# Using Inheritance Vectors to Impute Genotypes and Detect Genotyping Errors

Charles Y. K Cheung

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Ellen M Wijsman, Chair
Elizabeth A Thompson
Volodymyr Minin

Program Authorized to Offer Degree:
Biostatistics – Public Health

University of Washington

## Abstract

Using Inheritance Vectors to Impute Genotypes and Detect Genotyping Errors

Charles Y. K Cheung

Chair of the Supervisory Committee:
Professor Ellen M Wijsman
Department of Medicine

Recent emergence of the common disease-rare variant hypothesis has renewed interest in the use of large pedigrees for identifying rare causal variants. Genotyping of dense variants using technologies that include next-generation sequencing platforms is common in the search for such variants. In my dissertation, I developed and implemented computationally efficient approaches that are suitable for imputing genotypes and detecting Mendelian consistent (MC) genotyping errors of dense variants on large pedigrees.

I developed a pedigree-based approach to impute dense genotypes. By leveraging information from existing genotypes already assayed from previous studies, my approach can facilitate cost-effective use of sequence data for genetic analysis in the pursuit of rare causal variants, especially on large pedigrees. This approach is based on the use of inferred inheritance vectors (IVs). In this approach, I first sampled IVs by using a Markov chain Monte Carlo

sampler that can handle large pedigrees. A set of IVs is sampled using genotypes from a sparse set of markers that may consist of existing genotypes. Using sampled IVs, I imputed genotypes by estimating the probability distribution of genotypes for each individual and for each marker. I showed that my approach allows us to call alleles with high accuracy. Using a real pedigree, I showed that my approach is substantially more effective in calling rare alleles than BEAGLE, which is a population-based imputation approach. In addition, I evaluated my approach under different conditions, which include framework marker types, density of framework panel, threshold for calling genotypes, and population allele frequencies on calling genotypes.

I also developed a pedigree-based approach to detect MC genotyping errors. Detection of genotyping errors is a necessary step to minimize false results in genetic analysis and is especially important when the rate of genotyping errors is high, as has been reported in the current next-generation sequence data. Similar to the genotype imputation approach, this error detection approach is based on the use of sampled IVs. Using sampled IVs, I proposed two test statistics to detect MC genotyping errors. Unlike existing approaches, my approach enables error detection on large pedigrees with many markers. Using simulations, I showed that my approach effectively detects MC genotyping errors. In addition, I evaluated the effectiveness of my approach as a function of parameters, including the genotype observed pattern, density of framework markers, error rate, allele frequencies, and number of sampled inheritance vectors.

I concluded my dissertation by documenting some future directions of my research. In particular, one topic is about providing guidance for sequencing choices in pedigrees. Because of the current cost of sequencing, investigators may only have resource to sequence a few subjects per pedigree, so we need to carefully prioritize who to sequence. I provided some ideas about using a statistical framework to compare among design choices of subject selection and proposed

a method to select subjects. This work may facilitate improved and informed sequencing decisions.

# TABLE OF CONTENTS

Page

## LIST OF FIGURES

Figure Number Page

## ACKNOWLEDGEMENTS

I thank my parents for their unconditional support. My parents have made major sacrifice to go to Canada to raise my sister and me. I am grateful that their decision has allowed me to have opportunities that I would not have otherwise. Throughout my life, I was given the freedom to pursue my dreams. I was never swayed into pursuing a certain type of career that might not fit me. Furthermore, I was never once requested to fulfill any family obligations, which to say at least, is very rare among Chinese families. Amazingly, the only things that they continue to expect from me are to get enough rest, eat the right food, and follow their health tips. I am forever appreciative for their nurture. Without their unconditional support, I would not have been able to concentrate on doing things that makes me proud.

I thank Professor Ellen Wijsman. I recall that during my phone interview with Ellen six years ago, I expressed that one of my main objectives of going to the graduate school was to strengthen my communication skills. In my last six years, I am grateful that Ellen has continued to work with me to help me improve. I am grateful for her mentorship, patience, and trust. All the coaching has slowly helped me understand that most of the skills in life are indeed learnable as long as we continue to work hard. As I am about to enter the next stage of my life, I have developed trust and confidence in myself.

I thank both Professor Elizabeth Thompson and Professor Ellen Wijsman for being my role models. I am grateful of their examples as responsible scientists. I am grateful for them of

leading with fairness and integrity. I am grateful for them of setting the necessary standards but at the same time having patience with their students.

I thank my instructors for their teaching. I thank the Department of Biostatistics, including Gitana, for help and support. I also thank dear Ada for her positivity and emotional support, dear sister and close friends, including Gabriel and Chrissy, and others not listed, for their genuine wishes and for believing in me.

# DEDICATION

To my family

# Chapter 1

# INTRODUCTION

Why do some people have certain traits? Advances in molecular biology in past decades enable us to explore the answer. A major focus in human genetics is to determine the genetic basis of trait phenotypes. In my dissertation, I document the contributions that I made to this field. Specifically, I have developed an approach to impute genotypes and an approach to detect Mendelian consistent genotyping errors for dense variants that are particularly applicable to large pedigrees.

## 1.1 Basic Biology

The genetic instruction of life is passed down from one generation to another by a molecule called Deoxyribonucleic acid (DNA). DNA consists of a sequence of bases chained together. Each base is one of four types: guanine, adenine, thymine, and cytosine. In humans, DNA molecules are packaged into chromosomes. A typical human has about 3.4 billion DNA base pairs organized into 46 pairs of chromosomes: 22 pairs of autosomes and 1 pair of sex chromosomes. Each cell contains a copy of these chromosomes. The collection of inherited DNA, which includes these chromosomes as well as mitochondrial DNA, are referred to as the human genome.

Efforts have been made to catalogue the genetic variation in humans [Gibbs, et al. 2003]. Genetic variants are sites in the human genome where differences among individuals have been identified. Different forms of a variant are called alleles. There are different types of genetic variants. For instance, a single nucleotide polymorphism (SNP) is a variation at a single nucleotide in the genome. A microsatellite, or short tandem repeat (STR), is variation in the number of repeating short DNA sequences at a particular location. Other types of variations, including deletion and copy number variation (CNVs), also exist. Genetic variation exists in a population because of occurrence of mutations, or rare changes to DNA sequences, that are passed down to the next generation. Studying genetic variation can help us understand the differences in phenotypes among individuals.

Genotyping is the determination of the allelic types of genetic variants. Different biological assays at the DNA-level have been developed to genotype different types of genetic variants. Array-based technologies are widely used for genotyping SNPs that have already been catalogued in human populations. With current technology, a genotyping array has probes for millions of SNPs. More recently, sequencing technologies are increasingly used. Unlike array-based technology, sequencing discovers un-catalogued variants, which includes variants that are rarely observed in populations. Hence, sequencing allows us to detect variants that have not been catalogued.

## 1.2 Identifying the Genetic Cause of Diseases

A central hypothesis is that genetic variation in humans influences some phenotypes. A strong support for this hypothesis is that some phenotypes are heritable in families. We conduct experiments to try to find regions on the human chromosomes that are associated with those heritable phenotypes. To conduct these experiments, we need to collect DNA samples from individuals and measure their phenotypes. We then analyze the data to find evidence of relationship between genetic variation and phenotypes.

Strategies used to identify disease genes have evolved considerably over the past few decades. Pedigrees have long been central to the discovery of genes for simple Mendelian traits, leading to the identification of nearly 4500 disease genes by the end of 2011 [Amberger, et al. 2011]. The use of pedigrees, however, has not lead to findings that immediately explain the genetic cause of many of the more complex diseases. Since then, a shift towards the hypothesis that some common variants greatly elevate the risk of some complex diseases has led to large collaborative efforts to first identify those common variants using Genome Wide Association Studies (GWAS) of large population-based samples [Collins, et al. 1997] . While GWAS have yielded many candidate loci [Manolio, et al. 2008], common variants now appear to explain only a small percentage of heritability [Manolio, et al. 2009]. Empirical evidence [Bodmer and Bonilla 2008; Cohen, et al. 2004; Gorlov, et al. 2011; Leigh, et al. 2008; Li, et al. 2009; Sanna, et al. 2011] also suggests that some heritable complex diseases may more suitably be explained by rare variants. This hypothesis is leading to a resurgence in the use of large pedigrees, because the analysis of sequence data collected in large pedigrees is a particularly efficient design for identifying rare variants that affect disease risk [Cirulli and Goldstein 2010; Ott, et al. 2011].

The reason that the large pedigree design is efficient stems from enrichment of rare functional alleles in such pedigrees. In a pedigree in which multiple subjects have the heritable phenotype of interest, it is likely that these subjects inherited copies of the same functional allele. Also, pedigrees offer other advantages such as the control for genetic heterogeneity and population stratification, direct estimation of genetic contribution of different loci, and analysis of effects of parental original of alleles [Ott, et al. 2011].

## 1.3 Inheritance

My work focuses on pedigrees and requires a basic understanding of inheritance. In a pedigree, founders are individuals whose parents are not specified. Non-founders are individuals who are descendants of some founders. A genetic locus is a location on a chromosome. The plural form of locus is loci. The Mendel's First Law states that at a genetic locus, each parent passes exactly a copy of his or her two alleles to an offspring, and each of the two copies has an equal chance to be transmitted. We refer to the process through which there is transmission of a copy of DNA as meiosis.

A chromosome that an offspring inherits from a parent may be a combination of parental DNA strands (Figure 1.1). Because of the biological process called genetic recombination, an offspring can inherit the maternal copy of a parent's DNA at one position and the paternal copy of the parent's DNA at another position. Nevertheless, because crossovers do not occur frequently on a chromosome, the chance that two nearby loci originate from the same parental chromosome is high. Genetic linkage describes the phenomenon that two DNA segments near

each other tend to be inherited together. A haplotype refers to a specific combination of alleles at multiple loci on a chromosome.

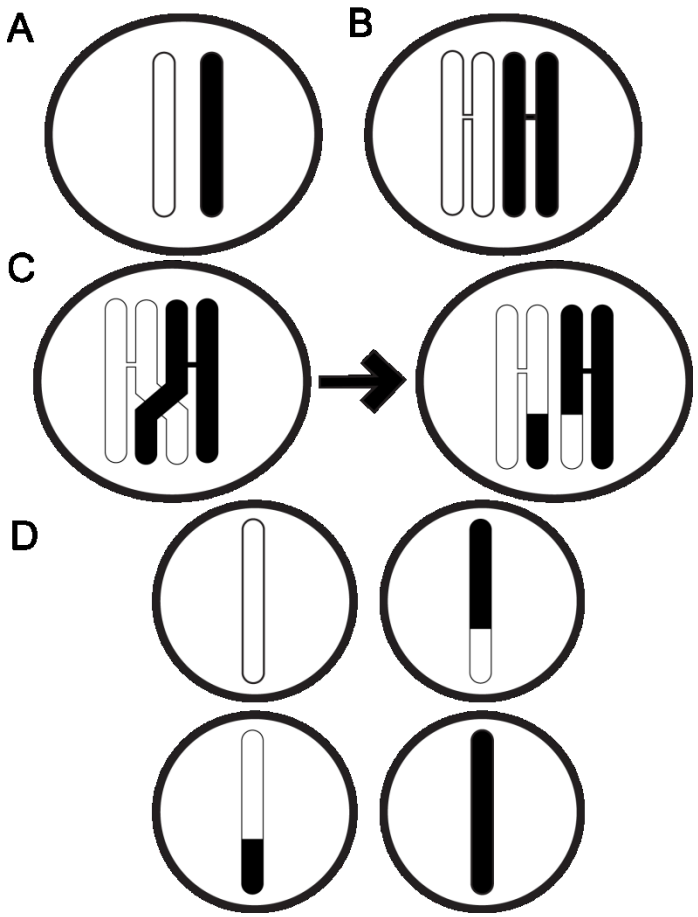



Figure 1.1 An example illustrating genetic recombination during meiosis: (A) a pair of homologous chromosomes; (B) each chromosome duplicates, producing two pairs of sister-chromatids; (C) exchange of genetic material between non-sister chromatids; D) 4 haploid gametes are produced, and 2 of the haploid gametes are recombinant chromosomes.

Here we define some terminology. A genetic marker is a genetic variant with known DNA location chosen for genetic analysis. Recombination fraction is the proportion of recombinants between two loci [Speed and Waterman 1996]. A map function relates

recombination fractions to distances in a genetic map, which specify the positions of genetic markers on a chromosome on an additive scale. A commonly used map function is the Haldane map function, which assumes that crossover events occur as a Poisson process, implying that recombination events occur independently across intervals. In other words, there is no interference between crossover events.

A concept central to inheritance is identity-by-descent (IBD). A pair of individuals is IBD at a locus if they inherit copies of the same ancestral chromosome. In the context of my work, the ancestral chromosome is carried by a founder in the pedigree. A pair of individuals either shares 0, 1, or 2 alleles IBD. For instance, a non-inbred parent and a child share 1 allele IBD, and a pair of siblings shares either 0, 1, or 2 copies of alleles IBD at a locus. Other more distantly related individuals may also be IBD at a locus. In the next chapter, we will discuss inheritance vectors (IVs), which represents IBD in a pedigree.

## 1.4 Overview of my Dissertation

In my dissertation, I first define IVs and an existing framework for inferring IVs using genetic data (Chapter 2). This framework enables me to develop an approach to impute missing genotypes (Chapter 3) and to detect genotyping errors (Chapter 4) in pedigrees. I discuss thoughts of how to select subjects for sequencing in pedigrees (Chapter 5). Finally, I conclude my dissertation by summarizing my contributions and discuss potential future directions (Chapter 6).

# Chapter 2

# INFERENCE OF INHERITANCE VECTORS

## 2.1 Overview of Chapter

Inheritance vectors (IVs) form the foundation of my work on genotype imputation and error detection. Therefore, in this chapter, we first define what IVs are. Then, we review an existing framework for inferring IVs probabilistically. This probabilistic framework is based on a hidden Markov model (HMM). Next, we briefly review an existing method that uses Markov chain Monte Carlo (MCMC) to infer IVs in large pedigree. Last, we discuss sampling of IVs at the positions of dense markers using IVs inferred **from** framework markers.

## 2.2 Terminology

IVs represent the transmission of DNA in a pedigree. Let $S_v$ denote the IV at locus $x$. $S_x = (S_{1x}, \dots, S_{Mx})$ is composed of a collection of segregation indicators $S_{mx}$, for $m = 1, \dots, M$, in a pedigree with $M$ meioses [Kruglyak, et al. 1996; Lander and Green 1987]. In a pedigree with $n_{nf}$ non-founders, $M = 2n_{nf}$. Segregation indicators are used to represent whether the maternal or paternal copy of DNA of a parent is transmitted from the parent to an offspring at a locus. Each segregation indicator $S_{mx}$ is a binary variable where

$S_{mx}$ = mat if DNA is transmitted from the parent's maternal DNA (the grand maternal DNA) at locus $x$ in meiosis $m$ and

$S_{mx}$ = pat if DNA is transmitted from the parent's paternal DNA (the grand paternal DNA) at locus $x$ in meiosis $m$.

## 2.3 The Hidden Markov Model Framework

Although the true IV at each locus is not observable, we can use observed genotype data at multiple loci to infer IVs using a HMM. A widely used algorithm to compute the likelihood of observed genotypes based on the HMM is the Lander-Green algorithm [Lander and Green 1987] (Figure 2.1). The Lander-Green algorithm is based on a model in which the observed data are the observed genotypes ($\boldsymbol{G} = (G_1, \ldots G_L)$) at locus $l = 1, \ldots, L$ and the hidden states are the IVs ($\boldsymbol{S} = (S_1, \ldots, S_L)$) at the corresponding loci. We specify the penetrance probability $P(G_l|S_l)$ as the probability of genotypes conditional on the IV at locus $l$. We also specify the transition probability $P(S_l|S_{l-1})$ as the probability of IV at locus $l$ conditional on the IV at the previous locus $l-1$. We use the first order Markov process to model IVs at the loci. Hence, the probability of the current event depends only on the most recent past event but not on any earlier events preceding the most recent past event.

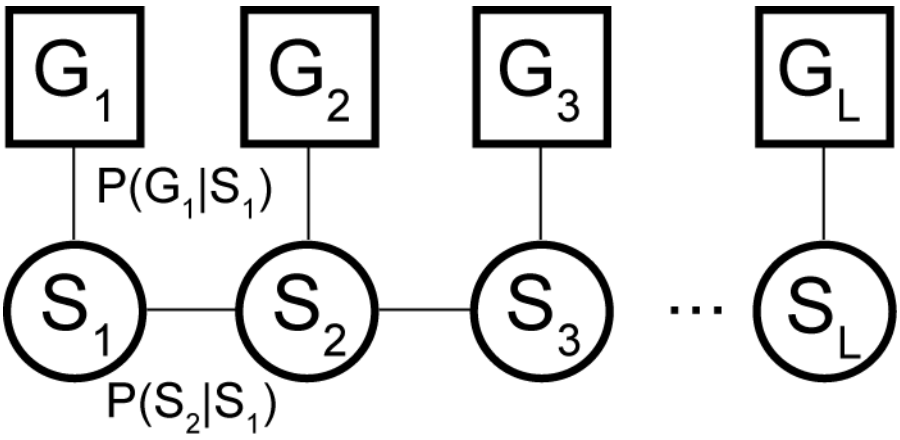



Figure 2.1 The HMM framework used in the Lander-Green algorithm

To calculate $P(G_l|S_l)$, we construct IBD-graphs using the $S_l$. First, we give a pair of distinct labels for each founder to represent their unique chromosomes. Collectively, we label founders with founder chromosomes labels from 1 to $2n_f$, where $n_f$ is the number of founders. We call these numerical labels founder genome labels (FGLs). Second, for each non-founder, from the top to the bottom generation, we label the transmitted chromosomes by FGL using information from $S_l$ to indicate which founding chromosomes are being transmitted. Third, we construct IBD-graphs. In each graph, the nodes are the FGLs. Each individual has a pair of FGLs. We create an edge by joining two FGLs for each individual who is observed for genotype at locus $l$. By joining FGLs, we form IBD-graphs**.** Let $ibdg_1$, $ibdg_2$,... denote the set of IBD-graphs that are constructed. Using these IBD-graphs, we can calculate $P(G_l|S_l)$ efficiently, as will be discussed below [Kruglyak, et al. 1996; Sobel and Lange 1996].

Before we discuss the calculation of $P(G_l|S_l)$, we first consider computing the unconditional probability of the genotypes in founders, as denoted by $P(G_l^F)$. Since the chromosomes of founders are assumed to be independent of each other, the alleles at a locus are

assumed to be sampled randomly from the population with probability equal to their allele frequencies denoted by $q_1, q_2, ....$ Therefore, $P(G_l^F)$ is just a product of the allele frequencies of the founder alleles.

The calculation of $P(G_l|S_l)$ is computationally efficient and is similar to the calculation of $P(G_l^F)$. Conditional on $S_l$, calculating the probability of observed genotypes only involves multiplying the allele frequencies of distinct copies of observed alleles. We convert $S_l$ into independent IBD-graphs $ibdg_z$, indexed by $z = 1,2, ...$, because the use of IBD-graphs enables us to identify which founder alleles are distinct. Let $G_l^z$ denote the set of genotypes of marker $l$ that are partitioned to $ibdg_z$. For each $ibdg_z$, there are at most 2 possible assignments, as denoted by $A_a^z$, $a = 1,2$, of how alleles can be matched to FGLs in the $ibdg_z$. For instance, in Figure 2.2, there is exactly 1 deterministic assignment of alleles to this IBD-graph because this IBD-graph contains a connected pair of nodes matching to a homozygous genotype. On the other hand, Figure 2.3 contains an IBD-graph in which 2 assignments of alleles to the IBD-graph are equally likely. Hence, for each $ibdg_z$, the probability of the genotypes partitioned to $ibdg_z$ and conditional on $A_a^z$, is equal to the product of population allele frequencies of the alleles assigned to the distinct FGLs. Hence,

$$P(G_l^z \,|ibdg_z, A_a^z) = \prod_j q_{jza},$$

where $q_{jza}$ is the allele frequency of allele $j$ belonging to $ibdg_z$ in assignment $A_a^z$.

