Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

____________________________ ________________________

Your Name Date

Novel Statistical Methods for Analyzing Next Generation Sequencing Data

By

Peizhou (Devin) Liao

Doctor of Philosophy

Biostatistics

Yijuan Hu, Ph.D.
Advisor

Glen A. Satten, Ph.D.
Advisor

Michael P. Epstein, Ph.D.
Committee Member

Zhaohui (Steve) Qin, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Novel Statistical Methods for Analyzing Next Generation Sequencing Data

By

Peizhou (Devin) Liao
M.S., Emory University, 2016
B.S., Nankai University, 2012

Advisors:
Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Abstract

Novel Statistical Methods for Analyzing Next Generation Sequencing Data

By

Peizhou (Devin) Liao

The recent advancement of next-generation sequencing (NGS) technologies and the rapid reduction of sequencing costs have led to extensive use of sequencing data in disease association studies and population genetic studies. New challenges arise from NGS data for statistical analysis, including genotype calling, inference of population structure, and design of sequencing studies, etc. In this dissertation, we propose some novel statistical methods for analyzing NGS data that can properly handle these issues.

A fundamental challenge in analyzing NGS data is to determine an individual's genotype correctly, as the accuracy of the inferred genotype is essential to downstream analyses. To improve the accuracy of called genotypes, in the first project, we propose a new likelihood-based genotype-calling approach that exploits all reads and estimates the per-base error rates by incorporating *phred* scores through a logistic regression model. The approach, which we call PhredEM, uses the expectation-maximization (EM) algorithm to obtain consistent estimates of genotype frequencies and logistic regression parameters. It also includes a simple, computationally efficient screening algorithm to identify loci that are estimated to be monomorphic, so that only loci estimated to be nonmonomorphic require application of the EM algorithm. PhredEM can be used together with a linkage-disequilibrium-based method such as Beagle, which can further improve genotype calling as a refinement step. We demonstrate the advantages of PhredEM over existing methods using both simulated data and real sequencing data from the UK10K project and the 1000 Genomes project.

Inferring population structure is important for both population genetics and genetic epidemiology. Principal components analysis (PCA) has been effective in ascertaining population structure with array genotype data but can yield biased conclusions when used with NGS data having sequencing properties that are systematically different across different groups of samples. To allow robust inference on population structure using PCA, in the second project, we provide an approach that is based on using sequencing reads directly without calling genotypes. Our approach is to adjust the data from different sequencing groups to have the same read depth and error rate so that PCA does not generate spurious components representing sequencing quality. To accomplish this, we have developed a subsampling procedure to match the depth distributions in different sequencing groups, and a read-flipping procedure to match the error rates. We average over subsamples and read flips to minimize loss of information. We demonstrate the utility of our approach using two datasets from 1000 Genomes, and further evaluate it using simulation studies.

We have recently developed TASER, an association test of rare variants with NGS data that allows systematic differences in sequencing qualities (e.g., depth and sequencing error rate) between cases and controls. However, it is unknown what is the optimal design of a case-control study that has a trade-off between number of samples and coverage of depth. In the third project, we conducted simulation studies to evaluate how the sequencing effort should be best allocated between sample size and depth based on factors including ancestry, sequencing error rate, and disease risk model. We found that the best power was generally achieved by sequencing as many samples as possible (while decreasing depth if necessary). We noted, however, when the sequencing platform had a very high error rate (e.g., 0.5%) and rarer variants incurred higher risks, the best power was then achieved with a medium (e.g., 10×) depth.

Novel Statistical Methods for Analyzing Next Generation Sequencing Data

By

Peizhou (Devin) Liao
M.S., Emory University, 2016
B.S., Nankai University, 2012

Advisors:
Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Acknowledgement

First and foremost, I would like to sincerely thank my advisors, Dr. Yijuan Hu and Dr. Glen, A. Satten, for their guidance, inspiration, and tremendous support over the past five years. It has been a great honor to work with them. They have taught me strong analytical and technical skills, and have constantly stimulated my thoughts. The passion and enthusiasm they have for their research was contagious and motivational for me, even during tough times in my Ph.D. studies. I am also grateful to them for helping me improve the ability to work independently as well as collaboratively that a successful statistician must have. I remember they always encourage me to think harder and work harder whenever I struggle with research projects. In my future endeavors, their supervision will never be forgotten.

I would also like to thank my dissertation committee members, Dr. Michael P. Epstein and Dr. Zhaohui (Steve) Qin, for all their contributions of time, thoughtful comments, and creative ideas. Their invaluable suggestions have led to substantial improvement in this dissertation.

I owe special thanks to Dr. Jeanie Park for providing me the opportunity to work as a biostatistician in her lab. Dr. Park has been supportive since the days I started working on her chronic kidney disease projects. The open discussions we have constitute one of the most enjoyable moments for me at Emory. I also have to thank Dr. Tianwei Yu, Dr. Hao Wu, and Dr. Qi Long, who gave me helpful advice on conducting research in biostatistics. I would like to extend my appreciation to all faculty and staff members in the Department of Biostatistics and Bioinformatics at Emory, and to all my friends, with whom I shared many precious moments over the past five years.

Finally, I wish to express my deepest gratitude to my whole family for their love and support. I cannot complete this work without their unconditional dedication. I hope this dissertation could be the perfect present to my entire family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Recent technological advances in next-generation sequencing (NGS) are producing massive amounts of sequencing data, which provide abundant information and extensive resources in disease association studies and population genetic studies. In current NGS methods, the whole genome or some targeted regions are subdivided into small fragments that get sequenced, and the sequencing reads are then aligned to the reference genome. The sequencing data can suffer from errors introduced in both the base-calling process and the alignment process. These errors cause considerable uncertainty in the downstream analyses based on the inferred single-nucleotide polymorphisms (SNPs) and genotypes. Moreover, making integration of samples sequenced at different depths or on different platforms can yield biased results, which partially explains the underutilization of NGS data. Finally, although sequencing costs are declining, performing whole-genome sequencing (WGS) at high depth in large cohort studies is still economically prohibitive, so that many NGS studies have adopted whole-exome sequencing (WES), or have kept the design of WGS but have chosen low or moderate depths. Given a fixed budget, it is critical to develop efficient study designs that consider the trade-off between the number of samples and the coverage depth, especially for detecting rare variant association. Under these circumstances, a variety of contexts in genetic studies, including genotype calling, ancestry estimation, disease mapping, and design of association studies etc., have become challenging, and have been the subject of extensive research. Therefore, this dissertation aims at developing novel statistical methods for improved genotype calling and robust inference of population structure. Another goal of this dissertation is to evaluate the statistical methods for testing rare variant association.

1.2 Research Topics

In the first project, we focuses on the fundamental challenge in analyzing NGS data, i.e., to determine an individual's genotype correctly. In the second project, we concern inference of population structure by combining NGS data with systematic differences in sequencing. In the third project, we explore the optimal designs for testing rare variant associations using TASER recently developed by our group. The methods for all three topics have undergone significant developments in recent years.

1.2.1 Genotype calling

In NGS studies, genotype calling refers to the determination of the actual genotype for each individual at each locus. It is a fundamental challenge in analyzing NGS data as the downstream analyses depend crucially on the accuracy of the inferred genotype. Basically, genotype calling relies on the number of reads (i.e., read depth T) and qualities of reads mapped to the locus. Genotypes covered by many reads can typically be called reliably. However, when a locus is covered by only a few reads, genotype calling is challenging because minor allele reads are indistinguishable from sequencing errors. The sequencing error rates of individual reads comprise both base-calling and alignment errors. The base-calling error rate ranges from a few tenths of a percent to several percent (Nielsen et al., 2011), can vary from base to base as a result of machine cycle and sequence context (Kircher et al., 2009), and also varies dramatically across different sequencing platforms. The *phred* score has been widely accepted as a measure of the base-calling error rate (Ewing et al., 1998; Ewing and Green, 1998). Nominally, the *phred* score is defined as

$$Q = -10 \log_{10} \Pr(\text{observed allele} \neq \text{true allele}). \tag{1.1}$$

Despite their widespread use, *phred* scores may not accurately reflect the true error rates in base calling because they fail to account for some important factors. For instance, the specific error pattern inherent in each nucleotide base (i.e., A, C, T and G) is not considered in *phred* scores (Li et al., 2004). Additionally, *phred* scores do not account for the position of the base within a read (DePristo et al., 2011). Since *phred* scores might be inaccurate representations of true base-calling error rates, methods have been developed to recalibrate base quality scores, such as the base quality score recalibration (BQSR) option in GATK (DePristo et al., 2011) and the base alignment quality (BAQ) option in SAMtools (Li, 2011). However, the effectiveness of recalibration highly depends on whether all important error predictors (e.g., machine cycle and dinucleotide context) are included in the recalibration model. In addition, the recalibration process can be computationally intensive (Yu et al., 2015). Compared with the base-calling error rate, the alignment error rate has less variability and a smaller magnitude.

A genotype-calling method generally uses a probabilistic framework, combining base-calling error rates and a marginal (population-level) distribution of genotype frequencies to provide an individual-level probability for each genotype (McKenna et al., 2010; Li et al., 2009a; Martin et al., 2010). Because the error rate plays a critical role in probabilistic genotype-calling algorithms, it is crucial that it be correctly specified, especially when sequencing depth is low to moderate.

In the first project, we propose a new genotype-calling approach which estimates base-calling error rates from the read data while incorporating the information in *phred* scores. We model an error rate as a logistic function of the *phred* score. The logistic regression model is readily integrated into a modification of the SeqEM likelihood which allows for a base-specific error probability. Like SeqEM (Martin et al., 2010), our approach uses the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Information from all individuals is used to estimate the unknown genotype frequencies and logistic regression parameters. We compute the probability of each latent genotype for each individual based on parameter estimates and use the empirical Bayes approach to assign the most likely genotype to each individual. We show that the logistic model fits real sequencing data well, and that the unknown parameters in our likelihood are consistently estimated. Because we allow separate logistic regression parameters at each locus, error predictors that are the same for all bases at a given locus (e.g., dinucleotide context) are automatically accounted for, as in SeqEM.

To minimize the effort of calling genotypes for the large majority of loci that are estimated to have no variation, we develop a simple, computationally efficient screening algorithm to identify loci that are estimated to be monomorphic and therefore do not require parameter estimation using the EM algorithm. Furthermore, we show that our approach can be used together with a linkage-disequilibrium (LD)-based method such as Beagle to improve genotype calling. Finally, we demonstrate through simulation studies and by comparison to gene array data that our approach is more accurate than both SeqEM and GATK. We illustrate our new approach through an application to two real sequencing datasets, one from the UK10K project and the other from the 1000 Genomes project.

1.2.2 Inference of population structure

Accurate estimation of ancestry remains an important topic in both population genetics and genetic epidemiology. Principal components analysis (PCA) is a powerful tool for inference of population structure, and has been effective in visualizing genetic data (Menozzi et al., 1978; Cavalli-Sforza et al., 1993), investigating population history and differentiation (Reich et al., 2008), and in adjusting for confounding due to population stratification in association studies (Price et al., 2006). It is known that the success of PCA depends on high-quality genotype data (Wang et al., 2014), such as the data generated from genotyping arrays.

NGS of DNA is replacing genotyping arrays, and is capable of probing the entirety of the human genome. However, sequencing protocols and platforms are highly variable in different studies. Systematic sequencing differences arise when samples sequenced at different depths or on different platforms are pooled for analysis. In population genetics, it is common to combine samples from different resources for a global study of population structure. In association mapping, some studies sequence cases at higher depth than controls by design, when the cases are unique and there is interest in identifying novel mutations (The UK10K Consortium, 2015). Some studies even sample only cases for sequencing and intend to compare them with publicly available sequenced controls such as the 1000 Genomes (The 1000 Genomes Project Consortium, 2010) or UK10K (The UK10K Consortium, 2015). In both settings, the controls typically have systematically different sequencing qualities, e.g., depth and/or base-calling error rate, from the cases. Even when their overall depths are similar, their depth in individual regions may be different; this can easily occur when different exome capture kits were used for cases and controls, and one kit captures a certain exonic region better than the other.

Traditional methods for performing PCA lead to incorrect differentiation of populations when applied to genotype calls from low or moderate coverage NGS data (Fumagalli et al., 2013). Such a problem becomes much worse if samples from multiple sequencing groups are all used to infer population structure. Recently, extensions of PCA have been made to utilize sequencing reads directly without calling genotypes. However, to date, no method exists to account for systematic sequencing differences to accurately estimate population structure.

In the second project, we provide a new approach to inferring population structure while explicitly accounting for the difference in read depth and error rates; it is based on sequencing reads directly without calling genotypes. The underlying approach is to adjust the data so that the sequencing quality appears to be equal among groups. We first describe a subsampling procedure to match the depth distributions in different sequencing groups, and a read-flipping procedure to adjust the data so that the error rates in different sequencing groups agree with the group having the largest error rate. Once the data are processed in this way, we calculate the variance-covariance matrix of the proportion of reads that are for the minor allele; this variance-covariance matrix does not have any spurious PCs corresponding to differences in sequencing quality. We then repeat the subsampling and allele-flipping procedures and average the resulting variance-covariance matrices, to minimize loss of information. We show that the information remaining is more than enough to make reliable inference of population structure. We demonstrate the performance of our method with two examples using data from the 1000 Genomes Project, one involving three discrete Asian populations and the other involving a continuous admixture of two populations. We further evaluated our method using simulation studies.

1.2.3 Sequencing design for rare variant association studies

NGS represents a powerful tool to fully understand the role of genetic variation underlying human diseases and traits. WGS or WES allow researchers to examine the contribution of variants across the full MAF spectrum in complex disease (Goldstein et al., 2013; Lee et al., 2014; Sham and Purcell, 2014), leading to great success in discovery of genes and causal variants over the past few years (Bamshad et al., 2011; Iossifov et al., 2014; Gilissen et al., 2014). We anticipate that NGS studies will continue to expand our understanding of complex trait architecture for some time to come.

Despite the falling cost of sequencing in recent years, it is still prohibitively expensive to conduct large-scale NGS studies using high-coverage sequencing. A key factor in the success of sequencing studies is the allocation of sequencing resource, in particular, how to divide the sequencing effort between the number of samples and the coverage depth (Sampson et al., 2011; Li et al., 2011; Sims et al., 2014). The efficient allocation of sequencing effort is essential for rare variant association studies because the accurate calling of rare variants inevitably requires each position being covered by a sufficient number of reads in the presence of the sequencing errors (Shen et al., 2011), while to have adequate power for detecting the rare variant association generally requires a large number of samples (Lee et al., 2014). Moreover, for the most commonly used case-control design in studying rare variant association, the total sequencing investment is not necessarily split equally between cases and controls. Indeed, controls are generally less interesting than cases so that controls may be sequenced at a much lower depth compared with cases (The UK10K Consortium, 2015). There are even NGS studies that sequence cases at lower depth than the public controls available (Luo et al., 2017).

In the third project, we systematically explore the power of rare variant association testing under the constraint of limited cost being available for sequencing controls while cases having been sequenced at a good coverage. We used TASER recently developed in Hu et al. (2016), which allows for systematic differences in sequencing between cases and controls, to perform the association test. Because the underlying disease model is generally unknown, we develop the omnibus TASER which combines multiple weight functions and maintains good power regardless of the true risk model. Via realistic simulations, we assess the impact of factors including ancestry, sequencing error rate, and disease risk model on the power. Our results show that, given a fixed budget, low-coverage sequencing of a large number of controls is generally preferred rather than moderate- to high-coverage sequencing of fewer controls. However, if the sequencing platform has high error rates and rarer variants incurred higher risks, the best power was then achieved with a moderate (e.g., 10×) depth.

1.3 Literature Review

1.3.1 Methods for calling genotypes

In early NGS studies, genotype calling proceeds by first filtering out reads of low *phred* scores, and then counting the number of alleles observed; if the number of minor allele reads (R), falls within some prespecified range, a homozygous or heterozygous genotype would be called (Hedges et al., 2009; Harismendy et al., 2009). This standard procedure works well with high-coverage sequencing data. However, the major disadvantage of this procedure is that by using fixed cutoff, it ignores the information about the allele frequency and the individual read quality. Another disadvantage is that this simple genotype calling method provides no quantification of uncertainty associated with the called genotype (Nielsen et al., 2011).

Most of recently developed genotype-calling methods use a probabilistic framework that provides posterior probabilities for potential genotypes by combining base-calling error rates and a prior distribution of genotype frequencies (McKenna et al., 2010; Li et al., 2009a; Martin et al., 2010). Specifically, at a particular locus, the read data $\boldsymbol{X}$, including T, R, and Q, is used to calculate the genotype likelihood $\Pr(\boldsymbol{X}|G)$ where G denotes the true genotype. In conjunction with a genotype prior, $\Pr(G)$, the posterior genotype probability is calculated as $\Pr(G|\boldsymbol{X}) \propto \Pr(\boldsymbol{X}|G)\Pr(G)$. These probabilistic methods generally differ in their approach to obtaining the error rates. For example, GATK uses error rates that are calculated directly from *phred* scores or recalibrated scores by applying equation (1.1), neither of which is precisely correct as discussed in 1.2.1. SAMtools obtains an error rate from the minimum of the *phred*-based error rate and the mapping error rate, so that the error rate is always adjusted downwards (Li, 2011). In addition, bases with low *phred* scores (e.g., $Q < 20$ or 30) are typically filtered out as part of quality control (QC) procedures. However, choosing a threshold for *phred* scores always involves a tradeoff: high thresholds may result in loss of useful information by eliminating bases that are correctly called, while low thresholds leave a large number of erroneously-called bases in the data, leading to false-positive variant calls. Instead of relying on *phred* scores, Martin et al. (2010) proposed SeqEM, a genotype-calling algorithm that estimates the error rate using the read data itself. However, the fundamental assumption of SeqEM that, at each locus, there is a uniform error rate for each read is generally not true, given the considerable variability in error rates implied by the variability in *phred* scores. Because SeqEM ignores *phred* scores entirely, the valuable information about errors encoded in *phred* scores is lost. Another difference among probabilistic methods is the approach to estimating the allele frequency. The majority of probabilistic methods such as GATK and SAMtools, estimate the allele frequency based on a single locus. Nielsen et al. (2012) presents a strategy to first estimate the Site Frequency Spectrum (SFS) jointly for all loci, and then use the resulting SFS to define better priors for calling genotypes.

The aforementioned approaches generally concern calling genotypes independently for each locus. However, it has been shown that utilizing the pattern of LD at nearby loci can further improve genotype calling accuracy, especially with low coverage sequencing data (Nielsen et al., 2011; Li et al., 2011). Several genotype imputation methods have been developed to infer genotypes by using the information at linked loci (Browning and Yu, 2009; Howie et al., 2009; Li et al., 2010; Marchini and Howie, 2010). The single-locus-based genotype calling approaches can be used together with LD-based imputation methods to incorporate LD information, which substantially improves the accuracy for genotype calling.

1.3.2 Methods for inferring population structure

PCA was initially introduced to analyze the genetic data in Menozzi et al. (1978), and has become the most common approach for inferring ancestry (Patterson et al., 2006;

Price et al., 2006; The Wellcome Trust Case Control Consortium, 2007; Yang et al., 2010). The top components explain the difference in genetic variation among the samples, which can be used to correct for confounding due to population stratification in a variety of ways (Price et al., 2006; Epstein et al., 2007; Luca et al., 2008). Construction of PCs based on genotype data is straightforward and computationally efficient (Jackson, 2003), which requires highly accurate genotype calls.

With the unprecedented volume of sequencing data being produced in recent years, methods have been developed for performing PCA utilizing the sequencing reads directly. Skoglund and Jakobsson (2011) obtained allele count data for PC calculation by randomly sampling one read from each individual at each position, in order to allow comparison between modern, high-quality data and the low-pass ancient data. Similarly, Malaspinas et al. (2014) developed a tool that samples a read at each position and compares the read count data with an existing reference panel of genotype data using multidimensional scaling. A major disadvantage of these methods is that it leads to great loss of information in the presence of sequencing errors. Fumagalli et al. (2013) proposed replacing the genotypic covariance matrix by its expected value with respect to the posterior genotype distribution given read data. Through explicit modeling of genotype probability distributions, the PCs can be accurately estimated when sequencing qualities are the same across the samples. However, the method of Fumagalli et al. (2013) does not take any measure to deal with differential sequencing qualities in terms of depth and error rates. Wang et al. (2014) proposed comparing each sequenced study sample to a set of reference individuals whose ancestral information is known and whose genome-wide array genotype data are available. This method seems to allow differences in depth but nevertheless assumes a constant error rate for all study samples. The same applies to the improved approach in Wang et al. (2015).

1.3.3 Design of NGS studies for testing rare variant associations

Over the past few years, the optimal design of sequencing-based rare variant association studies has been extensively explored, and the benefits of low-coverage sequencing are often highlighted (Li et al., 2011; Pasaniuc et al., 2012; Xu et al., 2016). However, all these studies focus on the scenario with equal sizes of independent case-control samples that are sequenced as part of the same experiment design. Because the sequencing qualities are the same, the conventional burden tests (Li and Leal, 2008; Madsen and Browning, 2009; Price et al., 2010) or variance component tests such as C-alpha (Neale et al., 2011) and SKAT (Wu et al., 2011) can be used for association testing based on genotype calls. Additionally, Skotte et al. (2012) proposed to replace genotype calls by their expected value given the observed sequencing data (i.e., genotype dosages), which results in higher power and better control of type I error than methods using called genotypes.

Recently, many discussions in the literature suggest that association testing using data with systematically different sequencing qualities (e.g., read depth and error rate) in case and control cohorts generates false signals if called genotypes or genotype dosages are used for the main effect (Derkach et al., 2014; Hu et al., 2016). To adjust for the confounding effect induced by genotype calls, Derkach et al. (2014) proposed a robust score test that uses genotype likelihoods whose differential variances in high- and low-depth samples are explicitly accounted for. One limitation of this method is that it requires correctly known locations of variants. Because the called genotypes are used to determine the variant locations, it still yields inflated type I error. Instead of calling genotypes, Hu et al. (2016) recently developed a likelihood-based burden test that directly models sequencing reads. We refer to Hu's method as TASER that is the name of their software. TASER includes a simple, computationally efficient screening algorithm to first identify a set of 'known' variants (i.e., estimated to be polymorphic). Then it computes the burden test statistic by adding up the score statistics at each ‘known’ variant with certain weights. Finally, TASER assesses the significance of the test statistic using bootstrap replications. One important feature of TASER is that it is at least as powerful as the standard genotype calling approach when the latter controls type I error.

Though methods have been developed to make valid inference when cases and controls are sequenced separately or at different depth, it remains unclear what is the most efficient study design that considers the trade-off between the sample size (n) and the coverage per sample (c), in order to maximize the power for detecting associations of rare variants. Thus, further work is required to select the optimal combination of n and c, especially for studies where, as is the common practice, cases are sequenced at high depth but a fixed budget is assigned to the sequencing of controls.

1.4 Outline

In Chapter 2, in Section 2.1, we introduce the PhredEM approach, the screening algorithm, and PhredEM with LD refinement. In Section 2.2, we report the results from simulation studies for comparing the performance of PhredEM to SeqEM, without and with LD refinement. In Section 2.3, we apply PhredEM to real sequencing data from the UK10K project and the 1000 Genomes project to illustrate the practical use of PhredEM. In Section 2.4, we provide a summary and some detailed discussions.

In Chapter 3, in Section 3.1, we present the subsampling and read-flipping procedure to adjust the sequencing data; we also describe two datasets from 1000 Genomes project and the simulation design. We report in Section 3.2 all results by comparing our method with some existing methods. We conclude our work with a brief discussion in Section 3.3.

In Chapter 4, in Section 4.1, we describe the omnibus TASER and our simulation design in details. We report the results from simulation studies in Section 4.2. We summarize our work and discuss some limitations in Section 4.3.

In the final chapter, we summarize all three projects and outline some possible topics as directions for future research.

Chapter 2

PhredEM: A Phred-Score-Informed Genotype-Calling Approach for Next-Generation Sequencing Studies

This Chapter is joint work with Dr. Yijuan Hu and Dr. Glen A. Satten. The paper has been published in *Genetic Epidemiology* (Liao et al., 2017).

2.1 Methods

2.1.1 PhredEM

We first consider one biallelic locus at a time. For the i-th individual, let G_i denote the underlying true genotype (coded as the number of minor alleles), T_i denote the total number of alleles that are mapped to the locus, and R_i ($R_i \leq T_i$) denote the number of mapped alleles that are called to be the minor allele. The *phred* scores are represented by $\boldsymbol{Q}_i = (Q_{i1}, \ldots, Q_{iT_i})'$, where Q_{ik} is the *phred* score associated with the k-th called allele and the prime ($'$) indicates the transpose of a vector. At each locus, values of T_i, R_i, and $\boldsymbol{Q}_i$ can be easily extracted from the pileup files produced by SAMtools. Let ϵ_{ik} be the true base-calling error rate of the k-th allele. We relate ϵ_{ik} to Q_{ik} through the logistic regression model

$$\log\left(\frac{\epsilon_{ik}}{1-\epsilon_{ik}}\right) = \beta_0 + \beta_1 Q_{ik}, \tag{2.1}$$

where β_0 and β_1 are unknown regression parameters that are locus specific. Let $\boldsymbol{\theta} = (\beta_0, \beta_1)'$ and $\epsilon_{ik}(\boldsymbol{\theta}) = \exp(\beta_0 + \beta_1 Q_{ik})/\{1 + \exp(\beta_0 + \beta_1 Q_{ik})\}$. Equation (2.1) is motivated by the fact that the *phred* score is a highly informative predictor of the base-calling error, even though (1.1) does not hold in the exact sense. In the Results section, we demonstrated that the logistic model fits the real sequencing data well.

