971.
[25] N. Breslow and N. E. Day, *Statistical Methods in Cancer Research*, vol. 1 of *The Analysis of Case-Control Studies*, International Agency for Research on Cancer Scientific Publications, Lyon, France, 1980.
[26] M. Y. Park and T. Hastie, "An L_1 regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society. Series B*, vol. 69, no. 4, pp. 659–677, 2007.
[27] M. Yeager, N. Orr, R. B. Hayes et al., "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24," *Nature Genetics*, vol. 39, no. 5, pp. 645–649, 2007.
[28] S. Garte, "Metabolic susceptibility genes as cancer risk factors: time for a reassessment?" *Cancer Epidemiology Biomarkers and Prevention*, vol. 10, no. 12, pp. 1233–1237, 2001.
[29] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.