Then, the probability of genotypes that are partitioned to $ibdg_z$ is

$$P(G_l^z \,|ibdg_z) = \sum_a P(G_l^z|ibdg_z, A_a^z)\, P(A_a^z).$$

Finally to calculate $P(G_l|S_l)$, we multiply the calculated probabilities of all partitions of the observed genotypes.

$$P(G_l|S_l) = \prod_z P(G_l^z \,|ibdg_z)$$

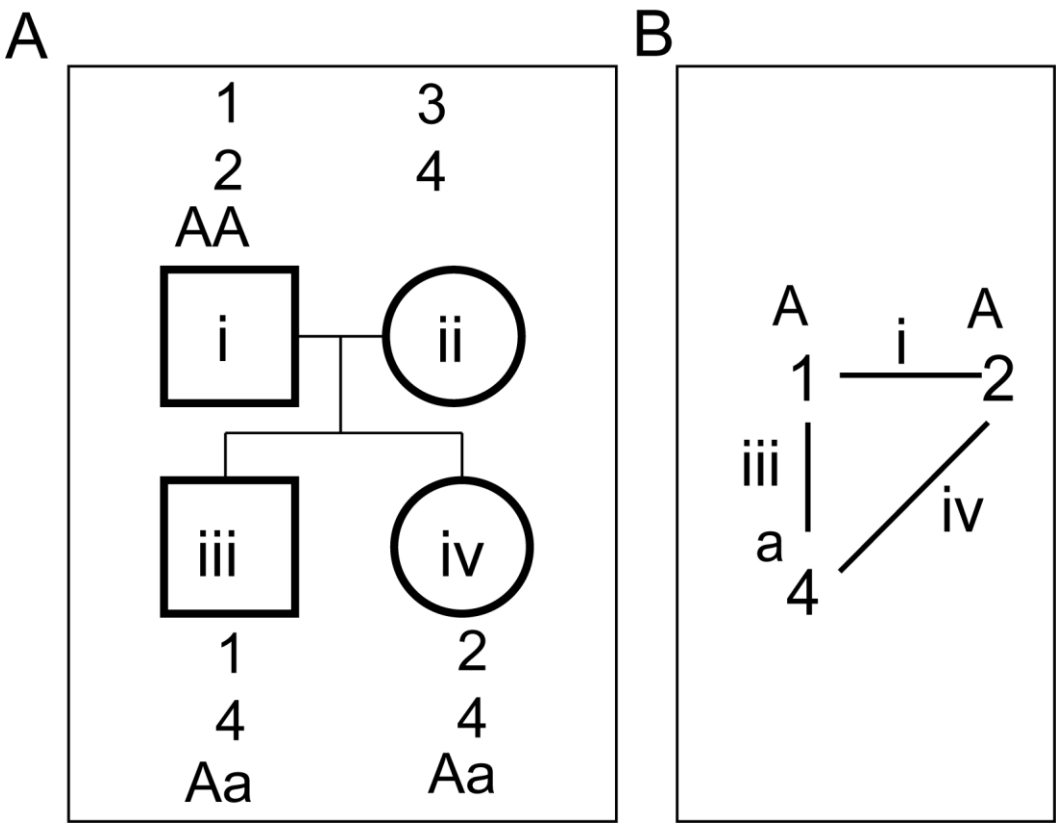



Figure 2.2 An example illustrating that only 1 assignment of alleles is consistent with this IBD-graph: A) In this pedigree, subjects are labeled by lower case roman numerals, FGLs are labeled by numbers, and observed genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A)

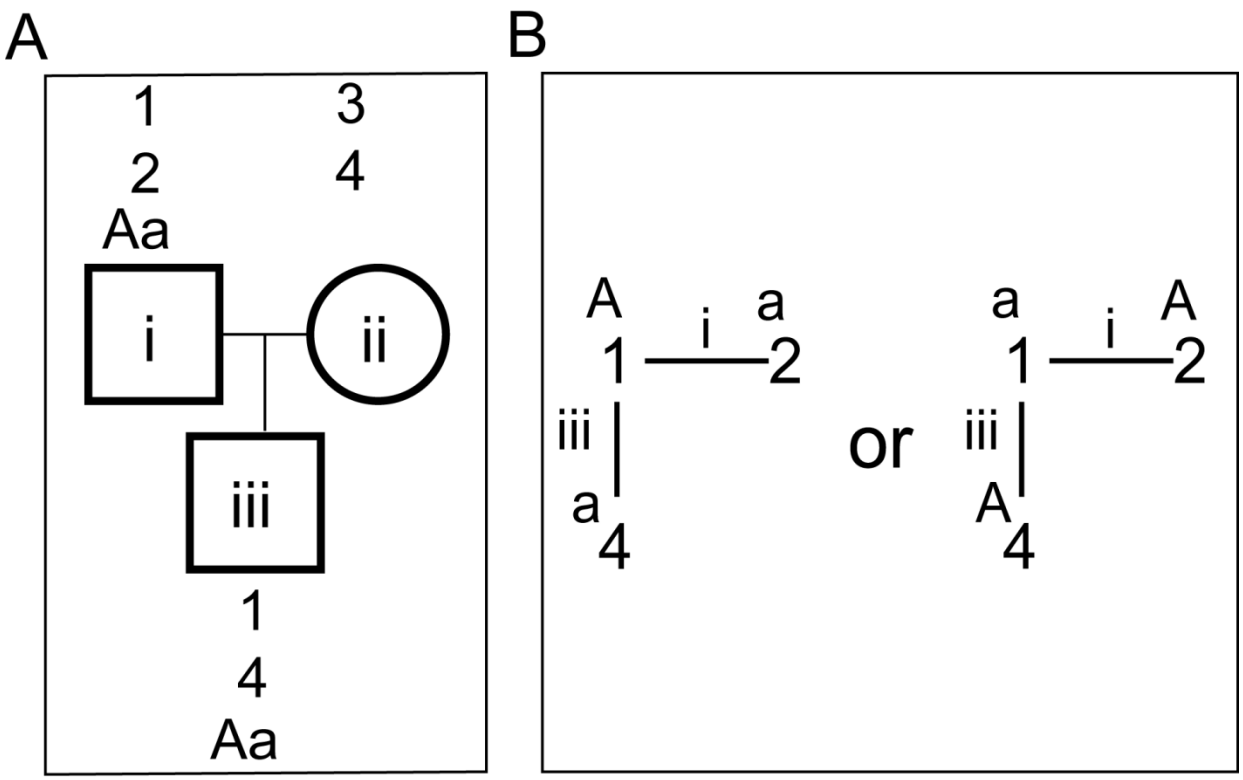



Figure 2.3 An example illustrating that 2 assignments of alleles are consistent with this IBD-graph: A) In this pedigree, subjects are labeled by lower case roman numerals, FGLs are labeled by numbers, and observed genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A)

We can also easily calculate $P(S_l|S_{l-1})$. Recall that $S_l = (S_{1,l}, \ldots, S_{M,l})$ for $M$ meioses. By Mendel's First Law, meioses are assumed to be independent. Therefore,

$P(S_l|S_{l-1}) = \prod_{m=1}^{M} P(S_{m,l}|S_{m,l-1})$.

$P(S_{m,l}|S_{m,l-1})$ is a function of the recombination rate in the interval between locus $l$ and $l-1$, as denoted by $\rho_l$. In my work, I assume that male and female meioses have the same recombination rates. Assuming the Haldane map function, the recombination rate between locus $l$ and locus $l-1$ as separated by the distance $d$ in Morgan units is $\rho_l = \frac{1}{2}(1 - e^{-2d})$.

$$P\left(S_{m,l} = s \middle| S_{m,l-1}\right) = 1 - \rho_l \qquad \text{if } s = S_{m,l-1}$$
$$= \rho_l \qquad \text{otherwise.}$$

The inference of IVs involves the evaluation of joint probabilities over loci conditional on genotypes. Fortunately, these calculations can be performed efficiently because the terms in the HMM can be factored and evaluated using the standard forward and backward algorithms [Baum, et al. 1970]. As a demonstration, we consider evaluating the likelihood $P(G)$.

I use the colon symbol as a symbol for inclusion in range. For example, let $G_{x:y} = (G_x, G_{x+1}, \dots, G_y)$ denote all the observed genotypes from locus $x$ to $y$. Also recall that for brevity, $\boldsymbol{G} = (G_1, \dots, G_L)$.

$$P(\boldsymbol{G}) = \sum_{S_{1:L}} P(G_{1:L}, S_{1:L} = s_{1:L})$$

$$= \sum_{S_{1:L}} P(S_1 = s_1) \prod_{l=2}^{L} P(S_l = s_l | S_{l-1} = s_{l-1}) \prod_{l=1}^{L} P(G_l | S_l = s_l)$$

$$= \sum_{S_{3:L}} \dots \sum_{S_2} P(S_3 = s_3 | S_2 = s_2) P(G_2 | S_2 = s_2) \sum_{S_1} P(S_2 = s_2 | S_1 = s_1) P(G_1 | S_1 = s_1)$$

$$P(S_1 = s_1)$$

In the equation above, we see that a seemingly cumbersome calculation can be factored into a calculation that depends on successively evaluating one summation at a time [Baum, et al. 1970]. In the factorization above, each summation consists of summing over $2^M$ operations, where $n_{nf}$ is the number of non-founders and $M = 2n_{nf}$ is the number of meioses in the pedigree. Since the Markov dependency is between a pair of loci and we have $L$ loci, we need to

perform computation on the order of $O(L2^M 2^M)$. Indeed, this calculation can be sped up by using a recursive divide-and-conquer algorithm on each meiosis to decrease the run time complexity to order $O(LM2^M)$ [Fishelson and Geiger 2004; Idury and Elston 1997].

## 2.4 Inferring IVs

Using the HMM, we can summarize inferred IVs in several ways: (1) the marginal probability distribution of IVs, (2) sampled set of IVs across loci, and (3) most likely IVs across loci.

(1) We calculate $P(S_l|G)$, the marginal probability distribution of IVs conditional on all observed genotypes. Here we discuss how to calculate such probability.

$$P(S_l|\boldsymbol{G}) = \frac{P(S_l, \boldsymbol{G})}{P(\boldsymbol{G})}$$

$$= \frac{P(G_{1:l}, S_l = s\ )P(G_{l+1:L}|G_{1:l}, S_l = s)}{P(\boldsymbol{G})}$$

$$= \frac{P(G_{1:l}, S_l = s\ )P(G_{l+1:L}|S_l = s)}{P(\boldsymbol{G})}$$

The last equality is true because of the conditional independence between $G_{1:l}$ and $G_{l+1:L}$ conditional on $S_l$. All three terms above are calculated similarly and efficiently using the forward or backward algorithm.

(2) We can also extend this calculation to calculate the joint distribution of IVs across loci, but because the number of joint IVs across loci is exponential to the number of loci, enumeration of all joint IVs states are not pursued. Instead, when the joint IVs are of interest, it is more practical to sample IVs jointly. In this HMM, we can obtain a Monte-Carlo sampling of IVs [Thompson 2000].

a. Calculate and store $P(S_l = s|G_{1:l})$ of all $s$ for $l = 2, \dots, L$ using forward computation

b. Calculate and sample successively from $P(S_l = s|\boldsymbol{G}, S_{l+1})$ from $l = L - 1$ to 1.

$$P(S_l = s|\boldsymbol{G}, S_{l+1}) = P(S_l = s|G_{1:l}, S_{l+1})$$

$$= \frac{P(S_l = s|G_{1:l})P(S_{l+1}|S_l = s)}{\sum_x P(S_l = x|G_{1:l})P(S_{l+1}|S_l = x)}$$

(3) In addition to sampling the IVs, the Viterbi algorithm [Viterbi 1967] can be used to find the most likely path of IVs across multiple loci. However, a disadvantage of using this approach is that it does not summarize uncertainty.

## 2.5 MCMC Sampling of IVs

We use the program gl_auto [Thompson 2011] in MORGAN to sample IVs at the positions of framework markers. The program samples IVs by using observed genotypes of the framework markers ($G_F^{ob}$) and population allele frequencies, in a manner similar to other pedigree-based linkage analysis methods [Abecasis, et al. 2002; Heath 1997; Kruglyak, et al. 1996; Lathrop, et al. 1984] . The program samples IVs from probabilities obtained by either exact or MCMC-based computation [Heath 1997; Thompson and Heath 1999; Tong and

Thompson 2008]. In sampling from exact probabilities, gl_auto uses the framework in the Lander-Green algorithm [Lander and Green 1987] to compute $P(S_l = s|G_{1:l})$, for each $s$ and each $l = 1, \ldots, L$. After that, gl_auto performs Monte Carlo sampling of IVs by the algorithm described in Section 2.4. However, this approach is computationally infeasible for large pedigrees. To handle large pedigrees, gl_auto uses a hybrid MCMC sampler based on both the Elston-Stewart [Elston and Stewart 1971] and Lander-Green algorithms, with components of this likelihood stored for subsequent efficient Monte Carlo sampling of IVs [Tong and Thompson 2008]. Evaluation of an older version of this hybrid sampler suggests that it outperforms SimWalk2, a widely used MCMC-based linkage analysis program, in terms of accuracy and computational speed when analyzing dense di-allelic markers typed on large and small pedigrees [Wijsman, et al. 2006]. Results have also shown that the current sampler in MORGAN performs even better than this older sampler [Tong and Thompson 2008]. We sample a set of IVs at the positions of the framework markers.

The MCMC sampler in MORGAN has been continuously refined over time. The original MCMC algorithm was a full-locus Gibbs sampler (L-sampler) that used reverse peeling (Heath, 1997). However, mixing can be poor for tightly linked markers [Daw, et al. 2005]. The meiosis sampler (M-sampler) was subsequently developed to sample IVs of a single meiosis jointly across loci [Thompson and Heath 1999]. This sampler was combined with the L-sampler to form the LM-sampler. The algorithm switches between the L and M samplers. This combined sampler performs better than either sampler and is irreducible. The multi-meiosis (MM) sampler allows the sampler to propose updates on several meioses simultaneously [Tong and Thompson 2008]. This scheme leads to substantial improvement on mixing when compared to the M-sampler. This sampler is combined with the L-sampler to form the LMM-sampler. The LMM sampler is

implemented in the program gl_auto from the MORGAN package [Thompson 2011].

## 2.6 From Sparse to Dense positions

To clarify what follows, we first present some terminology. We define a framework marker panel to be a relatively sparse set of markers that are used jointly to infer IVs along a chromosome of interest. The framework marker panel may consist of any marker types, including short tandem repeats (STRs), single nucleotide polymorphisms (SNPs), or a combination of these or other types of markers. These markers are assumed to be free of genotyping errors, even if in reality they are not. These framework markers are assumed to be in linkage equilibrium (LE) and may be genotyped on a high fraction of subjects in a pedigree. We define a dense marker panel to be additional markers that are not part of the framework panel.

In this section, I discuss a method to infer IVs at arbitrary dense positions. Because IVs are highly correlated across nearby positions, we can infer IVs sampled at arbitrary positions by using IVs sampled at the positions of framework markers. IVs sampled at dense positions can be well inferred without using dense markers because the use of a moderate number of framework markers can generally extract much of the information about the IVs in a pedigree [Wijsman, et al. 2006; Wilcox, et al. 2005].

### 2.6.1 Sampling IVs at the Positions of Framework Markers

We first infer IVs at sparse positions. Using gl_auto, we sample a set of IVs jointly across chromosomes at the positions of framework markers. These markers are relatively sparse and

assumed to be in linkage equilibrium.

Because framework markers are assumed to be sparse, we want to choose framework markers that are informative about meiotic transmission at the framework loci. Three factors increase the chance of picking informative markers. First, framework markers typed on a large number of subjects tends to be much more informative than markers typed on fewer subjects. Second, framework markers that are multi-allelic tend to be more informative than di-allelic markers. One measure for marker informativeness is the polymorphic information content [Botstein, et al. 1980]. Multi-allelic markers with equal allele frequencies are more informative than di-allelic markers with equal allele frequencies. Third, if framework markers are SNPs, markers with high minor allele frequencies tend to be more informative.

### 2.6.2 Sampling IVs at the Positions of Dense Markers

We first define some notation. Let $S_v$ denote the IV at the position of a dense marker $v$. The dense marker $v$ is flanked on the left by a framework marker $j$ and on the right by a framework marker $j+1$. Analogously, let $S_j$ denotes the IV at position $j$ and $S_{mj}$ denotes its segregation indicator for $m = 1, \ldots, M$.

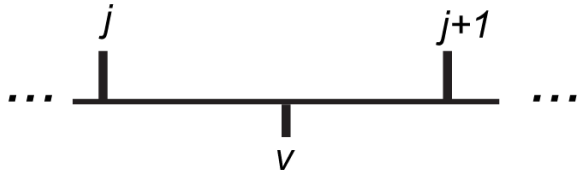



Figure 2.4 The dense marker  is flanked by two framework markers.

We seek to sample from $P(S_v = . \mid G_F^{ob})$, the probability distribution of IVs at the position of the dense marker $v$ conditional on observed framework markers. Let $F$ denote all framework markers. Since IVs are highly correlated across nearby positions, we infer IVs sampled at the position of the dense marker $v$ by only using IVs sampled at the positions of the framework markers. Given the Haldane Map function [Haldane 1919], IVs sampled at the positions of the closest flanking framework markers contain all information for the inference of IV at the position of the dense marker $v$ [Sobel, et al. 2001], i.e., $P(S_v = s_v | S_F = s_F) = P(S_v = s_v | S_j = s_j, S_{j+1} = s_{j+1})$, where $s_v$ and $s_F$ are configurations of IV at dense position $v$ and framework positions $F$. A set of IVs is sampled marginally for each dense marker $v$.

Since the $M$ meioses in a pedigree are independent, sampling $S_v$ corresponds to sampling each $S_{iv}$ independently. Under the Haldane map function,

$P(S_{mv} = s_{mv}, S_{m\,j+1} = s_{m\,j+1} | S_{mj} = s_{mj}) = P(S_{mv} = s_{mv} | S_{mj} = s_{mj}) P(S_{m\,j+1} = s_{m\,j+1} | S_{mv} = s_{mv})$, where $s_{mv}$ specifies whether the chromosome is inherited maternally or paternally at position $v$.

Since $P(S_{mv} = s_{mv} | S_{mj} = s_{mj}, S_{m\,j+1} = s_{m\,j+1}) = \frac{P(S_{mv} = s_{mv}, S_{m\,j+1} = s_{m\,j+1} | S_{mj} = s_{mj})}{P(S_{m\,j+1} = s_{m\,j+1} | S_{mj} = s_{mj})}$, it is straightforward to sample $S_{mv}$ conditional on the IVs at the flanking markers.

Each term is calculated easily given the Haldane map function. At position $v$, one IV ($S_v^k$) is sampled from jointly-sampled IVs obtained at positions $j$ and $j+1$. We repeat this process to sample a total of $n$ such IVs. This set of $S_v^k$, for $k = 1, \ldots, n$, provides an estimate of the probability $P(S_v = s_v | G_F^{ob})$, since

$$P\left(S_v = s_v \middle| G_F^{ob}\right) = \sum_{s_F} P\left(S_F = s_F \middle| G_F^{ob}\right) P\left(S_v = s_v \middle| S_F = s_F, G_F^{ob}\right)$$

$$= \sum_{s_F} P\left(S_F = s_F \middle| G_F^{ob}\right) P(S_v = s_v | S_F = s_F)$$

$$= \sum_{s_F} P\left(S_F = s_F \middle| G_F^{ob}\right) P(S_v = s_v | S_j = s_j, S_{j+1} = s_{j+1})$$

Then $\hat{P}\left(S_v = s_v \middle| G_F^{ob}\right) = \frac{1}{n}\sum_{k=1}^{n} P(S_v = s_v^k | S_j = s_j^k, S_{j+1} = s_{j+1}^k)$ is the required estimate, since $S_F$ is realized from $P\left(S_F = s_F \middle| G_F^{ob}\right)$.

# Chapter 3

# GENOTYPE IMPUTATION

In this chapter, I first discuss the background and motivation for imputing genotypes. Then, I review existing methods. Next, I present a computationally efficient approach for imputing dense genotypes on large pedigrees, which is implemented in the program GIGI (Genotype Imputation Given Inheritance). My MCMC-based approach uses only a sparse set of markers typed on most subjects, plus dense markers typed on a few subjects. By analyzing both simulated and real data, I demonstrate that my approach can impute many alleles accurately on many subjects in large pedigrees, even including some completely unobserved subjects. In addition, I evaluate parameters that affect imputation quality and demonstrate that our imputation approach is substantially more accurate for imputing rare alleles in a large pedigree than the use of the state-of-the-art population-based approach in BEAGLE.

## 3.1 Background and Motivation

While essential to the identification of causal variants, generation of very dense genotypes from platforms that include next-generation sequencing technologies is both expensive and challenging. First, the total cost of producing dense genotypes on many subjects remains expensive, especially for sequence data. Nevertheless, it can be important to carry out a deep and comprehensive analysis of all variants in a region of interest in order to reach a

conclusion about a causal locus [Musunuru, et al. 2010; Rosenthal, et al. 2011]. Second, it is not always possible to produce genotypes on all subjects because of the quality and quantity of available DNA. This issue is particularly acute in the case of high-throughput sequencing. Together, these two potential issues can inhibit optimal analyses. One solution is to genotype a subset of individuals and carry out genotype imputation to infer missing genotypes on unobserved subjects. Genotype imputation is a cost-effective approach to leverage existing genotype data, which are often available on many subjects, with new dense genotypes collected on just a few subjects.

## 3.2 Existing Methods for Genotype Imputation

Population-based and pedigree-based genotype imputation methods exist. Genotype imputation, as a general example of imputation [Little and Rubin 1987], typically infers missing data by borrowing information from correlated observations. Imputation in population-based samples leverages information from the correlation among dense markers due to linkage disequilibrium (LD) observed in outside reference samples of unrelated individuals [Browning and Browning 2007; Delaneau, et al. 2011; Howie, et al. 2009; Li, et al. 2010; Scheet and Stephens 2006; Stephens, et al. 2001]. In contrast, imputation in pedigrees uses the correlation of genotypes among relatives derived from sharing of genomic segments identical by descent (IBD) within pedigrees. For small pedigrees**,** Burdick and colleagues developed an imputation method and applied it to imputation of dense genotypes [Burdick, et al. 2006; Chen and Abecasis 2006]. In their study, sparse markers were available in all generations and dense markers were only available in the top two generations. They demonstrated that dense genotypes can be imputed in

the bottom generation. Rule-based long-range phasing methods, which detect long strings of non-conflicting homozygous genotypes to identify shared haplotypes between relatives, have also been developed [Daetwyler, et al. 2011; Kong, et al. 2008].

These existing genotype imputation methods have major limitations for use in large pedigrees. Population-based methods cannot impute genotypes on relatives who are completely unobserved for marker genotypes when used in the context of ignored pedigree structure. In addition, while high imputation accuracy can often be achieved when sufficient numbers of subjects in a reference panel are available [Browning and Browning 2009]**,** imputation of rare variants is particularly difficult [Krithika, et al. 2012; Li, et al. 2011]. Existing pedigree-based methods also have major limitations. Burdick's method cannot handle large pedigrees with many markers due to computational constraints. While existing rule-based methods [Daetwyler, et al. 2011; Kong, et al. 2008] can handle large pedigrees, they are *ad hoc*. In addition, they require high-quality dense genotype data on subjects for whom we want to impute data and do not account for recombination events.

## 3.3 My Approach to Impute Genotypes in Pedigrees

The dense markers to impute can be genotypes obtained from sequence data or from a dense SNP panel. These genotypes may be typed on fewer and even different subjects in the pedigree.

### 3.3.1 Overview

My goal is to impute genotypes of dense markers on the unobserved subjects. Here the

imputation relies on correlation due to inheritance in the pedigree. The inheritance of shared segments of chromosome is represented by inheritance vectors (IV) [Lander and Green 1987]. The imputation approach consists of four steps:

(1) Sample IVs at the positions of the framework markers conditional on the observed genotypes at the framework markers.

(2) Sample IVs at the positions of the dense markers conditional on the IVs sampled at the positions of the framework markers and the meiotic map.

(3) Estimate the probability distribution for each unobserved genotype at the dense marker positions conditional on all observed dense genotypes, known or estimated allele frequencies for the dense markers, and position-specific IVs corresponding to the dense markers.

(4) Call genotypes using the estimated probabilities and user-specified thresholds.

### 3.3.2 Details

Imputing dense genotypes

I estimate the probability distribution of the missing genotype of subject $p$ of dense marker $v$ ( $G_{pv}$), conditional on the observed genotypes of all framework markers ($G_F^{ob}$), the observed genotypes ($G_v^{ob}$) of dense marker $v$, and the allele frequencies of dense marker $v$. For each genotype configuration $g$, our estimator is based on the calculation:

$$P\left(G_{pv} = g \,\middle|\, G_F^{ob}, G_v^{ob}\right) = \sum_s P\left(G_{pv} = g \middle| S_v = s, G_F^{ob}, G_v^{ob}\right) P\left(S_v = s \middle| G_F^{ob}, G_v^{ob}\right)$$

$$\cong \sum_s P\left(G_{pv} = g \middle|\, S_v = s, G_F^{ob}, G_v^{ob}\right) P\left(S_v = s \middle| G_F^{ob}\right) \quad (1)$$

$$\cong \sum_s P\left(G_{pv} = g \middle|\, S_v = s, G_v^{ob}\right) P\left(S_v = s \middle| G_F^{ob}\right) \quad (2)$$

Equation (1) is an exact equality if dense marker $v$ is one of the framework markers: *ie.* $G_v^{ob} \subseteq G_F^{ob}$ . In general, equation (1) is a good approximation when $P\left(S_v = s \middle| G_F^{ob}, G_v^{ob}\right) \cong P\left(S_v = s \middle| G_F^{ob}\right)$, which says that the inference of IVs at the position of dense marker $v$ is not influenced much by the addition of the genotypes of dense marker $v$, given that we already observe the genotypes of the framework markers. Equation (2) is a good approximation when $P\left(G_{pv} = g \middle| S_v = s, G_F^{ob}, G_v^{ob}\right) \cong P\left(G_{pv} = g \middle|\ S_v = s, G_v^{ob}\right)$. Indeed, this approximation is an exact equality if the framework markers are in linkage equilibrium with dense marker $v$, as is assumed in the Lander-Green algorithm [Lander and Green 1987]. See Appendix A for further discussion. $P\left(G_{pv} = g \middle| S_v = s, G_v^{ob}\right)$ in equation (2) is calculated by

$$P\left(G_{pv} = g \middle| S_v = s, G_v^{ob}\right) = \frac{P\left(G_{pv}=g, G_v^{ob} \middle|\ S_v=s\right)}{\sum_k P\left(G_{pv}=k, G_v^{ob} \middle|\ S_v=s\right)}. \quad (3)$$

Each term in equation (3) can be computed efficiently [Kruglyak, et al. 1996; Sobel and Lange 1996]. The second term of equation (2) is estimated by the sampled IVs at position $v$. Since IVs are sampled conditionally on $G_F^{ob}$, equation (2) provides a Monte Carlo estimator for imputation of $G_{pv}$.

$$\hat{P}\left(G_{pv} = g \middle| G_F^{ob}, G_v^{ob}\right) = \frac{1}{n}\sum_{k=1}^{n} P\left(G_{pv} = g \middle| S_v^k, G_v^{ob}\right), \quad (4)$$

where $S_v^k$ is the IVs sampled at iteration $k$, for $k = 1, \ldots, n$. Equation (4) assumes that all $S_v^k$ are consistent with the observed genotypes of dense marker $v$. For practical purposes, we propose a modified estimator (5) that is based only on the sampled IVs that are consistent with the observed genotypes of marker $v$ :

$$\hat{P}\left(G_{pv} = g \middle| G_F^{ob}, G_v^{ob}\right) = \frac{1}{n^*}\sum_{k=1}^{n} P\left(G_{pv} = g \middle| S_v^k, G_v^{ob}\right), \quad (5)$$

where $n^* = \sum_{k=1}^{n} I(P(G_v^{ob} |\ S_v^k) > 0)$. $I(P(G_v^{ob} |\ S_v^k) > 0)$ is an indicator that the IVs sampled at

iteration $k$ is consistent with $G_F^{ob}$. Thus, $n^*$ is the number of sampled IVs that are consistent with the observed genotypes at dense marker $v$. A more thorough discussion of the estimators is presented in Appendix A.