Without loss of generality, we order the T_i alleles so that the first R_i alleles are called to be the minor allele and the rest the major allele. Assuming that the errors of the T_i alleles are independent of each other, the probability of observing R_i copies of the minor allele out of T_i alleles can be described as a sequence of independent Bernoulli trials. Specifically, given the true genotype G_i, the total number of alleles

T_i, and the *phred* scores $\boldsymbol{Q}_i$, the probability of observing R_i is written as

$$P_{\boldsymbol{\theta}}(R_i|G_i, T_i, \boldsymbol{Q}_i) = \begin{cases} \prod_{k=1}^{R_i} \epsilon_{ik}(\boldsymbol{\theta}) \prod_{k=R_i+1}^{T_i} \left\{1 - \epsilon_{ik}(\boldsymbol{\theta})\right\} & G_i = 0 \\ (0.5)^{T_i} & G_i = 1 \\ \prod_{k=1}^{R_i} \left\{1 - \epsilon_{ik}(\boldsymbol{\theta})\right\} \prod_{k=R_i+1}^{T_i} \epsilon_{ik}(\boldsymbol{\theta}) & G_i = 2. \end{cases} \tag{2.2}$$

Suppose that the sample consists of n unrelated individuals. Then the likelihood function takes the form

$$L_o(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{g=0,1,2} P_{\boldsymbol{\theta}}(R_i|g, T_i, \boldsymbol{Q}_i) P_{\boldsymbol{\pi}}(g), \tag{2.3}$$

where $P_{\boldsymbol{\pi}}(g)$ is the genotype frequency characterized by $\boldsymbol{\pi}$. Under Hardy-Weinberg Equilibrium (HWE), $\boldsymbol{\pi}$ consists of a single parameter π for the minor allele frequency (MAF). Then, $P_{\boldsymbol{\pi}}(0) = (1-\pi)^2$, $P_{\boldsymbol{\pi}}(1) = 2\pi(1-\pi)$, and $P_{\boldsymbol{\pi}}(2) = \pi^2$. Under Hardy-Weinberg Disequilibrium (HWD), $\boldsymbol{\pi} = (\pi, f)'$ where π and f are the MAF and the fixation index F_{st}, respectively. Then, $P_{\boldsymbol{\pi}}(0) = (1-f)(1-\pi)^2 + f(1-\pi)$, $P_{\boldsymbol{\pi}}(1) = 2\pi(1-\pi)(1-f)$, and $P_{\boldsymbol{\pi}}(2) = (1-f)\pi^2 + f\pi$.

The proposed likelihood is closely related to several existing methods. When $\beta_1 = 0$, the error rate is independent of the *phred* score, and expression (2.3) reduces to the likelihood of SeqEM. When $\beta_0 = 0, \beta_1 = -\log(10)/10$ and $\boldsymbol{\epsilon}$ is small, expression (2.1) is approximately equal to (1.1), and our model reduces to the Bayesian genotyper implemented in GATK. However, our likelihood fully exploits the read data and the *phred* scores, both of which could improve genotype-calling accuracy. Note that it is not necessary to filter out low-quality alleles, which still provide some information about $\boldsymbol{\theta}$. Because our model uses the read call data to adjust the relationship between *phred* scores and the error rate at each locus, it can be considered as a kind of *phred* score recalibration, except that the recalibration is done simultaneously with fitting other parameters to best fit the observed data. Like other multi-sample calling methods, our method also estimates the genotype frequencies and regression parameters by utilizing information across all individuals in the sample.

We may obtain the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ by maximizing the likelihood (2.3) via the EM algorithm described in the Appendix 2.6.1. However, if a locus has little variability (e.g., a monomorphic locus, singleton or doubleton) so that there are very few reads for the minor allele in the study sample, the MLE of β_1 based on (2.3) may be unreliable (Firth, 1993). To improve stability, we propose to modify the MLE of β_1 by leveraging information from other loci. Specifically, we introduce a Gamma distribution $\Gamma(-\beta_1; \kappa, \phi)$ as a penalty (or prior) for $-\beta_1$, where κ and ϕ are the shape and scale hyper-parameters, respectively. We first use the method of moments to obtain estimates $\widehat{\kappa}$ and $\widehat{\phi}$ based on the MLEs of β_1 from a set of loci that are either all or mostly estimated to be monomorphic; for loci that are estimated to be monomorphic, all reads for the minor allele can be treated as errors, and ordinary logistic regression can be used to estimate $\boldsymbol{\theta}$ at each locus. For genome- or exome-wide data, any region can be used as most loci are estimated to be monomorphic; the full EM algorithm only needs to be run for the few loci that are estimated to be polymorphic. We then obtain the maximum penalized likelihood estimators (MPLEs) by maximizing the penalized likelihood

$$L_o^*(\boldsymbol{\theta}, \boldsymbol{\pi}) = \Gamma(-\beta_1; \widehat{\kappa}, \widehat{\phi}) L_o(\boldsymbol{\theta}, \boldsymbol{\pi}). \tag{2.4}$$

Note that the MPLEs are asymptotically equivalent to the MLEs, as the Gamma penalty becomes negligible when the sample size n grows.

Denote the MPLEs by $\widehat{\boldsymbol{\pi}}$ and $\widehat{\boldsymbol{\theta}}$. We can estimate the probability distribution of the true genotype G_i for the i-th individual from their read count data T_i and R_i and their *phred* scores $\boldsymbol{Q}_i$ using the formula

$$\Pr(G_i = g | R_i, T_i, \boldsymbol{Q}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}) = \frac{P_{\widehat{\boldsymbol{\theta}}}(R_i | g, T_i, \boldsymbol{Q}_i) P_{\widehat{\boldsymbol{\pi}}}(g)}{\sum_{g'=0}^{2} P_{\widehat{\boldsymbol{\theta}}}(R_i | g', T_i, \boldsymbol{Q}_i) P_{\widehat{\boldsymbol{\pi}}}(g')}, \tag{2.5}$$

for $g = 0$, 1 and 2. At a single locus, genotype calls can be made by assigning each individual the genotype that their data assigns the highest estimated probability. Individuals with no read covering the locus are not assigned any genotype. Because the proposed method incorporates the *phred* scores and uses the EM algorithm, we refer to it as PhredEM.

2.1.2 Screening algorithm

The majority of loci in the human genome are monomorphic (The International SNP Map Working Group, 2001), and are as such of little interest in downstream analyses. To avoid running the full PhredEM algorithm at loci that are estimated to be monomorphic, we propose a simple and computationally efficient algorithm to identify and 'screen out' these loci; an earlier version of this screening algorithm that does not incorporate *phred* scores was first proposed in Hu et al. (2016). We assume HWE holds, as loci that might be called monomorphic must have either zero or extremely low MAFs. Then $\boldsymbol{\pi}$ contains only a single parameter π. We see that formula (2.5) assigns all mass to $G_i = 0$ when $\widehat{\pi} = 0$; thus loci with $\widehat{\pi} = 0$ would be called monomorphic if PhredEM was applied to obtain $\widehat{\pi}$. To determine whether $\widehat{\pi} = 0$ without fitting PhredEM, let $pl^*(\pi)$ denote the profile likelihood for π, namely,

$$pl^*(\pi) = \max_{\boldsymbol{\theta}} \log L_o^*(\boldsymbol{\theta}, \pi).$$

We show in the Appendix 2.6.2 that $pl^*(\pi)$ is a concave function of π, so that a negative value for the derivative of $pl^*(\pi)$ at $\pi = 0$ implies $\widehat{\pi} = 0$; in other words, we should screen out loci at which the derivative of $pl^*(\pi)$ at $\pi = 0$ is negative. At

$\pi = 0$, we can easily evaluate this derivative, because the part $L_o(\boldsymbol{\theta}, \pi)$ reduces to that of a logistic regression model in which we assign an outcome variable $Y_{ik} = 1$ to a minor allele read and $Y_{ik} = 0$ to a major allele read and regress Y_{ik} on Q_{ik}. Since our screening algorithm only involves fitting a standard logistic regression model plus a penalty term to solve for $\boldsymbol{\theta}$ and calculating a derivative function, it can significantly reduce the computing time that is needed to run PhredEM on whole exome or genome data.

A simple variant of the screening algorithm can also be used when estimating the parameters κ and ϕ for the gamma penalty term. If we first apply the screening algorithm using the *unpenalized* profile likelihood $pl(\pi) = \max_{\boldsymbol{\theta}} \log L_o(\boldsymbol{\theta}, \pi)$, we can easily find all loci having $\widehat{\pi} = 0$ without running the full EM algorithm to maximize (2.3) at all loci. If the MLE of π is zero, then β_0 and β_1 can be estimated using standard logistic regression since all minor allele reads are errors. The few loci for which $\widehat{\pi} > 0$ can either be excluded, or the full EM algorithm can be used to estimate β_0 and β_1.

2.1.3 PhredEM with LD refinement

Our approach does not use LD information. It is well known that use of LD patterns can substantially improve genotype calling for variants having moderate or high minor allele frequencies (Nielsen et al., 2011). However, we can easily incorporate LD information into our approach by calculating the genotype likelihood at each locus using (2.2), evaluated at the MPLE, and then using this genotype likelihood as input to Beagle (Browning and Yu, 2009).

2.2 Simulation Study

We conducted simulation studies to assess the performance of PhredEM (P) and PhredEM followed by Beagle (PB), relative to SeqEM (S) and SeqEM followed by Beagle (SB). We considered a sample size of 1,000 (results based on a sample size of 200 are reported in Supplemental Figure S2.1 and Supplemental Tables S2.1 and S2.2). In each replicate, for each individual we first generated a pair of haplotypes of European ancestry having length 100 kb using the coalescent simulator cosi (Schaffner et al., 2005). We then generated sequencing reads with fixed length 100 bp that mimic reads from the Illumina HiSeq 2000 single-end sequencing platform (Minoche et al., 2011). Specifically, for each read from an individual, we randomly selected one of the two haplotypes, randomly picked the starting position of the read along the haplotype, and simulated 100 *phred* scores from the empirical distribution observed in the UK10K data (Figure 2.2[a]). To incorporate the fact that base-calling errors occur at the end of the reads more frequently than at the beginning (Minoche et al., 2011), we rearranged the *phred* scores so that the last 15 bases of the read had the 15 lowest scores in a descending order; the first 85 bases thus received a random permutation of the remaining scores. Then, the base calls of the read were generated based on the underlying haplotype and error rates calculated from equation (1.1); we used (1.1) because it is more favorable to GATK than to our method. For each individual, we drew the total number of reads to be generated from a negative-binomial distribution with mean $1,000 \times c$ so as to achieve a pre-specified average read depth c. We considered three average depths: 6x, 10x, and 30x. In applying PhredEM and SeqEM, we first called genotypes with HWE and, if the estimated MAF was greater than 5%, we re-called genotypes with HWD (starting at parameter values obtained from HWE). The hyper-parameters for the Gamma prior for β_1 were estimated based on the MLEs of β_1 from the 100k loci in each replicate. All results reported here were based on 200 replicates of the entire process.

We first assessed the performance of PhredEM, SeqEM, PB, and SB in truly monomorphic loci. A monomorphic locus is mis-called if there is at least one call of the minor allele in the study sample. Figure 2.1(a) shows that, with or without LD refinement, PhredEM made fewer mistakes among monomorphic loci than SeqEM at all depths. In addition, LD-refinement has negligible improvement upon PhredEM at monomorphic loci.

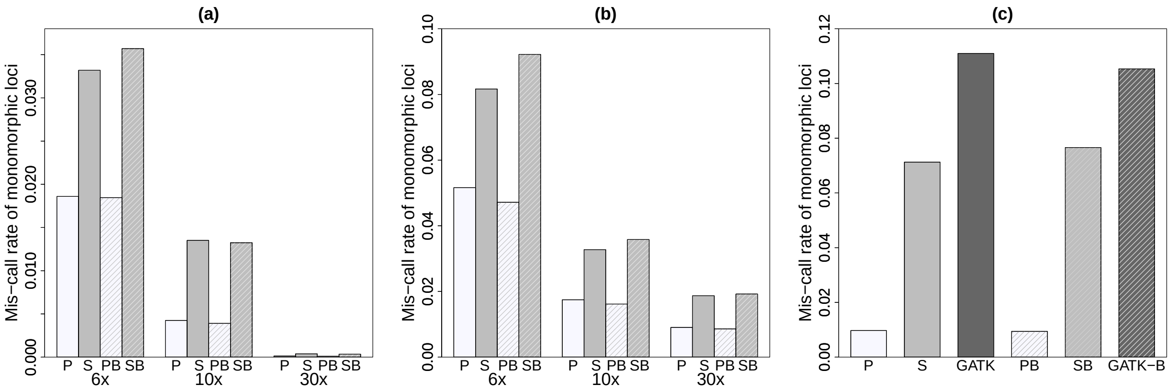

Figure 2.1: Mis-call rates at monomorphic loci in the analysis of (a) the simulated data, (b) the UK10K SCOOP data, and (c) the 1000 Genomes CEU data.
P and S represent PhredEM and SeqEM. PB, SB, and GATK-B represent PhredEM, SeqEM, and GATK, each followed by Beagle.

We then compared the four methods in calling genotypes for rare variants. We grouped variants into four categories based on the true minor allele counts (MACs): 1, [2, 10], [11, 20], and [21, 100], where MAC = 1 corresponds to singletons. As shown in Table 2.1, the overall number of mis-called genotypes obtained by PhredEM was less than that by SeqEM in all scenarios; for most cases, PhredEM reduced by almost one half the number of mis-called genotypes compared with SeqEM. For instance, when the MAC was between 11 and 20 and depth was 6x, SeqEM mis-called an average of 2.96 genotypes among 997 individuals whereas PhredEM mis-called 1.58. As expected, both methods became more accurate as the average read depth increased. Nevertheless, the performance of PhredEM was noticeably better than SeqEM at depth as high as 30x. We further examined the mis-called genotypes stratified by the underlying genotype. In both the strata of homozygotes ($G = 0$) and heterozygotes ($G = 1$), PhredEM mis-called fewer genotypes than SeqEM. Applying Beagle after PhredEM substantially improved the performance of PhredEM alone, except for singletons at which the two methods have comparable mis-call rates. The superiority of PhredEM over SeqEM remained after applying Beagle to both methods.

Table 2.1: Average number of mis-called genotypes per variant for rare variants in the simulation studies.

		Overall					Stratified									
							$G = 0$					$G = 1$				
MAC	Depth	N	P	S	PB	SB	N_0	P	S	PB	SB	N_1	P	S	PB	SB
1	6x	997.1	0.241	0.311	0.274	0.338	996.1	0.065	0.072	0.096	0.096	1	0.176	0.239	0.178	0.242
	10x	999.7	0.074	0.135	0.088	0.159	998.7	0.016	0.033	0.039	0.051	1	0.058	0.102	0.049	0.108
	30x	1000	0.001	0.004	0.002	0.006	999.0	0	0.001	0.001	0.003	1	0.001	0.003	0.001	0.003
$[2, 10]$	6x	997.1	0.525	0.845	0.439	0.691	993.5	0.106	0.112	0.162	0.193	3.6	0.417	0.730	0.275	0.496
	10x	999.7	0.191	0.315	0.142	0.243	996.2	0.049	0.060	0.063	0.082	3.5	0.140	0.253	0.079	0.161
	30x	1000	0.004	0.009	0.003	0.007	996.4	0.001	0.002	0.002	0.004	3.6	0.003	0.007	0.001	0.003
$[11, 20]$	6x	997.0	1.579	2.959	0.779	1.306	982.2	0.387	0.514	0.243	0.429	14.7	1.156	2.409	0.529	0.868
	10x	999.7	0.551	1.011	0.212	0.381	984.9	0.156	0.176	0.090	0.138	14.7	0.380	0.819	0.121	0.241
	30x	1000	0.011	0.026	0.005	0.010	985.1	0.004	0.007	0.003	0.005	14.8	0.007	0.019	0.002	0.005
$[21, 100]$	6x	997.0	4.197	7.633	1.416	2.217	947.8	0.667	2.108	0.347	0.696	48.5	3.136	5.131	1.051	1.489
	10x	999.7	1.457	2.722	0.361	0.603	949.9	0.347	0.606	0.126	0.210	49.1	1.002	1.998	0.230	0.381
	30x	1000	0.032	0.068	0.009	0.016	949.6	0.008	0.015	0.004	0.007	49.6	0.024	0.051	0.005	0.009

P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively. N, N_0 and N_1 are the average numbers of individuals covered by at least one read. G is the true genotype; the case $G = 2$ is omitted as it is barely seen for rare variants. MACs of 1, 10, 20, and 100 correspond to MAFs of 0.0005, 0.005, 0.01, and 0.05, respectively, given the sample size of 1,000.

For common variants, we stratified the results based on five MAF intervals. As shown in Table 2.2, PhredEM outperformed SeqEM in both the overall and stratified number mis-called. Overall, PhredEM correctly called 3–4 more genotypes than SeqEM at depth $\leq$ 10x. The number mis-called by PhredEM increases as the MAF increases because the information in the *phred* scores is not used when $G = 1$, which can be seen from (2.2). Furthermore, minor allele homozygotes are more likely to be mis-called than major allele homozygotes due to the smaller prior probability of the former. As expected, applying Beagle after PhredEM substantially improved genotype calling by PhredEM alone for common variants, and the improvement was most profound for heterozygotes ($G = 1$). This marked improvement was also shown in Supplemental Table S2.3 where the error rates are reported given the called variants instead of the true variants as in Table 2.2.

Table 2.2: Average number of mis-called genotypes per variant for common variants in the simulation studies.

		Overall					Stratified														
							$G = 0$					$G = 1$					$G = 2$				
MAF	Depth	N	P	S	PB	SB	N_0	P	S	PB	SB	N_1	P	S	PB	SB	N_2	P	S	PB	SB
$(0.05, 0.1]$	6x	997.0	11.47	16.94	1.97	2.85	857.9	0.95	4.61	0.44	0.77	133.5	8.80	10.58	1.49	2.01	5.6	1.72	1.75	0.04	0.07
	10x	999.7	3.62	6.28	0.49	0.71	858.9	0.59	1.61	0.14	0.22	135.4	2.59	4.18	0.33	0.47	5.4	0.44	0.49	0.02	0.02
	30x	1000	0.07	0.15	0.01	0.02	858.9	0.01	0.03	0	0.01	135.5	0.05	0.11	0.01	0.01	5.6	0.01	0.01	0	0
$(0.1, 0.2]$	6x	997.1	22.66	28.31	2.15	2.96	725.3	1.14	5.67	0.45	0.78	249.2	17.81	18.78	1.60	2.05	22.6	3.71	3.86	0.10	0.13
	10x	999.7	6.38	9.93	0.64	0.80	731.7	0.72	2.45	0.17	0.21	246.2	4.94	6.55	0.44	0.54	21.8	0.72	0.93	0.03	0.05
	30x	1000	0.11	0.24	0.01	0.02	727.1	0.02	0.05	0	0.01	250.2	0.08	0.17	0.01	0.01	22.7	0.01	0.02	0	0
$(0.2, 0.3]$	6x	997.1	35.53	40.59	2.44	3.33	568.4	1.17	5.51	0.48	0.87	367.3	29.49	29.72	1.79	2.19	61.4	4.87	5.36	0.17	0.27
	10x	999.7	9.64	13.71	0.80	0.95	562.6	0.75	2.86	0.17	0.22	373.3	8.03	9.43	0.57	0.65	63.8	0.86	1.42	0.06	0.08
	30x	1000	0.15	0.34	0.01	0.03	564.7	0.03	0.08	0	0.01	372.2	0.11	0.22	0.01	0.02	63.1	0.01	0.04	0	0
$(0.3, 0.4]$	6x	997.0	45.56	49.99	2.66	3.74	423.2	1.09	4.57	0.46	0.96	450.1	40.19	40.11	1.94	2.32	123.7	4.28	5.31	0.26	0.46
	10x	999.7	11.68	15.91	0.89	1.01	426.6	0.69	2.80	0.16	0.21	451.8	10.11	11.28	0.64	0.69	121.3	0.88	1.83	0.09	0.11
	30x	1000	0.18	0.38	0.02	0.03	425.0	0.03	0.07	0.01	0.01	452.7	0.13	0.25	0.01	0.02	122.3	0.02	0.06	0	0
$(0.4, 0.5]$	6x	997.1	50.63	54.88	2.72	4.02	305.0	1.22	3.83	0.39	0.90	491.6	45.53	45.46	2.01	2.41	200.5	3.88	5.59	0.32	0.71
	10x	999.7	12.90	17.17	0.98	1.11	302.6	0.63	2.46	0.14	0.18	493.2	11.38	12.38	0.71	0.76	203.9	0.89	2.33	0.13	0.17
	30x	1000	0.19	0.41	0.02	0.03	302.7	0.03	0.08	0.01	0.01	494.3	0.14	0.26	0.01	0.02	203.0	0.02	0.07	0	0

P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively. N, N_0, N_1, and N_2 are the average numbers of individuals covered by at least one read. G is the true genotype.

We further examined the *phred* scores at loci having genotypes that are called differently by PhredEM and SeqEM. In Table 2.3, we displayed the average *phred* score associated with major and minor alleles at such loci, stratified by the underlying genotype (G) and genotypes called by PhredEM (G_{P}) and SeqEM (G_{S}). At loci with $(G_{\mathrm{P}}, G_{\mathrm{S}}) = (0, 1)$, regardless of the value of G, the major alleles tend to have high *phred* scores whereas the minor alleles tend to have low scores, explaining why PhredEM called these loci major allele homozygotes; the average *phred* scores for minor alleles are consistently lower under $G = 0$ than that under $G = 1$, because in the former case the minor alleles are all errors and in the latter case the minor alleles are a mixture of errors and true alleles. Similarly, for loci with $(G_{\mathrm{P}}, G_{\mathrm{S}}) = (2, 1)$, the major alleles tend to have low scores, which are even lower under $G = 2$ than those under $G = 1$. In other cases when PhredEM called heterozygous genotypes, we observe high average *phred* scores for both major and minor alleles. These patterns of *phred* scores confirm that PhredEM worked as expected. While the results in Table 2.3 pertain to common variants, those for rare variants are similar and are shown in Supplemental Table S2.4.

Table 2.3: Average *phred* scores associated with major (M) and minor (m) alleles at loci that are called differently by PhredEM and SeqEM in the simulation studies for common variants.

		$G = 0$				$G = 1$								$G = 2$			
		$(0,1)$		$(1,0)$		$(0,1)$		$(1,0)$		$(1,2)$		$(2,1)$		$(1,2)$		$(2,1)$	
MAF	Depth	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m
$(0.05, 0.1]$	6x	37.2	9.4	37.1	36.3	37.2	14.4	37.1	38.6	39.0	34.9	9.2	37.4	37.5	34.8	7.5	37.4
	10x	37.2	9.7	37.1	36.8	37.0	15.2	37.1	38.6	38.4	36.7	10.3	37.4	36.4	35.7	8.3	37.5
	30x	37.1	10.5	36.9	37.0	36.7	17.2	37.2	38.3	39.3	37.5	18.1	38.2	37.1	37.2	7.7	37.5
$(0.1, 0.2]$	6x	37.2	9.2	37.1	36.2	37.1	13.4	37.1	38.6	39.0	34.5	9.2	37.5	36.2	34.2	7.7	37.3
	10x	37.3	9.4	37.1	37.0	37.1	14.3	37.1	38.6	38.5	36.9	10.8	37.2	36.2	36.1	8.6	37.6
	30x	37.2	10.4	37.2	37.1	37.1	16.7	37.1	38.3	38.5	37.3	13.2	37.6	36.2	37.3	9.2	37.2
$(0.2, 0.3]$	6x	37.2	9.1	37.1	36.2	37.2	12.5	37.1	38.6	38.6	35.3	9.6	37.4	35.9	33.3	8.0	37.2
	10x	37.2	9.3	37.2	36.9	37.2	13.5	37.1	38.6	38.6	36.9	11.1	37.1	36.9	36.7	8.8	36.8
	30x	37.3	10.0	37.1	37.7	36.9	14.8	37.2	38.5	38.5	37.0	14.4	37.4	37.2	36.9	9.6	37.1
$(0.3, 0.4]$	6x	37.4	8.8	37.3	36.6	37.2	12.0	37.1	38.6	38.8	36.1	10.3	37.2	35.2	33.3	8.3	37.0
	10x	37.0	9.2	37.1	36.7	37.1	13.2	37.1	38.6	38.5	37.4	11.7	37.2	36.6	37.0	8.9	37.2
	30x	37.2	10.2	36.8	37.5	37.4	14.8	37.1	38.4	38.4	37.0	14.9	36.6	37.3	37.3	10.2	36.8
$(0.4, 0.5]$	6x	37.1	8.5	36.3	36.4	37.2	11.4	37.0	38.6	38.7	36.9	10.9	37.2	35.8	35.8	8.5	37.2
	10x	37.3	9.1	37.1	36.7	37.1	12.7	37.1	38.6	38.6	37.1	12.3	37.1	36.3	37.1	9.0	37.1
	30x	37.2	10.3	37.2	37.1	36.9	14.6	37.0	38.5	38.3	37.3	15.1	37.7	37.5	37.3	10.3	37.2

G is the true genotype. $(G_{\mathrm{P}}, G_{\mathrm{S}}) = (0,1)$, $(1,0)$, et al. represent loci that are called to be G_{P} and G_{S} by PhredEM and SeqEM, respectively.