Calling Genotypes

While I can leave the imputed results as estimated probabilities, we can also call genotypes. Using a confidence-based genotype-calling approach, I call both alleles if $\hat{P}\left(G_{pv} = g \middle| G_F^{ob}, G_v^{ob}\right) > t_1$, where $t_1$ is a user-defined threshold. In allele-calling, I first use genotype-calling. If we cannot call the complete genotype, we call one of the two alleles if $\hat{P}(G_{pv} = a/. | G_F^{ob}, G_v^{ob}) > t_2$, where $a/.$ denotes that the genotype contains an $a$ allele. While this second threshold $t_2$ can be arbitrary, we set $t_2 = t_1 + \frac{1-t_1}{2}$. A reason for this choice is that for a di-allelic marker, the algorithm will select the more probable genotype configuration when the estimated probability of the heterozygous configuration is equal to $t_1$. Besides the confidence-based genotype-calling approach, we can alternatively call the most-probable genotype. In this approach, a genotype call is always made.

Note:

To achieve computational efficiency, I use the approximation that $P\left(S_v = s \middle| G_F^{ob}, G_v^{ob}\right) \cong P\left(S_v = s \middle| G_F^{ob}\right)$. Making this approximation allows us to sample IV by using only framework markers. This approximation states that the knowledge of $G_v^{ob}$does not dramatically influence the inference of $S_v$ given that $G_F^{ob}$ is already observed. This approximation is reasonable because meiotic events on a chromosome do not occur frequently and the use of a moderately sparse set of markers can often extract much of the information of the IV in a pedigree [Wijsman, et al. 2006; Wilcox, et al. 2005].

## 3.4 Evaluating Imputation Performance using Simulation

### 3.4.1 Measuring Quality

I used three metrics to evaluate imputation quality. Call rate measures the percentage of alleles called, accuracy measures the percentage of alleles called correctly among the alleles called, and consistency measures the percentage of IVs that are consistent with the observed genotypes at a marker locus. In real data, these metrics were calculated by averaging over all marker loci and across all subjects. In simulated data, these metrics were further averaged over all simulation replicates. In addition, I summarized the call rate by subject.

### 3.4.2 Simulated Data

*Pedigrees*

I simulated data on a 5-generation pedigree of 52 subjects (Figure 3.1). While this pedigree is beyond the limit of exact computational methods for multipoint computation, the use of gl_auto's MCMC option enabled computation on this large pedigree. I used simulated descent patterns from a previous study [Wijsman, et al. 2006] to obtain genotypes in non-founders after simulating genotypes in founders. I analyzed several replicates with different descent patterns. Results from the first 10 replicates gave consistent interpretation and were therefore deemed a sufficient sample size.

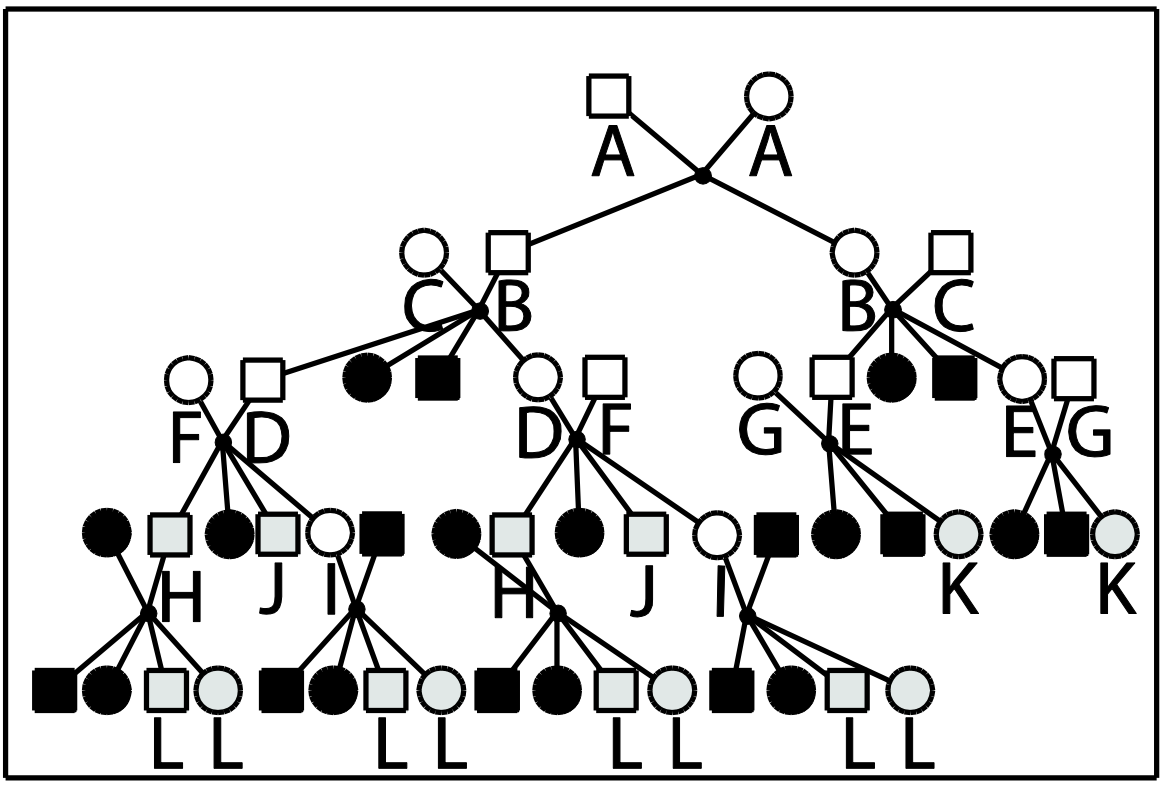



Figure 3.1 Simulated pedigree of 52 subjects. Different designs of subjects observed for genotypes are indicated by different shading schemes: all subjects (shaded or not shaded); many subjects (any shaded); and few subjects (black shaded). Classes of subjects are indicated by letters.

*Markers*

I simulated both framework and dense markers on a chromosome of 100 centimorgan (cM). I simulated two types of framework markers: di-allelic and 4-allelic. The di-allelic markers with uniform allele frequencies spaced uniformly at one marker per 0.5 cM represented a SNP linkage panel. The 4-allelic markers with uniform allele frequencies spaced uniformly at one marker per 4 cM represented a STR marker panel. The 4-allelic markers represent what may exist in a region where there has been some follow-up genotyping. To examine a density of STR

markers that is more commonly available in legacy samples in initial genome scans, I also thinned the STRs to a density of one marker per 8 cM. To test the effect of framework marker density and observed data patterns, I also thinned the framework SNP markers and varied the number of observed subjects, as described below. In addition to the framework markers, I simulated 25,000 uniformly-spaced di-allelic dense markers at a density of one marker per .004 cM with population allele frequencies simulated from the uniform [0, 1] distribution. These markers approximate markers from SNP-chip or variants that might be available from high-throughput sequencing. After generating the complete marker dataset by simulating the founder alleles and gene-dropping through the descent patterns as described above, I retained dense SNPs in only 22 subjects, as indicated by the unlabeled subjects in Figure 1. I imputed genotypes on the 30 labeled subjects.

### 3.4.3 Analysis of Simulated Data

I varied the number of subjects typed and choices of framework marker panels (Table 3.1). I considered three designs where all (52), many (36), or few (22) subjects were genotyped for framework markers (Figure 3.1). The design where all subjects were genotyped for framework markers ($S_a$) was unrealistic but provided a benchmark for optimal inference of IVs. The other designs captured more realistic situations where only subjects from the more recent generations were available. Besides the number of subjects observed for genotypes, I considered different types of framework marker panels: SNP-only (S), STR-only (M), and hybrid SNP and STR panels of various marker densities (Table 3.1). The hybrid panels ($S_fM^4_m$ and $S_fM^8_m$) captured the situation where STR markers observed on many subjects from older studies were combined with newly collected dense markers observed on fewer subjects.

Table 3.1. Designs of framework marker panels used in simulation

| **Designs** | **Number of Subjects typed** | |
|---|---|---|
| | **SNPs** | **STRs** |
| $\mathbf{S_a}$ | 52 | - |
| $\mathbf{S_m}$ | 36 | - |
| $\mathbf{S_f}$ | 22 | - |
| $\mathbf{S_f M^4_m}$ | 22 | 36 |
| $\mathbf{M^4_m}$ | - | 36 |
| $\mathbf{S_f M^8_m}$ | 22 | 36 |
| $\mathbf{M^8_m}$ | - | 36 |

SNPs (S) were spaced at 0.5cM apart; superscript indicates marker density of STRs (M) in cM per marker; subscript refers to whether all (a), many (m), or few (f) subjects were typed. Refer to Figure 3.1 for the different designs.

I also evaluated four other conditions that might affect the imputation quality. I varied the density of the framework markers, as the density of markers might affect the inference of IVs [Wijsman, et al. 2006]. For this purpose, I thinned the original 0.5 cM spaced framework SNPs to obtain 1 cM and 2 cM spaced SNPs. I varied the call threshold on accuracy and call rate by varying call thresholds ranging from $t_1$=0.5 to 0.999999 (~1) while fixing $t_2$ midway between $t_1$ and 1. I refer to the case where the call threshold was ~1 as "practically deterministic". Unless

otherwise stated, the default $t_1$=0.8 and $t_2$=0.9 call thresholds were used. I investigated the effect of minor allele frequencies (MAF) on imputation accuracy by binning markers into MAF bins of size 0.01. Finally, I investigated the effect of the distance of dense markers from the closest framework markers under a STR-only framework panel ($M^8_m$), again by binning dense markers by the distance from their closest framework markers.

The meiotic map for the SNP markers was obtained by converting sequence position to Haldane map position by linear interpolation using the positions of the STRs from the Rutger map [Matise, et al. 2007] and the sequence positions of the STR markers with map positions determined by the Haldane map function. The population allele frequencies of the SNP markers were estimated using this pedigree along with 3 other large pedigrees with similar European ethnicity. The estimation was performed using Loki ver 2.4.6 [Heath 1997] in order to account for the pedigree structures.

### 3.4.4 Results and Interpretations

Data Patterns and Framework Marker Panels

High call rates were obtained in most subjects from multiple branches in the simulated large pedigree (Figure 3.2). In the $S_fM^8_m$ framework panel, subjects descended from the central pedigree, who tended to share more alleles with relatives, had higher call rates than married-in spouses (96.1% vs. 88.7% for Group D vs. F; 95.4% vs. 81.2% for E vs. G). In addition, high call rates were observed in subjects from the bottom generation who had multiple relatives typed for dense markers but were not themselves typed for dense markers (95.8% in Group L). Also,

high call rates were observed even in some subjects who were not typed for either sparse framework markers or dense markers ( >95% in Group D, E, and I).

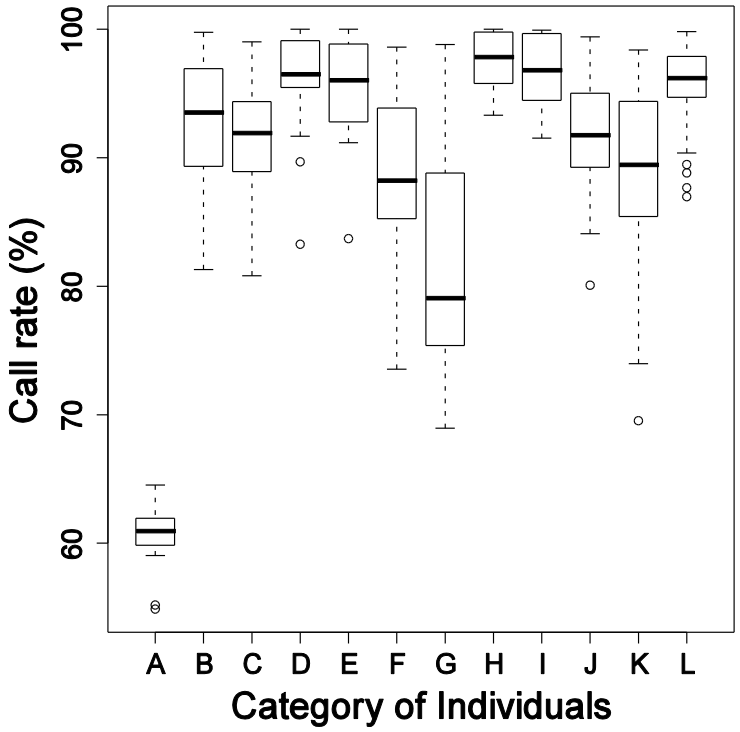



Figure 3.2 Call rate across classes of subjects: we used the $S_fM^8_m$ framework panel from simulation. Refer to Figure 3.1 for the description of the different classes of subjects.

The call rate depended much more on the number of subjects typed than on the density of framework markers. Among different framework panels considered (Table 3.2), the design where only a few subjects were typed for framework SNPs ($S_f$) gave the lowest call rate (78.8%). Regardless of the type of panel, having more subjects typed for the framework panel increased the call rate to 89.1-92.1%, for $S_m$ and all STR panels. Genotyping the majority of subjects for the framework panel (92.1% for $S_m$) is nearly as beneficial as genotyping all subjects (93.5% for

$S_a$). In contrast, altering marker density did not strongly influence the call rate. Doubling the density of STR markers increased the call rate only slightly (89.1% vs. 90.7% for $M^8_m$ vs. $M^4_m$). Similarly, increasing density by adding SNP markers on a few subjects to an existing STR panel only slightly improved the call rate, both when the STR panel was sparse (89.1% vs. 90.9% for $M^8_m$ vs. $S_fM^8_m$) or dense (90.7% vs. 91.5% for $M^4_m$ vs. $S_fM^4_m$).

Table 3.2. The effect of different framework panels on imputation quality evaluated using simulated data

| **Quality Metric (%)** | **Panels with only SNPs** | | | **Panels with STRs** | | | |
|---|---|---|---|---|---|---|---|
| | $S_a$ | $S_m$ | $S_f$ | $S_fM^4_m$ | $M^4_m$ | $S_fM^8_m$ | $M^8_m$ |
| **Called** | 93.5 (91.7,95.8)[a] | 92.1 (90.1,94.7) | 78.8 (77.8,79.6) | 91.5 (89.7,93.1) | 90.7 (87.2,93.1) | 90.9 (89.5,92.8) | 89.1 (86.3,90.2) |
| **Accuracy** | 99.6 (99.4,99.7) | 99.2 (97.8,99.6) | 98.7 (97.0,99.5) | 99.2 (98.6,99.6) | 99.2 (98.9,99.4) | 98.6 (98.0,99.4) | 98.6 (98.1,99.0) |
| **Consistency** | 93.6 | 92.1 | 90.2 | 92.4 | 71.0 | 91.6 | 54.2 |

a. range across 10 runs: (low, high)

When I call a genotype, it is highly accurate across all conditions considered (Table 3.2). Among SNP-only panels, the accuracy was the lowest in the $S_f$ design (98.7%). Typing more subjects for SNPs ($S_m$) only slightly increased the accuracy (99.2%). Doubling the density of STR markers also only slightly increased the accuracy (98.6% vs. 99.2% for $M^8_m$ vs. $M^4_m$). In addition, the 4cM spaced STR panel typed on many subjects ($M^4_m$) was similar in accuracy to the denser but di-allelic SNP panel typed on many subjects ($S_m$). Unlike the call rate, increasing density by adding SNP markers on a few subjects to an existing STR panel did not improve accuracy at all, whether the STR panel was sparse (98.6% for $M^8_m$ and $S_fM^8_m$) or dense (99.2% for $M^4_m$ and $S_fM^4_m$).

Both the call rate and accuracy increased only slightly when the density of the SNP framework panel increased (Table 3.3). Among SNP-only panels typed on many subjects, the call rate increased slightly when doubling the SNP density from one marker per 2 cM (90.5%) to one marker per 1cM (91.7%) and again when doubling density from one marker per 1 cM to one marker per 0.5 cM (92.1%). Similarly, while the accuracy increased slightly when doubling SNP density from one marker per 2cM (98.9%) to one marker per 1cM (99.2%), it did not further increase when doubling from one marker per 1 cM to one marker per 0.5 cM. Both gains, however, were modest, as the call rate and the accuracy were high even at the 2 cM density. Overall, these marginal increases in both call rate and accuracy were consistent with the previous results from increasing the STR marker density (Table 3.2).

Table 3.3. The effect of different marker density on imputation quality evaluated using the $S_m$ panel in simulated data

| **Quality Metric (%)** | **Spacing of SNPs in Framework Panel (cM) typed on $S_m$** | | |
|---|---|---|---|
| | 0.5 | 1 | 2 |
| **Called** | 92.1 (90.1, 94.7)[a] | 91.7 (90.3, 93.8) | 90.5 (88.6, 93.5) |
| **Accuracy** | 99.2 (97.8, 99.6) | 99.2 (97.8, 99.5) | 98.9 (98.1, 99,4) |
| **Consistency** | 92.1 | 84.5 | 69.5 |

a. range across 10 runs: (low, high)

Unlike call rate and accuracy, consistency depended strongly on the density of framework markers (Table 3.3). All panels which contained the 0.5cM spaced SNPs had high consistency (> 90.2%) (Table 3.2). However, consistency decreased as the density of framework markers decreased. As the marker spacing in SNP-only panels decreased from 0.5cM to 2cM, consistency decreased from 92.1% to 69.5% (Table 3.3). Similarly, as the marker spacing in STR-only panels decreased from 4cM to 8cM, consistency also decreased from 71.0% to 54.2% (Table 3.2). Even though the 8cM spaced STR panel ($M^8_m$) had the lowest consistency (54.2%), the call rate and the accuracy were still high.

Other Parameters

Call thresholds affected both call rate and accuracy but in different directions. The use of a more stringent call threshold decreased the call rate (Figure 3.3). For instance, under the design $S_fM^8_m$, the call rate decreased from 95.8% to 81.0% as the call threshold increased from $t_1$=0.6 to $t_1$=0.99. In contrast, the use of a more stringent threshold increased the accuracy: accuracy increased from 97.8% to 99.9% as the call threshold increased from $t_1$=0.6 to $t_1$=0.99. However, the change in accuracy was less dramatic than that of the call rate, since accuracy was already high at a liberal call threshold (97.8% for $t_1$=0.6). In this particular simulation, a reasonable balance between call rate and accuracy was achieved at the call threshold of $t_1$=0.8.

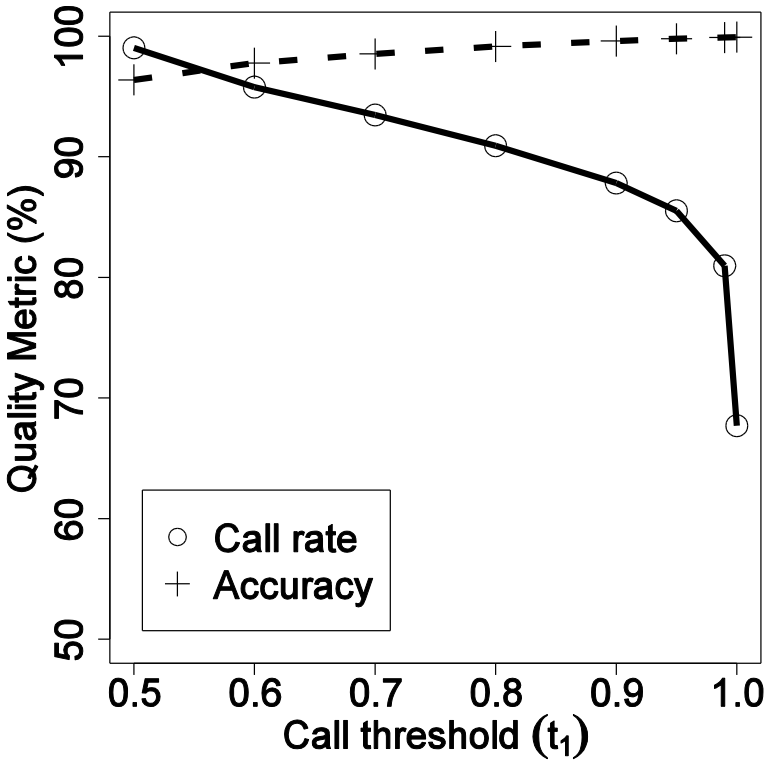



Figure 3.3 Call rate and accuracy as a function of call threshold. We used the $S_fM^8_m$ framework panel from simulation.

The minor allele frequency (MAF) of dense markers also affected quality metrics. At the default $t_1$=0.8 call threshold, the call rate decreased as the MAF increased (Figure 3.4A). Also, there was a sudden drop in the call rate at MAF=0.2. Examination of the call rate when I varied the call thresholds suggested that the MAF coinciding with the location of the sudden increase was at $1 - t_1$ (not shown). The reason is most likely because the call threshold directly determines where the imputation algorithm relies primarily on using allele frequencies to make calls. Besides the call rate, the accuracy decreased as the MAF increased from 0 to 0.2 but was approximately constant near 99.1% for frequencies above 0.2. I also called alleles using the

"practically deterministic" threshold (Figure 3.4B). Similar to the call rate with $t_1$=0.8, the call rate with this threshold continued to show a decreasing trend as the MAF increased. However, the imputation accuracy was almost perfect regardless of the MAF.