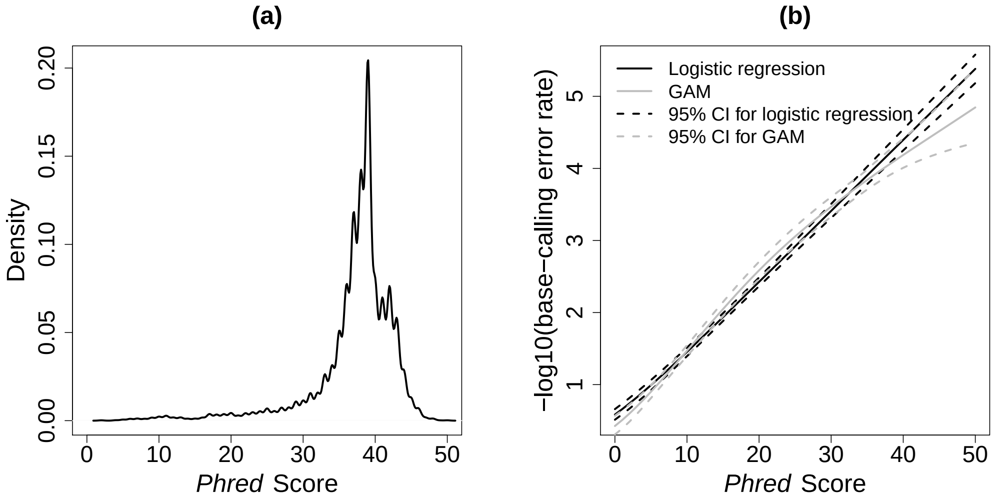

Figure 2.2: UK10K SCOOP data. (a) Distribution of *phred* scores. (b) Logistic regression model and generalized additive model (GAM) fit to the sequencing data at loci that were identified as monomorphic.

2.3 Application to the UK10K SCOOP Data

To confirm that the results from our simulations hold when analyzing real sequencing data, we analyzed data from the Severe Childhood Onset Obesity Project (SCOOP) cohort sequenced as part of the UK10K project. The sequenced SCOOP cohort consists of 784 UK Caucasian patients with severe early onset obesity, who were whole-exome sequenced using the Illumina HiSeq 2000 platform with an average depth of ∼60x. We first used SAMtools to generate pileup files from BAM files, filtering out reads that are PCR duplicates, have mapping score ≤ 30, or have improperly mapped mates. From the pileup files, we extracted read count data and *phred* scores. The distribution of the *phred* scores is shown in Figure 2.2(a).

Using the SCOOP sequencing data, we checked the fit of the logistic regression model in (2.1). First, we applied our screening algorithm to identify loci that were estimated to be monomorphic (i.e., $\widehat{\pi} = 0$). At such loci, we could reliably treat all minor allele reads as errors. Assigning $Y = 1$ and 0 for minor allele reads and major allele reads, respectively, we can determine the relationship between $\Pr(Y =$

1) and the corresponding *phred* scores Q. To create a subset of such data that is computationally manageable, we randomly selected 1,000 monomorphic loci from each of the 22 chromosomes and randomly picked one individual from each locus, forming a dataset of 22,000 (Y, Q) pairs. Then, we fit the logistic regression model in [2.1] and, as a gold standard, fit a smooth spline function of *phred* scores using the generalized additive model (GAM) (Wood, 2006). Figure 2.2(b) shows the fitted curves and pointwise 95% confidence intervals from the two models. The logistic regression fit always fell within the 95% confidence region of the GAM. Thus, we conclude that over the range of *phred* scores found in real data, the logistic model adequately describes the relationship between *phred* scores and base-calling error rates well.

To facilitate the evaluation of PhredEM and especially the comparison with SeqEM, we first selected a set of genotypes that can serve as gold standard. Specifically, we downloaded from the UK10K website the VCF files for the SCOOP cohort, which contained genotypes called by SAMtools and filtered by GATK. In addition, we excluded a variant if its average depth across samples is less than 20. We excluded a genotype whose genotype likelihood (on the *phred* scale) was ≤ 20 (i.e., nominal genotyping error rate ≥ 0.01) and excluded a variant completely if it has more than 20% of genotypes with likelihood ≤ 20. These exclusion criteria ensured that all selected genotypes were called with particularly high quality. We thus refer to these genotypes as 'true' genotypes. After applying the exclusion criteria, there remain 416,402 loci in the entire exome. Since the loci with true genotypes were selected towards having high read depth, both PhredEM and SeqEM would perform well if applied to the original data. To create sequencing data with low or median depth, we then subsampled the observed reads with equal probability.

We based the estimation of hyper-parameters κ and ϕ on 100k random loci that were reliably estimated to be monomorphic (i.e., with coverage $>$ 60x and the MLE of the MAF π is zero); these 100k loci mimic real sequencing data in which the vast majority of loci are monomorphic whereas the 416,402 loci extracted from the VCF files are mostly polymorphic. We then applied PhredEM and SeqEM to call genotypes assuming HWE at first and, if the estimated MAF was over 5%, we re-called genotypes assuming HWD. The computation time for PhredEM to call the subsampled UK10K data depends on the average depth. For example, it took ∼5 h on an Intel Xeon E5-2660 machine with 2.60 GHz and 6.4 GB memory to call genotypes at the 416,402 loci in the 6x dataset.

The numbers of mis-called genotypes, averaged over all variants on chromosomes 1–22 and stratified by MAF ranges, are displayed in Table 2.4. For rare variants (MAF ≤ 0.05), the pattern in the number of mis-called genotypes by PhredEM and SeqEM agreed well with the results in the simulation section, with PhredEM generally producing more accurate genotype calls. The biggest difference occurred when the variants were relatively rare, i.e., MAF $\in (0.001, 0.01]$; when the average read depth was ∼6x, PhredEM generated an average of 1.9 more correct genotypes out of 757 individuals than SeqEM for loci with MAFs in this range. For common variants (MAF > 0.05), the differences between the two methods were smaller, possibly because *phred* scores at heterozygous loci are not informative; this also explains the increase in genotype-calling error rates with increasing MAF found in Table 2.4. As seen in the simulation results, applying Beagle after PhredEM improved the performance of PhredEM alone for all variants except for the very rare ones (e.g., MAF $\in (0, 0.001]$). The *phred* scores at loci with differently called genotypes by PhredEM and SeqEM are summarized in Supplemental Table S2.6. These results exhibited the same patterns seen in the simulated data. The mis-call rate at monomorphic loci (Figure 2.1 [b]) also show the same pattern seen in the simulated data (Figure 2.1 [a]).

Table 2.4: Average number of mis-called genotypes per variant in analysis of the UK10K SCOOP data (subsampled to achieve different depths).

		Overall					Stratified														
							$G = 0$					$G = 1$					$G = 2$				
MAF	Depth	N	P	S	PB	SB	N_0	P	S	PB	SB	N_1	P	S	PB	SB	N_2	P	S	PB	SB
$(0, 0.001]$	6x	762.4	0.30	0.98	0.33	1.23	761.4	0.11	0.74	0.13	0.99	1.0	0.19	0.24	0.20	0.24	0	0	0	0	0
	10x	778.7	0.26	0.94	0.30	1.02	777.7	0.08	0.73	0.12	0.82	1.0	0.18	0.21	0.18	0.20	0	0	0	0	0
	30x	783.6	0.16	0.77	0.21	0.91	782.6	0.05	0.63	0.11	0.75	1.0	0.11	0.14	0.10	0.16	0	0	0	0	0
$(0.001, 0.01]$	6x	757.0	1.77	3.68	1.65	3.43	752.4	0.32	2.21	0.53	2.10	4.5	1.36	1.38	1.05	1.25	0.1	0.09	0.09	0.07	0.08
	10x	776.1	1.64	3.32	1.43	2.95	771.4	0.30	1.92	0.43	1.91	4.6	1.26	1.32	0.93	0.97	0.1	0.08	0.08	0.07	0.07
	30x	782.2	1.02	2.25	0.81	1.90	777.5	0.27	1.06	0.35	1.31	4.6	0.69	1.13	0.42	0.54	0.1	0.06	0.06	0.04	0.05
$(0.01, 0.05]$	6x	751.1	10.45	11.52	7.30	8.84	713.1	1.09	3.21	0.85	1.68	37.3	8.87	7.80	6.22	6.91	0.7	0.49	0.51	0.23	0.25
	10x	772.3	8.22	9.17	6.35	7.67	733.4	0.95	2.84	0.66	1.39	38.2	6.88	5.91	5.50	6.08	0.7	0.39	0.42	0.19	0.20
	30x	779.1	1.41	2.33	0.89	1.37	739.9	0.44	1.03	0.28	0.63	38.5	0.78	1.10	0.48	0.60	0.7	0.19	0.20	0.13	0.14
$(0.05, 0.1]$	6x	749.5	19.52	20.28	11.27	12.25	646.8	1.35	2.14	1.58	1.94	98.4	15.87	15.78	8.89	9.48	4.3	2.30	2.36	0.80	0.83
	10x	772.4	11.99	12.76	7.54	8.13	666.7	1.20	1.73	1.11	1.40	101.3	9.33	9.48	5.89	6.12	4.4	1.46	1.55	0.54	0.61
	30x	779.8	2.34	2.52	1.15	1.47	673.3	0.68	0.72	0.44	0.63	102.0	1.29	1.40	0.43	0.51	4.5	0.37	0.40	0.28	0.33
$(0.1, 0.2]$	6x	748.5	38.06	38.54	13.47	13.90	546.4	2.12	2.47	1.74	1.95	184.9	28.69	28.81	9.82	9.94	17.2	7.25	7.26	1.91	2.01
	10x	772.3	21.36	21.65	7.95	8.30	563.9	1.91	2.28	1.33	1.52	190.6	15.56	15.44	5.41	5.56	17.8	3.89	3.93	1.21	1.22
	30x	779.4	3.43	3.55	1.10	1.13	569.4	0.84	0.86	0.46	0.50	191.9	1.77	1.87	0.28	0.26	18.1	0.82	0.82	0.36	0.37
$(0.2, 0.3]$	6x	747.3	62.54	62.93	14.70	15.21	423.7	2.76	3.15	1.75	1.79	276.6	46.20	46.15	10.69	11.05	47.0	13.58	13.63	2.26	2.37
	10x	771.8	33.96	34.41	8.51	8.72	437.7	2.58	2.85	1.37	1.49	285.5	24.84	24.94	5.78	5.81	48.6	6.54	6.62	1.36	1.42
	30x	779.6	4.70	4.86	1.22	1.31	442.3	1.17	1.22	0.45	0.50	288.0	2.48	2.57	0.43	0.45	49.3	1.05	1.07	0.34	0.36
$(0.3, 0.4]$	6x	748.3	81.03	81.28	15.37	15.91	317.9	2.99	3.30	1.94	2.03	338.2	62.50	62.37	11.02	11.39	92.2	15.54	15.61	2.41	2.49
	10x	772.1	42.04	42.40	9.07	9.21	328.1	2.74	3.02	1.40	1.48	349.0	32.34	32.33	6.15	6.20	95.0	6.96	7.05	1.52	1.53
	30x	780.5	5.43	5.51	1.55	1.63	331.9	1.23	1.27	0.48	0.51	352.1	3.03	3.04	0.65	0.70	96.5	1.17	1.20	0.42	0.42
$(0.4, 0.5]$	6x	747.3	95.93	96.22	15.82	16.40	221.1	4.41	4.58	2.16	2.21	378.7	75.66	75.69	11.19	11.61	147.5	15.86	15.95	2.47	2.58
	10x	771.4	49.78	50.13	9.47	9.68	228.3	3.39	3.56	1.53	1.58	390.5	39.19	39.35	6.32	6.39	152.6	7.20	7.22	1.62	1.71
	30x	778.7	6.88	7.15	1.82	1.93	230.8	1.34	1.39	0.51	0.60	393.4	4.32	4.46	0.88	0.87	154.5	1.22	1.30	0.43	0.46

G is the true genotype. N, N_0, N_1, and N_2 are the average numbers of individuals covered by at least one read. P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively.

To gain more insights into the mechanisms of PhredEM and SeqEM, we listed in Table 2.5 the raw data at eight loci (from the subsampled dataset at 6x) that were called differently by PhredEM and SeqEM. Generally, base calls with low *phred* score are error-prone, and PhredEM treats these unreliable calls as likely errors when calling the genotype. By contrast, SeqEM depends heavily on the proportion of minor allele reads among the total reads and ignores the quality measure of each allele. For example, at Locus 1, the six major alleles were of high quality while the two minor alleles were likely to be errors. In this case, PhredEM distinguishes between alleles of different qualities and produced the correct genotype but SeqEM, which cannot account for low quality alleles, calls the incorrect genotype.

Table 2.5: Eight example loci in the UK10K SCOOP data (subsampled to 6x).

	Reads		*Phred* scores		Genotype		
Locus	M	m	M	m	True	P	S
1	6	2	21 36 37 38 39 42	9 16	0	0	1
2	6	1	18 18 27 36 39 40	33	0	1	0
3	4	1	20 34 34 36	15	1	0	1
4	5	1	25 32 32 34 39	37	1	1	0
5	1	5	35	20 25 38 40 40	1	1	2
6	1	5	14	33 37 38 38 40	1	2	1
7	1	4	32	30 34 37 39	2	1	2
8	2	5	11 17	30 34 35 36 39	2	2	1

M and m represent major and minor alleles, respectively. True is the true genotype. P and S represent the called genotypes by PhredEM and SeqEM, respectively.

2.4 Application to the 1000 Genomes CEU Data

To compare PhredEM to GATK, we considered data from the CEU samples in the 1000 Genomes project. It is hard to make this comparison using simulated data, since it is difficult to construct BAM files for the simulated data, and because the 100KB region we simulated is to short to train the BQSR model used in GATK. It is also hard to make this comparison using the UK10K SCOOP data, as BAM files for the subsampled data are not easily available. In the CEU cohort, 99 unrelated individuals were whole-genome sequenced with an average depth of ∼7.3x. We adopted the same filters for the reads as in the analysis of UK10K SCOOP data. As the 99 CEU samples have also been genotyped on the Illumina Omni 2.5 array, we treated these array genotypes as the gold standard. We excluded array SNPs at which ≥5% of the samples have missing array genotypes or are not covered by any reads. We also removed 11,119 array SNPs where the genotypes called using sequencing data for all three methods (SeqEM, PhredEM and GATK) indicated a MAF that differed by more than 0.2 from the MAF based on the array genotypes. After these exclusions, there were 1,842,422 array SNPs available for comparison.

We estimated the hyper-parameters for PhredEM based on a random subset of 100k array SNPs that are called as monomorphic using the genotype array in the 99 CEU cohort. In addition to PhredEM and SeqEM, we also applied GATK, using the base quality score recalibration step implemented in BQSR (GATK version 3.6) and a genotype calling step by UnifiedGenotyper with default options. It took 3.4 days for BQSR and 1.3 days for UnifiedGenotyper to run; in contrast, it took a total of 1.7 days for PhredEM to call the same set of genotypes.

PhredEM performed better than SeqEM and GATK in general. Figure 2.1(c) shows that, at monomorphic loci (i.e., no polymorphism in the array genotypes of the 99 samples), PhredEM has the smallest mis-call rate with or without LD refinement whereas GATK has the highest mis-call rate. Table 2.6 displays the numbers of mis-called genotypes at polymorphic loci, stratified by the ‘true’ MAFs (i.e., based on array genotypes). In most strata, the numbers for PhredEM are smaller than that for GATK, with or without Beagle. The results stratified on the estimated MAF by each method are presented in Table S2.5, which shows similar patterns. All results consistently indicate that GATK tends to call too many heterozygotes at rare variants and monomorphic loci. Table S2.7 compares the sensitivity (i.e., the probability of calling a minor allele given a minor allele is truly present) and specificity (probability of calling the major allele given the major allele is truly present) for the methods we consider in Table S2.7. We find that PhredEM with the LD refinement has the highest specificity (although the differences are tiny, they are significant and when amplified to the genome-wide scale can represent a meaningful difference). PhredEM with LD refinement has the best sensitivity at very low MAF by a considerable amount (0.828, compared to 0.748 for GATK with LD refinement); for higher MAFs, GATK with LD refinement outperforms PhredEM with LD refinement by smaller amounts (e.g., 0.954 for PhredEM with LD vs. 0.958 for GATK with LD). When evaluating the importance of the differences reported in Table S2.7, it is worth noting that the number of truly polymorphic alleles with low MAF is much smaller than the number of monomorphic alleles, so that a small difference in specificity results in more mis-calls than a larger difference in sensitivity. This explains how GATK with LD can have a higher sensitivity but a lower accuracy as reported in Table 2.6.

Table 2.6: Average number of mis-called genotypes per variant in the analysis of the 1000 Genomes CEU data.

MAF	N	P	S	GATK	PB	SB	GATK-B
$(0, 0.01]$	98.07	0.185	0.203	0.808	0.197	0.220	0.701
$(0.01, 0.05]$	98.08	0.546	0.562	0.716	0.285	0.306	0.326
$(0.05, 0.1]$	98.04	1.330	1.334	1.541	0.451	0.482	0.445
$(0.1, 0.2]$	98.02	2.553	2.519	2.781	0.724	0.749	0.685
$(0.2, 0.3]$	98.03	3.716	3.889	3.919	0.727	0.794	0.742
$(0.3, 0.4]$	98.02	4.648	4.827	4.733	0.835	0.923	0.865
$(0.4, 0.5]$	98.01	5.189	5.380	5.118	0.886	0.979	0.915

MAF is the minor allele frequency observed in the array genotype data. P and S represent PhredEM and SeqEM, respectively. PB, SB, and GATK-B represent PhredEM, SeqEM, and GATK followed by Beagle. N is the average number of individuals covered by at least one read.

2.5 Discussion

In this project, we have developed a *phred*-score-informed genotype-calling approach for NGS studies, called PhredEM. We also proposed a simple and computationally efficient screening algorithm to identify loci that would be called as monomorphic. PhredEM improves the accuracy of genotype-calling by estimating base-calling errors from both read data and *phred* scores, and by using all sequencing reads available without setting a *phred*-score-based quality threshold. PhredEM is closely related to the SeqEM approach, which can be viewed as a special case of PhredEM. We showed that the logistic model relating *phred* score to base-calling error rate used in PhredEM fits real sequencing data well. The software program implementing PhredEM, also called PhredEM, is freely available at `http://web1.sph.emory.edu/users/yhu30/software.html`. The webpage also contains a link to utility programs that process raw BAM files for use as inputs to PhredEM.

In our logistic regression model (2.1), the *phred* score is the only predictor for the base-calling error. Other important predictors for base-calling quality could also be included. One interesting factor is the position in the read (Brockman et al., 2008), although it is unclear whether this has an independent effect once the *phred* score is accounted for. We did not consider the mapping score as a possible covariate because there is little variability in mapping scores (Li et al., 2008) (see Supplemental Figure S2.2). However, we recommend that PhredEM should be applied after excluding alignments with mapping scores less than 30.

Our approach is similar in spirit to GATK with BQSR because we allow the relationship between error and *phred* score to be determined by fit to the data, but our approach is more accurate and computationally more tractable. Because we allow a separate set of error parameters at each locus, we automatically account for any covariates that are locus-dependent such as the actual alleles at each locus. We could also consider adding other predictors of error that are included in BQSR that vary across reads.

We recommend using PhredEM with the HWE assumption first, because most loci have low MAFs and HWE has a minimal effect for them. If the estimated MAF is greater than 5%, a second pass of PhredEM could easily be made using the model assuming HWD, which is more robust. Our numerical studies (not shown) suggest that at medium or high read depth ($\geq$10x), the estimated genotype frequencies based on the calls from PhredEM converged rapidly to their true values with increasing sample size even when assuming HWD.

PhredEM is based on several simplifying assumptions. First, the sample should consist of independent, unrelated individuals; this is essential to the likelihood in expression (2.3). A version of PhredEM could be constructed for trio data by modeling the joint genotypes of parents and offspring, for example, using the conditional-on-parental genotypes (CPG) approach of Schaid and Sommer (1993). We also assume that errors are symmetric, i.e. that the probability of a read for the major allele being mis-called as the minor allele is the same as the probability of the minor allele being mis-called as the major allele. Further, PhredEM assumes that all variants are biallelic. The biallelic assumption is reasonable because only a small fraction of SNPs have been verified to carry three or more alleles (Hodgkinson and Eyre-Walker, 2010). In analyzing the UK10K and 1000 Genomes data, we deleted in advance all calls for bases that differed from the two most frequent bases at every locus.

LD information is helpful in identifying monomorphic loci and calling genotypes for both rare and common variants. Therefore, we recommend always using Beagle in conjunction with PhredEM when calling genotypes for NGS data.

In summary, we developed PhredEM, an improved genotype caller which reduces the genotype-calling errors for NGS data. We also proposed a simple and computationally inexpensive algorithm for screening out loci that are estimated to be monomorphic. We showed in simulations that the proposed approach generates fewer incorrect calls than SeqEM regardless of the average read depth and sample size. Using the UK10K and 1000 Genomes sequencing data, we demonstrated the capability of PhredEM to improve the genotype-calling accuracy over SeqEM and GATK in real sequencing data.

2.6 Appendix

2.6.1 EM algorithm

In the EM algorithm, G_i $(i = 1, \ldots, n)$ is treated as missing. The complete-data log-likelihood has the form

$$l_c(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{g=0}^{2} I(G_i = g) \big\{ \log P_{\boldsymbol{\theta}}(R_i | g, T_i, \boldsymbol{Q}_i) + \log P_{\boldsymbol{\pi}}(g) \big\}.$$

Let $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\pi}^{(k)}$ be the parameter values after the kth iteration. In the E-step of the $(k+1)$th iteration, we evaluate $E\{I(G_i = g) | R_i, T_i, \boldsymbol{Q}_i\}$ for $g = 0, 1, 2$, which can be written as

$$\omega_{ig}^{(k)} \equiv \frac{P_{\boldsymbol{\theta}^{(k)}}(R_i | g, T_i, \boldsymbol{Q}_i) P_{\boldsymbol{\pi}^{(k)}}(g)}{\sum_{g'=0}^{2} P_{\boldsymbol{\theta}^{(k)}}(R_i | g', T_i, \boldsymbol{Q}_i) P_{\boldsymbol{\pi}^{(k)}}(g')}.$$

In the M-step, we maximize $l_c(\boldsymbol{\theta}, \boldsymbol{\pi})$ with $I(G_i = g)$ replaced by $\omega_{ig}^{(k)}$. Specifically, under HWE we update π by a closed form $\pi^{(k+1)} = (2n)^{-1} \sum_{i=1}^{n} (2\omega_{i2}^{(k)} + \omega_{i1}^{(k)})$, or under HWD we update π by the same $\pi^{(k+1)}$ and update f by $f^{(k+1)} = 1 - \sum_{i=1}^{n} \omega_{i1}^{(k)} / \big\{ 2n\pi^{(k+1)} (1 - \pi^{(k+1)}) \big\}$. We use a one-step Newton-Raphson iteration to update $\boldsymbol{\theta}$. We iterate between the E-step and M-step until the changes in the parameter estimates are negligible.

2.6.2 Proof of concavity of $pl^*(\pi)$

First, we prove that, for fixed $\boldsymbol{\theta}$, the function $h(\pi) = \log\big\{\sum_{g=0,1,2} P_{\boldsymbol{\theta}}(R|g,T,\boldsymbol{Q})P_{\boldsymbol{\pi}}(g)\big\}$ is concave. Under HWE, we write $h(\pi) = \log\big\{a\pi^2 + b(1-\pi)^2 + 2c\pi(1-\pi)\big\}$, where $a = P_{\boldsymbol{\theta}}(R|G=2,T,\boldsymbol{Q})$, $b = P_{\boldsymbol{\theta}}(R|G=0,T,\boldsymbol{Q})$, and $c = (0.5)^T$. The second derivative of $h(\pi)$ is

$$h''(\pi) = -\frac{2\big\{(a+b-2c)\pi + (c-b)\big\}^2 + 2(c^2-ab)}{\big\{a\pi^2 + b(1-\pi)^2 + 2c\pi(1-\pi)\big\}^2}.$$

Because $ab = \prod_{k=1}^T \epsilon_k(\boldsymbol{\theta})\big\{1-\epsilon_k(\boldsymbol{\theta})\big\} \le (0.25)^T = c^2$, we obtain $h''(\pi) \le 0$ and thus $h(\pi)$ is a concave function of π.

Because the sum of concave functions is still concave, $\log L_o(\boldsymbol{\theta},\pi)$ is concave in π for fixed $\boldsymbol{\theta}$. It follows that $\log L_o^*(\boldsymbol{\theta},\pi) = \log\Gamma(-\beta_1;\widehat{\kappa},\widehat{\phi}) + \log L_o(\boldsymbol{\theta},\pi)$ is also concave in π for fixed $\boldsymbol{\theta}$. Because the pointwise supremum over $\boldsymbol{\theta}$ preserves the concavity [Boyd and Vandenberghe, 2004], $pl^*(\pi)$ is concave.

2.7 Supplemental Materials

Table S2.1: Average number of mis-called genotypes per variant for rare variants in the simulation studies when $n = 200$.

		Overall					Stratified									
							$G=0$					$G=1$				
MAC	Depth	N	P	S	PB	SB	N_0	P	S	PB	SB	N_1	P	S	PB	SB
1	6x	199.4	0.196	0.276	0.211	0.318	198.4	0.019	0.063	0.029	0.096	1	0.177	0.213	0.182	0.222
	10x	199.9	0.065	0.096	0.077	0.125	198.9	0.010	0.020	0.019	0.036	1	0.055	0.076	0.058	0.089
	30x	200	0.001	0.003	0.002	0.004	199.0	0	0.001	0.001	0.001	1	0.001	0.002	0.001	0.003
$[2,4]$	6x	199.4	0.358	0.587	0.294	0.508	196.8	0.036	0.137	0.058	0.172	2.6	0.318	0.446	0.235	0.334
	10x	199.9	0.113	0.192	0.095	0.159	197.3	0.017	0.040	0.032	0.055	2.6	0.095	0.151	0.062	0.103
	30x	200	0.002	0.005	0.002	0.005	197.4	0	0.002	0.001	0.002	2.6	0.002	0.003	0.001	0.003
$[5,20]$	6x	199.4	1.007	1.663	0.428	0.784	189.2	0.084	0.459	0.107	0.308	10.1	0.845	1.127	0.316	0.467
	10x	199.9	0.340	0.601	0.143	0.219	189.5	0.047	0.148	0.053	0.090	10.2	0.267	0.425	0.088	0.127
	30x	200	0.007	0.016	0.003	0.006	189.4	0.001	0.003	0.001	0.002	10.4	0.005	0.012	0.002	0.003

P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively. N, N_0, and N_1 are the average numbers of individuals covered by at least one read. G is the true genotype; the case $G = 2$ is omitted as it is barely seen for rare variants. MACs of 1, 4, and 20 correspond to MAFs of 0.0025, 0.01, and 0.05, respectively, given the sample size of 200.

Table S2.2: Average number of mis-called genotypes per variant for common variants in the simulation studies when $n = 200$.

		Overall					Stratified														
							$G = 0$					$G = 1$					$G = 2$				
MAF	Depth	N	P	S	PB	SB	N_0	P	S	PB	SB	N_1	P	S	PB	SB	N_2	P	S	PB	SB
$(0.05, 0.1]$	6x	199.4	2.63	3.60	0.61	1.09	170.8	0.16	0.87	0.14	0.40	27.5	2.18	2.43	0.46	0.67	1.1	0.29	0.30	0.01	0.02
	10x	199.9	0.80	1.31	0.22	0.31	171.7	0.09	0.35	0.07	0.11	27.1	0.64	0.88	0.14	0.19	1.1	0.07	0.08	0.01	0.01
	30x	200	0.01	0.03	0	0.01	171.5	0	0.01	0	0	27.4	0.01	0.02	0	0.01	1.1	0	0	0	0
$(0.1, 0.2]$	6x	199.4	4.97	5.87	0.85	1.34	144.6	0.21	1.03	0.15	0.46	50.4	3.99	4.03	0.66	0.82	4.4	0.77	0.81	0.04	0.06
	10x	199.9	1.40	2.07	0.27	0.37	145.1	0.11	0.50	0.07	0.12	50.4	1.13	1.36	0.19	0.23	4.4	0.16	0.21	0.01	0.02
	30x	200	0.03	0.05	0.01	0.01	145.6	0	0.01	0	0	50.0	0.02	0.03	0.01	0.01	4.4	0	0.01	0	0
$(0.2, 0.3]$	6x	199.4	7.58	8.42	1.28	1.58	112.6	0.23	0.97	0.16	0.47	74.5	6.30	6.34	1.01	0.99	12.3	1.05	1.11	0.11	0.12
	10x	199.9	2.07	2.82	0.38	0.48	112.8	0.12	0.58	0.08	0.15	74.7	1.74	1.93	0.26	0.28	12.4	0.21	0.31	0.04	0.05
	30x	200	0.03	0.06	0.01	0.01	112.6	0.01	0.01	0	0	75.1	0.02	0.04	0.01	0.01	12.3	0	0.01	0	0
$(0.3, 0.4]$	6x	199.4	9.49	10.17	1.19	1.78	84.5	0.23	0.81	0.16	0.44	90.3	8.31	8.23	0.91	1.13	24.6	0.95	1.13	0.12	0.21
	10x	199.9	2.61	3.32	0.45	0.57	84.1	0.12	0.54	0.08	0.16	91.6	2.30	2.41	0.32	0.33	24.2	0.19	0.37	0.05	0.08
	30x	200	0.04	0.08	0.01	0.01	85.5	0.01	0.02	0	0	89.6	0.03	0.05	0.01	0.01	24.9	0	0.01	0	0
$(0.4, 0.5]$	6x	199.4	10.73	11.25	1.40	1.83	60.0	0.33	0.77	0.14	0.37	98.9	9.58	9.45	1.05	1.16	40.5	0.82	1.03	0.21	0.30
	10x	199.9	2.90	3.53	0.52	0.62	59.7	0.12	0.48	0.07	0.15	99.6	2.58	2.61	0.36	0.36	40.6	0.20	0.44	0.09	0.11
	30x	200	0.04	0.08	0.01	0.01	60.6	0.01	0.01	0	0	98.2	0.03	0.06	0.01	0.01	41.2	0	0.01	0	0

P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively. N, N_0, N_1, and N_2 are the average numbers of individuals covered by at least one read. G is the true genotype.

Table S2.3: Average number of mis-called genotypes per locus in the simulation studies when $n = 1,000$.

	Depth	P	S	PB	SB
MAC = 0	6x	0.001	0.002	0.001	0.002
	10x	0	0.001	0	0.001
	30x	0	0	0	0
1	6x	0.842	0.916	0.841	0.921
	10x	0.505	0.772	0.484	0.768
	30x	0.028	0.081	0.024	0.072
$[2, 10]$	6x	0.971	1.191	0.927	1.329
	10x	0.235	0.478	0.228	0.532
	30x	0.004	0.008	0.003	0.009
$[11, 20]$	6x	1.625	3.237	0.799	1.305
	10x	0.549	1.018	0.215	0.375
	30x	0.010	0.025	0.004	0.008
$[21, 100]$	6x	4.441	8.078	1.425	2.223
	10x	1.464	2.761	0.361	0.599
	30x	0.033	0.068	0.009	0.015
MAF $\in (0.05, 0.1]$	6x	12.077	17.401	1.981	2.852
	10x	3.645	6.328	0.491	0.701
	30x	0.074	0.148	0.012	0.018
$(0.1, 0.2]$	6x	23.552	28.964	2.148	2.957
	10x	6.399	9.960	0.637	0.798
	30x	0.107	0.236	0.013	0.022
$(0.2, 0.3]$	6x	36.315	41.182	2.441	3.326
	10x	9.660	13.713	0.806	0.952
	30x	0.156	0.338	0.020	0.027
$(0.3, 0.4]$	6x	46.074	50.301	2.656	3.738
	10x	11.698	15.930	0.892	1.005
	30x	0.186	0.378	0.020	0.029
$(0.4, 0.5]$	6x	50.899	54.762	2.732	4.031
	10x	12.901	17.172	0.985	1.104
	30x	0.187	0.405	0.021	0.027

MAC and MAF are the number of minor alleles and minor allele frequency based on called genotypes of each method. P, S, PB, and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, SeqEM followed by Beagle, respectively.

Table S2.4: Average *phred* scores associated with called major (M) and minor (m) alleles at loci that are called differently by PhredEM and SeqEM in the simulation studies for rare variants when $n = 1,000$.

		$G = 0$				$G = 1$			
		$(0, 1)$		$(1, 0)$		$(0, 1)$		$(1, 0)$	
MAC	Depth	M	m	M	m	M	m	M	m
1	6x	37.1	11.2	37.7	38.9	37.5	21.5	37.0	38.6
	10x	38.4	10.6	37.4	35.0	37.6	21.7	36.5	37.5
	30x	37.2	11.8	37.1	35.9	38.5	22.2	37.3	37.8
$[2, 10]$	6x	37.1	10.6	37.1	38.0	37.4	20.1	37.1	39.0
	10x	37.6	11.0	37.3	38.4	36.9	20.9	36.8	38.5
	30x	37.4	11.6	36.8	36.3	37.6	24.2	37.1	37.9
$[11, 20]$	6x	37.1	9.9	37.1	37.2	36.9	17.7	37.1	38.9
	10x	36.9	10.7	37.8	37.9	36.8	18.2	37.1	38.8
	30x	36.5	11.2	36.8	35.5	37.6	21.6	36.5	38.1
$[21, 100]$	6x	37.2	9.6	37.1	36.8	37.2	15.8	37.1	38.7
	10x	37.5	10.0	37.2	37.5	36.8	16.6	37.4	39.2
	30x	37.2	10.8	37.3	36.8	37.6	19.4	37.0	38.3

G is the true genotype. $(G_P, G_S) = (0, 1)$, $(1, 0)$, etc. represents loci that are called to be G_P and G_S by PhredEM and SeqEM, respectively. MACs of 1, 10, 20, and 100 correspond to MAFs of 0.0005, 0.005, 0.01, and 0.05, respectively, given the sample size of 1,000.

Table S2.5: Average number of mis-called genotypes per locus in the analysis of the 1000 Genomes CEU data.

	P	S	GATK	PB	SB	GATK-B
MAC = 0	0.304	0.323	0.505	0.297	0.339	0.498
MAF $\in (0, 0.01]$	0.455	0.586	0.163	0.312	0.452	0.157
$(0.01, 0.05]$	0.649	0.724	0.684	0.298	0.425	0.376
$(0.05, 0.1]$	1.384	1.396	2.042	0.407	0.469	0.963
$(0.1, 0.2]$	2.501	2.519	3.516	0.550	0.612	1.338
$(0.2, 0.3]$	3.752	3.778	4.446	0.712	0.784	1.228
$(0.3, 0.4]$	4.611	4.621	4.814	0.809	0.904	0.928
$(0.4, 0.5]$	5.071	5.068	5.095	0.862	0.962	0.883

MAC and MAF are the number of minor alleles and minor allele frequency based on called genotypes of each method. P, S, PB, SB, and GATK-B represent PhredEM, SeqEM, PhredEM followed by Beagle, SeqEM followed by Beagle, and GATK followed by Beagle, respectively.

Table S2.6: Average *phred* scores associated with called major (M) and minor (m) alleles at loci that are called differently by PhredEM and SeqEM in analysis of the UK10K SCOOP data (subsampled to achieve different depths).

		$G = 0$				$G = 1$								$G = 2$			
		$(0, 1)$		$(1, 0)$		$(0, 1)$		$(1, 0)$		$(1, 2)$		$(2, 1)$		$(1, 2)$		$(2, 1)$	
MAF	Depth	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m
$(0, 0.001]$	6x	36.2	13.0	38.2	38.8	34.4	16.9	38.0	39.2	37.6	38.7	NA	NA	NA	NA	NA	NA
	10x	32.4	13.2	37.3	38.2	34.8	16.5	37.8	39.1	38.4	21.3	NA	NA	38.5	37.2	38.7	39.4
	30x	32.1	13.3	35.6	34.2	31.8	16.2	36.9	37.5	38.4	18.5	NA	NA	NA	NA	NA	NA
$(0.001, 0.01]$	6x	36.7	13.7	37.6	36.9	38.9	16.6	39.1	38.7	38.8	24.2	NA	NA	37.9	37.5	38.3	38.6
	10x	32.0	13.7	36.6	35.9	34.0	16.3	37.6	38.4	38.1	19.2	32.7	37.6	38.4	37.0	29.3	36.3
	30x	30.9	13.8	34.7	29.8	32.1	16.3	35.7	34.8	38.0	16.2	23.7	30.4	37.8	32.1	20.4	36.4
$(0.01, 0.05]$	6x	35.0	13.6	37.2	36.7	38.0	16.3	37.8	36.4	38.2	20.7	37.5	37.6	39.8	37.4	33.2	37.1
	10x	32.1	13.6	36.7	36.5	33.9	15.9	37.0	35.5	38.1	29.7	35.7	36.0	39.2	36.9	28.4	38.0
	30x	30.5	13.7	35.2	31.1	32.0	15.7	35.0	30.5	37.2	19.9	22.4	29.7	39.6	37.2	20.6	35.4
$(0.05, 0.1]$	6x	36.6	11.0	37.3	39.4	38.0	16.2	37.5	36.3	38.8	32.5	36.3	36.9	38.5	37.0	32.4	36.7
	10x	35.1	10.9	37.2	38.4	35.2	15.6	36.9	36.1	38.7	34.6	34.4	35.4	39.1	37.1	24.4	36.2
	30x	32.4	12.3	37.1	38.7	33.4	16.2	35.8	33.9	36.2	26.7	32.7	32.1	39.5	37.5	20.9	34.3
$(0.1, 0.2]$	6x	36.9	10.1	37.6	39.9	39.2	16.1	37.2	38.7	38.5	36.5	36.2	36.9	38.8	36.7	29.3	36.2
	10x	36.4	10.2	37.4	39.4	37.1	16.6	36.9	39.0	38.2	36.6	31.4	36.3	39.5	37.1	25.6	35.9
	30x	33.7	11.5	37.4	39.6	35.0	18.4	36.4	39.0	39.4	36.6	18.5	35.5	39.7	37.4	15.1	33.8
$(0.2, 0.3]$	6x	36.4	10.4	37.5	39.7	38.1	15.9	37.2	39.0	38.2	36.4	33.5	36.8	38.4	36.6	29.6	36.3
	10x	36.1	10.4	37.3	39.0	37.0	16.1	37.0	38.9	38.5	36.6	34.1	36.6	39.5	37.1	24.5	36.0
	30x	33.7	11.9	37.1	39.6	35.5	18.8	36.4	39.2	38.6	36.4	18.0	35.5	39.5	37.3	13.9	34.1
$(0.3, 0.4]$	6x	36.8	11.5	37.5	39.3	37.8	16.5	37.3	39.8	38.7	36.6	34.5	37.3	38.8	36.7	31.3	36.7
	10x	36.1	11.3	37.4	39.6	37.0	17.2	37.0	39.3	38.6	36.8	30.3	36.8	39.4	37.1	26.9	35.9
	30x	32.9	13.0	37.2	39.6	31.7	18.5	36.3	37.4	39.7	36.8	20.2	34.9	39.6	37.3	12.7	34.6
$(0.4, 0.5]$	6x	36.5	18.3	37.0	39.2	37.1	22.5	38.8	39.8	39.2	37.6	24.2	37.9	38.2	38.0	16.8	36.3
	10x	35.9	17.5	37.3	39.7	36.8	24.2	38.1	39.4	39.6	37.9	31.6	37.6	39.8	37.3	19.8	35.9
	30x	34.2	13.3	37.6	39.5	34.2	25.7	38.0	37.9	38.3	38.2	32.9	37.6	39.6	37.1	11.7	34.2

G is the true genotype. $(G_{\text{P}}, G_{\text{S}}) = (0, 1)$, $(1, 0)$, etc. indicates loci that are called to be G_{P} and G_{S} by PhredEM and SeqEM, respectively.

Table S2.7: Specificity and Sensitivity in the analysis of the 1000 Genomes CEU data.

	Specificity						Sensitivity					
MAF	P	S	GATK	PB	SB	GATK-B	P	S	GATK	PB	SB	GATK-B
$(0, 0.01]$	1.0	1.0	0.995	1.0	0.999	0.995	0.849	0.845	0.699	0.828	0.843	0.748
$(0.01, 0.05]$	0.999	0.999	0.998	0.999	0.999	0.999	0.901	0.901	0.898	0.954	0.957	0.958
$(0.05, 0.1]$	0.999	0.999	0.997	0.999	0.999	0.998	0.903	0.903	0.904	0.973	0.972	0.976
$(0.1, 0.2]$	0.998	0.997	0.996	0.998	0.997	0.998	0.902	0.905	0.910	0.977	0.978	0.980
$(0.2, 0.3]$	0.997	0.996	0.995	0.997	0.996	0.996	0.915	0.912	0.924	0.986	0.986	0.987
$(0.3, 0.4]$	0.996	0.995	0.993	0.996	0.995	0.995	0.925	0.923	0.933	0.988	0.987	0.989
$(0.4, 0.5]$	0.990	0.988	0.992	0.995	0.993	0.993	0.939	0.936	0.943	0.990	0.989	0.990

P and S represent PhredEM and SeqEM, respectively. PB, SB, and GATK-B represent PhredEM, SeqEM, and GATK followed by Beagle. MAF is the minor allele frequency observed in the array genotype data. Let n_0, n_1, and n_2 denote the average number of mis-called genotypes for $G = 0, 1$, and 2. Specificity $= 1 - n_0/N_0$ and Sensitivity $= 1 - (n_1 + 2n_2)/(N_1 + 2N_2)$.

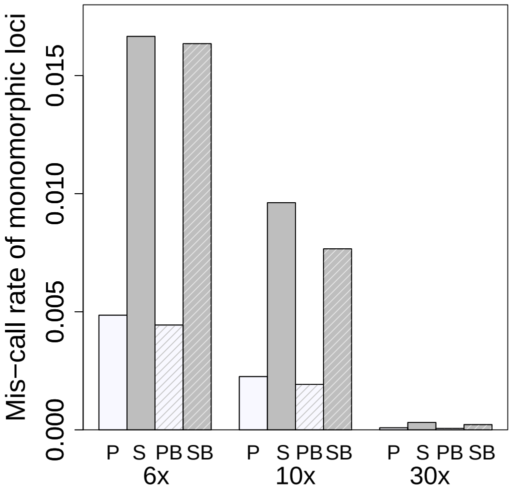

Figure S2.1: Mis-call rates of monomorphic loci when $n = 200$.

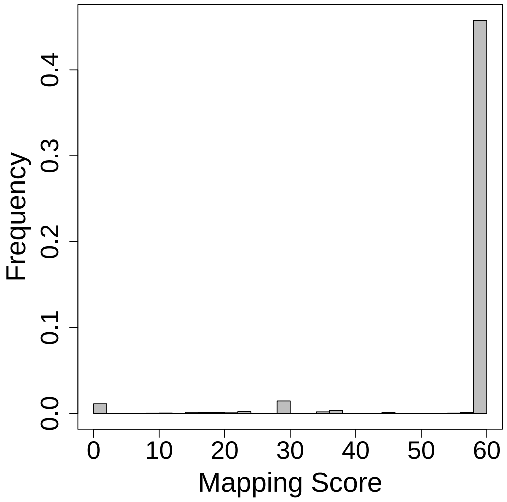

Figure S2.2: Distribution of mapping scores in the UK10K SCOOP data.

Chapter 3

Robust Inference of Population Structure from Next-Generation Sequencing Data with Systematic Differences in Sequencing

This Chapter is joint work with Dr. Yijuan Hu and Dr. Glen A. Satten. The manuscript is currently under review in *Bioinformatics*.

3.1 Methods

3.1.1 Estimating the per-base error rate

We consider biallelic single-nucleotide polymorphisms (SNPs). Let G denote the unknown true genotype (coded as the number of minor alleles) of an individual at a SNP, T denote the number of reads mapped to the SNP, and R ($R \leq T$) denote the number of reads carrying the minor allele. Similar to SAMtools (Li et al., 2009a), GATK (DePristo et al., 2011), and SeqEM (Martin et al., 2010), we assume that R given T and G follows a binomial distribution

$$P_\epsilon(R|T,G) = \begin{cases} \text{Binomial}(T,\epsilon) & \text{if } G=0 \\ \text{Binomial}(T,0.5) & \text{if } G=1 \\ \text{Binomial}(T,1-\epsilon) & \text{if } G=2, \end{cases} \tag{3.1}$$

where ϵ is the probability that a read allele is different from the true allele and is referred to as the error rate. The "errors" here comprise both base-calling and alignment errors. We treat ϵ as a free parameter that is locus-specific and estimate it from the read data using SeqEM, which is a multi-sample, single-locus genotyper (although we do not use its genotyping results). Because PCA typically uses common variants, which often do not follow Hardy-Weinberg equilibrium (HWE) in the presence of population stratification, we adopt the model allowing for Hardy-Weinberg disequilibrium in SeqEM. Suppose there are M sequencing groups of samples with potentially differences in sequencing qualities, referred to as groups 1, 2, . . . , and M. Then, we obtain separate error estimates by applying SeqEM independently in each group at each SNP.

3.1.2 Pruning SNPs and picking ancestry informative markers

The genome (and even the exome) has far more SNPs than are necessary for accurate ancestry assignment using PCA. Even with genome chip data, it is common practice to perform LD-based SNP pruning to generate a subset of nearly independent SNPs. Thus, we propose an initial pruning step to find SNPs that have low LD and also have enhanced chance of being ancestry-informative markers (AIMs). Because it is not necessary to remove artifacts related to differences in sequencing quality at this stage, we use simple methods that can be easily applied on a large scale. In particular, we ignore differences in sequencing depth and use a simple correction for sequencing error.

Given the model in (3.1), it is possible to show that

$$\mathcal{G} = \frac{R/T - \epsilon}{0.5 - \epsilon}$$

is an unbiased estimator for the true genotype G, by noting that model (3.1) implies $E(R/T|T) = (0.5 - \epsilon)G + \epsilon$ and then marginalizing over T. For each sequencing group, we calculate $\mathcal{G}$ at each SNP by replacing ϵ with its estimator.

For SNP pruning, we then calculate the pairwise Pearson correlation coefficient based on the values of $\mathcal{G}$, and apply standard LD-based pruning (Purcell et al., 2007). After pruning, we may use all remaining SNPs to infer population structure; alternatively, we may restrict to a panel of AIMs that have maximum allele frequency differences between predefined populations. If the underlying populations are unknown, we can pick AIMs by selecting those SNPs having the highest variance of $\mathcal{G}$. The ability to estimate genetic ancestry strongly depends on the number of AIMs used. Because we employ a non-specific strategy to identify AIMs, we recommend using at least 10K AIMs (Pardo-Seco et al., 2014).

3.1.3 Handling systematic differences in sequencing

Once we have selected a set of SNPs for calculating PCs, we adjust the data so that the sequencing quality is the same across sequencing groups. First, we subsample read counts so that the depth distribution is the same in each sequencing group. To illustrate our algorithm, we first consider the case where each sequencing group has the same sample size. In this case, at each SNP we sort the observations by depth and then match the observations in each group having the same rank order of depth (we randomize the order among observations having the same depth in each group). At a given SNP, each matched set thus has one observation from each sequencing group. For each matched set, we then sample the reads from each observation (without replacement) to equal the smallest depth found in that set; the data from the observation having the lowest depth is not changed. After subsampling each matched set, the depth for each sequencing group at the given SNP is the same. Repeating this procedure at each SNP results in a dataset for which all sequencing groups have the same depth at each SNP. Details on the algorithm we use when sample sizes of the sequencing groups differ is found in the Appendix 3.4.1.

Once the depth is equal across sequencing groups, we then equalize the error rate across sequencing groups using a read-flipping procedure. Specifically, let $\widehat{\epsilon}_1$, $\widehat{\epsilon}_2$,. . . ,$\widehat{\epsilon}_M$ be the estimated error rates for the M groups at a SNP. Suppose that group M has the highest error rate, i.e., $\widehat{\epsilon}_M = \max_{m=1,\ldots,M}\{\widehat{\epsilon}_m\}$. We then flip each read allele (i.e., change a minor allele read to a major allele read, or *vice versa*) in group m $(m = 1, \ldots, M - 1)$ with probability $(\widehat{\epsilon}_M - \widehat{\epsilon}_m)/(1 - 2\widehat{\epsilon}_m)$ to achieve the same error rate as in group M. Justification for this choice is found in Appendix 3.4.2.

After subsampling and read-flipping, we then compute the variance-covariance matrix of R/T (use of $\mathcal{G}$ is no longer required as the error rates are now the same in each sequencing group). Denote the (centered and scaled) matrix of R/T for the b-th subsampled data by $\boldsymbol{X}_b$. Note that to the extent that we have correctly matched the depth and error rates, $\boldsymbol{X}_b\boldsymbol{X}_b^T$ does not have any PCs that correspond to differences in sequencing quality. To minimize loss of information, we repeat the subsampling and allele flipping procedures B times and aggregate the variance-covariance matrices by averaging to obtain $\mathbb{X} = \sum_{b=1}^{B} \boldsymbol{X}_b\boldsymbol{X}_b^T/B$. Finally, we calculate PCs from $\mathbb{X}$, which preserves the unbiasedness of individual $\boldsymbol{X}_b\boldsymbol{X}_b^T$'s. We recommend using $B = 100$, which we have found to achieve good accuracy of PCs at an affordable computational cost in our numerical studies.

3.1.4 Application to stratified and admixed populations from 1000 Genomes

To evaluate our approach and compare with existing methods, we constructed data for a population having three similar but distinct subpopulations. The subpopulations were based on samples from three Asian populations in the 1000 Genomes Project: 103 Han Chinese from Beijing, China (CHB), 104 Japanese from Tokyo, Japan (JPT), and 99 Kinh from Ho Chi Minh City, Vietnam (KHV). All samples had high-depth whole-exome sequencing (WES) data with an average depth of 39.5× and low-depth whole-genome sequencing (WGS) data with an average depth of 7.0×, as well as genotype data from the Illumina Omni2.5 array. To explore the effect of systematic differences in sequencing quality on population genetic inference, we assumed data from two sequencing groups (e.g., two studies), one having WES data (called group 1) and the other having WGS data (called group 2). To vary the subpopulation frequencies by group, we randomly sampled 75% of CHB, 50% of JPT, and 25% of KHV to form group 1 and the remaining to form group 2. For some analyses we also thinned the depth of the WGS data to ~4× to examine the performance of different approaches with lower depth in group 2.