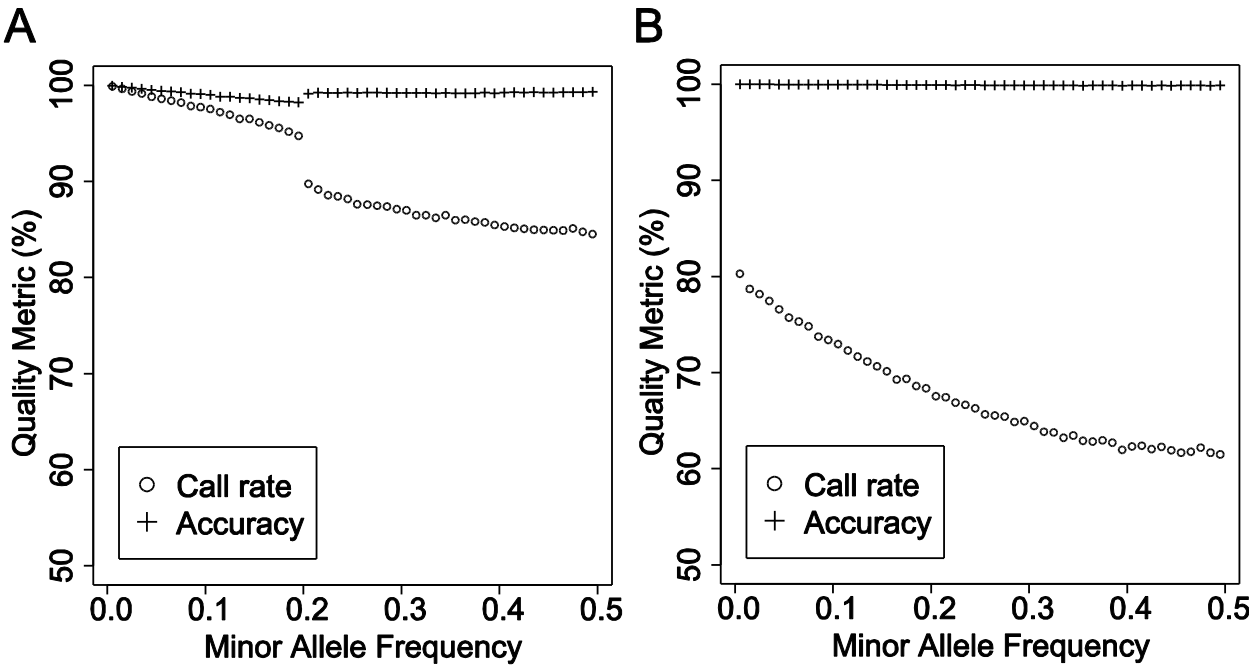



Figure 3.4 Call rate and accuracy as a function of minor allele frequency. I used the $S_fM^8_m$ framework panel from simulation. Different call thresholds were used: A) $t_1$ =0.8, $t_2$ =0.9 and B) practically deterministic.

The distance between dense genotypes and their respective nearest framework markers affected consistency much more than the call rate and accuracy (Figure 3.5). Under the $M^8_m$ panel, consistency decreased substantially as dense genotypes were farther from the nearest framework markers, e.g., from ~63% to ~45% as the map distance increased from 0 cM to 4 cM.

On the other hand, the accuracy and call rate did not drop much as the map distance increased, even though decreasing trends were observed.

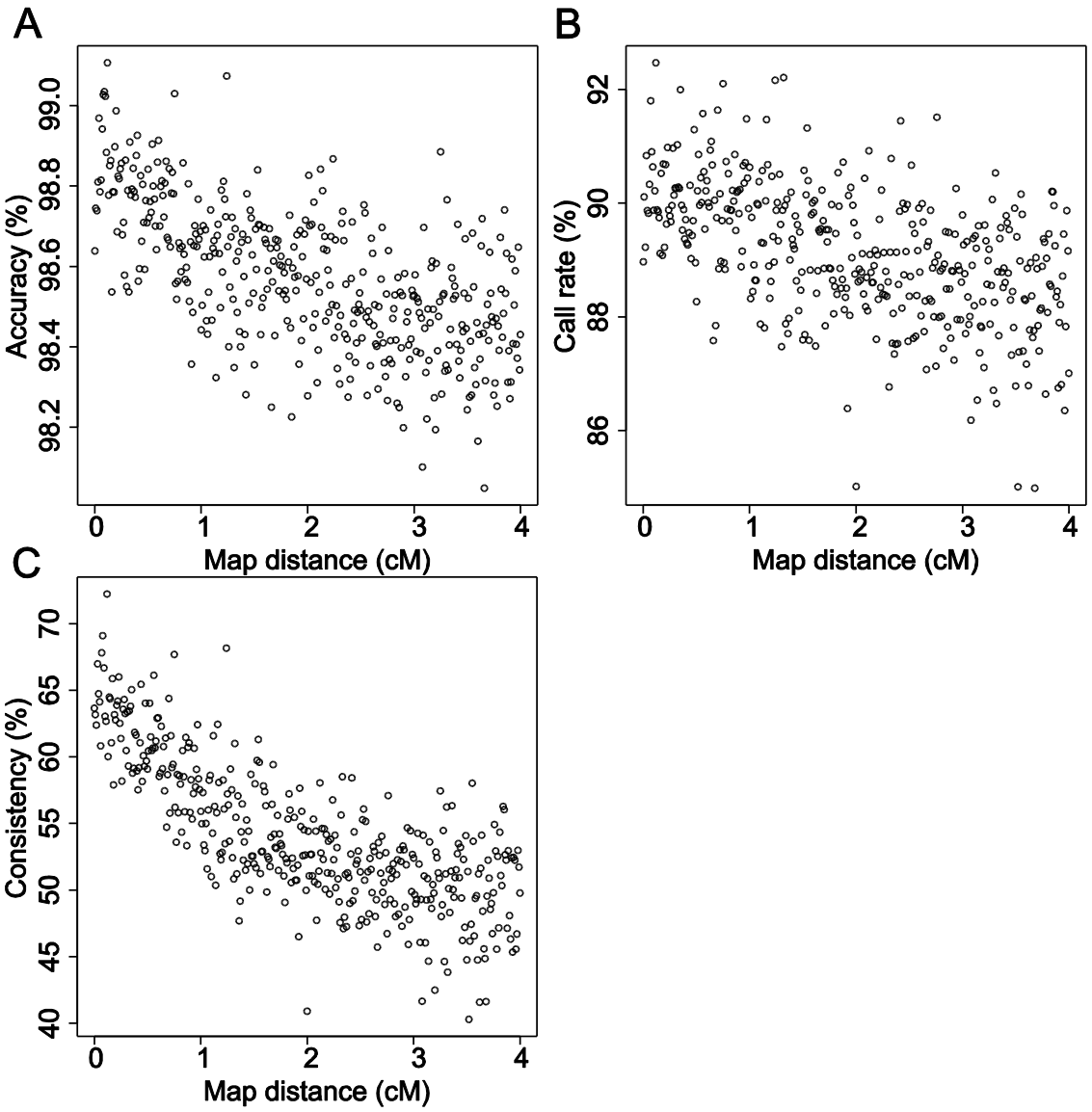



Figure 3.5 Impact of distance from the nearest framework marker. We used the $M^8_m$ framework panel from simulation. We measured the A) accuracy, B) call rate, and C) consistency.

## 3.5 Evaluating Imputation Performance using Real Data

### 3.5.1 Analysis of Real Data

Analysis of a real dataset allowed evaluation of my approach in data that contains complexities not captured well in simulated data. These include, but are not limited to, variable marker informativeness, potential misspecification of the genetic map and allele frequencies, undetected genotyping errors, and LD between markers. I used a 5-generation pedigree of 95-members [Wijsman, et al. 1998; Wijsman, et al. 2010], with some branches from the original pedigree omitted because they contained neither sparse nor dense marker data. Of the subjects retained, the average sibship sizes in the second, third, fourth, and fifth generations were 3, 3.8, 2, and 1.4. This pedigree also included one large sibship of size 9. I focused on imputing SNPs in a ~50 cM interval defined as a region of interest for a cardiovascular trait [Wijsman, et al. 2010]. My original dataset contained 60 subjects observed for 323 SNPs and 64 subjects observed for 21 STRs in the lowest 4 generations. Most of these SNPs were tightly linked with a few adjacent SNPs.

I performed an analysis that resembled the situation where we had legacy genome scan marker data and now collected new denser markers on a few subjects (Table 3.4). I retained SNP genotypes on 13 subjects scattered throughout different branches of the pedigree, and I masked SNP genotypes on the other 47 subjects. To infer IVs, I used a framework panel composed of all available STRs and a subset of 29 SNPs spaced out across the region of interest from the 13 subjects. Then, I used the sampled IVs and observed SNPs on the 13 subjects to impute missing genotypes of 294 SNP markers on the other subjects. The masked genotypes from the 47 subjects allowed us to evaluate the accuracy of imputation. The meiotic map for the SNP markers was

obtained by linear interpolating from Haldane map position of STR markers[Matise, et al. 2007] with sequence positions of both STRs and SNPs. The population allele frequencies of the SNP markers were estimated by Loki (version 2.4.6) [Heath 1997] using this pedigree along with 3 other large pedigrees with similar European ethnicity.

Table 3.4: Number of markers and number of subjects typed in the real data analysis with GIGI

| | | Markers | Subjects |
|---|---|---|---|
| Framework panel | STRs<br>SNPs | 21<br>29 | 64<br>13 |
| Dense genotypes | SNPs | 294 | 13 |
| Missing genotypes to impute | SNPs | 294 | 82[a] |

a. These 82 subjects are completely unobserved for SNPs, but some of them are observed for STRs

### 3.5.2 Results and Interpretations

High imputation accuracy and call rate were also obtained on the real data. In the real pedigree, the method called 68% of alleles among the 47-subjects that could be validated and achieved an accuracy of 97.6% using the default threshold (Figure 3.6). Relaxation of the call threshold to $t_1$=0.6 increased the call rate to 85% but with a decline to 93% in the accuracy. Similar to the simulated data, allele call rate was inversely related to the population allele frequency (not shown).

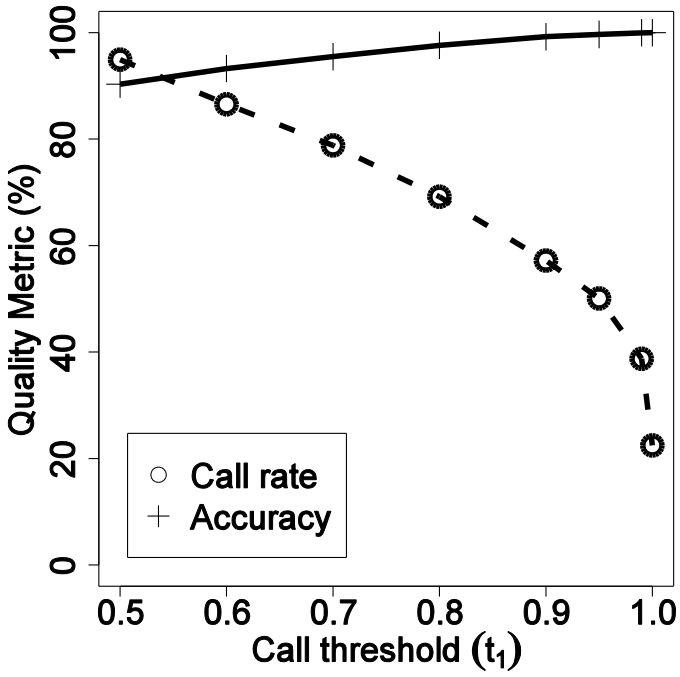



Figure 3.6 Call rate and accuracy as a function of call threshold in real pedigree

## 3.6 Comparison with BEAGLE

I compared GIGI to BEAGLE, a state of the art population-based genotype imputation approach [Browning and Browning 2009; Browning and Browning 2007]. BEAGLE can be used by ignoring the pedigree structure. Since BEAGLE uses information from population-level LD while not incorporating the pedigree structure, I sought to understand how the use of different sources of information may affect genotype imputation in a pedigree. I used the same real-data pedigree and region of interest. I computed the accuracy and call rate over all genotypes, as well as separately, over rare SNPs. I defined a rare allele to be the minor allele of a SNP with minor allele frequency less than 0.05. The accuracy of imputing rare alleles is especially important

because the primary motivation for using such pedigrees may be to identify rare variants that affect disease risk or phenotypic variation.

I evaluated GIGI and BEAGLE 3.3 (Table 3.5) using the real dataset. Since BEAGLE does not output genotype probabilities if I use both STRs and SNPs, I modified the previous analysis to perform a SNP-only analysis (Figure 3.7: $n_1 = 13$, $n_2 = 47$; $n_3 = 35$). Similar to the previous analysis, the same 13 subjects were given the complete genotype data. In BEAGLEs terminology, these genotype data were the outside reference samples used to infer haplotypes of dense markers. In the other 47 subjects, we kept genotypes of 35 approximately evenly-spaced SNPs and masked genotypes of the remaining 288 SNPs. Under this setup (Design FW), I compared GIGI and BEAGLE based on the imputed results of the masked SNPs on the 47 subjects.

Table 3.5 Number of subjects in different SNP-based designs used for the comparison between GIGI and BEAGLE

| | | Program | | | | |
|---|---|---|---|---|---|---|
| | | GIGI | BEAGLE | | | |
| Experiment[a] | | Framework | Framework | | Leave-One-Out | |
| Design | | FW | FW | FWO | L1 | L1O |
| SNPs [b] | Sparse | 47 | 47 | 47 | 0 | 0 |
| | Dense | 13 | 13 | 13+202 | 13+47 | 13+47 +202 |

a. FW=Framework Panel; L1=Leave-One-Out; O=includes the same SNPs on 202 subjects from 3 other pedigrees as a reference panel. See text.

b. Sparse = 35 SNPs; Dense=all 323 SNPs, except in the Leave-One-Out analysis which omits 1 SNP at a time from the 47 subjects. See text.

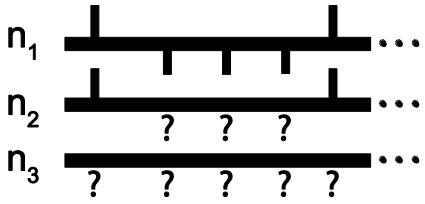



Figure 3.7 Different subjects have different levels of genotypes. Some subjects ($n_1$ of them) had observed genotypes for both framework markers (top ticks) and dense markers (bottom ticks); $n_2$ of the subjects had observed genotypes for framework markers but had missing genotypes (symbol ?) for dense markers; $n_3$ of the subjects were completely unobserved for both framework and dense markers.

### 3.6.1 Design of Experiment

I also evaluated BEAGLE under other designs (Table 3.4). When markers are tightly linked and when sample size of the outside reference is large, BEAGLE is more likely to perform well. In Design L1, we supplied BEAGLE with more markers by using a Leave-One-Out analysis where we imputed one SNP at a time, based on all other SNPs. In this Leave-One-Out analysis, genotypes of each SNP in the 47 subjects were omitted sequentially and were subsequently imputed back. In Design FWO, we added genotypes from 202 subjects to the outside reference panel (Figure 3.7: $n_1 = 13+202$, $n_2 = 47$; $n_3 = 35$). These 202 subjects were derived from 3 other pedigrees of similar ethnic background and who were typed on the same SNP platform. In Design L1O, I supplied both dense markers and additional outside reference samples. In each design, I called genotypes using both the most probable genotype calling and the threshold based approaches. To evaluate the performance of imputing rare alleles, I performed a subgroup analysis in genotypes that contained at least one copy of rare allele.

### 3.6.2 Results and Findings

GIGI called genotypes containing rare alleles with substantially higher accuracy than BEAGLE (Table 3.6). Under design FW and using the most probable genotype calling, GIGI called these genotypes with an accuracy of 62.9%, in contrast to BEAGLE, which only achieved an accuracy of 4.5%. Increasing the number of dense markers and providing more subjects in the reference panel (design L1, and L1O) improved BEAGLE's accuracy in calling genotypes containing rare alleles (up to 28.1%), but the accuracy was still much lower than that for GIGI. In addition, GIGI called 46.2% more rare genotypes for relatives who were completely untyped.

These genotypes were not called by BEAGLE because BEAGLE did not impute genotypes on completely unobserved subjects. Using the confidence-based calling with the default threshold (Table 3.6), the same trends were observed.

Table 3.6. Comparison between GIGI and BEAGLE under various designs of the real data

| | | | **Program** | | | | |
|---|---|---|---|---|---|---|---|
| | | | GIGI | BEAGLE | | | |
| **Call Choice** | **Group[a]** | **Metric[b]** | **FWK[c]** FW | **FWK[c]** FW | FWO | **Leave-One-Out[c]** L1 | L1O |
| Most Probable | Rare | A | 62.9 | 4.5 | 4.5 | 28.1 | 14.6 |
| | Overall | A | 79.7 | 70.2 | 73.3 | 82.5 | 95.4 |
| Threshold | Rare | A | 67.1 | 5.2 | 4.5 | 15.7 | 17.6 |
| | | C | 82.0 | 86.5 | 100 | 57.3 | 76.4 |
| | Overall | A | 96.4 | 88.8 | 91.5 | 95.1 | 98.0 |
| | | C | 47.7 | 47.1 | 43.8 | 54.1 | 93.6 |

a. Rare = among genotypes containing at least 1 copy of rare allele

b. A=accuracy (%), C =call rate (%). Under the most probable genotype calling approach, C=100%, and therefore has been omitted.

c. Design. FWK=Framework. Refer to Table 3.5 for the description of the designs.

I also compared the overall genotype accuracy and genotype call rate in GIGI and BEAGLE (Table 3.6). Under the design FW and using the most probable genotype calling, GIGI called genotypes with higher accuracy than BEAGLE (79.7% vs. 70.2%). However, the

availability of outside reference (FWO) or dense framework marker panel (L1) improved both accuracy and call rate in BEAGLE. In particular, the joint use of dense framework SNPs and outside references (L1O) improved the imputation accuracy of BEAGLE substantially (95.4%). The accuracy of imputing genotypes using GIGI could increase substantially to 96.4% when using the default threshold. However, the tradeoff was that a considerable fraction of genotypes was not called (52.3%).

## 3.7 Discussion

I have introduced an approach for carrying out genotype imputation in potentially large pedigrees. By harnessing existing computational tools that combine exact computation with MCMC-based sampling, my imputation approach can be used in pedigrees that range from small to very large with a range of possible missing data that may include founders. My results demonstrate that imputed genotypes can be accurate and obtained with high call rate. Results from analysis of the simulated data suggest that the number of subjects genotyped for a framework panel has a higher influence on the quality of imputed genotypes than does marker density. Also, results from the analysis of real data show that my genotype imputation approach has higher accuracy than the population-based approach as implemented in the state-of-the-art BEAGLE in imputing rare alleles.

My imputation approach can efficiently incorporate new collections of very dense markers, including high-throughput sequencing data, into studies involving existing genome scan data. Results obtained here suggest that an existing framework panel does not need to have high density to infer IVs as needed for genotype imputation. This is consistent with both theory

[Boehnke 1994] and past results [Wijsman, et al. 2006; Wilcox, et al. 2005] that show diminishing gains in determining inheritance at a particular position with increasing marker density. My results suggest that markers from existing genome scans can be leveraged to allow genotype imputation of dense markers on many individuals when these existing marker genotypes are coupled with dense markers typed on some subjects.

Results from the comparison between GIGI and BEAGLE have several implications. First, GIGI provides much higher accuracy in calling rare alleles than does BEAGLE. This is an expected outcome, since explicitly modeling the transmission of genomic segments using the pedigree structure allows rare alleles on such segments to be reliably called. In contrast, BEAGLE is not accurate in calling rare alleles, a result which agrees with other studies [Krithika, et al. 2012; Li, et al. 2011]. When rare alleles are segregating in only one pedigree, increasing the reference sample size is unlikely to help impute such rare alleles. Such pedigree-specific rare or ultra-rare alleles may be typical, especially for causal risk alleles, as is suggested by the very large number of alleles known for some disease loci[Leigh, et al. 2008]. Second, conditioning on the pedigree structure together with marker data in the pedigree allows imputation of genotypes in relatives who are unobserved for any genotypes. Both of these considerations are important when the motivation is to identify rare causal variants in pedigrees. Third, BEAGLE excels at imputing common variants when both dense markers and an adequate number of reference samples are present. Under these conditions, BEAGLE imputes common alleles quite accurately and calls them with high confidence; however, the availability of dense framework markers typed on many subjects is not always guaranteed in pedigree studies. A potential future direction would be to integrate the use of information from population LD that BEAGLE uses into GIGI to improve the call rate for common variants.

Two notable features of our approach allow efficient imputation in large pedigrees. First, our approach separates the inference of IVs from imputation of dense genotypes. One advantage of this strategy is that it circumvents the linkage equilibrium assumption between markers that is needed for application of the Lander-Green algorithm. This is an advantage because the estimated probability of IVs may be incorrect if the linkage equilibrium assumption is violated, which may lead to an increase in false-positive linkage signals [Huang, et al. 2004; Schaid, et al. 2002]. Another advantage is computational efficiency, which is achieved because IVs only needed to be sampled once using sparse framework markers. This approach is in contrast to the computationally intensive approach used in MERLIN to incorporate LD through the use of haplotype blocks [Abecasis, et al. 2002]. Second, our approach uses a state-of-the-art MCMC sampler for analysis of large pedigrees. This allows us to sample IVs to enable analyses that are otherwise computationally intractable on large pedigrees.

Genetic analyses can be performed with dense imputed genotypes to identify variants that affect traits. In large pedigrees, it may be fruitful to limit the initial search space to regions where there is positive evidence for linkage with the trait, since only here is there sufficient joint segregation of trait and markers to provide strong confidence in any implicated variants. In these regions, we can then search for causal variants using different approaches. One approach is to perform a measured genotype approach on imputed SNPs, treating them as covariates to adjust out a linkage signal [Almasy and Blangero 2004] in, *e.g.*, a variance component analysis. Another approach is to perform a family-based association test that is suitable for small [Lunetta, et al. 2000] or large pedigrees [Bourgain, et al. 2003]. Yet another approach is to perform exploratory analyses using simple filters to correlate disease status with rare variants. Since many types of analysis require genotypes on many subjects, the use of imputed genotypes will enable these

types of analyses. In any case, where imputation is used, the most significant results should be checked with direct genotyping, just as is standard for population-based studies [Thomas, et al. 2009].

Misspecifying population allele frequencies can affect imputations. When some founders are not observed, the use of inaccurate population allele frequencies in framework markers may affect the inference of IVs. Then, the use of inaccurately inferred IVs may decrease imputation accuracy. Misspecifying the allele frequencies of the dense markers also affects imputation. When an IV suggests that a founder allele is not observed, my genotype imputation approach uses the population allele frequency to infer this missing allele. Misspecifying the population allele frequency can, therefore, affects imputation accuracy. Nevertheless, if an allele is called using constraint from the pedigree structure or IV, the call will not be affected by the population allele frequencies. If I set a higher call threshold, less calls will be made using the population allele frequencies, so misspecifying the population allele frequencies will matter less.

There were two most obvious explanations of why wrong imputation calls were made on some subjects even under the practically deterministic threshold. First, the sampled IVs might be inaccurate. For instance, if framework genotypes contain Mendelian consistent genotyping errors, the sampled IVs could be biased. Also, if the framework genotypes are not informative, sampled IVs might not be similar to the true IV. If the entire set of the finite number of sampled IVs are not representative of the true distribution of IVs, imputation calls could be wrong even under the deterministic threshold. Second, issues with MCMC mixing could also affect genotype imputation, preventing IVs from being sampled at the correct distribution from all allowable states. This could also lead to bias in the estimated probabilities of genotypes in imputation.

## 3.8 Software Availability

The imputation method is implemented in the Genotype Imputation Given Imputation (GIGI) software and is written in C++. It is publicly available at http://faculty.washington.edu/wijsman/software.shtml.

## 3.9 Concluding Remark

My genotype imputation approach, as implemented in GIGI, can facilitate cost-effective genetic analyses, including but not limited to the identification of rare causal variants in complex traits. Since rare alleles affecting traits can be enriched in pedigrees, the use of large pedigrees is an efficient design to detect signals that are statistically significant. Such pedigrees are emerging as an important class of data used to identify rare causal variants. Statistical analyses of such large pedigrees using imputed dense genotypes may benefit from increased power.