To explore the effect of systematic differences in sequencing quality on association testing in a genetic epidemiologic study, we next considered group 1 to be a set of cases and group 2 to be a set of controls in an association study. Because cases and controls have different subpopulation composition, we expect to see confounding by population stratification unless the effect of ancestry is correctly accounted for; in truth no SNPs are associated with case-control status after adjustment for population stratification. To evaluate the success of PCA methods, we construct quantile-quantile (Q-Q) plots for testing association with all the 1,138,558 common SNPs (i.e., minor allele frequency [MAF] ≥ 0.05) on the genotyping array, using the score test for logistic regression model which used the array genotypes for the main effect and the top 10 PCs as covariates. Here we used the array genotypes as the true genotypes for the main effect in order to focus on the impact of different methods for calculating the PCs.

To evaluate our approach and compare with existing methods in a situation with continuous admixture, we also considered estimation of the proportion of African ancestry for the 55 Americans of African ancestry in southwest USA (ASW) from the 1000 Genomes Project. To this end, we used WES data ($\sim$40×) from the CEU (99 Utah residents with northern and western European ancestry) and YRI samples (108 Yoruba in Ibadan, Nigeria) but assumed only WGS data ($\sim$6.5×) were available for the ASW samples. We calculated PCs for all three populations together. We then estimated the proportion of African ancestry for each individual in ASW by the ratio of the distance between the individual's PC1 and the centroid of PC1 in CEU and the distance between the centroids of YRI and CEU (Ma and Amos, 2012).

In processing the sequencing data, we first used SAMtools to generate pileup files from BAM files, restricting to exonic regions and filtering out reads that are PCR duplicates, have mapping scores < 30, and have improperly mapped mates. From the pileup files, we extracted read count data for each locus (i.e., base pair), excluding reads with *phred* base-quality scores < 20 at this locus. Additionally, we filtered out individual-level read count data (i.e., setting T to 0) that do not fit the binomial model (3.1) by a read-based quality-control (QC) procedure (Hu et al., 2016). We excluded a locus altogether if more than 5% of samples in either group have $T = 0$. We focused on SNPs with MAF $\geq$ 5%, which can be easily and accurately identified from the read data using the MAF estimated by SeqEM. We also pruned these SNPs for pairwise correlation at a threshold of $r^2 = 0.5$. To facilitate comparison with the truth, we further restricted to SNPs whose array genotypes are available.

3.1.5 Simulation design

To further assess the impact of different ways of calculating PCs on the power of association tests, we conducted simulation studies based on the stratified data example using three Asian populations from the 1000 Genomes Project described previously. We generated allele frequencies for three populations (called populations 1, 2, and 3) using the approach described by Fumagalli et al. (2013). Population-specific MAFs are sampled based on F_{ST} values that differentiate the subpopulations; details can be found in Appendix 3.4.3. Based on Tian et al. (2008) we set F_{ST}=0.0065 to differentiate between CHB and JPT/KHV and then set F_{ST}=0.011 to differentiate between JPT and KHV. We assumed 100K common SNPs (overall MAF $\geq$ 0.05) in each replicate data set. We treated these SNPs as independent of each other and simulated the genotypes assuming HWE within each population.

To generate the disease (case-control) status, we started with a general population in which 1/6, 1/3, and 1/2 of individuals are from populations 1, 2, and 3, respectively. Under the null hypothesis of no genetic association, we generated disease status D_i for the i-th individual using the risk model $\log\{P(D_i = 1)/P(D_i = 0)\} = \alpha - \log(3)I_{\{\mathcal{P}_i=2\}} - 2\log(3)I_{\{\mathcal{P}_i=3\}}$, where α was chosen to achieve a disease rate of $\sim$1% in population 1 and $\mathcal{P}_i$ is the population that the i-th individual belongs to. We then sampled until we had obtained an equal number of cases and controls. We considered designs with 150 cases and 150 controls to mimic the sample size of our stratified 1000 Genomes dataset, as well as designs with 1000 cases and 1000 controls that represent a more typical genetic association study from a single study center. As with our original stratified data, this sampling scheme yielded an approximately equal number of members from each of the three populations in the case-control study population. Further, the cases in the study population comprised about ~75% of the study population from population 1, ~50% of the study population from population 2, and ~25% of the study population from population 3, which again matches the compositions of populations in sequencing group 1 in the stratified 1000 Genomes dataset. To calculate power we assumed an alternative hypothesis in which each copy of the first 10 SNPs increases the odds of disease by a common log odds ratio β, so that risk model becomes $\log\{P(D_i = 1)/P(D_i = 0)\} = \alpha + \beta \sum_{j=1}^{10} G_{ij} - \log(3) I_{\{\mathcal{P}_i=2\}} - 2\log(3) I_{\{\mathcal{P}_i=3\}}$, where G_{ij} is the genotype of the j-th SNP. Again, we sampled until an equal number of cases and controls were drawn to form the study population.

We simulated read count data rather than raw sequencing reads for sake of computational efficiency; this is reasonable as each SNP was generated independently. We fixed the average depth at 39.5× and set the average error rate to 0.17% (as observed in the 1000 Genomes WES data) for cases and varied the average depth between 7× and 4× and the average error rate between 0.1% (as observed in the 1000 Genomes WGS data) and 1% (Nielsen et al., 2011) for controls. For more details about locus-specific depth and error rate, see Appendix 3.4.4. Finally, we sampled R_i given (G_i, T_i, ϵ) according to model (3.1). The whole process was repeated to generate 100 replicate data sets.

3.2 Results

3.2.1 Inference on a stratified population from 1000 Genomes

After applying QC filters to the original WES and WGS data, we obtained 41,672 SNPs exome wide, of which 27,688 SNPs remained for calculating PCs after LD pruning. We applied the proposed method, the PCA method based on called genotypes in which the genotypes were called by SeqEM, and the PCA method based on array genotypes which serves as the gold standard. We refer to these methods as New, CG, and TG. In addition, we applied the Fumagalli method. We first evaluated the methods for their ability to differentiate subpopulations. We focused on scatter plots of PC1 versus PC2, because the first two PCs are expected to capture the majority of genetic variability given that there are three discrete populations. The upper panel of Figure 3.1 shows that PCs calculated using New, CG and Fumagalli inferred the same structure as PCs calculated using TG. To investigate if the $\sim 7\times$ WGS data accounted for this, we thinned the depth of the WGS data to $\sim 4\times$, resulting in 25,158 SNPs, of which 19,114 SNPs remained for calculating PCs after LD pruning. The lower panel of Figure 3.1 shows that New still provided a similar estimation of population structure as TG. In contrast, both Fumagalli and CG caused group 2 to shift away from group 1 within each subpopulation. In particular, Fumagalli shrank group 2 towards the origin relative to group 1; this is not unexpected, because the sequencing depth affects the accuracy with which the posterior genotype probabilities used in Fumagalli are calculated. A two-sample t-test comparing the distance measure $\text{PC1}^2 + \text{PC2}^2$ between samples from group 1 and samples from group 2 confirmed that the distances were significantly different for Fumagalli (p-value < 0.001) and CG (p-value $= 0.002$) but not for New (p-value $= 0.306$) and TG (p-value $= 0.439$) for the WGS data thinned to $\sim 4\times$ depth.

Figure 3.2 shows Q-Q plots for tests of association calculated at each SNP. Because

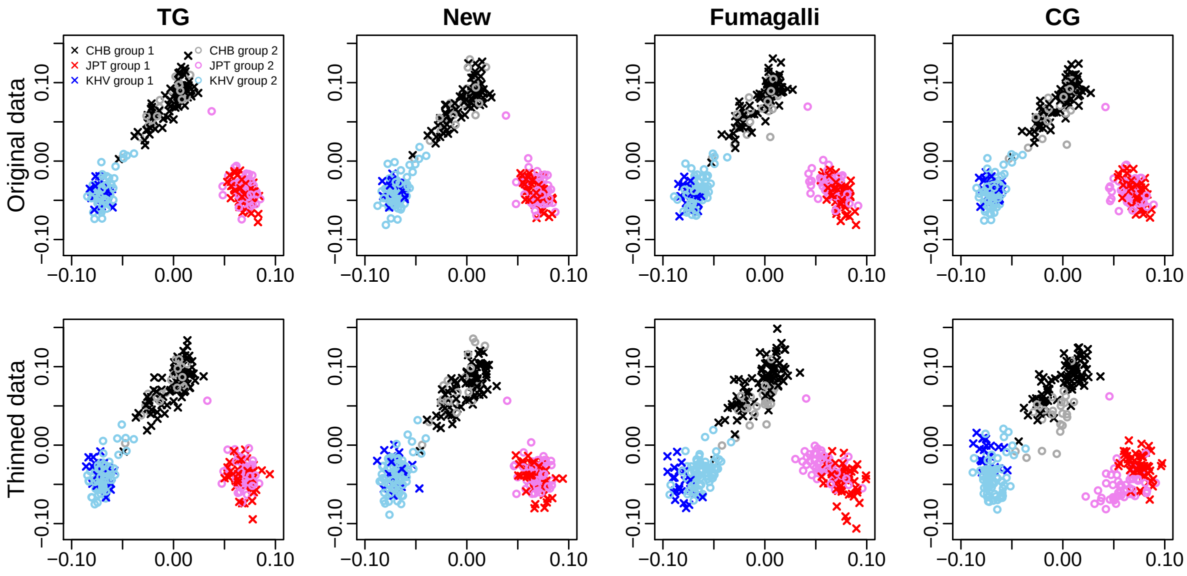

Figure 3.1: Scatter plots of PC1 (x-axis) versus PC2 (y-axis) in the analysis of the stratified 1000 Genomes data with three discrete Asian populations

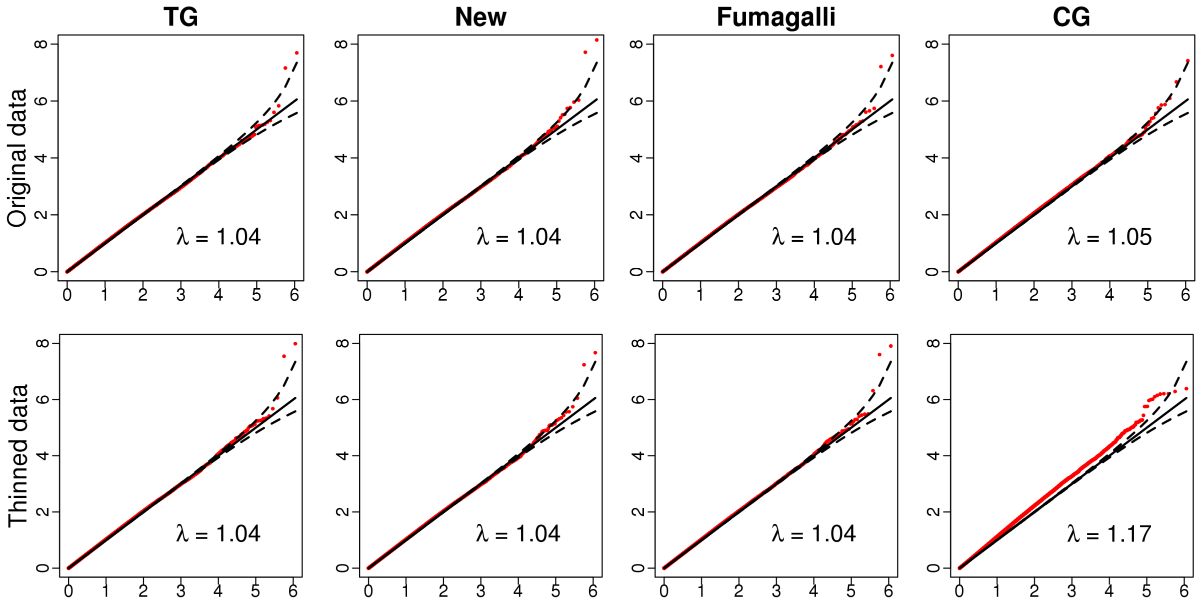

Figure 3.2: Q-Q plots of $-\log_{10}$ (observed p-values) (y-axis) versus $-\log_{10}$ (expected p-values) (x-axis) in the analysis of the stratified 1000 Genomes data with three discrete Asian populations.
The solid black line represents the reference line of global null hypothesis of no association. The dashed curves represent a 95% point-wise confidence band.

there is no true association, the extent to which these plots track the 45^o line is a measure of how well the PCs calculated by each method control for population stratification. For CG, the departure of the observed p-values from the global null hypothesis of no association may be too subtle to be seen clearly with the original data at ∼7×; it is much more pronounced with the thinned data at ∼4×, with genomic control $\lambda = 1.17$. New, Fumagalli, and TG all led to good control of type I error ($\lambda = 1.04$); note that the type I error would be highly inflated ($\lambda = 1.22$, Supplemental Figure S3.2) without adjusting for any PCs.

3.2.2 Inference on an admixed population from 1000 Genomes

Applying QC filters to the admixed population data resulted in 43,503 SNPs exome wide; LD pruning reduced this number to 34,563 SNPs. After thinning the depth of WGS to ∼4×, 24,939 SNPs remained, of which 21,215 SNPs were available for calculating PCs after LD pruning. We calculated PCs using New, Fumagalli, CG, and TG and compared the estimated proportions of African ancestry from each method with that calculated using TG; these results are shown in Figure 3.3 (for values in $[0.55, 0.9]$) and Supplemental Figure S3.3 (for all values). The estimates using New agreed closely with those using TG for either the original or the thinned data. By contrast, the estimates using Fumagalli and CG were biased; this bias was more severe with the thinned data. We also quantified the deviation from the values obtained using TG by calculating the sum of squared differences (SS) between estimates obtained using each method and those obtained by TG. New had the lowest value of SS, which was very close to zero and much lower than those of Fumagalli and CG; values of SS are shown in Figure 3.3.

Table 3.1: Type I error (divided by the nominal significance level of 0.05)

			5K SNPs				50K SNPs				100K SNPs				10K AIMs			
n	c_0	ϵ_0	TG	New	F	CG	TG	New	F	CG	TG	New	F	CG	TG	New	F	CG
300	4×	1%	1.09	1.09	1.08	1.24	1.09	1.09	1.07	0	1.09	1.08	1.07	0	1.09	1.09	1.07	1.44
		0.1%	1.09	1.09	1.08	1.21	1.09	1.09	1.07	0	1.09	1.09	1.07	0	1.09	1.09	1.07	1.34
	7×	1%	1.09	1.08	1.08	1.11	1.09	1.08	1.08	1.37	1.08	1.08	1.08	0	1.09	1.08	1.08	1.12
		0.1%	1.09	1.09	1.08	1.10	1.09	1.09	1.08	1.16	1.09	1.08	1.08	1.78	1.09	1.09	1.08	1.10
2000	4×	1%	1.01	1.01	1.02	1.09	1.01	1.01	1.02	0	1.01	1.01	1.03	0	1.01	1.01	1.01	1.12
		0.1%	1.01	1.01	1.02	1.09	1.01	1.01	1.01	0	1.01	1.01	1.01	0	1.01	1.01	1.01	1.11
	7×	1%	1.01	1.01	1.02	1.03	1.01	1.01	1.02	1.05	1.01	1.01	1.01	1.10	1.01	1.01	1.01	1.06
		0.1%	1.01	1.02	1.01	1.03	1.01	1.01	1.01	1.04	1.01	1.01	1.01	1.09	1.01	1.01	1.01	1.05

n is the sample size. c_0 and ϵ_0 are the average depth and average error rate in controls. New is the proposed method. F is the Fumagalli method. TG and CG are the PCA methods based on true and called genotypes, respectively. The results are based on 10M tests.

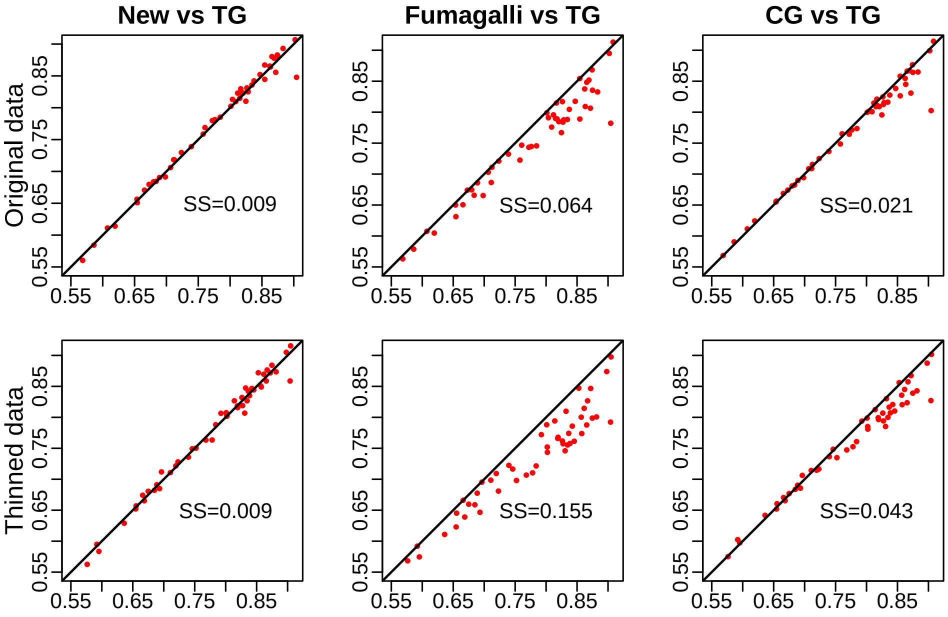

Figure 3.3: Agreement between estimated proportions of African ancestry calculated using each method (y-axis) and TG (x-axis) for the analysis of an admixed population from the 1000 Genomes Project.

SS is the sum of squared difference between the displayed method and TG. The axes are restricted to $[0.55, 0.9]$ to show more detail for the majority of samples; plots showing the full range $[0, 1]$ can be found in Supplemental Figure S3.3.

3.2.3 Simulation studies

For calculating PCs in each replicate, we considered a random set of 5K, a random set of 50K, and all 100K SNPs, as well as a common practice of using 10K AIMs. We applied the New, Fumagalli, CG, and TG methods. Note that TG now refers to the PCA method based on true genotypes.

We first verified the type I error rate for association testing in the presence of population stratification in a wider range of scenarios than seen in the stratified (3-subpopulation) 1000 Genomes dataset. For each simulation replicate, we tested for association using each of the 100K SNPs using the score test for logistic regression. We used 100 data replicates, so that the type I error results displayed in Table 4.2 are based on 10M tests. In all scenarios, New and Fumagalli had similar type I error rates as TG (though TG had slightly inflated size when the sample size was 300). The performance of CG was mixed; the type I error rate was sometimes inflated but sometimes zero, depending on the scenario. Note that the scenarios in the second and fourth rows under "5K SNPs" mostly resembled the first 1000 Genomes dataset and had similar results.

To get more insights into the mixed performance of CG, we examined the PCs using one replicate of data. In the scatter plots (Figure 3.4) of PC1 versus PC2 with 4× average depth and 1% average error rate in controls, we observed that CG caused controls to shift away from cases within each population and the shift became a complete separation when the number of SNPs used for calculating PCs was increased. With 10K AIMs, the shift was less severe but could still be seen. With an average depth of 7× and a much smaller average error rate (0.1%), CG still resulted in a slight shift when calculated using a large number of SNPs (Supplemental Figure S3.4). Because we simulated three populations, we expected PC1 and PC2 to capture all genetic variability; thus we expect any information in PC3 and PC4 to be related to differential sequencing. In Figure 3.5 and Supplemental Figure S3.5 we see that PC3 calculated using New or TG have no information about subpopulation or case-control status. However, PC3 calculated using CG can be highly informative about the case-control status, indicating a signal arising from differential sequencing quality; this may explain why CG sometimes had zero size. PC1–PC4 calculated using Fumagalli exhibited differential shrinkage; interestingly, this pattern seemed to have no impact on the type I error.

In Figure 3.6 we report the effect of PC calculation method on the power to measure a true association; we omit CG because it did not control type I error. New achieved almost the same power as TG in all scenarios. Fumagalli resulted in substantial loss of power when 50K or 100K SNPs were used for inferring PCs and the controls had 1% average error rate. We found that PC3 by Fumagalli had a large difference in mean between cases and controls in scenarios that Fumagalli lost power (Figure 3.7). Since no difference was expected for PCs other than PC1 and PC2, this PC3 effect was likely the cause of the power loss.

3.3 Discussion

We have presented a robust approach to inferring population structure that is based on analyzing raw sequencing reads directly, without calling genotypes. Our subsampling and read-flipping procedures ensure that the sequencing qualities are matched among all sequencing groups. As a result, the PCs generated from our method do not capture any difference in sequencing qualities, unlike existing methods.

In evaluating our method, we considered discrete populations as well as admixed populations. We have focused on two groups with differential sequencing qualities in the main text, but our method also worked well for studies having three sequencing groups (see Appendix 3.4.5 and Supplemental Figure S3.6).

In our simulation studies, we considered different numbers of SNPs for calculating

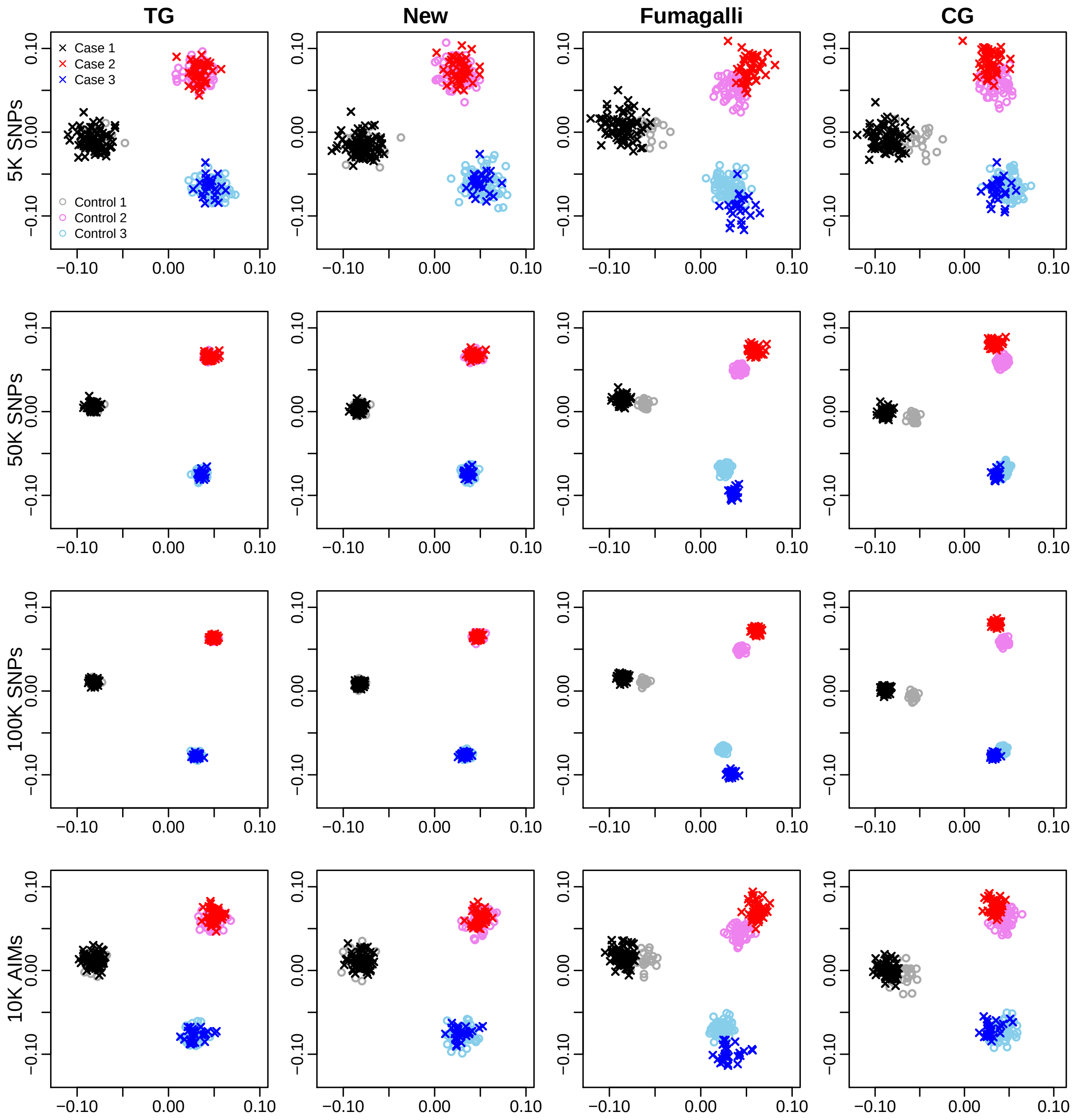

Figure 3.4: Scatter plots of PC1 (x-axis) versus PC2 (y-axis) in simulation studies with 4× average depth and 1% average error rate in controls.
The plots are based on a single replicate data set generated under the null hypothesis of no association and have data from 150 cases and 150 controls.