# Chapter 4

# DETECTION OF MENDELIAN CONSISTENT GENOTYPING ERRORS

In this chapter, I first discuss the background and motivation for detecting genotyping errors. Second, I survey existing approaches. Third, I present my approach that enables effective detection of Mendelian consistent genotyping errors of dense markers typed on pedigrees. In a simulated large pedigree, I show that my method can detect a high percentage of Mendelian consistent genotyping errors. Fourth, I evaluate parameters and conditions that affect the performance of error detection. Fifth, I compare my error detection approach to some existing approaches. Finally, I present a preliminary investigation of potential extension to my approach that can reduce false positives in error detection.

## 4.1 Background and Motivation

A major goal in human genetics is to understand how human biology functions. By first identifying mutations in human chromosomes that lead to diseases, further investigations can help delineate the molecular mechanisms that affect normal body functions. In this pursuit, analyzing families using approach such as linkage analyses followed by fine-mapping methods has successfully led to the identification of over 4500 genes associated with diseases [Amberger, et al. 2011]. Genetic analyses require the use of genotypes identified from genetic variants, and validity of analyses ultimately depends on the accuracy of genotype measurements. However,

from older technologies used in Single Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs) to newer technology used in sequence data, none of these genotyping technologies are immune to genotyping errors.

Genotyping errors affect the validity of results in genetic studies. Genotyping errors affect the quality of genetic maps [Buetow 1991; Goldstein, et al. 1997; Hackett and Broadfoot 2003], haplotype reconstruction [Kirk and Cardon 2002], and measures of Linkage-Disequilibrium [Akey, et al. 2001; Hackett and Broadfoot 2003]. They also reduce power in association analyses [Abecasis, et al. 2001] and affect linkage analysis [Cherny, et al. 2001; Lebrec, et al. 2008]. For linkage analysis, moderately low genotyping error rate can either decrease power [Abecasis, et al. 2001; Douglas, et al. 2000; Walters 2005] or lead to false evidence of linkage [Douglas, et al. 2002]. Genotyping errors continue to impact results from analysis performed using current genotyping technologies. For instance, genotyping errors inflate the type I error rates in the transmission disequilibrium test [Gordon, et al. 2001; Mitchell, et al. 2003] and haplotype-based tests [Knapp and Becker 2004]. Furthermore, in genotype imputation, genotyping errors can additionally propagate to imputed genotypes, which can lead to problems in downstream analysis. Hence, genotyping errors should first be removed if an analysis does not explicitly account for them.

In pedigrees, genotyping errors are either Mendelian inconsistent (MI) or Mendelian consistent (MC). A MI genotyping error is a genotyping error that is detected because the observed genotypes are not consistent with the transmission pattern as specified by Mendel's First Law. Specialized program such as PedCheck [O'Connell and Weeks 1997] and many programs to perform linkage analyses [Abecasis, et al. 2002; Lathrop, et al. 1983; Sobel, et al. 2002] can be used to detect MI errors. When a marker is flagged as Mendelian Inconsistent (MI),

this marker most likely has genotyping error or a *de-novo* mutation. The fraction of errors that escape the Mendelian Inconsistency check depends largely on a particular pedigree structure [Mukhopadhyay, et al. 2004] and can be high [Douglas, et al. 2002; Gordon, et al. 1999], especially when a considerable numbers of subjects are unobserved for genotypes. If a genotyping error is not MI, it is MC and cannot be detected using MI checks. The detection of genotyping errors in SNPs is more difficult than in STRs because genotyping errors in di-allelic SNPs can escape MI checks more easily[Douglas, et al. 2002]. Thus, a tool to detect MC genotyping errors are needed for analyses that uses dense di-allelic markers.

### 4.1.1 MC errors in Di-allelic Markers can Affect Linkage Analysis

While studies have investigated the problem that genotyping errors in sparse markers can affect linkage results [Abecasis, et al. 2001; Douglas, et al. 2000], here I demonstrate that MC genotyping errors can also affect linkage results in modern genotype data involving di-allelic markers. Using the same 52-member pedigree as previously described (Figure 3.1), I selected 5 simulated descent patterns on a 100cM chromosome [Wijsman, et al. 2006] in which a trait with risk allele D with allele frequency of 0.2 was in the middle of the chromosome. These simulated descent patterns were chosen because there were modest maximum lod scores. Using these descent patterns, I simulated clean genotypes of framework markers with a density of 1 marker per 0.5 cM and retained marker genotypes on 34 subjects, which is a subset of subjects identical to the design in Figure 3.1 2, except with 2 subjects removed from the 3$^{rd}$ generation. In each run, I simulated 2 MC genotyping errors by changing the allelic state of randomly selected genotype and only retaining errors that are MC in framework markers between 40 cM and 60 cM

on the 100 cM chromosome. I repeated this process 30 times for each chosen descent patterns to obtain a total of 150 genotype-marker datasets. Finally, I performed linkage analyses using lm_multiple from MORGAN. I used the same trait model as was previously used [Wijsman, et al. 2006] with incomplete penetrance parameters: dd=0.05, Dd=0.7, and DD=0.8. Then, I plotted the lod scores at the positions of the framework markers for each dataset.

Results from these simulations showed that even one or two MC errors can affect the results of linkage analysis on a marker panel with SNPs spaced at a density of 1 marker per 0.5 cM. In the first example, MC error attenuated the linkage signal at the position of the trait locus (Figure 4.1) and therefore deflected the position with strongest evidence of linkage. This could mislead an investigator into discounting the precise positions of the trait locus. In the second example, MC error inflated the maximum lod score (Figure 4.1).

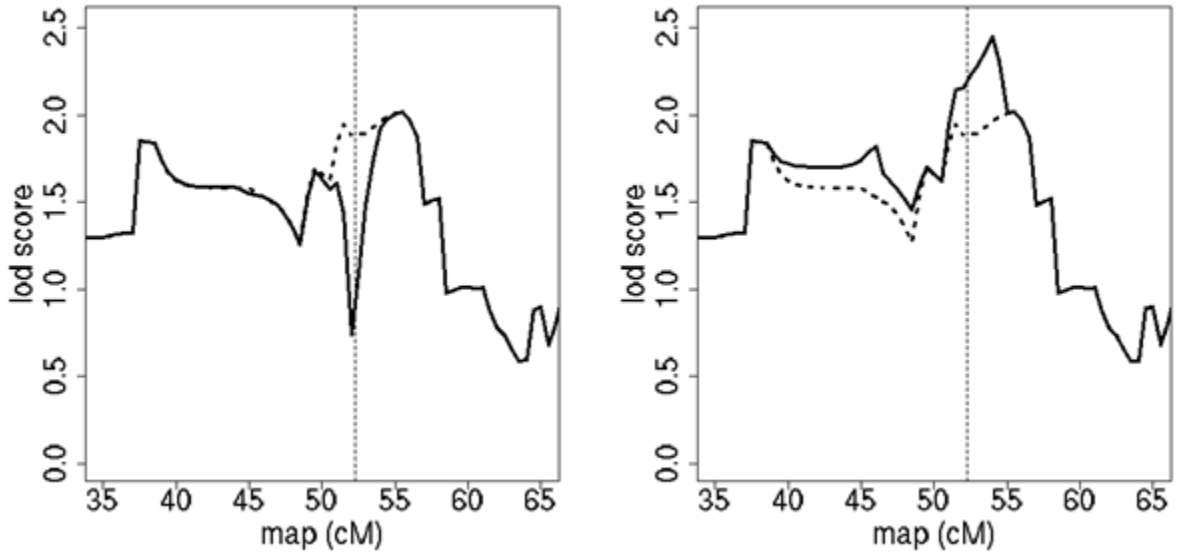



Figure 4.1 The presence of one or two MC genotyping errors in SNPs can affect the result of linkage analyses: dash line: no error; solid line: with 2 genotyping errors; vertical line: position of the trait locus.

Genotyping errors will likely present major challenges to studies based on genotypes generated from next-generation sequence. At the present, the rate of genotyping errors in Next-Generation sequence data is still quite high [Zhi, et al. 2012]. Therefore, traditional analyses performed using these new sequence data inherit and may even exaggerate all existing problems that are associated with the analyses using SNP genotypes from array-based platforms. Moreover, genotyping errors likely hinder the identification of real rare functional variants. An important concern is that rare alleles that are identified can be artifact of genotyping errors.

## 4.2 Survey of Existing Methods

Inheritance vectors (IVs) [Lander and Green 1987], which represent the inheritance of chromosomes in related individuals, provide a framework to understand existing MI and MC error detection approaches in pedigrees. MI checks implemented in PedCheck [O'Connell and Weeks 1997] are single-marker approaches that evaluate markers one-by-one, flagging a marker if the likelihood is 0. This requirement is equivalent to checking that the observed genotypes are incompatible with all possible configurations of IVs. MC error detection approaches are an extension to the MI check. These approaches perform multi-point computation that use genotypes of neighboring markers to probabilistically model the transmission of chromosomes. The transmission of chromosomes in the pedigree can be represented by IVs. Using the probability distribution of IVs at each marker, MC error detection approaches then identify potential errors by using summary statistics such as the posterior probability of error.

While MC genotyping error detection approaches are useful, existing approaches are not suitable for use on large pedigrees with many dense markers, such as variants identified in sequence data. One class of approaches aims to detect "excess recombinants". While effective for detection of MC genotyping errors, this class of approaches is either limited to use in sibling-pairs [Douglas, et al. 2000], trios [Sieberts, et al. 2002], small pedigrees [Abecasis, et al. 2002], or are not computationally inefficient for large pedigrees that have a large number of markers [Sobel, et al. 2002]. Another class of approaches detects MC genotyping errors by using information from Linkage Disequilibrium [Becker, et al. 2006; Kennedy, et al. 2008; Zou, et al. 2003]. These approaches used observed genotypes from multiple unrelated subjects or families to estimate haplotype frequencies and detected errors by computing test statistics based on likelihood ratios. However, this approach also has many limitations, which includes (1) the requirement for externally phased haplotypes that may not be available, (2) a requirement for tightly linked markers, (3) limitation to use in trios and nuclear families, and (4) potentially poor performance for identifying errors in rare variants because of the lack of reference samples that have those rare variants. Thus, development of a method is needed to detect MC genotyping errors suitable for dense markers such as variants identified from sequence data.

## 4.3 A New Approach to Detect Mendelian Consistent Genotyping Errors

### 4.3.1 Overview

IVs can be used to identify errors in markers that are Mendelian consistent. If we had the true IV, it could be used to detect MC genotyping errors by checking whether the true IV is

consistent with the observed genotypes. An IV is consistent with the observed genotypes if there is at least one configuration in which alleles of the observed genotypes can be assigned to the descent information specified by the IV. If the genotypes of markers are free of errors, the true IV at the positions of these markers would be consistent with the observed genotypes. On the other hand, inconsistency between the observed marker genotypes and the true IV is an indicator for genotyping errors. Figure 4.2 illustrates the use of an IV to detect a MC error. In this figure, the observed genotypes are consistent with the pedigree structure. However, inconsistency between observed genotypes and IV suggests a genotyping error. In reality, any biological phenomenon that leads to a departure from Mendel's First Law (eg. mutations or deletions) also results in inconsistency, but in this work we would ignore this possibility.

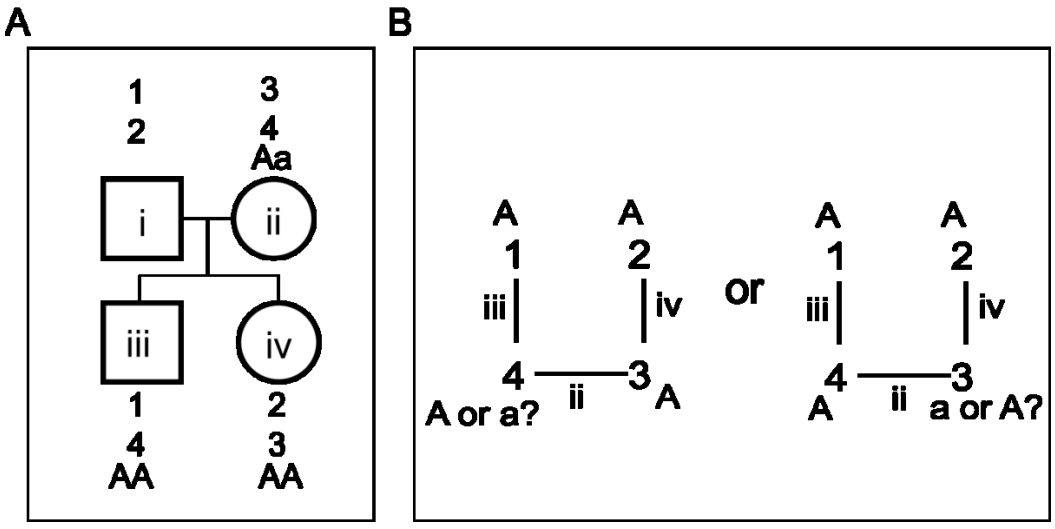



Figure 4.2 An example illustrating that MC genotyping error can be detected by IV: A) In this pedigree, subjects are labeled by roman numerals, FGLs are labeled by numbers, and observed genotypes are labeled by letters. B) Incompatibility between the observed genotypes and IBD-graph will result from any order of which subjects used to construct the IBD-graph. Order of subjects used: (left) iii, iv, then ii (right): iii, ii, then iv

While the true IV is not observable, we can infer IVs probabilistically to detect probable genotyping errors. We can sample a set of IVs at a position between framework marker positions by using the approach described in Chapter 2. While the genotypes of framework markers must be consistent with the set of sampled IVs at framework positions, markers genotyped at other positions are not guaranteed to be consistent with the sampled IVs. If a sampled IV between the position of two framework markers is informative of the true pattern of chromosomal descent, inconsistency between a sampled IV at that position and genotypes of a dense marker at that position provides evidence for a genotyping error. The presence of inconsistency in a single sampled IV is not by itself an indicator of error, but if we assume that at least a reasonable percentage of the sampled IVs are informative, a very low percentage of consistent sampled IVs suggests the presence of a genotyping error.

My approach consists of four steps.

(1) Sample IVs at the positions of the framework markers conditional on genotypes of the framework markers

(2) Sample IVs at the positions of the dense markers conditional on the IVs sampled at the positions of the framework markers and the meiotic map

(3) Flag a dense marker as having genotyping error if its test statistic is less than a pre-defined threshold (e.g. 5%). Two test statistics are proposed:

- S1. Percent consistency: percentage of sampled IVs that are consistent with the observed genotypes of the dense marker
- S2. Posterior probability of no error

(4) If a genotyping error is flagged by S2, identify the subject that most likely has the genotyping error

### 4.3.2 Details of the Approach

Detecting Mendelian consistent genotyping errors

I first define some terminology. Let $G_v$ be the observed genotypes, $X_v$ be the true genotypes that may be different from $G_v$, and $S_v^k$ be the k[th] sampled IV at dense marker $v$. In addition, let $F$ be the framework markers used to infer IVs and $G_F$ be the observed genotypes of the framework markers. I assume that $G_F$ does not contain genotyping errors.

I propose two test statistics to detect MC genotyping errors in dense marker $v$:

S1: Percent consistency

$$\%consistency = \frac{1}{n}\sum_{k=1}^{n} I(P(G_v|S_v^k) > 0)$$

S2: Estimated posterior probability of the unobserved genotype configuration

$$\hat{P}(X_v = x|G_F, G_v) = \frac{1}{n^*}\sum_{i=1}^{n} \frac{P(X_v=x|S_v^k)P(G_v|X_v=x)}{\sum_w P(X_v=w|S_v^k)P(G_v|X_v=w)} \quad (1),$$

where $n^* = \sum_{i=1}^{n} I(\sum_w P(X_v = w|S_v^i)P(G_v|X_v = w) > 0)$. The term $I(\sum_w P(X_v = w|S_v^k)P(G_v|X_v = w) > 0)$ is an indicator variable that is 1 if $S_v^k$ is consistent with at least one configuration of $X_v$ and 0 otherwise. There is no guaranteed that $n^* = n$ because my error model does not assume the presence of multiple genotyping errors and because the sampled IVs are not necessarily the true IV. In each of S1 and S2, the approach flags a marker if the test statistic is below a pre-defined threshold $c$.

S1 and S2 offer different strengths and weaknesses. S1 has two main advantages. First, computing S1 is fast, which is ideal for scanning for errors in a large number of markers. Second, the calculation of S1 does not require the use of allele frequencies. Even though S1 contains the term $P(G_v|S_v^k)$, evaluating whether the quantity is greater than 0 does not involve the use of

allele frequencies, which is an advantage because allele frequencies may not be easily available. On the other hand, S2 also has an unique advantage. While S2 is computationally more intensive than S1, S2 can be used to help identify which subject likely contains the error, as will be discussed below.

S2 is derived as follows:

$$
\begin{aligned}
P(X_v = x|G_F, G_v) &= \sum_s P(X_v = x|G_v, S_v = s, G_F) P(S_v = s|G_F, G_v) \\
&= \sum_s \frac{P(X_v = x, G_v|S_v = s\,, G_F)}{P(G_v|S_v = s, G_F)} P(S_v = s|G_F, G_v) \\
&\cong \sum_s \frac{P(X_v = x, G_v|S_v = s, G_F)}{P(G_v|S_v = s, G_F)} P(S_v = s|G_F) \\
&\cong \sum_s \frac{P(X_v = x, G_v|S_v = s)}{P(G_v|S_v = s)} P(S_v = s|G_F) \\
&= \sum_s \frac{P(X_v = x|S_v = s) P(G_v|X_v = x, S_v = s)}{\sum_k P(X_v = k|S_v = s) P(G_v|X_v = k, S_v = s)} P(S_v = s|G_F) \\
&= \sum_s \frac{P(X_v = x|S_v = s) P(G_v|X_v = x)}{\sum_k P(X_v = k|S_v = s) P(G_v|X_v = k)} P(S_v = s|G_F)
\end{aligned}
$$

Refer to section 3.3.2 for a discussion of the inequality in the third and fourth line of the above equation.

Each $S_v^k$ is not guaranteed to be Mendelian consistent with some $x$ unless dense marker $v$ is a framework marker. Inconsistency between $S_v^k$ and a particular joint genotype configuration occurs because either $S_v^k$ is not the true IV or the joint genotypes at the dense marker $v$ contains an error. Here, I assume that if $S_v^k$ is not consistent with any $x$, then $S_v^k$ is a sampled IV of low

quality that should not be used. Therefore, I use the modified estimator:

$$\hat{P}(X_v = x|G_F, G_v) = \frac{1}{n^*}\sum_{i=1}^{n} \frac{P(X_v=x|S_v^k)P(G_v|X_v=x)}{\sum_w P(X_v=w|S_v^k)P(G_v|X_v=w)},$$

where $n^* = \sum_{k=1}^{n} I(\sum_w P(X_v = w|S_v^k)P(G_v|X_v = w) > 0)$.

S2 estimates the posterior probability of a joint genotype configuration being the truth. When $x = G_v$, $\hat{P}(X_v = x|G_F, G_v)$ estimates the posterior probability that there is no genotyping error. When $x \neq G_v$, $\hat{P}(X_v = x|G_F, G_v)$ estimates the posterior probability of other joint genotype configurations. Thus, these estimated probabilities can be used to identify the subject who most likely has genotyping error. Calculating S2 is straightforward because $P(X_v = x|S_v^k)$ in Equation (1) can be efficiently computed [Kruglyak, et al. 1996; Sobel and Lange 1996].

Calculating the posterior probability in S2 requires an error model in the $P(G_v|X_v = g)$ term of Equation (1). First, the model assumes that genotyping errors are independent at different markers and in different individuals with error probability ε. Hence, if $m$ subjects are observed at dense marker $v$, $P(G_v|X_v = G_v) = (1 - \varepsilon)^m$. Second, if there is a genotyping error, only 1 of the 2 alleles is incorrect and the 2 alleles have an equal chance to be incorrect. Third, in a marker with $t$ allelic types, the true allele of the incorrect allele is any of the $t - 1$ other allelic types with an uniform probability $\frac{1}{t-1}$. For instance, if the true joint genotypes $x$ and observed joint genotypes $G_v$ differ only at subject $p$, as denoted by the true genotype $x_p$ and the observed genotype $G_{pv}$,

$P(G_v|X_v = x) = (1 - \varepsilon)^{m-1}\, \varepsilon\, P(G_{pv}|x_p; G_{pv} \neq x_p)$. For di-allelic SNP,

$P(G_{pv} = AA|x_p = Aa; G_{pv} \neq x_p) = \frac{1}{2}$ and $P(G_{pv} = Aa|x_p = AA; G_{pv} \neq x_v) = 1$. This simple error model allows efficient computation because it restricts the number of possible true genotypes.

To further achieve computational efficiency, the model assumes that the pedigree has at most 1 genotyping error at dense marker $v$. This strategy avoids the task of enumerating all potential true genotyping configurations exhaustively, which includes the unlikely configurations with multiple genotyping errors. Thus, this model evaluates at most $1 + 2m(t-1)$ instead of up to $t^{2m}$ configurations, which grows exponentially with the number of observed genotypes.

## 4.4 Evaluation using Simulated Data

### 4.4.1 Simulation

Pedigrees

I used the same 52 member simulated pedigree as previously described (Figure 3.1) to evaluate my error detection approach. I also used the same 3 designs for which subjects are observed for genotypes to test the ability to detect MC genotyping errors when varying the completeness of the observed genotype data.

Markers

I simulated both framework and dense markers on a chromosome of 100 centi-Morgans (cM) (Figure 4.3). I simulated 200 di-allelic framework markers spaced uniformly at a density of 1 marker per 0.5 cM. In addition to the framework markers, I simulated 25,000 di-allelic dense markers spaced uniformly at a density of 1 marker per 0.004 cM with population allele frequencies simulated from the uniform 0-1 distribution. For each run, I generated the complete genotype marker dataset by simulating founder alleles and gene-dropping through the descent patterns. Then, I introduced errors on top of the simulated genotypes. For each combination of

subjects observed for dense markers, genotyping errors were introduced independently at pre-defined allelic error rate $\varepsilon$ as defined below.

Figure 4.3 Schematic diagram of the framework (top ticks) and dense markers (bottom ticks) in a region.

### 4.4.2 Measuring Performance

I used several quality metrics to evaluate the error detection approach. Under each experimental condition, I computed sensitivity, specificity, and positive predictive value (PPV). Sensitivity is the percentage of markers flagged at markers that contained MC genotyping error. Specificity is the percentage of markers not being flagged at markers that did not have genotyping errors. Positive predictive value is the percentage of flagged markers that indeed contained a MC genotyping error. In S2, I also computed the accuracy of identification of the correct subject who had MC errors in markers that in reality did have MC errors. To compare the performance of using the test statistics S1 and S2 over a range of detection thresholds, I also plotted the Receiver Operating Characteristic curve, which plots sensitivity against specificity as a function of the detection threshold. Unless otherwise stated, S1 was the only calculated test statistic.

### 4.4.3 Analysis

For each experimental condition, error detection followed a sequence of steps. First, the approach sampled IVs by using framework markers obtained with gl_auto. The approach sampled a total of 50,000 IVs but kept every $50^{th}$ IV for error detection to obtain a set of IVs that

were relatively uncorrelated. Second, I removed MI genotyping errors of dense markers using gl_auto's Mendelian Inconsistency check option. Third, I detected MC errors by using the test statistic S1 on the remaining markers that passed the MI check.