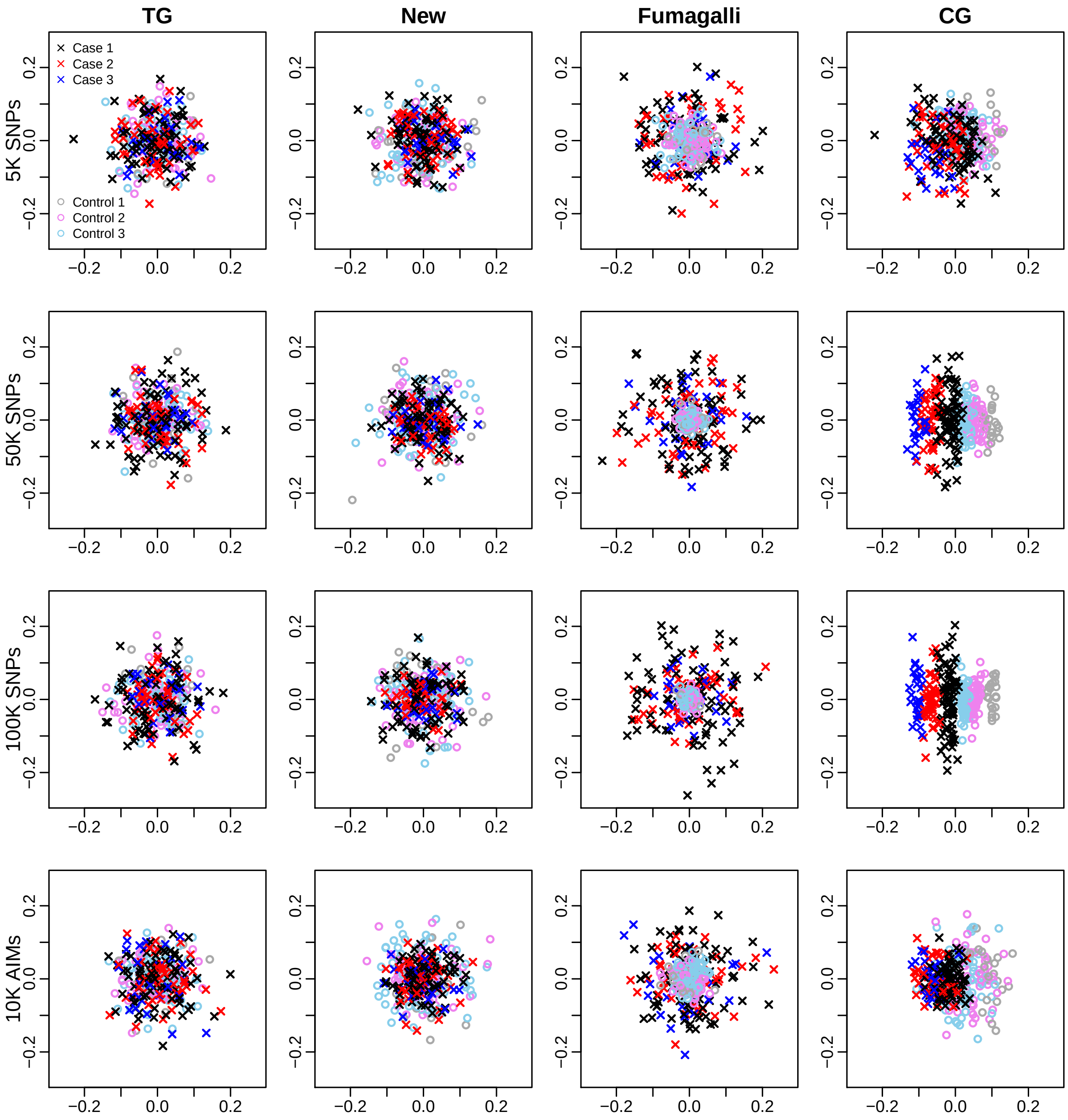

Figure 3.5: Scatter plots of PC3 (x-axis) versus PC4 (y-axis) in simulation studies with 4× average depth and 1% average error rate in controls.
The plots are based on a single replicate data set generated under the null hypothesis of no association and have data from 150 cases and 150 controls.

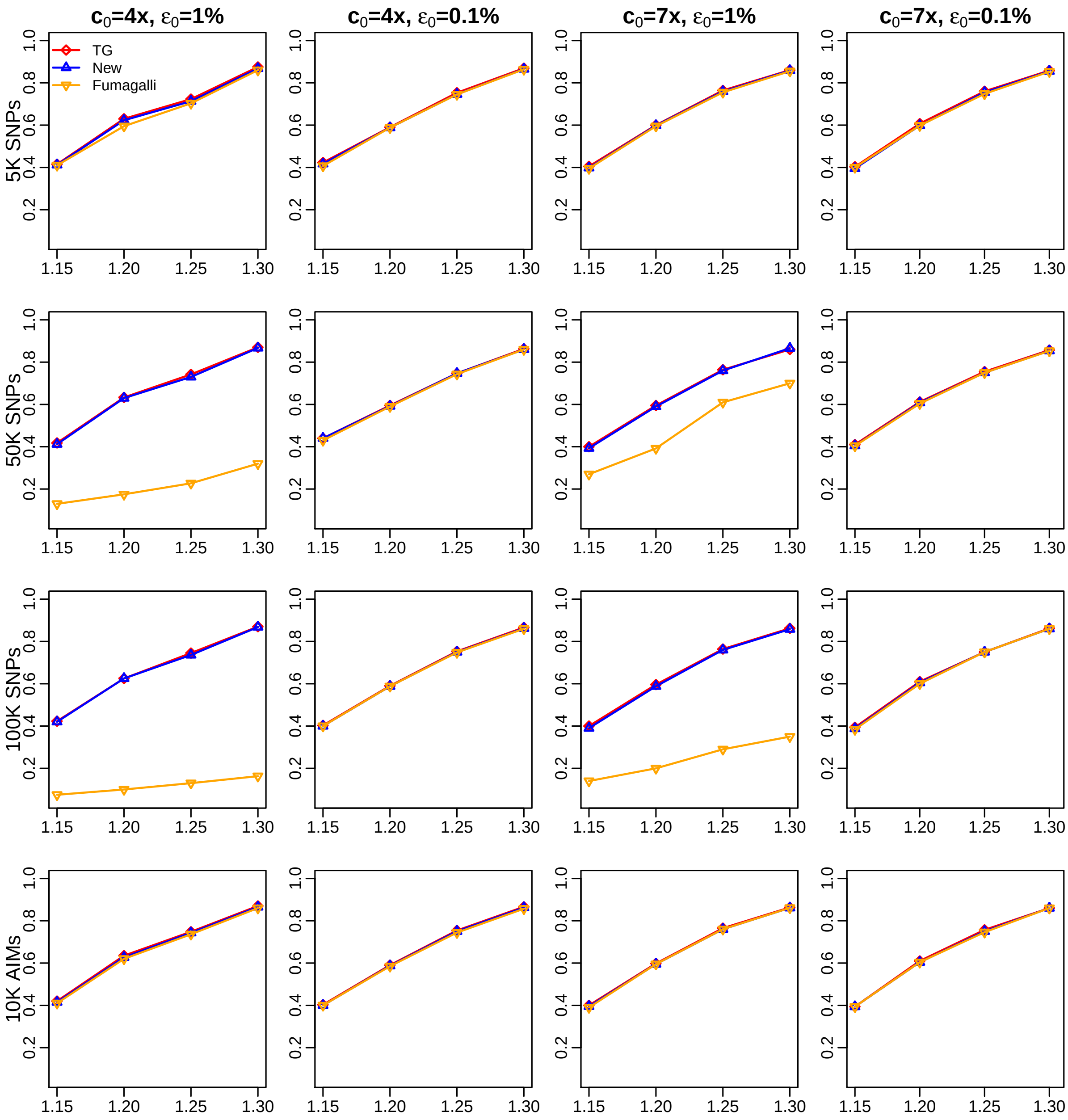

Figure 3.6: Power (y-axis) at the nominal significance level of $\alpha = 0.05$ over different ORs (x-axis) based on 1000 cases and 1000 controls.
The results are based on 1000 tests (10 disease-susceptibility SNPs per replicate and 100 replicates).

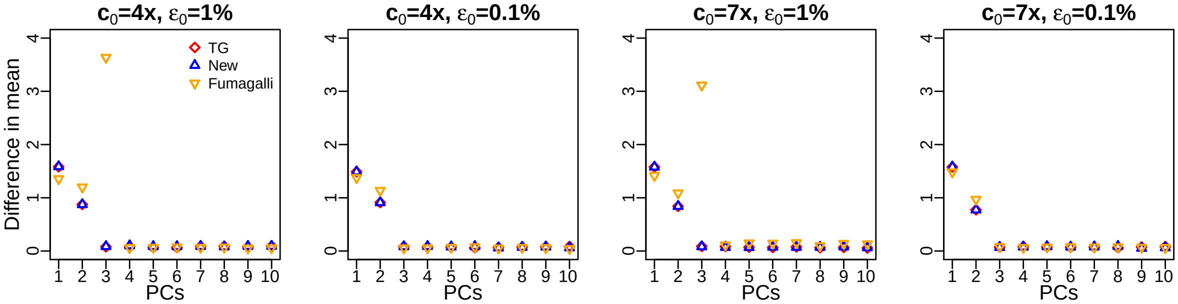

Figure 3.7: Mean differences (×100) of top ten PCs between cases and controls in simulation studies.
The y-axis represents the absolute difference in mean. The PCs were calculated using 100K SNPs in 100 replicate data sets each having 1000 cases and 1000 controls. The odds ratio $\exp(\beta)$ was 1.3.

PCs. In practice, the number of SNPs needed to accurately infer population structure and correct for population stratification relies on the degree of population differentiation (Price et al., 2006). Therefore, we recommend using as many SNPs as possible after filtering out rare SNPs, pruning for strong correlations, and prioritizing AIMs. If this is computationally prohibitive, we recommend using AIMs and increasing the number of AIMs used until diagnostic plots of PCs stabilize.

In association testing, we used array genotypes in the analysis of 1000 Genomes data (or true genotypes in the simulation studies) for the main effect in the logistic regression. In practice, array genotypes are not always available. More importantly, sequencing studies offer many more SNPs than those on genotyping arrays. We are currently developing methods for association analysis that is based on sequencing reads directly (for both the main effect and the ancestry correction), without using array genotypes or calling genotypes. In this work, we used array genotypes to ensure that the effects we reported were due to the way PCs were calculated, rather than (possibly differential) error in the genotype used to fit the risk model.

Our approach was based on some simplifying assumptions. First, we assumed that the error rate at a SNP was the same across samples in a sequencing group, which was directly estimated from the read count data without using any information on *phred* scores and mapping scores. In the analysis of the 1000 Genomes data, we filtered out reads with mapping scores < 30 and removed bases with *phred* scores < 20, so that the assumption of uniform error rate is reasonable. Second, we assumed that errors were symmetric, i.e. that the probability of a read for the major allele being mis-called as the minor allele was the same as the probability of the minor allele being mis-called as the major allele.

The proposed methods are implemented in the C/C++ program `TASER-PC`, which is publicly available at `http://web1.sph.emory.edu/users/yhu30/software.html`. The software program `TASER-PC` is readily scalable to genome-wide analysis. For example, with 1000 cases at $\sim$39.5$\times$ and 1000 controls at $\sim$4$\times$, it took $\sim$7 hours on a single thread of an Intel Xeon X5650 machine with 2.67 GHz to calculate the PCs based on 100K SNPs and 100 repeats of the subsampling and read flipping procedures. When only 10K AIMs were used, it took only $\sim$0.7 hours. In general, the computation increases linearly with the number of individuals, the number of SNPs, and the read depth. Additionally, our program allows parallelization of the multiple repeats of subsampling and read flipping on multiple machines.

Our method requires the read count data at each SNP, which is readily available in the raw variant call format (VCF) files that are generated from GATK (with field name ‘AD’) or SAMtools (with field name ‘DP4’). Unfortunately, many publicly available VCF files have been trimmed so that the read count information is lost (e.g., the VCF files on the 1000 Genomes website). We have shown here, and in other settings (Hu et al., 2016), that the information in this single field can allow for adjustment of differential sequencing quality. For this reason, we advocate that this information be kept in future VCF files. Alternatively, we provide software to extract the read count data from raw BAM files.

3.4 Appendix

3.4.1 Matching the read distributions in different sequencing groups when sample sizes of the sequencing groups differ

We denote the sample sizes in sequencing groups 1, 2, $\ldots$, M by n_1, n_2, $\ldots$, n_M, respectively. We assume that group 1 has the largest sample size, i.e., $n_1 = \max_{m=1,\ldots,M}\{n_m\}$. The subsampling procedure proceeds as follows:

1. For group m $(m = 2, \ldots, M)$, draw $(n_1 - n_m)$ "pseudo" samples and add them back to the group so as to match the sample size n_1 in group 1. Specifically, write $n_1 = n_m \times s_m + t_m$. The "pseudo" samples consist of $(s_m - 1)$ repeats of the entire original samples in group m and t_m random samples drawn from group m without replacement. The data (T, R) across the whole genome are copied from the original samples to the "pseudo" samples.

2. At each SNP, sort the observations by T and then match the observations in each group having the same rank order of T into n_1 sets so that each matched set has one observation from each group. For each matched set, we then sample the reads from each observation (without replacement) to equal the lowest T found in that set; the data from the observation having the lowest T is not changed.

3. Discard the $(n_1 - n_m)$ "pseudo" samples in group m $(m = 2, \ldots, M)$.

3.4.2 Choosing the read-flipping probability

At a given SNP, let ϵ_1, ϵ_2,$\ldots$,ϵ_M be the error rates for the M sequencing groups. Suppose that group M has the highest error rate, i.e., $\epsilon_M = \max_{m=1,\ldots,M}\{\epsilon_m\}$. Our goal here is to find the probability p to flip each read allele in group m ($m = 1, \ldots, M-1$) so as to achieve the error rate ϵ_M. Let A_{b} and A_{a} be the alleles before and after flipping, respectively, and t and e be the true allele and the erroneous allele, respectively. Because $P(A_{\mathrm{a}} = e) = P(A_{\mathrm{a}} = e|A_{\mathrm{b}} = t)P(A_{\mathrm{b}} = t) + P(A_{\mathrm{a}} = e|A_{\mathrm{b}} = e)P(A_{\mathrm{b}} = e)$, we have $\epsilon_M = p(1-\epsilon_m) + (1-p)\epsilon_m$. Solving for p, we obtain $p = (\epsilon_M - \epsilon_m)/(1-2\epsilon_m)$, which can be estimated by plugging in the estimated error rates.

3.4.3 Sampling MAFs for three populations

At a SNP, we first generated the overall ancestral population minor allele frequency (MAF) π from a uniform distribution (Pritchard and Donnelly, 2001; Price et al., 2006) over $[0.05, 0.95]$. Using the Balding-Nichols model (Balding and Nichols, 1995), we sampled the MAF for population 1, π_1, and the ancestral MAF for populations 2 and 3, π_{23}, independently from the beta distribution with parameters $\pi(1-F_{ST})/F_{ST}$ and $(1-\pi)(1-F_{ST})/F_{ST}$, where $F_{ST} = 0.0065$ as estimated between CHB and JPT/KHV (Tian et al., 2008). Then, we sampled the MAFs for populations 2 and 3, π_2 and π_3, independently from the same beta distribution with parameters $\pi_{23}(1 - F_{ST,23})/F_{ST,23}$ and $(1-\pi_{23})(1-F_{ST,23})/F_{ST,23}$, where $F_{ST,23} = 0.011$ as estimated between JPT and KHV (Tian et al., 2008).

3.4.4 Simulating read count data

Given an average read depth, we generated the depth T at a locus for an individual by the same two-step strategy as described in Hu et al. (2016). Specifically, when the average depth was 39.5×, we generated the locus-specific mean depth c using Beta$(1.2, 6.0)$, which was then re-scaled to achieve the mean 39.5; given c, we generated the individual T's from NB$(c, 0.24)$. This mimicked the level of variability seen in the WES data of 1000 Genomes. When the average depth was 7×, we generated c using Beta$(20.2, 22.9)$ (and then re-scaled to achieve the mean 7) and T using

NB($c, 0.20$), which mimicked the WGS data of 1000 Genomes. When the average depth was 4×, we generated c using Beta($22.2, 8.2$) (and then re-scaled) and T using NB($c, 0.16$), which mimicked the thinned WGS data of 1000 Genomes. Supplemental Figure S3.1 displays the distributions of locus-specific mean depths observed in the 1000 Genomes data and those generated in our simulation studies; the latter closely resembles the former.

Given an average error rate, we sampled the locus-specific error rate from beta distributions. To achieve the average error rates of 0.17% and 0.1%, we used Beta($0.81, 450$) and Beta($0.25, 250$), respectively, as observed in the WES and WGS data of 1000 Genomes. To achieve the average error rates of 1%, we used Beta($88, 8755$) as used in Hu et al. (2016).

3.4.5 A simulation study assuming three groups with differential sequencing qualities

We generated three populations in the same manner as for the two-group simulation studies. Then we assumed three sequencing groups, group 1 having 75, 50, and 25 individuals from populations 1, 2, and 3, respectively, group 2 having 50, 100, and 50, and group 3 having 25, 50, and 75. Group 1 has an average depth of 4× and an average error rate of 1%, group 2 has 10× and 0.17%, and group 3 has 39.5× and 0.1%. The top two PCs are displayed in Supplemental Figure S3.6, which shows similar patterns to Figure 3.4 (for the two-group simulation studies) for all methods.

3.5 Supplemental Materials

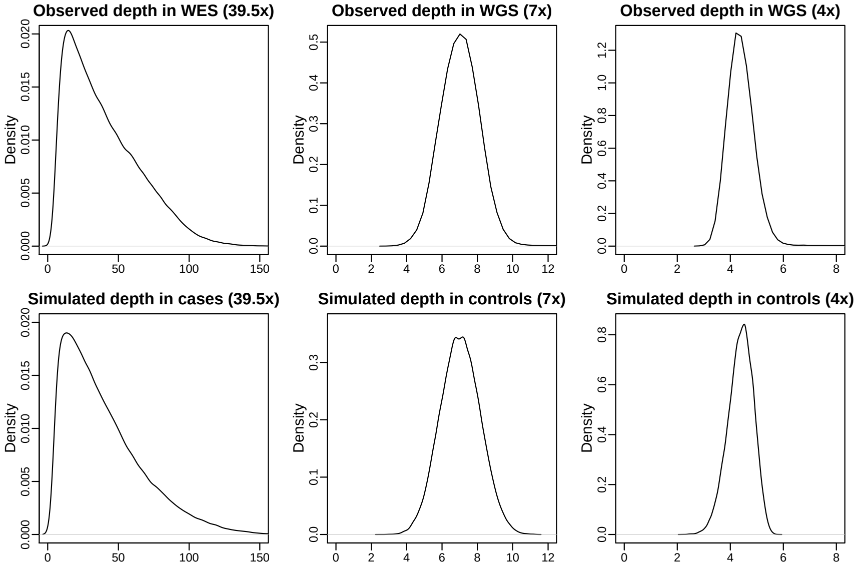

Figure S3.1: Distributions of locus-specific mean depths observed in the 1000 Genomes data (top panel) and generated in the simulation studies (bottom panel).

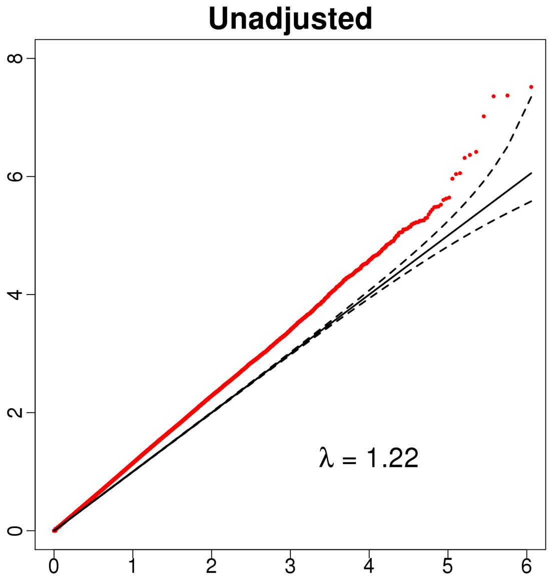

Figure S3.2: Q-Q plots of $-\log_{10}$(observed p-values) (y-axis) versus $-\log_{10}$(expected p-values) (x-axis) in the analysis of the stratified 1000 Genomes data with three discrete Asian populations without adjusting for PCs.

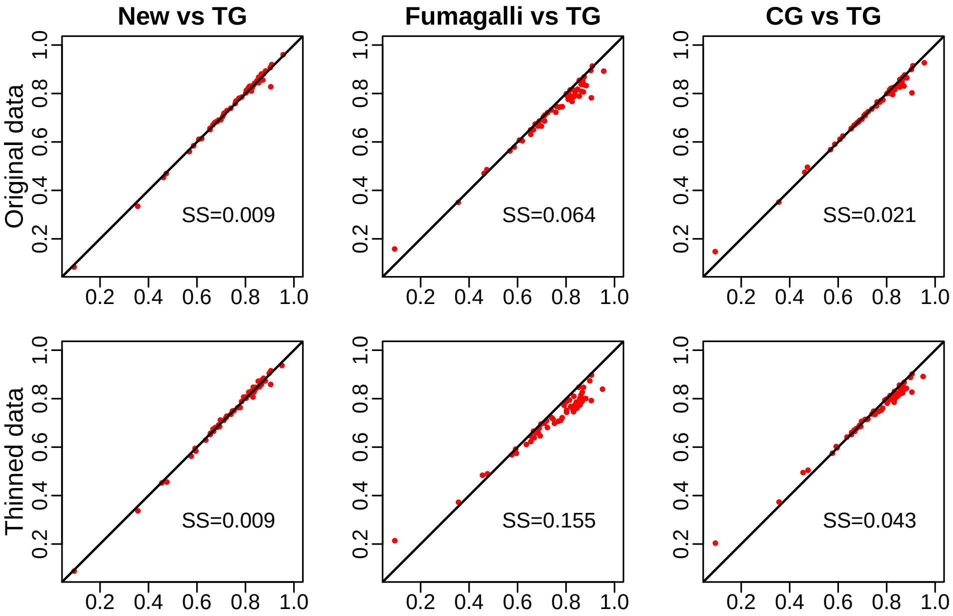

Figure S3.3: Agreement between estimated proportions of African ancestry calculated using each method (y-axis) and TG (x-axis) for the analysis of an admixed population from the 1000 Genomes Project.
SS is the sum of squared difference between the displayed method and TG.

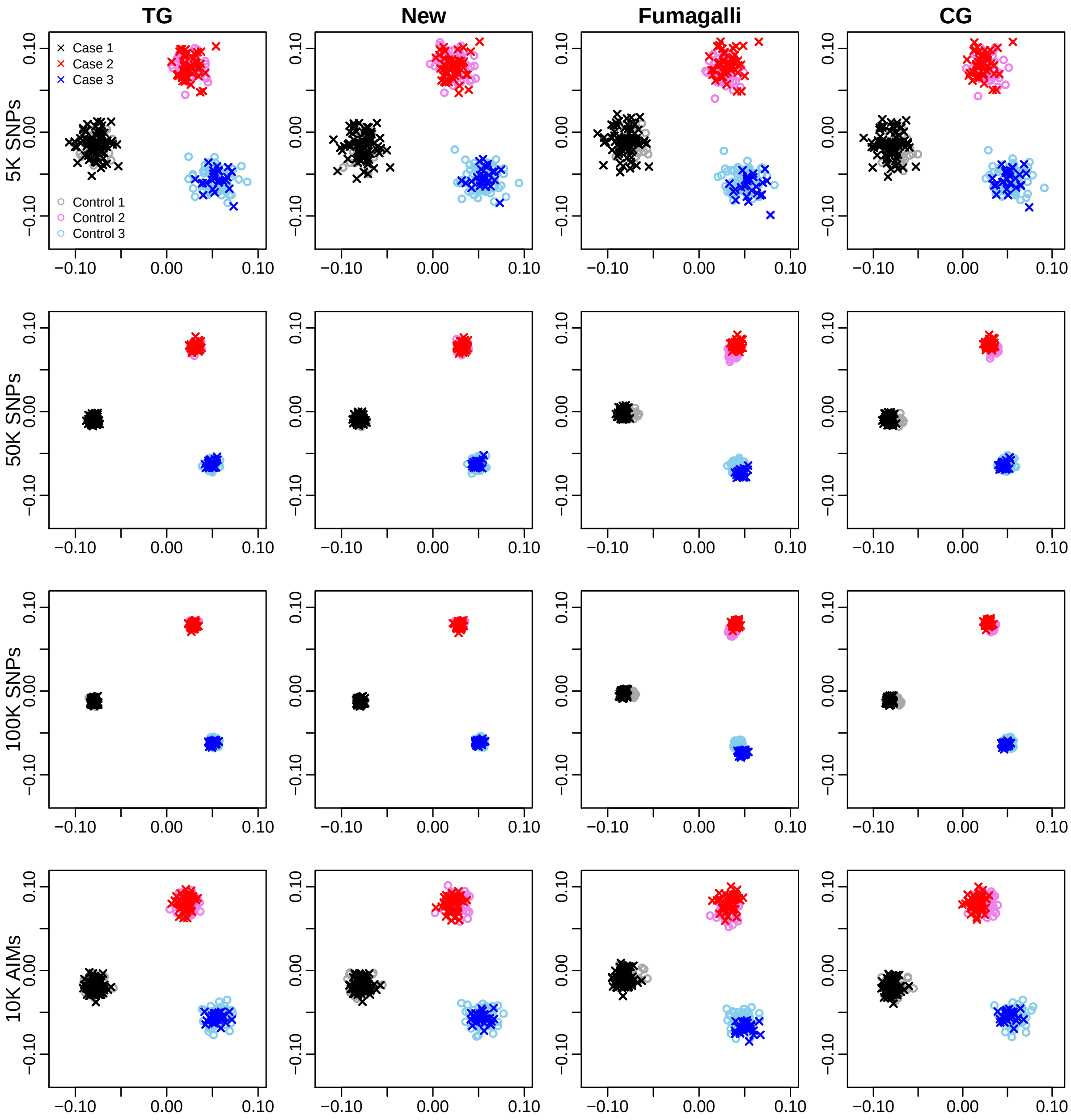

Figure S3.4: Scatter plots of PC1 (x-axis) versus PC2 (y-axis) in simulation studies with 7× average depth and 0.1% average error rate in controls.
The plots are based on a single replicate data set generated under the null hypothesis of no association and have data from 150 cases and 150 controls.