Default Experimental Condition

I set up a default experimental condition. In this default setting, both framework and dense genotypes were typed on the design with "many" subjects (Figure 3.1). Errors were simulated at a rate of $\varepsilon$=0.25%. In the error check, the approach computed both test statistics S1 and S2. The default error detection threshold is 0.05. If S2 flagged an error, the approach attempted to identify the subject who most likely had the MC error. As a benchmark, I also used the simulated true IVs at dense positions to detect MC errors. The percentage of markers inconsistent with the true IVs is the percentage of markers with error that we hope to detect when my method uses inferred IVs instead of the true IV. The default summaries were reported using results from run #1, but results from other runs were also presented whenever there was a need to illustrate the variability of the results. Since the quality metrics are not as meaningful when summarized as averages, results were reported per run.

I evaluated our error detection approach under default experimental conditions. First, I compared the performance of using S1 and S2, as S2 might perform better because it used an error model. Second, I varied the threshold of error detection from the most stringent threshold of 0% to a relaxed threshold of 20% to evaluate the impact of relaxing the threshold. Third, I varied the number of IVs used to determine if the use of more IVs would improve error detection. Fourth, I stratified the markers by bins of minor allele frequencies (MAF) to investigate if MAF of markers affected the quality metrics. In particular, it is of interest to

determine if my approach can effectively detect genotyping errors for markers with low MAF.

Evaluation of the Effect of Other Parameters

I evaluated several other parameters that might affect the performance of error detection. First, I varied the number of subjects typed for dense SNP markers by using the same 3 observed designs patterns (i.e. all (52) subjects observed, most (36) subjects observed, and few (22) subjects observed), as were used in Chapter 3 (Figure 3.1) to evaluate how different levels of pedigree constraints affected error detection. Second, I varied the number of subjects typed for framework markers to evaluate whether the use of less informative IVs negatively affected error detection. Third, I thinned the framework markers into a marker per 1 cM and a marker per 2cM to test the effect of varying framework marker spacing. Fourth, I varied $\varepsilon$ at levels 0.005%, 0.05%, 0.25%, and 1%. This range of error rates encompasses the error rates of different genotyping platforms including SNP platforms, STR genotyping platform, and other platforms that resembles current sequencing technologies that have been reported to be more error-prone.

### 4.4.4 Results and Interpretations

Genotyping error detection with the default condition

My approach detected most MC genotyping errors. Under the default experimental condition ($D_mF_m$), my approach achieved high sensitivity (88.9%) that matched very closely the sensitivity obtained by using the true IV (90.8%) (Table 4.1). In addition, the specificity (99.0%) and PPV (93.5%) were high (Table 4.1). The worst sensitivity, specificity, and PPV across runs were greater than 83%, 97%, and 83%, respectively. In most of the runs, the ROC graph showed that the test statistics S1 and S2 performed similarly in the range of the thresholds considered,

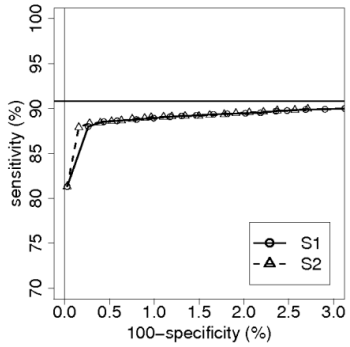
but in run #6, S2 has higher specificity at similar sensitivity under the threshold of 0.01 (Figure 4.4). In addition, S1 and S2 intersected at the threshold of 0%. These observations were consistent across runs (not shown). Also, S2 accurately identified subjects with error (84.2%) in run #1.

Table 4.1: Effect of the number of subjects typed for dense markers and framework markers

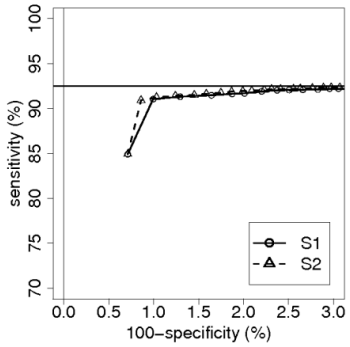
| | | Design[1] | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $D_aF_a$ | $D_aF_m$ | $D_aF_f$ | $D_mF_a$ | $D_mF_m$[2] | $D_mF_f$ | $D_fF_a$ | $D_fF_m$ | $D_fF_f$ |
| Run #1 | Sens[3] | 88.7 | -[4] | - | 89.8 | 88.9 | - | 66.6 | 65.4 | 65.0 |
| | Spec[3] | 98.9 | - | - | 99.2 | 99.0 | - | 99.7 | 99.6 | 99.5 |
| | PPV[3] | 91.9 | - | - | 94.4 | 93.5 | - | 95.8 | 95.3 | 94.0 |
| | True IV Sens[3] | 89.7 | - | - | 90.8 | 90.8 | - | 68.0 | 68.0 | 68.0 |
| Run #10 | Sens[3] | 95.7 | -[4] | - | 88.9 | 89.1 | - | 59.6 | 59.3 | 56.8 |
| | Spec[3] | 99.1 | - | - | 99.2 | 98.5 | - | 99.7 | 99.5 | 99.7 |
| | PPV[3] | 93.7 | - | - | 94.5 | 90.3 | - | 96.0 | 93.2 | 95.6 |
| | True IV Sens[3] | 96.7 | - | - | 90.0 | 90.0 | - | 60.8 | 60.8 | 60.8 |

1. Design with subjects typed for dense (D) and framework markers (F) on all (a), many (m), or few (f) subjects
2. Default experimental setting: dense markers on many subjects and Framework markers on many subjects
3. Sens=Sensitivity; Spec=Specificity; PPV=Positive Predictive Value; True IV Sens=Sensitivity obtained using the true IV to detect MC genotyping errors
4. - = these experimental condition were not considered

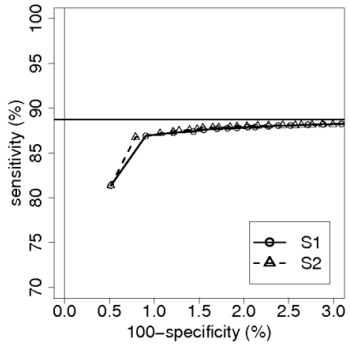
Run #1



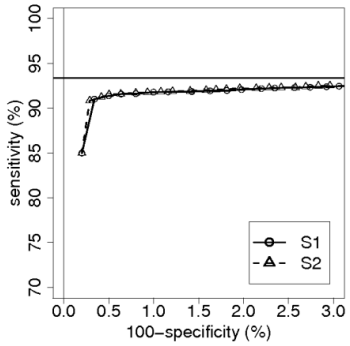

Run #2



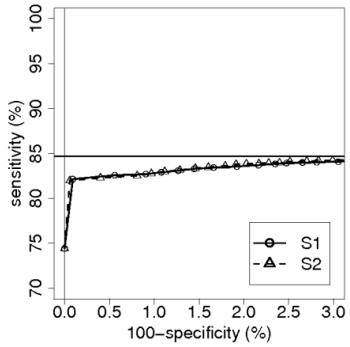

Run #3




Run #4




Run #5




Run #6

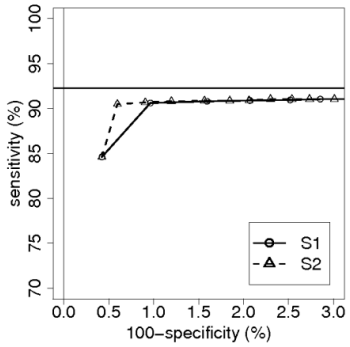



Figure 4.4 Receiver Operating Characteristic curves comparing the performance of error detection using different summary statistic as a function of thresholds. Evaluation at different thresholds, as indicated by points, increased at an increment of 1% from the rightmost 0%

Increasing the threshold of detection increased the sensitivity but decreased the specificity and the PPV (Figure 4.4). In run #1, as the threshold relaxed slightly from the most stringent level of 0% to 1%, sensitivity increased markedly (81.3% to 88.0%). As the threshold continued to relax from 1% to 20%, sensitivity increased only slightly (88.0% to 90.1%). This qualitative finding was observed in all runs. In contrast, as the threshold relaxed from 0% to 20%, specificity decreased gradually from 100% to 96.3%. A small decrease in specificity influenced the PPV substantially, as was evident by a large drop in PPV by 20.1% corresponding to a decrease of 3.7% in specificity. Even at the threshold of 0%, specificity was not 100% (eg. 99.3% for run #3 in Figure 4.4)**,** which suggests that even the most stringent threshold did not guarantee perfect specificity.

The use of more IVs also improved the ability to detect MC genotyping errors (Table 4.2). Specificity increased from 97.1% to 99.0% as the number of IVs increased from 10 to 1000, which led the PPV to increase from 83.2% to 93.5%. In contrast, sensitivity stayed roughly constant when the number of IVs used increased from 10 to 1000 (89.3% and 88.9%). Across runs, the magnitude of improvement with increasing the number of IVs varied. For instance, in run #3, PPV only increased minimally from 87.7% to 89.4% as the number of IVs used increased from 10 to 1000.

Table 4.2: Effect of varying the number of IVs used to detect MC genotyping errors

| | | Number of inheritance vectors | | | | |
|---|---|---|---|---|---|---|
| | Metric | 10 | 50 | 100 | 500 | 1000 |
| Run #1 | Sensitivity | 89.3 | 89.0 | 89.1 | 88.9 | 88.9 |
| | Specificity | 97.1 | 98.2 | 98.0 | 99.0 | 99.0 |
| | PPV | 83.2 | 88.6 | 87.6 | 93.2 | 93.5 |
| Run #3 | Sensitivity | 87.4 | 87.6 | 87.6 | 87.7 | 87.7 |
| (lowest | Specificity | 98.0 | 98.2 | 98.2 | 98.3 | 98.3 |
| diff. PPV) | PPV | 87.7 | 89.0 | 88.9 | 89.4 | 89.4 |
| Run #4 | Sensitivity | 91.8 | 91.7 | 91.8 | 91.8 | 93.5 |
| (highest | Specificity | 97.9 | 98.9 | 98.8 | 99.0 | 99.0 |
| diff. PPV) | PPV | 87.4 | 92.9 | 92.7 | 93.8 | 98.7 |

*scoring statistic S1 was used

The MAF of dense markers affected the quality metrics (Table 4.3). Higher MAF did not lead to an increase or decrease sensitivity. However, specificity dropped gradually from 100% in the MAF=[0-0.05] bin to 98.3% in the MAF=(0.40-0.50] bin, which led to a decrease of PPV from 99.7% to 89.9%. At very low MAF, the error detection approach continued to be effective at detecting genotyping errors.

Table 4.3: Quality metrics as a function of the range of minor allele frequencies of dense markers

| | | Minor Allele Frequency | | | | | |
|---|---|---|---|---|---|---|---|
| | Metric | [0-0.05] | (0.05-0.10] | (0.10-0.20] | (0.20-0.30] | (0.30-0.40] | (0.40-0.50] |
| Run #1 | Sensitivity | 88.3 | 87.5 | 87.7 | 88.9 | 90.9 | 89.1 |
| | Specificity | 100 | 99.7 | 99.2 | 99.2 | 98.5 | 98.3 |
| | PPV | 99.7 | 97.7 | 95.0 | 94.7 | 90.2 | 89.9 |

*scoring statistic S1 was used

Experiments

Typing more subjects for dense markers significantly improved the sensitivity but decreased specificity (Table 4.1). Sensitivity increased substantially when the number of subjects typed for dense markers increased (66.6% for $D_fF_a$ design vs. 89.8% for $D_mF_a$ design), when all subjects were typed for framework markers. This observation was consistent across runs (not shown). In run #1, as well as in most other runs, sensitivity did not increase much when the number of subjects typed for dense markers increased from many to all (89.8% for $D_mF_a$ design vs. 88.7% for $D_aF_a$ design), suggesting that most constraints between dense genotypes and descent patterns needed to detect MC errors were present when many subjects were typed for dense markers. However, in run #10, sensitivity continued to increase when the number of subjects typed for dense markers increased from many to all subjects (88.9% for $D_mF_a$ design vs. 95.7% for $D_aF_a$ design) (Table 4.1). In run #1, unlike sensitivity, specificity (99.7% for $D_fF_a$ , 99.2% for $D_mF_a$, and 98.9%, for $D_aF_a$) and PPV (95.8% for $D_fF_a$, 94.4%, for $D_mF_a$ and 91.9% for

$D_aF_a$) decreased as the number of subjects typed increased (Table 4.1). The decrease in specificity reflects an increase in false detections. An increase in false detections likely occurs because typing more subjects for dense markers increases the ability to distinguish that many inferred IVs are not the true IV, which is an alternative explanation for inconsistency.

In contrast, typing more subjects for framework markers did not greatly improve performance. Sensitivity, specificity, and PPV only increased slightly as more subjects were typed for framework markers (Table 4.1). For instance, when only a few subjects were typed for the dense markers, increasing the number of subjects typed for framework markers from few up to all subjects ($D_fF_f$ design to $D_fF_a$ design) only slightly increased sensitivity from 65% to 66.6%, specificity from 99.5% to 99.7%, and PPV from 94% to 95.8% for run #1 (Table 4.1). Even though these increases seemed trivial, these increasing trends were observed in all runs (not shown).

While the number of subjects typed for framework markers did not substantially affect the quality of error detection, the spacing of framework markers substantially affected specificity and PPV (Table 4.4). The specificity decreased moderately as the spacing of framework panel increased from a marker per 0.5 cM to a marker per 1 cM (99% to 96.1%) but decreased substantially as the spacing of framework panel decreased from a marker per 1 cM to a marker per 2 cM (96.1% to 82.8%). Even though specificity dropped merely by 2.9% when density decreased from the 1 marker per 0.5 cM to 1 marker per 1 cM, PPV decreased substantially from 93.5% to 78.2%. At 1 marker per 2 cM, PPV was only 45%. A histogram of the percentage of IVs consistent with observed dense genotypes suggested that a higher percentage of markers had very low consistency as the framework density decreased (Figure 4.5). In some other runs, such as run #6, the PPV did not decrease (83.1% to 83.6%) when the density decreased from a marker

per 0.5 cM to a marker per 1 cM. Nevertheless, the PPV did dramatically decrease as the density decreased from a marker per 1 cM to a marker per 2 cM.

Table 4.4: Effect of varying the density of framework markers

| | | Framework Marker Density (cM between markers) | | |
|---|---|---|---|---|
| | Metric | 0.5 | 1 | 2 |
| Run #1 | Sensitivity | 88.9 | 86.5 | 87.8 |
| | Specificity | 99.0 | 96.1 | 82.8 |
| | PPV | 93.5 | 78.2 | 45.0 |
| Run #6 | Sensitivity | 91.0 | 89.4 | 88.1 |
| | Specificity | 97.1 | 97.3 | 85.8 |
| | PPV | 83.1 | 83.6 | 48.9 |

**Run #6 - smallest diff. in PPV between density 0.5 and 1

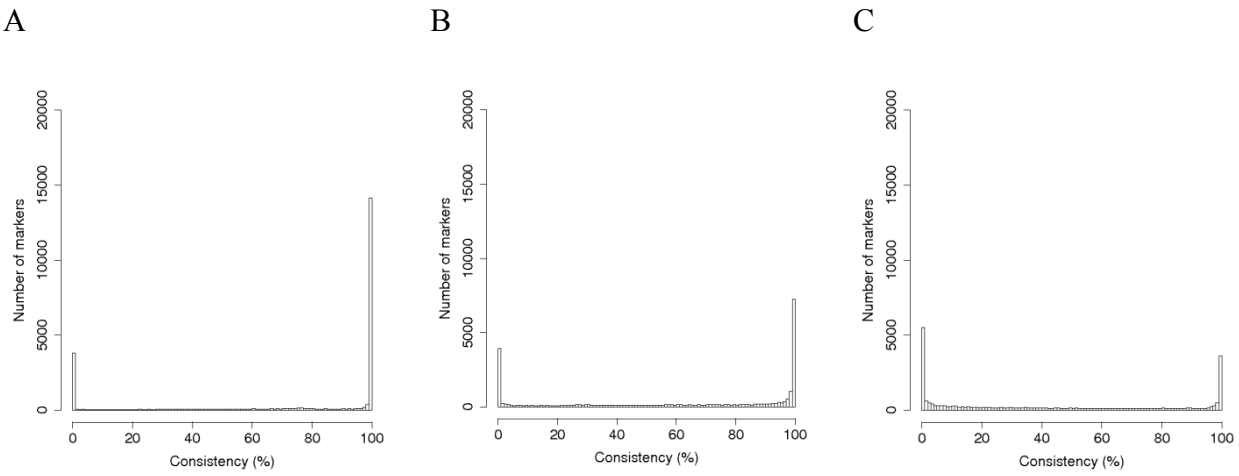



Figure 4.5 Varying the density of the framework panel affects the distribution of the percentage of sampled IVs that are consistent with the observed genotypes across markers, as illustrated using results from run #1: A) 0.5 cM; B) 1 cM; C) 2 cM between adjacent framework markers

A higher error rate greatly increased the PPV but not the sensitivity and specificity (Table 4.5). Whereas the PPV was only 22.8% when the allelic error rate was 0.005%, the PPV increased to 98.7% when the allelic error rate increased to 1%. Because sensitivity and specificity were similar across all error rates, this increase in PPV was mainly because at higher rates there were more errors to detect.

Table 4.5: Effect of different allelic error rates

| | | Allelic Error Rate | | | |
|---|---|---|---|---|---|
| | Metric | 0.005% | 0.05% | 0.25% | 1% |
| Run #1 | Sensitivity | 87.7 | 86.9 | 88.9 | 91.5 |
| | Specificity | 99.0 | 99.0 | 99.0 | 99.0 |
| | PPV | 22.8* | 72.8 | 93.5 | 98.7 |

* the total number of errors was low

When fewer numbers of subjects were typed for dense di-allelic markers, a higher percentage of the errors were MC (Table 4.6). For instance, at an allelic error rate of 0.25%, the percentage of MC errors was 45.5% when all subjects were observed for dense markers. The percentage of MC errors increased to 94.8% when only a few subjects were observed for dense markers, respectively. This result illustrates that especially when only a few subjects are observed, most genotyping errors will remain undetected by only using a MI check.

Table 4.6. Effect of the number of subjects typed for dense markers on the percentage of genotyping errors that are MC– run #1: Total number of errors (%MC)

| Dense marker design | Allelic error rate | | | |
|---|---|---|---|---|
| | 0.005% | 0.05% | 0.25% | 1% |
| All | 122 (38.5%) | 1240 (49.7%) | 5783 (45.5%) | 16291 (36.9%) |
| Many | 95 (85.3%) | 24132 (81.7%) | 4175 (80.0%) | 12856 (75.1%) |
| Few | 59 (94.9%) | 542 (93.0%) | 2617 (94.8%) | 9020 (93.2%) |

### 4.4.5 Discussion

I highlight a few results. First, my simulated results show that my error detection approach can achieve high sensitivity, specificity, and PPV. Second, when a true genotyping error is detected, the approach can identify with high accuracy a subject who has a genotyping error. Third, the test statistic S1 has almost identical sensitivity as the test statistic S2. The incorporation of an error model and allele frequencies into S2 does not improve the ability to globally detect genotyping errors, which suggests that the constraint from sampled IVs largely determines the ability of our approach to detect errors.

My error detection approach excels in detecting MC genotyping errors in rare variants. The results from simulations show that specificity and PPV increase as the minor allele frequency of the marker increases, while sensitivity stays roughly constant. A simple explanation for this increasing trend in specificity is that clean markers with lower MAF contain a higher percentage of homozygous alleles for the major allele in observed genotypes, so they are consistent with a larger set of IVs when the genotypes do not have errors.

The quality of sampled IVs affects the ability to detect genotyping errors. Evaluation suggests that the use of inferred IVs is almost as effective as the use of the true IV to identify errors. However, specificity of error detection increases if I use IVs that are closer to the truth, since the use of the theoretical true IV always yields perfect specificity. An important consideration is that my detection approach requires the use of a moderate number of framework markers to infer IVs to increase specificity to an acceptable level, a result which agrees with a general earlier remark that multipoint error detection methods are effective only when the marker density is relatively high [Chang, et al. 2006]. More precisely, if the framework markers are not informative about the descent pattern at the dense markers, a high fraction of IVs will likely be inconsistent even with clean observed dense genotypes. Under such a scenario, the specificity of error detection would be low because clean markers would also be flagged as containing error. This result is in contrast to genotype imputation which does not require stringent marker density to achieve high imputation accuracy (Section 3.4). Also, contrary to imputation, for error detection marker density is more crucial than who is typed for framework markers. Regardless, informative markers should still be used, and the use of other informative framework markers such as multi-allelic markers may improve the inference of IVs.

If we are not confident that our framework markers are free of errors, it is beneficial to run the error detection procedure using a few sets of framework markers. Framework markers that contain genotyping errors may lead to biased sampling of IVs [Lebrec, et al. 2008; Markus, et al. 2011]. To reduce the number of false error calls, we may consider using a few sets of framework markers to detect genotyping errors.

We can greatly improve the computational time. Our results from simulations suggest that the use of a smaller set of IVs decreases specificity but does not substantially decreases

sensitivity. Because of these properties, we may first run the error detection program on a small number of IVs (e.g. 10 IVs) as an initial scan with a more lenient threshold to screen out most markers that are free of errors. Then, on a much reduced set of markers that are enriched for errors, we can re-run the error detection program using a full set of IVs. In addition, because S1 is much faster than S2, and because the two scores give comparable performance, S1 should be used in the initial scan. Then, on a subsequent scan, S2 can be used to additionally identify the subjects that most likely have genotyping errors.

The gl_auto program has a practical limit on the number of framework markers that we can use. Because of the potential issue in MCMC mixing and the violation of the assumption that markers are not in linkage disequilibrium as part of the Lander-Green algorithm, framework markers can not be too closely spaced. For instance, MCMC mixing issue could lead to biased sampling of IVs that are not similar to the true IV, so observed genotypes could be inconsistent with all those finite number of sampled IVs. However, if at some point we can integrate even more markers into the inference of IVs, a more accurate and precise set of IV may be obtained. With such set of IV, the approach described here would likely show improvement in specificity and PPV.

The results from evaluating the effect of varying the allelic error rates and observed data pattern for dense genotypes have three implications. First, known information about error rate should be taken into consideration when choosing a threshold for error detection. If the error rate is low, a higher fraction of flagged markers will likely to be false positives, so a more stringent threshold may be appropriate. Second, when only a few subjects are typed for the dense markers, my approach, as in other existing MC error detection approaches, has reduced ability to detect MC genotyping errors. Third, when only a few subjects are typed for dense markers, a high

percentage of errors will be MC. This result illustrates that the use of only the MI check alone to detect genotyping errors is inadequate.

## 4.5 Comparison with Existing Methods

In this section, I compared several existing approaches for detecting MC genotyping errors. I considered Sibmed [Douglas, et al. 2000], Eclipse [Sieberts, et al. 2002], Mendel [Sobel, et al. 2002], Simwalk2 [Sobel, et al. 2002], Merlin [Abecasis, et al. 2002], and my approach, all of which belong to the same class of approach that uses descent patterns to detect MC genotyping errors. I first describe their similarities and differences. Then, I evaluate and compare these approaches by simulation.

### 4.5.1 Model Comparison

Most of these approaches have constraints on the structure of pedigrees. Sibmed handles sibling-pairs only. Eclipse handles various configurations of three individuals. Merlin, Mendel, and my approach handle small general pedigrees. Merlin and Mendel, however, cannot handle large pedigrees because the computational burden increases exponentially with the number of non-founders. The use of MCMC methodology enables my approach and Simwalk2 to handle large pedigrees. Because the basic framework of Mendel and Simwalk2 are identical, I will omit Simwalk2 for brevity. Table 4.7 summarizes the different approaches.

Table 4.7 Summary of approaches to detect MC genotyping errors.

| | My approach | Mendel/ Simwalk 2 | Merlin | Sibmed | Eclipse |
|---|---|---|---|---|---|
| Features: | | | | | |
| Pedigree | Large | Small/ Large | Small | Sibpair | Three subjects |
| Suitable for many dense markers | Yes | No | Yes? | Maybe | Maybe |
| Identify subject with error | Yes | Yes | Yes? | No | No |
| Need to supply allele frequencies | No in S1; Yes in S2 | Yes | Yes | Yes | Yes |
| Detect errors in all markers | No[A] | Yes | Yes | Yes | Yes |
| Model: | | | | | |
| HMM Likelihood evaluation | Exact or MCMC | Exact/ MCMC | Exact | Exact | Exact |
| Deal with markers in LD | Yes | No | Option? | No | No |
| Use an error model | Yes[B]/No | Yes | No | Yes | Yes |
| Assume error in at most 1 genotype | Yes | No | Yes | No | No |
| Detect errors using | Statistic/ Prob. | Prob. | Statistic | Prob. | Prob. |

A. We assume that framework markers are clean- can check framework markers by using different framework panels

B. No in S1, Yes in S2: See Figure 4.7

All these approaches use a HMM. The underlying hidden states of the four approaches are identical. While the hidden states in Sibmed and Eclipse are the number of shared alleles identical-by-descent among individuals at each locus, these hidden states can also be expressed as IVs, as in Mendel, Merlin, and my method. Merlin uses the basic Lander-Green algorithm and without an error model (Figure 2.1). In contrast, other approaches extend the Lander-Green algorithm to model errors.

Sibmed, Eclipse, Mendel, and my approach model genotyping errors. All of these methods assume that error in each genotype occurs independently with an error rate $\varepsilon$. Nevertheless, these approaches model errors differently. Sibmed simply models errors by distinguishing whether the observed genotypes at a locus are completely correct or not. If the observed genotypes of the sibings contain an error, Sibmed calculates the probability of observed genotypes as if they are randomly selected from the population: i.e. $P(G_k|S_k; error) = P(G_{1k})P(G_{2k})$ at the k$^{th}$ locus, where $G_{1k}$ and $G_{2k}$ are the genotypes of the first and second siblings, respectively. Specifically, Sibmed calculates the probability of observed genotypes conditional on the alleles that are not IBD. Unlike Sibmed, Eclipse, Mendel, and my approach model the true genotype at the level of each observed subject. Eclipse and Mendel explicitly model the true genotypes within the HMM framework at each locus (Figure 4.6). The true genotypes at each locus are hidden states which may be different from the observed true genotypes. The calculation of the likelihood of the observed genotypes involves enumerating different possibilities of what the true genotypes are. While Sibmed, Eclipse and Mendel model genotyping error at each marker, my approach assumes that framework markers have 0% error (clean). The second test statistic of my method only models error at each test position (Figure 4.7).

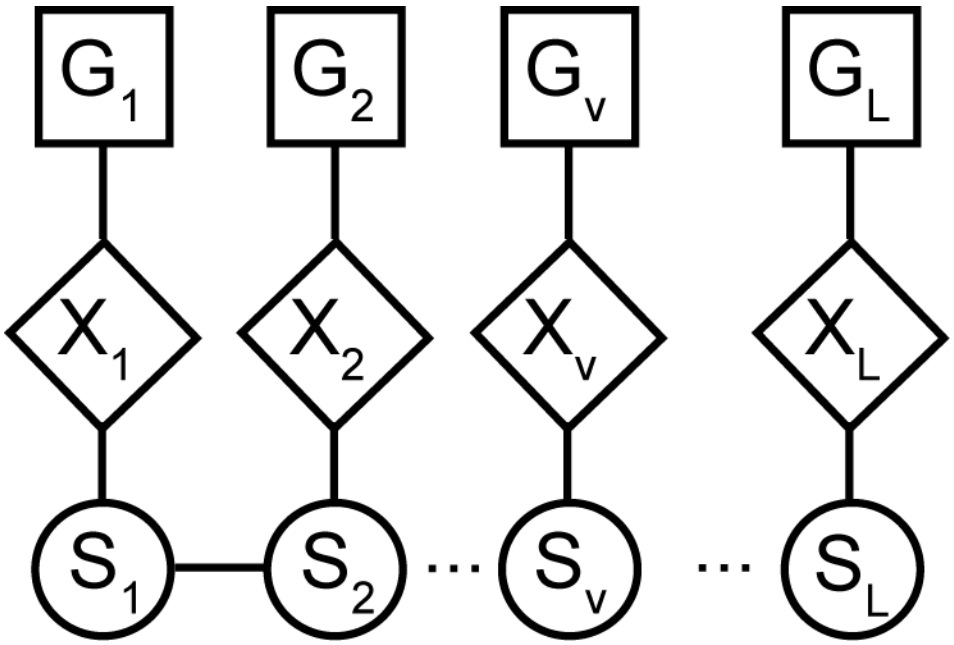



Figure 4.6 HMM with error model at every locus in Eclipse and Mendel . Let S, X, G, and subscript denote the IV, true genotypes, observed genotypes, and locus index, respectively.

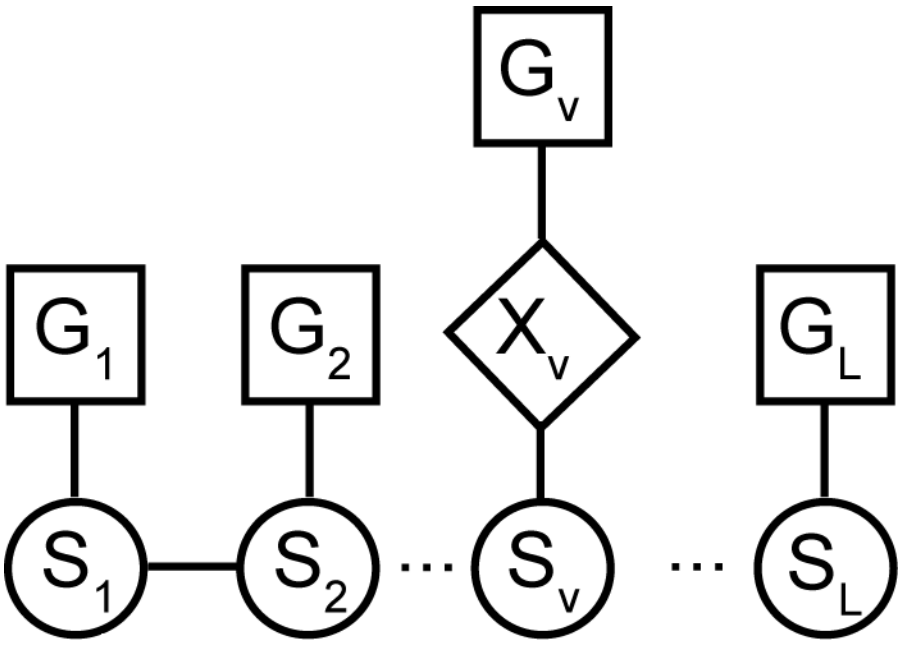



Figure 4.7 HMM with error model only at the position of test position in my approach. Let S, X, G, and subscript denote the IV, true genotypes, observed genotypes, and locus index, respectively.

Eclipse, my approach, and Mendel use different error models. Eclipse and my approach assume that at most one of the two alleles in a genotype contains error. If a genotyping error is present, the true allele is equally likely to be of any other allelic types. This restriction alleviates the need of having to enumerate all possibilities of what are the true genotypes. In my approach, the model further restricts one genotyping error per locus. Mendel has options for different error models. The first error model is identical to that of Eclipse. Another error model integrates population allele frequencies of different alleles. I will consider this as the default error model in Mendel. Mendel also implements other error models that are based on empirical error rates, but they are not discussed here.

These approaches use different test statistics to detect genotyping errors. Sibmed, Eclipse, and Mendel calculate the posterior probability of no error. Because Mendel considers configurations of other possible true genotypes, Mendel can also calculate the posterior probability of other true genotypes to suggest the most likely true genotype configuration. Such feature was not implemented in Eclipse. My approach calculates either the percent consistency or the posterior probability of genotyping configuration as test statistics. Like Mendel, the latter statistic can be used to identify the most likely error genotype, and thus the most likely subject who contains error. The error detection procedure implemented in Merlin does not use posterior probability and instead computes a test statistic for each observed genotype omitted one at a time, thus enabling Merlin to inform us which genotype most likely contains an error [Abecasis, et al. 2002; Mukhopadhyay, et al. 2004]. For each genotype $g$, Merlin computes a genotype mistyping score $\gamma$. A large value of $\gamma$ suggests the presence of genotyping error in genotype $g$. This test statistic $\gamma = r_{unlinked}/r_{linked}$. The term $r_{linked} = \frac{L(G\backslash g|\theta)}{L(G|\theta)}$, where $L(G|\theta)$ and $L(G\backslash g|\theta)$ are the likelihood of all observed genotypes and the likelihood of all observed genotypes but omitting genotype $g$, respectively, computed using the Lander-Green algorithm with recombination fractions between markers denoted collectively by $\theta$. The term $r_{unlinked} = \frac{L(G\backslash g|\theta=\frac{1}{2})}{L(G|\theta=\frac{1}{2})}$ are the ratio of analogous likelihoods calculated assuming that all markers are unlinked. For ease of comparison, I used the inverse of the test statistic, so the test flags genotyping error if the test statistic is below a threshold. In a marker, a test statistic was calculated by omitting each observed genotype. I used the minimum of the test statistics as the test statistic to detect errors for each marker.

### 4.5.2 Evaluation

I simulated data on small pedigrees to test the performance of these methods. I simulated a 5-member nuclear family (2 parents+ 3 siblings) with all subjects observed for genotypes. I simulated a small chromosome of 10 cM. In each of the 1000 datasets, I simulated 30 SNPs with MAF of 0.5. For simplicity, I simulated the clean data by dropping alleles through randomly simulated descent pattern that did not contain recombination. In each dataset, I created an error dataset by changing an allele of the $15^{th}$ marker (middle marker) of one randomly selected subject and a double-error dataset by changing an allele of the $16^{th}$ marker (neighboring marker) on another randomly selected subject.

I compared the performance of the different approaches. First, I checked and discarded any datasets that failed the Mendelian Inconsistency check. In the MC datasets, I checked for genotyping errors using my own implementations of various approaches: Sibmed, Eclipse, Mendel, Merlin, and my approach. The implementations were all based on exact calculation under the HMM framework and without any stochastic component. I extended Sibmed (Sibmed-e) and Eclipse (Eclipse-e) for use in 5 observed individuals. Sibmed-e, Eclipse-e, and Mendel are only different because of their error models, as were already discussed. I calculated the sensitivity and specificity of detecting errors in the middle marker and used these summaries as the main metric to evaluate performance. In addition, I calculated the sensitivity and specificity of detecting errors in the middle marker when the neighboring marker has a MC genotyping error to investigate the robustness of various methods to detect error given error at a neighboring marker. I flagged an error if each test statistic in Sibmed, Eclipse, Mendel, and my approach is below 0.10 and flagged an error if the test statistic in Merlin is smaller than 0.01. As would be

seen in the results section, any sensible choice of thresholds as those that I have chosen would yield similar interpretation.

I also investigated the effect of not simultaneously using genotypes of all observed subjects. Since the original algorithm of Sibmed is restricted to the use in sibling pairs without parents, I split the pedigree into three sibling pairs. On each dataset, I ran Sibmed on each sibling pair and flagged genotyping error if the test statistic in any pair was below 0.10. Like Sibmed, the original algorithm of Eclipse does not use parental genotypes. Analogously, I ran Eclipse on three siblings to evaluate the loss of not using parental genotypes.

### 4.5.3 Results of Evaluation

The approaches performed equally well when all 5 subjects were observed. All methods achieved a high sensitivity of 89.0% among datasets with MC genotyping errors. All methods achieved perfect specificity (100%). In probability-based methods, the test statistics of error data were either high or extremely low (Figure 4.8), while the test statistics of clean data was all quite high (Figure 4.9). Since the test statistics never dropped below 0.8, any threshold between 0.1 and 0.8 would achieve the same sensitivity and sensitivity.

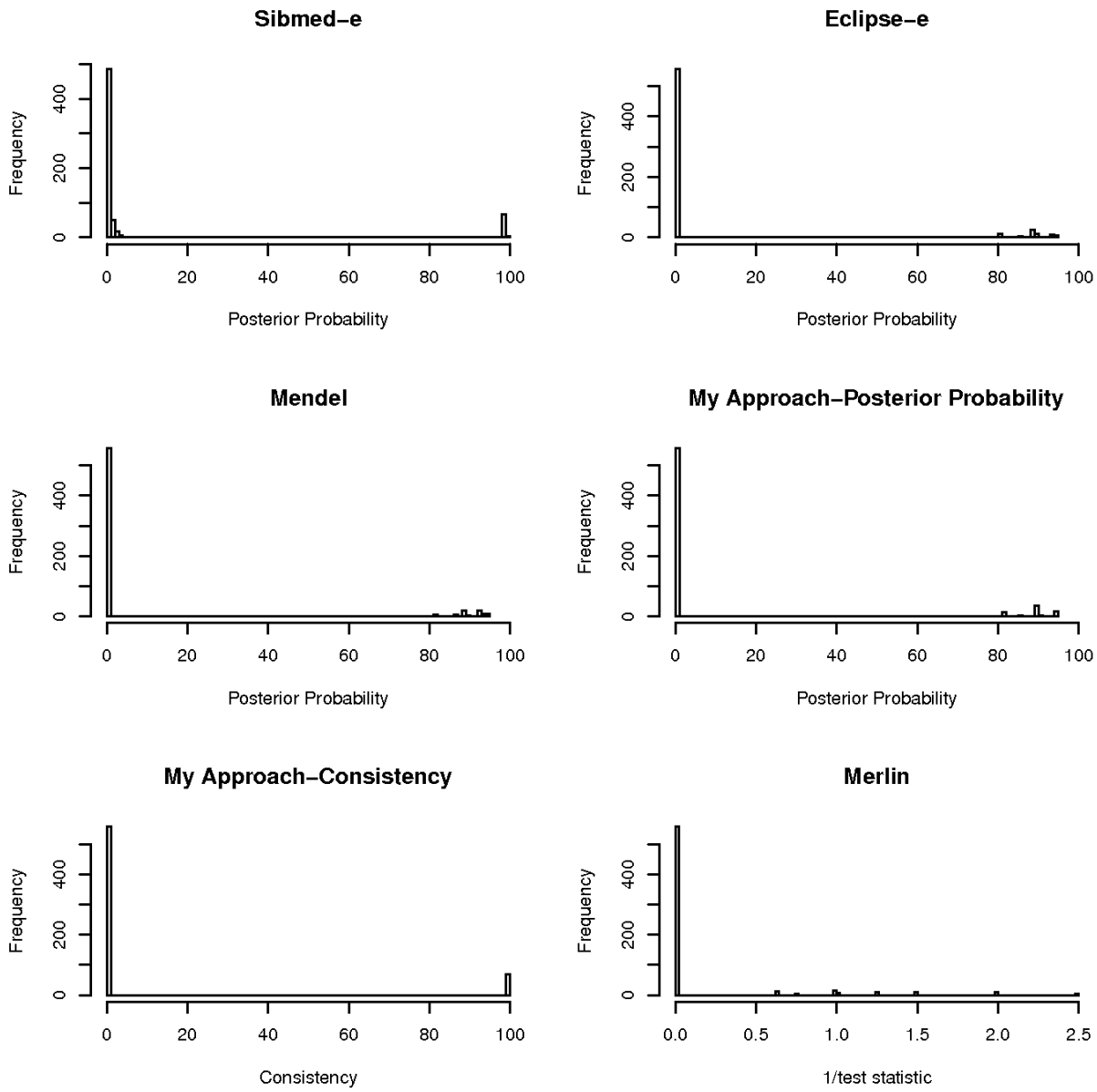



Figure 4.8 Sensitivity analysis: test statistics at the middle marker in datasets that contain errors in the middle marker.

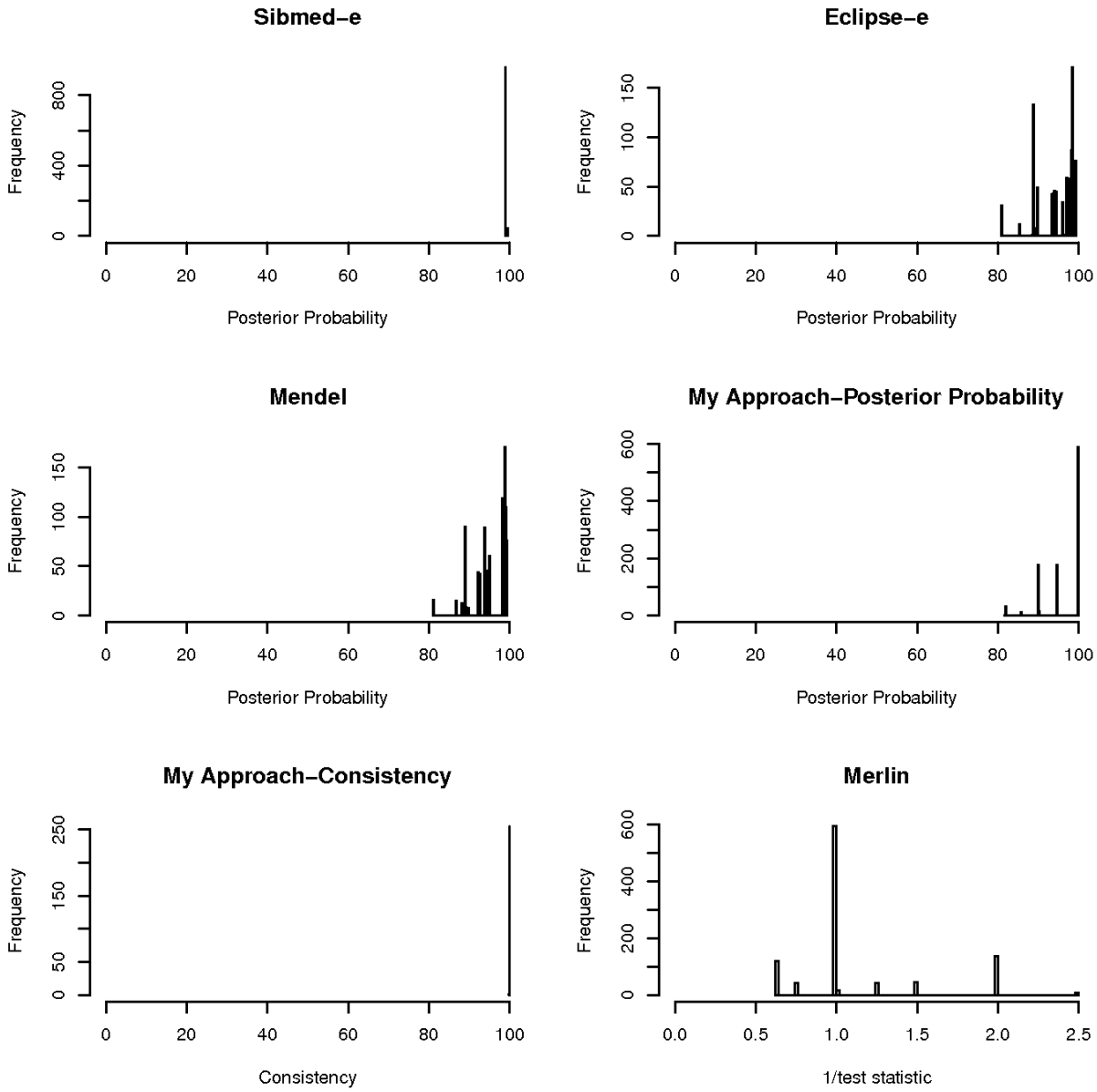



Figure 4.9 Specificity analysis: test statistics computed at the middle marker in clean datasets.

The presence of MC genotyping error at the neighboring marker affected the performance of some but not all approaches (Table 4.8). Sensitivity was the lowest in Merlin (60.3%) and my approach (63.8% and 62.8%). Sibmed-e, which treated all errors as one group, also had lower sensitivity compared to Mendel and Eclipse-e (68.7% in Sibmed-e vs. 89.1% in Eclipse-e and Mendel). In fact, Sibmed-e and Eclipse-e did not suffer from genotyping errors at the neighboring marker. At the threshold of 0.1, my approach with the consistency statistic had a specificity of 99.75%, while all other approaches continued to achieve 100% specificity. In approaches that had lower sensitivity, a fraction of the test statistics were dispersed towards the direction of no detection (Figure 4.10). In my approach, error at the neighboring marker shifted a fraction of the test statistics towards low values, which is the direction of detection, in datasets that did not contain error in the middle marker (Figure 4.11). This result suggested that in my approach, I should choose a low threshold to ensure high specificity when I use percent consistency as the test statistic to minimize false positives. Another interesting observation is that when the framework marker has an error, the use of posterior probability as the test statistic seems to be more robust in protecting specificity than the use of percent consistency (Figure 4.11).

Table 4.8. Performance of testing for MC genotyping errors at the middle marker in the presence of MC genotyping error at the neighboring marker.

| | Sibmed - extended | Eclipse - extended | Merlin | Mendel | My method (Consistency /Posterior Probability) | My method– no error at neighboring marker |
|---|---|---|---|---|---|---|
| Sensitivity | 68.7 | 89.1 | 60.3 | 89.1 | 63.8/62.8 | 89.1 |
| Specificity | 100 | 100 | 100 | 100 | 99.75/100 | 100 |

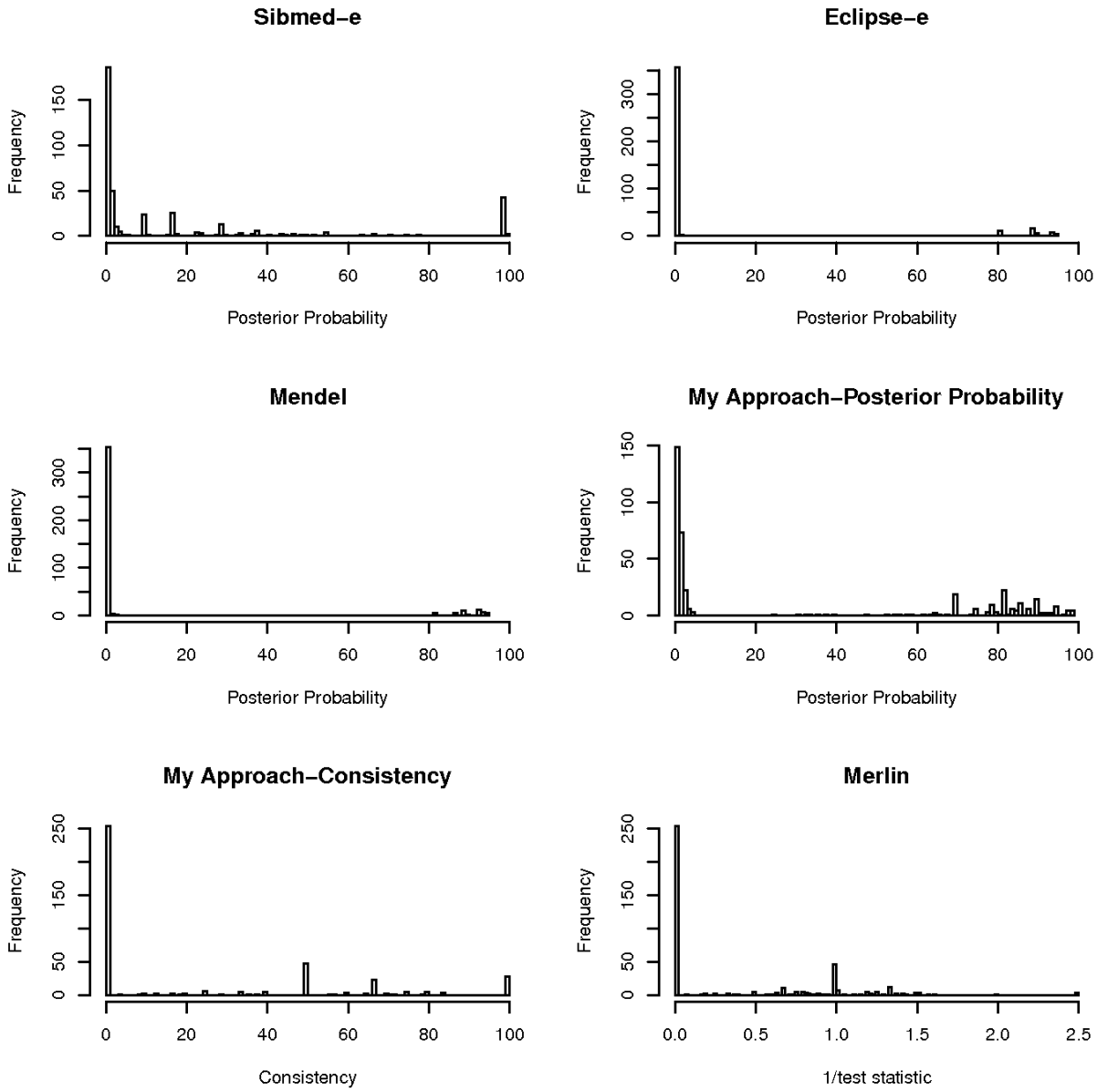



Figure 4.10 Sensitivity analysis: Test statistics at the middle marker in datasets with MC genotyping errors in both the middle marker and neighboring marker.