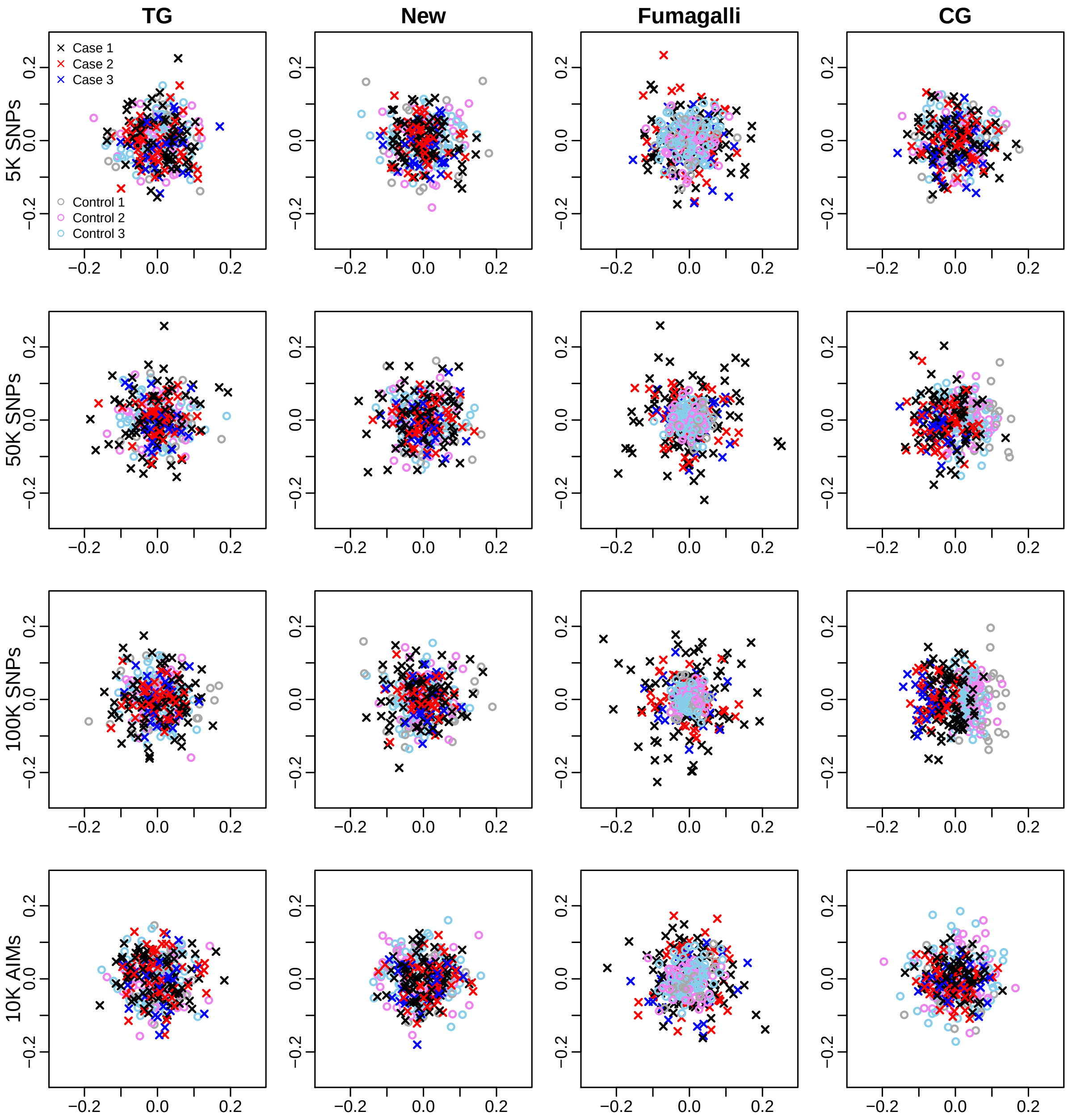

Figure S3.5: Scatter plots of PC3 (x-axis) versus PC4 (y-axis) in simulation studies with 7× average depth and 0.1% average error rate in controls.
The plots are based on a single replicate data set generated under the null hypothesis of no association and have data from 150 cases and 150 controls.

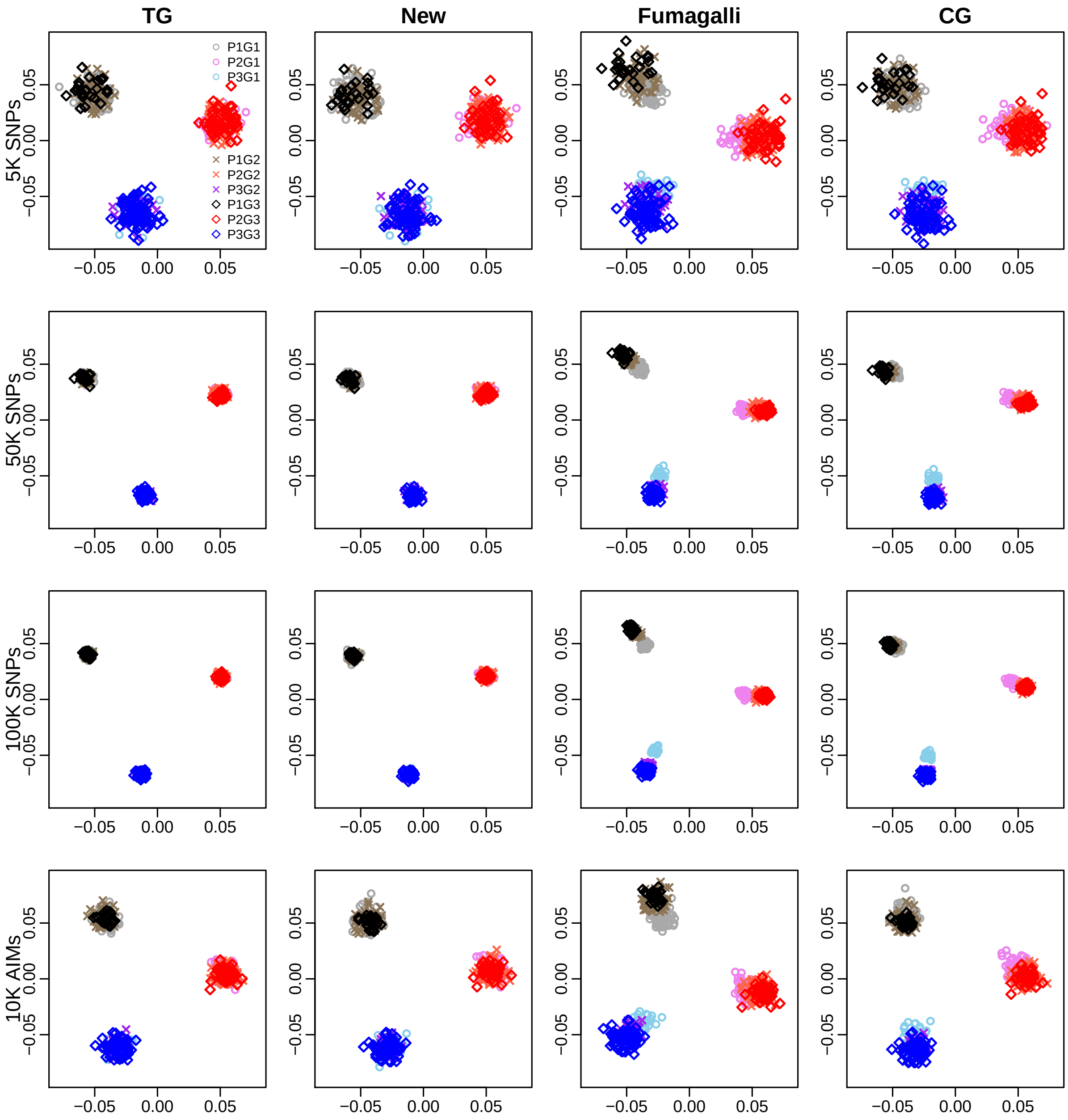

Figure S3.6: Scatter plots of PC1 (x-axis) versus PC2 (y-axis) in simulation studies with three sequencing groups.
P1G1 means population 1 in group 1, etc. The plots are based on a single replicate data set.

Chapter 4

Optimal Design of Next-Generation Sequencing Studies for Testing Rare Variant Associations

This Chapter is joint work with Dr. Yijuan Hu and Dr. Glen A. Satten. The manuscript is currently in preparation.

4.1 Methods

4.1.1 Omnibus TASER

TASER uses a weighted burden test statistic, which is the sum of the score statistic of all ‘screen-in’ rare variants within a gene (or a function unit), to assess the association between a gene and the trait of interest. Although TASER has well controlled type I error for any weight functions, better choice of weights can lead to improved power. However, the optimal prechosen weights depends on how the rare variants influence the trait, which is generally unknown. To consider multiple possible weights simultaneously, we extend TASER by adopting a multistage permutation strategy that combines different weights and achieves good power regardless of the nature of the underlying association (Tang et al., 2016). This omnibus TASER treats the minimal p-value across weights as the test statistic, and a permutation procedure is used to assess the significance of the test statistic. Like TASER, we adopt a sequential stopping rule to minimize the computational cost (Besag and Clifford, 1991).

4.1.2 Simulation design

4.1.2.1 Generating European and African haplotypes

In order to make the simulations more realistic and informative, we considered a gene which has 4 exons that are separated by 3 introns, and contains 2,419 loci in exons and 18,029 loci in introns. This gene is very ‘typical’ in the sense that its gene coding length corresponds to the 50% percentile of gene coding length among all genes. To examine the effect of the site frequency spectrum (SFS) on association tests (Lin and Tang, 2011; Moutsianas et al., 2015), we considered European and African populations separately because African population contains a larger number of rare variant sites than European population (Zawistowski et al., 2014). For each population, we used Cosi2 to generate a base population of 100,000 haplotypes with the exact gene length that mimics LD pattern, local recombination rate and population history through a coalescent model (Shlyakhter et al., 2014). In practice, researchers often filter variants by their functional consequence using bioinformatics tools, in order to identify a set of "probably" damaging variants (Lee et al., 2014). To construct a set of deleterious variants, we simply sampled a proportion of exonic rare variants (MAF$\leq$ 0.01) within each population. Specifically, we drew 53.4% of extremely rare variants having MAF <= 0.002, 46.1% of moderately rare variants having MAF $\in (0.002, 0.005]$, and 40% of less rare variants having MAF $\in (0.005, 0.01]$, because the fraction of deleterious variants declines with the increase in their MAFs (Subramanian, 2012). The characteristics of the damaging rare variants in the two base populations are summarized in Table 4.1. According to Table 4.1, more harmful rare variants exist in the African population than those in the European population, and the total allele frequency in the African population almost doubles that in the European population.

Table 4.1: Characteristics of deleterious rare variants in the base populations

		MAF		
Population	Total MAFs	$(0, 0.002]$	$(0.002, 0.005]$	$(0.005, 0.01]$
European	0.039	59	2	2
African	0.057	64	1	7

Variants with MAF $\leq$ 0.01 are defined as rare variants.

4.1.2.2 Generating individual genotypes and phenotypes

To generate individual genotypes, we sampled from the 100,000 haplotypes assuming HWE and allowing recombination in introns (but not in exons). To generate disease outcomes, we considered two additional risk models, one assuming equal attributable risk (AR) for each deleterious variant:

$$\log\{P(D=1)/P(D=0)\} = \alpha + \sum_{j=1}^{m} G_j \log(1 + \text{AR}/2\pi_j)$$

, and the other assuming equal odds ratio (OR) for each deleterious variant:

$$\log\{P(D=1)/P(D=0)\} = \alpha + \sum_{j=1}^{m} G_j \log(\mathrm{OR}),$$

where m is the total number of deleterious variants in exons, G_j and π_j are the genotype and MAF of the j-th deleterious variant, and α was set to -3 to achieve a disease rate of $\sim 5\%$. The first risk model implies that rarer variants have a larger effect on disease while the second implies that all variants have the same effect size regardless of the MAF.

4.1.2.3 Generating sequencing read count data

At a locus, we denote the number of reads mapped to the locus and the number of reads identical to the minor allele as T and R. We first draw the locus-specific error rate ϵ from a beta distribution that yields the pre-specified average rate. Then the per-sample depth T was simulated using the same two-step strategy as outlined in Hu et al. (2016). For more details about sampling error rate and depth, refer to Appendix 4.4.1. Note that at each locus, ϵ and T were sampled independently for cases and controls which allows different sequencing designs as in real studies. At last, similar to SeqEM, we simulate R given T and ϵ from a binomial distribution

$$P_\epsilon(R|T,G) = \begin{cases} \mathrm{Binomial}(T,\epsilon) & \text{if } G=0 \\ \mathrm{Binomial}(T,0.5) & \text{if } G=1 \\ \mathrm{Binomial}(T,1-\epsilon) & \text{if } G=2. \end{cases}$$

4.1.2.4 Sequencing designs

We considered the study scenarios with 500 cases sequenced at high depth and with varying number of controls that receive the fixed total investment of sequencing capacity. For cases, we fixed the average depth to 30× which is a typical good coverage depth for NGS studies (Telenti et al., 2016); we also set the average error rate at 0.1% (Lou et al., 2013) as cases of interest are often sequenced using state-of-the-art sequencing technology. For simplicity, we measured the sequencing capacity by the cost of total coverage for controls. We considered a total sequencing effort of approximately 10,000× per locus. Given the total sequencing effort, we sequenced different number of controls (n_0) by choosing the average depth (c_0) accordingly. To examine low- to high-coverage sequencing designs, we considered three settings including 1666@6× (sequencing 1666 controls at average depth 6), 1000@10×, and 334@30×. We also varied the average error rate between 0.1% and 0.5% (Li et al., 2011) to evaluate the impact of the accuracy of sequencing technology.

4.2 Results

When applying TASER on the simulated data, we considered two different weight functions: $w_j = 1/\sqrt{\pi_j(1-\pi_j)}$ and $w_j = 1$ ($j = 1, 2, \ldots, m$), and further obtained the p-value for the omnibus TASER. We first evaluated the type I error of TASER using the weighted burden test (WT) and the unweighted burden test (UT), and the omnibus TASER (OT), and summarized the results in Table 4.2. Because TASER use a score test, any choice of weight functions can control the type I error. Table 4.2 shows that UT tends to be slightly conservative for the scenario 334@30× and 1000@10× in both European and African populations. The type I error rate is slightly inflated for WT when the depth is low (i.e., 6×) and the error rate is high (i.e., 0.5%). OT has the correct type I error in all scenarios considered.

Figures S4.1 and S4.2 shows the power results for WT, UT, and OT with different combinations of n_0 and c_0 in European population using $\epsilon_0 = 0.1\%$ and $\epsilon_0 = 0.5\%$, respectively. Compared with the optimum test (i.e., WB under equal AR and UB

Table 4.2: Empirical type I errors for TASER with different weights.

		European			African		
n_0@c_0	ϵ_0	WT	UT	OT	WT	UT	OT
334@30×	0.1%	0.008	0.010	0.008	0.009	0.009	0.008
	0.5%	0.008	0.009	0.008	0.010	0.009	0.009
1000@10×	0.1%	0.009	0.009	0.008	0.010	0.008	0.008
	0.5%	0.009	0.008	0.008	0.010	0.008	0.008
1666@6×	0.1%	0.011	0.010	0.010	0.010	0.010	0.009
	0.5%	0.013	0.011	0.011	0.013	0.010	0.011

The nominal significance level is 0.01. Each entry is based on 10,000 replicates.

under equal OR), OT lost some power but was able to maintain power very close to that of the optimum test, indicating that OT is robust to poor weight choice.

Figure 4.1 contrasts the power of OT for different sequencing designs in controls for European population. With fixed sequencing effort (total 10,000× per locus), if the average error rate is low (i.e., 0.1%), power increases significantly when the number of controls increases from 334 to 1666 (the depth decreasing from 30× to 6×). However, if the average error rate is too high (i.e., 0.5%), low-coverage sequencing of more controls may reduce power compared with sequencing fewer controls at a moderate depth. As shown in Figure 4.1, under equal AR, the scenario 1666@6× yielded lower power than the scenario 1000@10× while under equal OR, the scenario 1666@6× still achieves the best power. In fact, with low depth in controls, erroneous reads are difficult to be distinguished from minor allele reads which thereby inflates the MAF in controls, especially for extremely rare variants. Consequently, the true association signals coming from extremely rare variants will be substantially reduced under equal AR. For equal OR, because the power to detect association depends on the total MAF—the higher the total MAF, the higher the power, hence, high sequencing error rates do not decrease the overall power much using controls that are poorly covered. The power plots for African populations are displayed in Figure 4.2, which exihibits the similar pattern as in Figure 4.1.

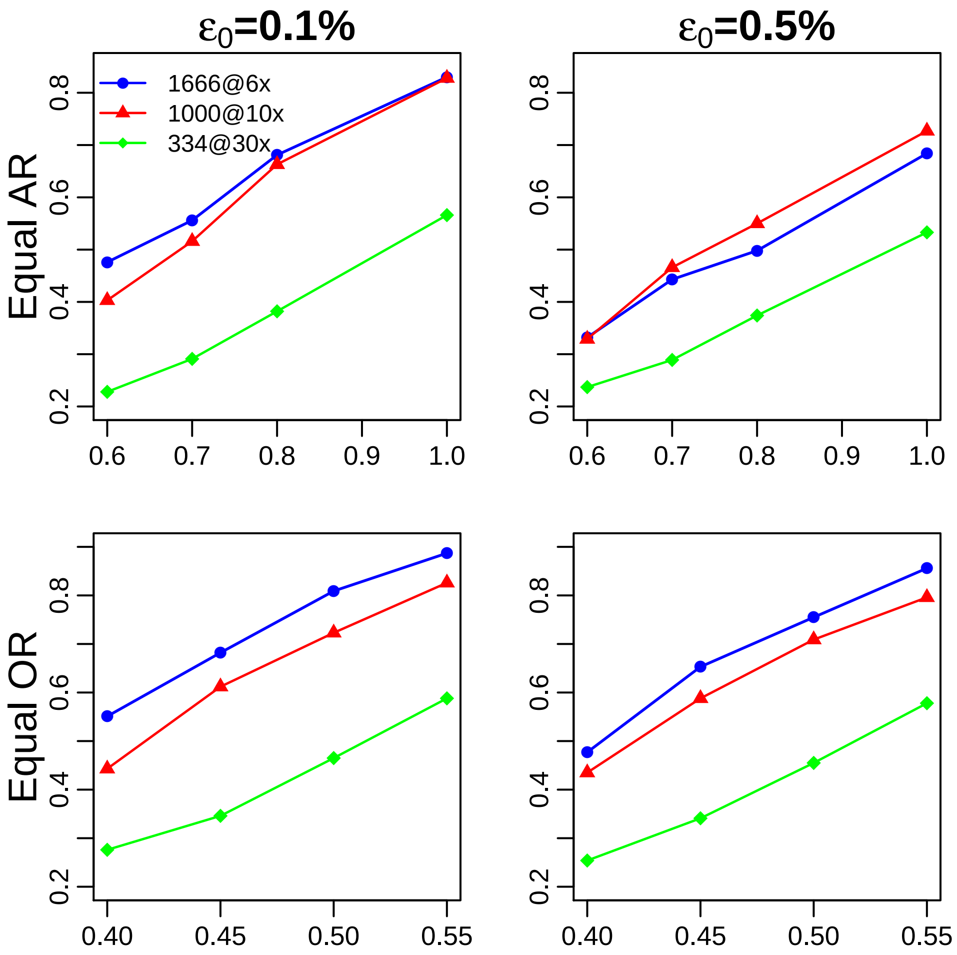

Figure 4.1: Power of the omnibus TASER for different combinations of n_0 and c_0 for European population.

4.3 Discussion

In this project, using TASER, we conducted a comparative analysis of low- to high-coverage sequencing in the design of rare variant association studies where cases are sequenced at high depth while sequencing controls face constant budget constraints. We found that, to maximize the power for detecting associations, deploying low-coverage sequencing to a large control cohort is generally more efficient than moderate- or high-coverage sequencing of a small control cohort. However, if the sequencing error

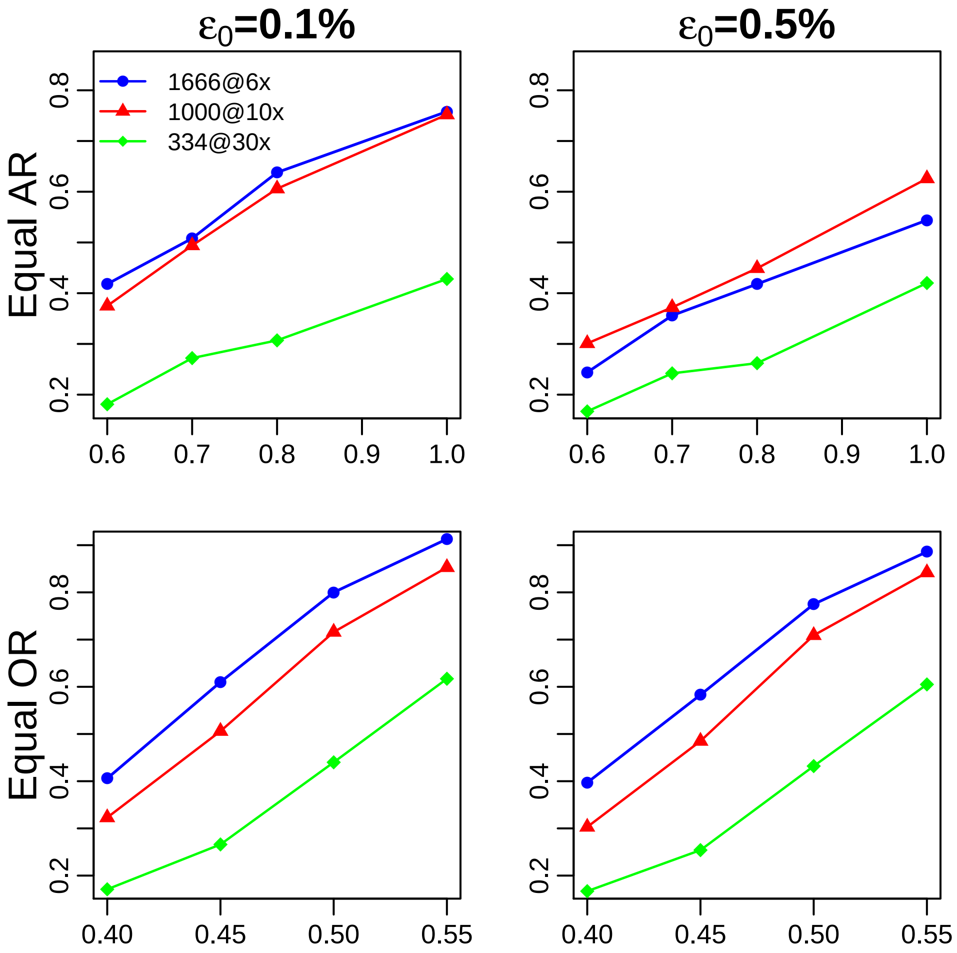

Figure 4.2: Power of the omnibus TASER for different combinations of n_0 and c_0 for African population.

rates of controls are very high and more rare variants are believed to have greater impact on the disease, the depth of controls cannot go too low. In this situation, sequencing controls at moderate coverage (e.g., 10×) is the most powerful in detecting associations.

Our simulated haplotypes preserve the realistic linkage disequilibrium (LD) structures though we do not make use of the LD information. TASER assumes independence across rare variants when the bootstrap replicates are generated. Because rare variants are not usually in high LD with other variants, the exploitation of LD information is not likely to change our conclusions.

Genetic architecture of diseases and traits differs and may be unclear. Here we considered two commonly used disease models to provide the insight into the influence of true risk models. A comprehensive understanding of the effect of genetic architecture on association power requires further research. For the omnibus TASER, we restricted the possible weight functions for rare variants on two special cases: equal weights and weights that are the inverse of the variance of the estimated MAF. Some more sophisticated weighting functions (Lin and Tang, 2011) can be implemented to improve the robustness of the omnibus TASER. However, it should be noted that combining more weight functions tends to reduce the power of the omnibus TASER.

Our simulation study has a number of limitations. First, the sequencing investment is simply represented by the cost of total coverage. In reality, sequencing costs consist of per-sample preparation costs, direct sequencing cost, and possibly the sample recruitment cost. However, we can aggregate all specific cost to estimate the per-unit sequencing cost, which can still be used to identify optimal designs. Second, we investigated European and African populations separately because the current TASER approach does not allow confounders such as principal components for ancestry. Future work should explore the optimal sequencing design using populations of various ethnicities or admixed populations. Finally, we did not include any protective variants in the genetic region under investigation. Burden tests have been shown to suffer from a dramatic loss of power when variants influence the disease in different directions or there is an excess of neutral variants. To focus on the key factors such as site frequency spectra and disease models, we do not consider the scenario where the gene carries both deleterious and protective variants.

4.4 Appendix

4.4.1 Simulating read count data

Given an average read depth, we generated the depth T at a locus for an individual by the same two-step strategy as described in Hu et al. (2016). Specifically, when the average depth was 30×, we generated the locus-specific mean depth c using Beta$(2.1, 4.1)$, which was then re-scaled to achieve the mean 30; given c, we generated the individual T's from NB$(c, 0.24)$. When the average depth was 10× or 6×, we generated c using Beta$(20.2, 22.9)$ (and then re-scaled to achieve the mean) and T using NB$(c, 0.2)$.

Given an average error rate, we sampled the locus-specific error rate from beta distributions. To achieve the average error rates of 0.1%, we used Beta$(8.76, 8755)$. To achieve the average error rates of 0.5%, we used Beta$(44, 8755)$ as used in Hu et al. (2016).

4.5 Supplemental Materials

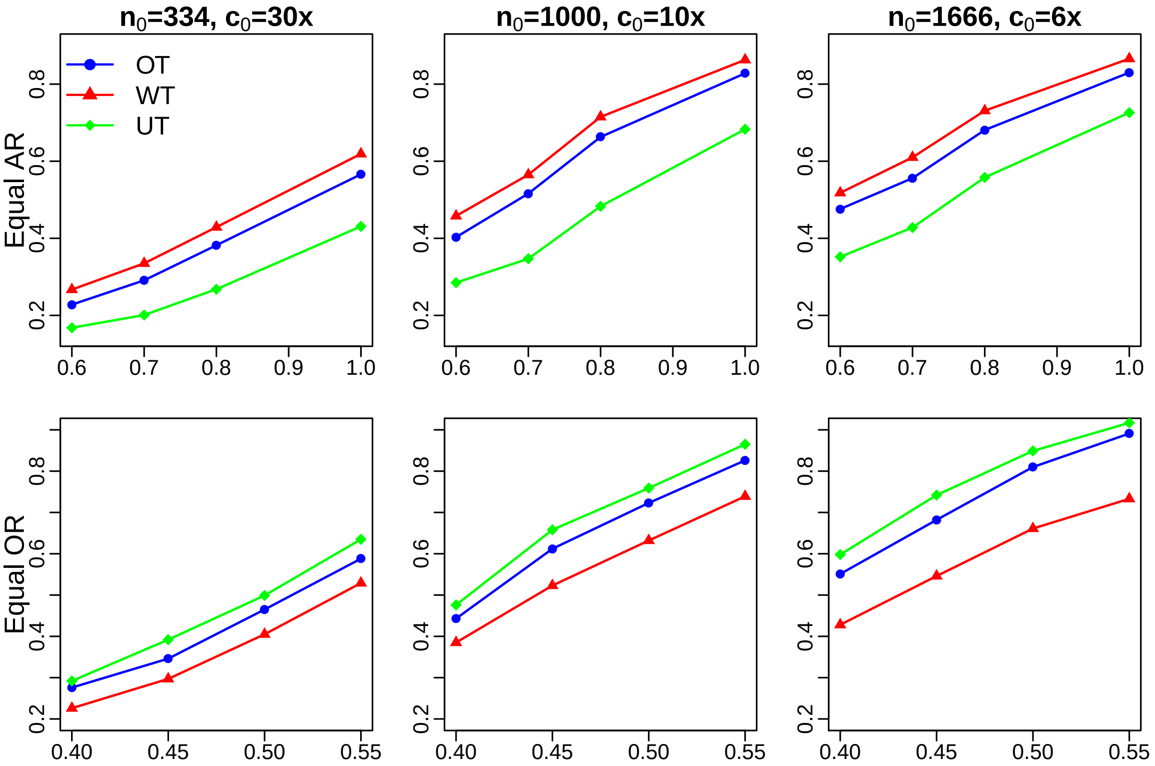

Figure S4.1: Power of TASER with different weights for different combinations of n_0 and c_0 for European population using $\epsilon_0 = 0.1\%$.

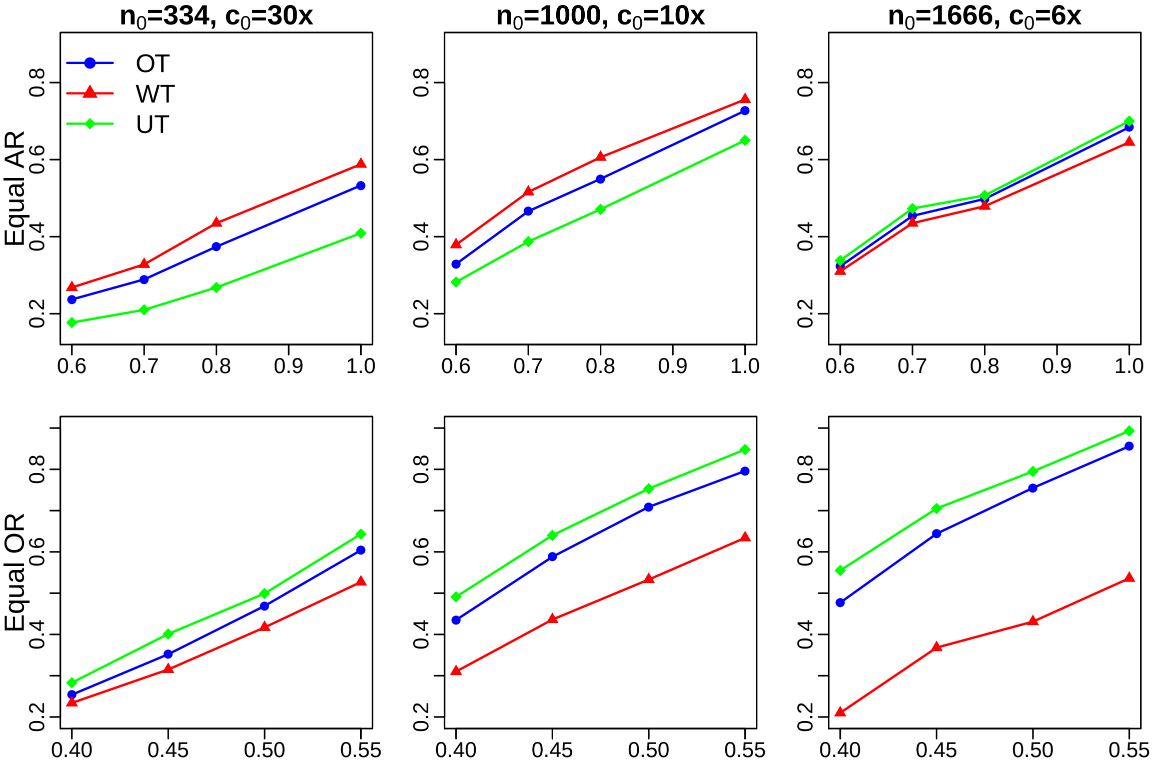

Figure S4.2: Power of TASER with different weights for different combinations of n_0 and c_0 for European population using $\epsilon_0 = 0.5\%$.

Chapter 5

Summary and Future Work

5.1 Summary

Recent advances in NGS allow researchers to examine the roles of variants across the full MAF spectrum, leading to great success in disease association studies and population genetic studies. The unique data structure and the diverse study designs present new challenges to statistical analysis. This dissertation aims to provide novel statistical approaches for some of the major challenges.

We first focused on the fundamental challenge in the analysis of NGS data, i.e., determining an individual's genotype correctly. To obtain accurate genotype calls, we have developed a *phred*-score-informed genotype-calling approach for NGS studies, named PhredEM. We also proposed a simple and computationally efficient screening algorithm to identify loci that are estimated to be monomorphic. The PhredEM approach improves the accuracy of genotype-calling by estimating base-calling errors from both read data and *phred* scores, and by using all sequencing reads available without removing reads according to an arbitrary threshold for the *phred* score. Simulation studies show that PhredEM performs better than SeqEM, in terms of the genotype-call error rate. We apply PhredEM to SCOOP data from the UK10K project and CEU data from the 1000 Genomes project to illustrate its advantages over both GATK and SeqEM.

Next, we focused on inferring population structure in the presence of differential sequencing qualities among different groups. We have developed a subsampling procedure and a read-flipping procedure to account for the systematic differences in read depth and error rates. To minimize loss of information, we repeat the subsampling and allele-flipping procedures and average the resulting variance-covariance matrices. We demonstrate that the PCs generated from our method do not capture any difference in sequencing qualities, unlike existing methods, with two examples using data from the 1000 Genomes Project, one involving three discrete Asian populations and the other involving a continuous admixture of two populations. The simulation studies further show the better performance of our method.

Finally, we explored the optimal design of NGS studies for testing rare variant associations using TASER developed by our group, which properly handles systematic differences in sequencing between cases and controls. We used realistic and informative simulation design to investigate how to effectively allocate sequencing resource between the sample size and read depth. We found that the best power was generally achieved by sequencing as many samples as possible (while decreasing depth if necessary). We noted, however, when the sequencing platform had a very high error rate (e.g., 1%) and rarer variants incurred higher risks, the best power was then achieved with a medium (e.g., 10x) depth.

5.2 Future Work

Some potential research topics related to the two projects are listed here.

1. **Accounting for population stratification in testing rare-variant association without calling genotypes**

   Hu et al. (2016) proposes a likelihood-based approach to test associations for rare variants that directly models sequencing reads without calling genotypes. Our proposed PCs in the second project, which are also based on sequencing reads, can be readily included in the likelihood-based association test. We can adopt a stratification-score-based strategy (Epstein et al., 2007; Allen and Satten, 2011; Epstein, Duncan, Jiang, Conneely, Allen and Satten, 2012), which is attractive due to the fact that the stratification score only needs to be calculated once and can be used for all regions. We use the stratification scores to weight the contribution of each case individual to the score function to obtain the score function that we would have seen if the confounding variable follows that distribution observed in controls. The next step is to generate bootstrap datasets that have the same amount of confounding as the original data. To this end, we carry out two separate analyses, one using weights that standardize the case-control population to the control population and one using weights that standardize to the case population, to obtain the distributions of allele frequencies at rare variants that we would have seen if the confounding variable of all subjects follows the distribution observed in controls and cases, respectively. By drawing case and control genotypes from their respective distributions, we build into bootstrap datasets the amount of confounding that is explained by stratification scores. The weighted approach may lead to a loss of power. Alternatively, we can model the confounding effect parametrically by specifying the allele frequency given the confounding variable.

2. **Incorporating *phred* scores in testing rare-variant association without calling genotypes**

   The methodology described in Hu et al. (2016) estimates error rates directly from the read data for each locus independently. Thus, considerable information regarding errors gained through the base-calling and alignment process is lost. It is possible to model the variability in error rates that is explained by base-calling and alignment quality scores. Our logistic model to relate error rates to *phred* scores in the first porject can replace the binomial model which assumes a constant error rate. We can also take other predictors of error rates into account in the logistic model such as the alignment error rates and the position of reads. This new approach, by exploiting more information, has the potential to further improve the statistical power for detecting rare variant associations.

3. **Testing for association for rare variants in case-parent trio studies**

   The methods developed in Hu et al. (2016) can be extended to trio studies. We consider developing methods that incorporate the family structure and characterize the association in terms of transmissions, while allowing the distribution of parental genotypes to be nonparametric.

Bibliography

Allen, A. S. and Satten, G. A. (2011), 'Control for confounding in case-control studies using the stratification score, a retrospective balancing score', American Journal of Epidemiology **173**(7), 752–760.

Balding, D. J. and Nichols, R. A. (1995), 'A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity', Genetica **96**, 3–12.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. and Shendure, J. (2011), 'Exome sequencing as a tool for Mendelian disease gene discovery', Nature Review Genetics **12**(11), 745–755.

Besag, J. and Clifford, P. (1991), 'Sequential monte carlo p-values', Biometrika **78**(2), 301–304.

Boyd, S. and Vandenberghe, L. (2004), Convex Optimization, Cambridge University Press, New York.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C. and Jaffe, D. B. (2008), 'Quality scores and SNP detection in sequencing-by-synthesis systems', Genome Research **18**, 763–770.

Browning, B. L. and Yu, Z. (2009), 'Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies', The American Journal of Human Genetics **85**, 847–861.

Cavalli-Sforza, L., Menozzi, P. and Piazza, A. (1993), 'Demic expansions and human evolution', Science **259**(5095), 639–646.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **39**, 1–38.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. (2011), 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', Nature Genetics **43**, 491–498.

Derkach, A., Chiang, T., Gong, J., Addis, L., Dobbins, S., Tomlinson, I., Houlston, R., Pal, D. K. and Strug, L. J. (2014), 'Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic', Bioinformatics **30**(15), 2179–2188.

Devlin, B. and Roeder, K. (1999), 'Genomic control for association studies', Biometrics **55**, 997–1004.

Epstein, M. P., Allen, A. S. and Satten, G. A. (2007), 'A simple and improved correction for population stratification in case-control studies', The American Journal of Human Genetics **80**(5), 921–930.

Epstein, M. P., Duncan, R., Broadaway, K. A., He, M., Allen, A. S. and Satten, G. A. (2012), 'Stratification score matching improves correction for confounding by population stratification in case-control association studies', Genetic Epidemiology **36**(3), 195–205.

Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S. and Satten, G. A. (2012), 'A permutation procedure to correct for confounders in case-control studies, including tests of rare variation', The American Journal of Human Genetics **91**(2), 215–223.

Ewing, B. and Green, P. (1998), 'Base-calling of automated sequencer traces using phred. II. Error probabilities', Genome Research **8**(3), 186–194.

Ewing, B., Hillier, L., Wendl, M. and Green, P. (1998), 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment', Genome Research **8**(3), 175–185.

Firth, D. (1993), 'Bias reduction of maximum likelihood estimates', Biometrika **80**, 27–38.

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A. and Nielsen, R. (2013), 'Quantifying population genetic differentiation from next-generation sequencing data', Genetics **195**(3), 979–992.

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B. A., Kleefstra, T., Brunner, H. G., Vissers, L. E. L. M. and Veltman, J. A. (2014), 'Genome sequencing identifies major causes of severe intellectual disability', Nature **511**(7509), 344–347.

Goldstein, D. B., Allen, A., Keebler, J., Margulies, E. H., Petrou, S., Petrovski, S. and Sunyaev, S. (2013), 'Sequencing studies in human genetics: design and interpretation', Nature Review Genetics **14**(7), 460–470.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J. and Levy, S. (2009), 'Evaluation of next generation sequencing platforms for population targeted sequencing studies', Genome Biology **10**(3), R32.

Hedges, D. J., Burges, D., Powell, E., Almonte, C., Huang, J., Young, S., Boese, B., Schmidt, M., Pericak-Vance, M. A., Martin, E., Zhang, X., Harkins, T. T. and Züchner, S. (2009), 'Exome sequencing of a multigenerational human pedigree', PLoS ONE **4**(12), e8232.

Hodgkinson, A. and Eyre-Walker, A. (2010), 'Human triallelic sites: evidence for a new mutation mechanism', Genetics **184**, 233–241.

Howie, B. N., Donnelly, P. and Marchini, J. (2009), 'A flexible and accurate genotype imputation method for the next generation of genome-wide association studies', PLoS Genetics **5**, e1000529.

Hu, Y., Liao, P., Johnston, R. H., Allen, A. S. and Satten, G. A. (2016), 'Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls', PLoS Genetics **12**(5), doi:10.1371/journal.pgen.1006040.

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., Sullivan, C. A., Walker, M. F., Waqar, Z., Wei, L., Willsey, A. J., Yamrom, B., Lee, Y.-h., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M. C., Ye, K., McCombie, W. R., Shendure, J., Eichler, E. E., State, M. W. and Wigler, M. (2014), 'The contribution of de novo coding mutations to autism spectrum disorder', Nature **515**(7526), 216–221.

Jackson, J. E. (2003), A User's Guide to Principal Components, John Wiley & Sons, New York.

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J. and Nielsen, R. (2011), 'Estimation of allele frequency and association mapping using next-generation sequencing data', BMC Bioinformatics **12**, 231.

Kircher, M., Stenzel, U. and Kelso, J. (2009), 'Improved base calling for the Illumina genome analyzer using machine learning strategies', Genome Biology **10**, R83.

Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. (2014), 'Rare-variant association analysis: study designs and statistical tests', The American Journal of Human Genetics **95**(1), 5–23.

Li, B. and Leal, S. M. (2008), 'Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data', The American Journal of Human Genetics **83**(3), 311–321.

Li, H. (2011), 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', Bioinformatics **27**(21), 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009a), 'The Sequence Alignment/Map format and SAMtools', Bioinformatics **25**(16), 2078–2079.

Li, H., Ruan, J. and Durbin, R. (2008), 'Mapping short DNA sequencing reads and calling variants using mapping quality scores', Genome Research **18**, 1851–1858.

Li, M., Nordborg, M. and Li, L. M. (2004), 'Adjust quality scores from alignment and improve sequencing accuracy', Nucleic Acids Research **32**, 5183–5191.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M. and R., A. G. (2011), 'Low-coverage sequencing: implications for design of complex trait association studies', Genome Research **21**(6), 940–951.

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010), 'MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes', Genetic Epidemiology **34**(8), 816–834.

Liao, P., Satten, G. A. and Hu, Y. (2017), 'PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies', Genetic Epidemiology **00**, 1–13.

Lin, D. and Tang, Z. (2011), 'A general framework for detecting disease associations with rare variants in sequencing studies', The American Journal of Human Genetics **89**(3), 354–367.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G. and Bustamante, C. D. (2008), 'Proportionally more deleterious genetic variation in european than in african populations', Nature **451**(7181), 994–997.

Lou, D. I., Hussmann, J. A., McBee, R. M., Acevedo, A., Andino, R., Press, W. H. and Sawyer, S. L. (2013), 'High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing', Proceedings of the National Academy of Sciences **110**(49), 19872–19877.

Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H.-E., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., Devlin, B., Roeder, K. and Trucco, M. (2008), ‘On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants’, The American Journal of Human Genetics **82**(2), 453–463.

Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., Lamb, C. A., McCarthy, S., Ahmad, T., Edwards, C., Serra, E. G., Hart, A., Hawkey, C., Mansfield, J. C., Mowat, C., Newman, W. G., Nichols, S., Pollard, M., Satsangi, J., Simmons, A., Tremelling, M., Uhlig, H., Wilson, D. C., Lee, J. C., Prescott, N. J., Lees, C. W., Mathew, C. G., Parkes, M., Barrett, J. C. and Anderson, C. A. (2017), ‘Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7’, Nature Genetics **49**(2), 186–192.

Ma, J. and Amos, C. I. (2012), ‘Principal components analysis of population admixture’, PLoS ONE **7**(7), 1–12.

Madsen, B. E. and Browning, S. R. (2009), ‘A groupwise association test for rare mutations using a weighted sum statistic’, PLoS Genetics **5**(2), 1–11.

Malaspinas, A.-S., Tange, O., Moreno-Mayar, J. V., Rasmussen, M., DeGiorgio, M., Wang, Y., Valdiosera, C. E., Politis, G., Willerslev, E. and Nielsen, R. (2014), ‘bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS)’, Bioinformatics **30**(20), 2962–2964.

Marchini, J. and Howie, B. N. (2010), ‘Genotype imputation for genome-wide association studies’, Nature Review Genetics **11**, 499–511.

Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S. and Morris, R. W. (2010), ‘SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies’, Bioinformatics **26**, 2803–2810.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010), 'The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data', Genome Research **20**, 1297–1303.

Menozzi, P., Piazza, A. and Cavalli-Sforza, L. (1978), 'Synthetic maps of human gene frequencies in Europeans', Science **201**(4358), 786–792.

Minoche, A. E., Dohm, J. C. and Himmelbauer, H. (2011), 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems', Genome Biology **12**, R112.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., Consortium, G., McVean, G., Boehnke, M., Altshuler, D. and McCarthy, M. I. (2015), 'The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease', PLoS Genetics **11**(4), 1–24.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K. and Daly, M. J. (2011), 'Testing for an unusual distribution of rare variants', PLoS Genetics **7**(3), 1–8.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. and Wang, J. (2012), 'SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data', PLoS ONE **7**(7), 1–10.

Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. (2011), 'Genotype and SNP calling from next-generation sequencing data', Nature Review Genetics **12**, 443–451.

Pardo-Seco, J., Martinón-Torres, F. and Salas, A. (2014), 'Evaluating the accuracy of AIM panels at quantifying genome ancestry', BMC Genomics **15**(1), 543.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., Sullivan, P. F., Bergen, S., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S. M., Haas, D. W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P. I. W., Reich, D. and Price, A. L. (2012), 'Extremely low-coverage sequencing and imputation increases power for genome-wide association studies', Nature Genetics **44**(6), 631–635.

Patterson, N., Price, A. L. and Reich, D. (2006), 'Population structure and eigenanalysis', PLoS Genetics **2**(12), e190. doi:10.1371/journal.pgen.0020190.

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J. and Sunyaev, S. R. (2010), 'Pooled association tests for rare variants in exon-resequencing studies', The American Journal of Human Genetics **86**(6), 832–838.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', Nature Genetics **38**(8), 904–909.

Pritchard, J. K. and Donnelly, P. (2001), 'Case–control studies of association in structured or admixed populations', Theoretical Population Biology **60**(3), 227 – 237.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000), 'Association mapping in structured populations', The American Journal of Human Genetics **67**, 170–181.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. and Sham, P. C. (2007), 'Plink: a tool set for whole-genome association and population-based linkage analyses', The American Journal of Human Genetics **81**(3), 559–575.

Reich, D., Price, A. L. and Patterson, N. (2008), 'Principal component analysis of genetic data', Nature Genetics **40**(5), 491–492.

Sampson, J., Jacobs, K., Yeager, M., Chanock, S. and Chatterjee, N. (2011), 'Efficient study design for next generation sequencing', Genetic Epidemiology **35**(4), 269–277.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J. and Altshuler, D. (2005), 'Calibrating a coalescent simulation of human genome sequence variation', Genome Research **11**, 1576–1583.

Schaid, D. J. and Sommer, S. S. (1993), 'Genotype relative risks: methods for design and analysis of candidate-gene association studies', The American Journal of Human Genetics **53**, 1114–1126.

Sham, P. C. and Purcell, S. M. (2014), 'Statistical power and significance testing in large-scale genetic studies', Nature Review Genetics **15**(5), 335–346.

Shen, Y., Song, R. and Pe'er, I. (2011), 'Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association', Bioinformatics **27**(14), 1995–1997.

Shlyakhter, I., Sabeti, P. C. and Schaffner, S. F. (2014), 'Cosi2: an efficient simulator of exact and approximate coalescent with selection', Bioinformatics **30**(23), 3427–3429.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. (2014), 'Sequencing depth and coverage: key considerations in genomic analyses', Nature Review Genetics **15**(2), 121–132.

Skoglund, P. and Jakobsson, M. (2011), 'Archaic human ancestry in East Asia', Proceedings of the National Academy of Sciences **108**(45), 18301–18306.

Skotte, L., Korneliussen, T. S. and Albrechtsen, A. (2012), 'Association testing for next-generation sequencing data using score statistics', Genetic Epidemiology **36**(5), 430–437.

Subramanian, S. (2012), 'Quantifying harmful mutations in human populations', European Journal of Human Genetics **20**(12), 1320–1322.

Tang, Z.-Z., Chen, G. and Alekseyenko, A. V. (2016), 'PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances', Bioinformatics **32**(17), 2618–2625.

Telenti, A., Pierce, L. C. T., Biggs, W. H., di Iulio, J., Wong, E. H. M., Fabani, M. M., Kirkness, E. F., Moustafa, A., Shah, N., Xie, C., Brewerton, S. C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B. A., Och, F. J., Turpaz, Y. and Venter, J. C. (2016), 'Deep sequencing of 10,000 human genomes', Proceedings of the National Academy of Sciences **113**(42), 11901–11906.

The 1000 Genomes Project Consortium (2010), 'A map of human genome variation from population-scale sequencing', Nature **467**, 1061–1073.

The International SNP Map Working Group (2001), 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', Nature **409**(6822), 928–933.

The UK10K Consortium (2015), 'The UK10K project identifies rare variants in health and disease', Nature **526**(7571), 82–90.

The Wellcome Trust Case Control Consortium (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', Nature **447**(7145), 661–678.

Tian, C., Kosoy, R., Lee, A., Ransom, M., Belmont, J. W., Gregersen, P. K. and Seldin, M. F. (2008), 'Analysis of East Asia genetic substructure using genome-wide SNP arrays', PLoS One **3**(12), e3862. doi:10.1371/journal.pone.0003862.

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., Branham, K. E., Heckenlively, J., The FUSION Study, Fulton, R., Wilson, R. K., Mardis, E. R., Lin, X., Swaroop, A., Zöllner, S. and Abecasis, G. R. (2014), 'Ancestry estimation and control of population stratification for sequence-based association studies', Nature Genetics **46**, 409–415.

Wang, C., Zhan, X., Liang, L., Abecasis, G. R. and Lin, X. (2015), 'Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation', The American Journal of Human Genetics **96**(6), 926 – 937.

Wood, S. N. (2006), Generalized Additive Models: An Introduction with R, Chapman & Hall/CRC.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011), 'Rare-variant association testing for sequencing data with the sequence kernel association test', The American Journal of Human Genetics **89**(1), 82–93.

Xu, C., Wu, K., Zhang, J.-G., Shen, H. and Deng, H.-W. (2016), 'Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study', Genetic Epidemiology **41**(3), 187–197.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. and Visscher, P. M. (2010), 'Common SNPs explain a large proportion of the heritability for human height', Nature Genetics **42**(7), 565–569.

Yu, Y. W., Yorukoglu, D., Peng, J. and Berger, B. (2015), 'Quality score compression improves genotyping accuracy', Nature Biotechnology **33**(3), 240–243.

Zawistowski, M., Reppell, M., Wegmann, D., St Jean, P. L., Ehm, M. G., Nelson, M. R., Novembre, J. and Zollner, S. (2014), 'Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests', European Journal of Human Genetics **22**(9), 1137–1144.