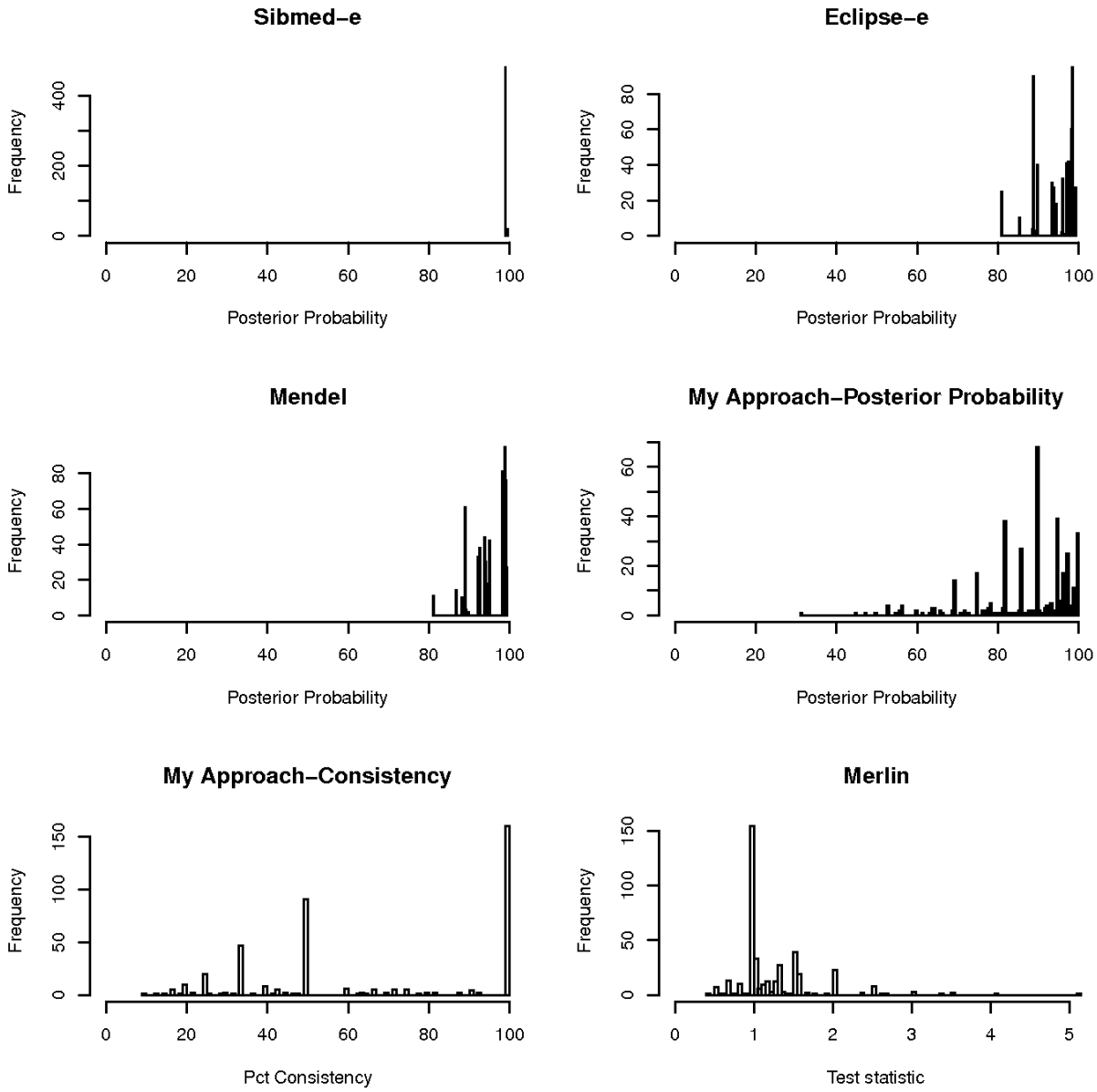



Figure 4.11 Specificity analysis: Test statistics of the middle marker in datasets with MC genotyping errors in the neighboring marker.

Not simultaneously using all observed subjects decreased the power to detect MC genotyping errors (Table 4.9). Splitting the trio of children into 3 pairs decreased sensitivity to detect MC errors in children slightly from 53.2 to 48.8%. In the trio of children, Eclipse, which

used a more complicated error model, slightly outperformed Sibmed (53.7% vs. 53.2%). Availability of genotypes from parents substantially increased sensitivity to detect MC errors in children (100%).

Table 4.9. Different use of genotype data affect sensitivity (%) in detecting MC genotyping errors

| | Pairwise children-Sibmed | Children trio – extended-Sibmed | Children trio – Eclipse | Parents and children – Eclipse/all other methods |
|---|---|---|---|---|
| Children | 48.8 | 53.2 | 53.7 | 100 |
| Parents | 0 | 0 | 0 | 73.7 |
| Overall | 28.4 | 30.9 | 31.3 | 89.0 |

### 4.5.4 Discussion

What detection threshold should we choose in probability-based methods? It was suggested that the choice should depend on the pedigree structure and observed patterns [Mukhopadhyay, et al. 2004]. My result from previous section shows the test statistics are either very low or quite high in markers that contain genotyping errors, which suggests that some MC errors are detectable while others are not. This observation is consistent with the hypothesis that since information to detect MC errors comes primarily from inconsistency between observed alleles and the most likely inferred IVs, MC genotyping errors that are detectable should lead to low posterior probabilities that the joint genotypes do not contain an error. We observe that error at the neighboring marker inflates a portion of probabilities in error data and deflates a portion of

probabilities in otherwise clean data. This outcome is also likely in other suboptimal settings where IVs are not precisely or accurately inferred. Hence, if we need to test many markers, we may wish to choose a low threshold to ensure high specificity to reduce the number of false positives. Ultimately, the type of downstream genetic analysis we want to perform can influence the choice of threshold.

The simulation explored the question of whether modeling errors is useful. The fact that the use of percent consistency as a test statistic used in my method and the ratio statistic used in MERLIN detects as many MC genotyping errors as do methods that uses error models suggests that information to detect MC errors comes primarily from inconsistency between observed alleles and the most likely inferred IVs. Nevertheless, in the presence of multiple genotyping errors at nearby loci, result suggests that having accurately inferred IVs affects the quality of error detection. Explicitly modeling errors at every locus improves the inference of IVs. Results also show that the error model in Sibmed is less adequate than those in other error models, which model error for each observed genotype. The error model in Sibmed was originally developed for handling sibling pairs, so it was not the ideal choice for handling more general pedigree structures. The error models in Eclipse and Mendel work well, which shows that incorporating errors at the level of per genotype is beneficial. These error models would likely work well for other pedigree structures and likely will not benefit much from increasing the model complexity, since most information to detect errors come from the compatibility between most likely states of IVs and observed genotypes, as was argued in Section 4.4.

My approach does not model errors at every locus. Rather, it assumes that we already have clean framework markers to infer IVs. These framework markers can be derived from the use of a subset of dense markers. Because framework markers are relatively sparse, we can

expect that even if genotyping errors are present in the framework panel, the number of genotyping errors will be much fewer by chance. If genotyping error in framework markers is a concern, we can use multiple framework panels to detect genotyping errors to minimize the problem associated with errors in framework markers.

My error detection approach has a few unique features that enable efficient computation for detection of errors in many markers on large pedigrees. First, I separate the inference of IVs from error detection. This feature prevents wasteful computation associated with using markers that do not add much power and prevents the inference of IVs from violating the assumption that markers are in Linkage Equilibrium, which are inherent in all existing methods. In principle, when inferred IVs are informative of the true descent pattern, error detection on dense markers should be high in sensitivity and specificity. Second, I sample IVs at dense positions by using IVs sampled at framework positions. This feature allows me either to use percent consistency (S1) as a test statistic or to calculate posterior probability of true genotype (S2) at each test position. The percent consistency test is very computationally efficient because unlike approaches that use an error model, it does not require enumeration over possibilities of true genotypes. Third, in test statistic S2, in which I incorporate an error model, I restrict genotyping error to only one subject per marker. Unless the error rate is high or the pedigree is quite large, genotyping errors will rarely occur in multiple individuals. Thus, this implementation speeds up the calculation at each test position. With slight modification, less restricted enumeration of genotype configurations can be achieved to allow for, e.g., two genotyping errors at a locus, if desired.

I used idealistic simulation conditions. The simulation assumes that there is no LD between markers, markers are relatively close to each other, markers have high MAF and tend to

be more informative, there is no recombination in this region, and all 5 subjects are observed. These conditions enable all these HMM-based approaches to infer IVs well, which is a requirement for all these methods to work well. Since the simulation conditions are ideal, my simulation also illustrates that some MC genotyping errors are not detectable even if IVs are well-inferred.

## 4.6 Visualizing Genotyping Errors and the Joint Consistency Test

Using my approach for error detection, I tested for genotyping errors on a 62-member large pedigree (Figure 4.12). In a ~30cM region of a chromosome of interest, 4333 SNPs were previously cleaned by MI checks. To detect MC genotyping errors, I computed the percent consistency for each SNP.

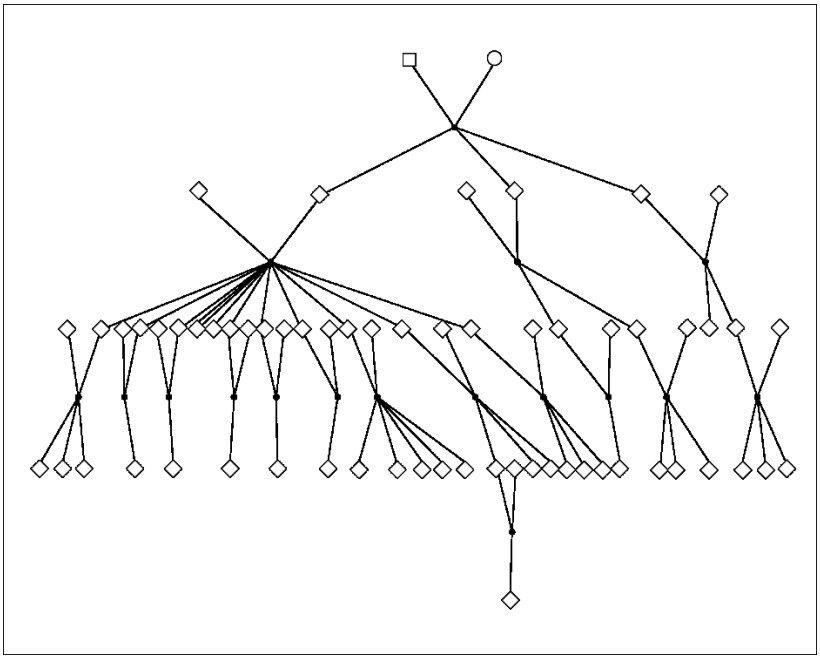


Figure 4.12 Detection of Mendelian consistent genotyping errors is performed on a 62-subjects 5-generation real pedigree.

### 4.6.1 Using Sampled IVs to Detect Mendelian Consistent Genotyping Errors

The percent consistency varied across SNPs. While most SNPs had a high percent consistency, some SNPs had very low percent consistency (Figure 4.13). A total of 36 SNPs had genotypes that were inconsistent with all sampled IVs at their respective loci. These markers were flagged as candidates for having MC genotyping errors.

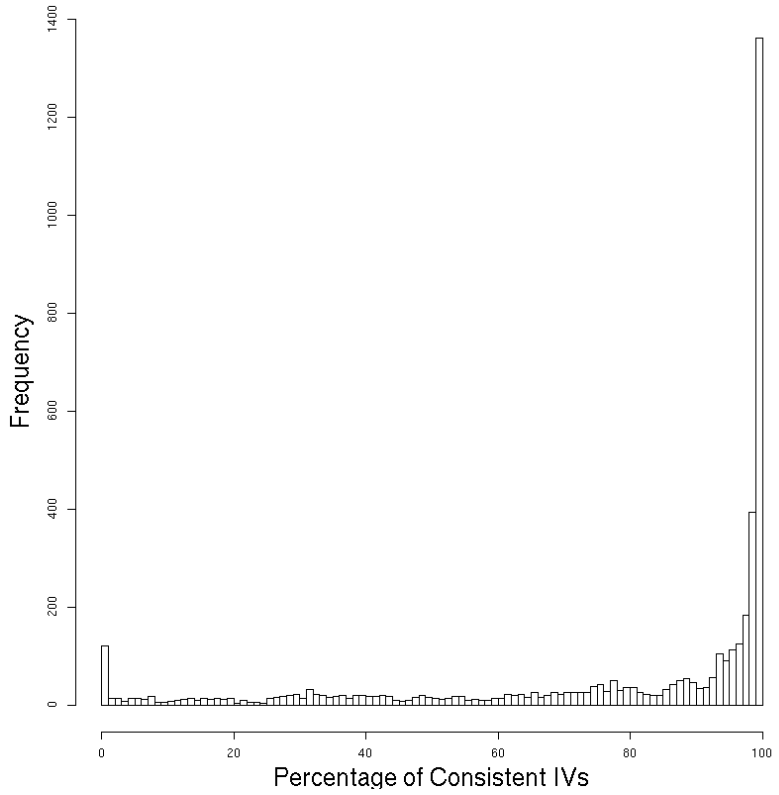



Figure 4.13 Histogram of the percentage of consistency IVs of SNPs in the R-pedigree data

I also plotted the percent consistency across SNPs (Figure 4.14). It contained a few interesting features. First, there were some markers with low consistency that could be candidates for having genotyping errors. Second, the percent consistency often fluctuated greatly

from one SNP to another. Third, there was correlation in the percent consistency between nearby SNPs. For instance, between marker 3000 and 3200, the percent consistency was either at the top or moving on a gradually increasing baseline. In addition, there was huge fluctuation between high percent consistency and low percent consistency between marker 2450 and 2900 and between marker 3950 and the last marker.

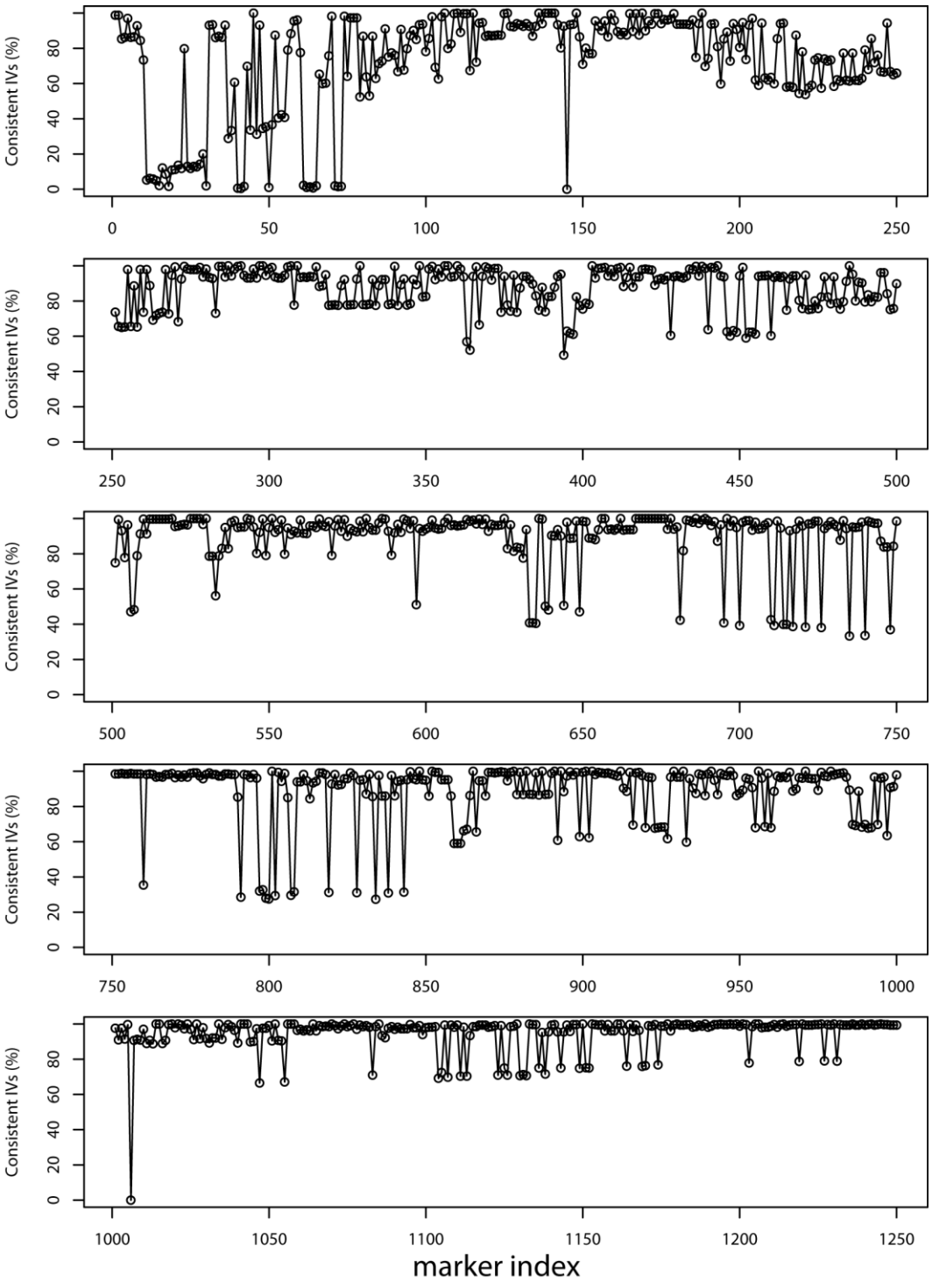
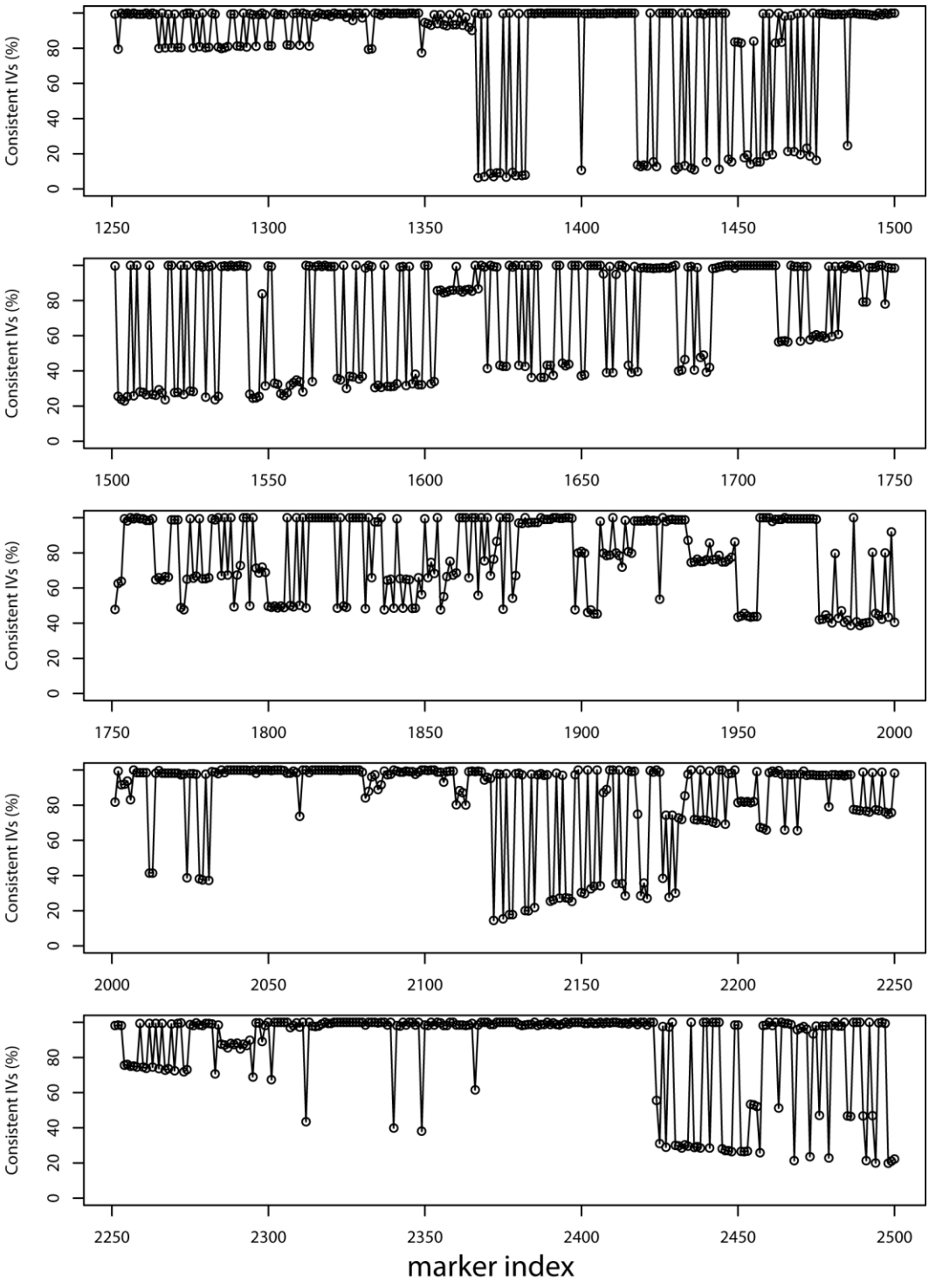
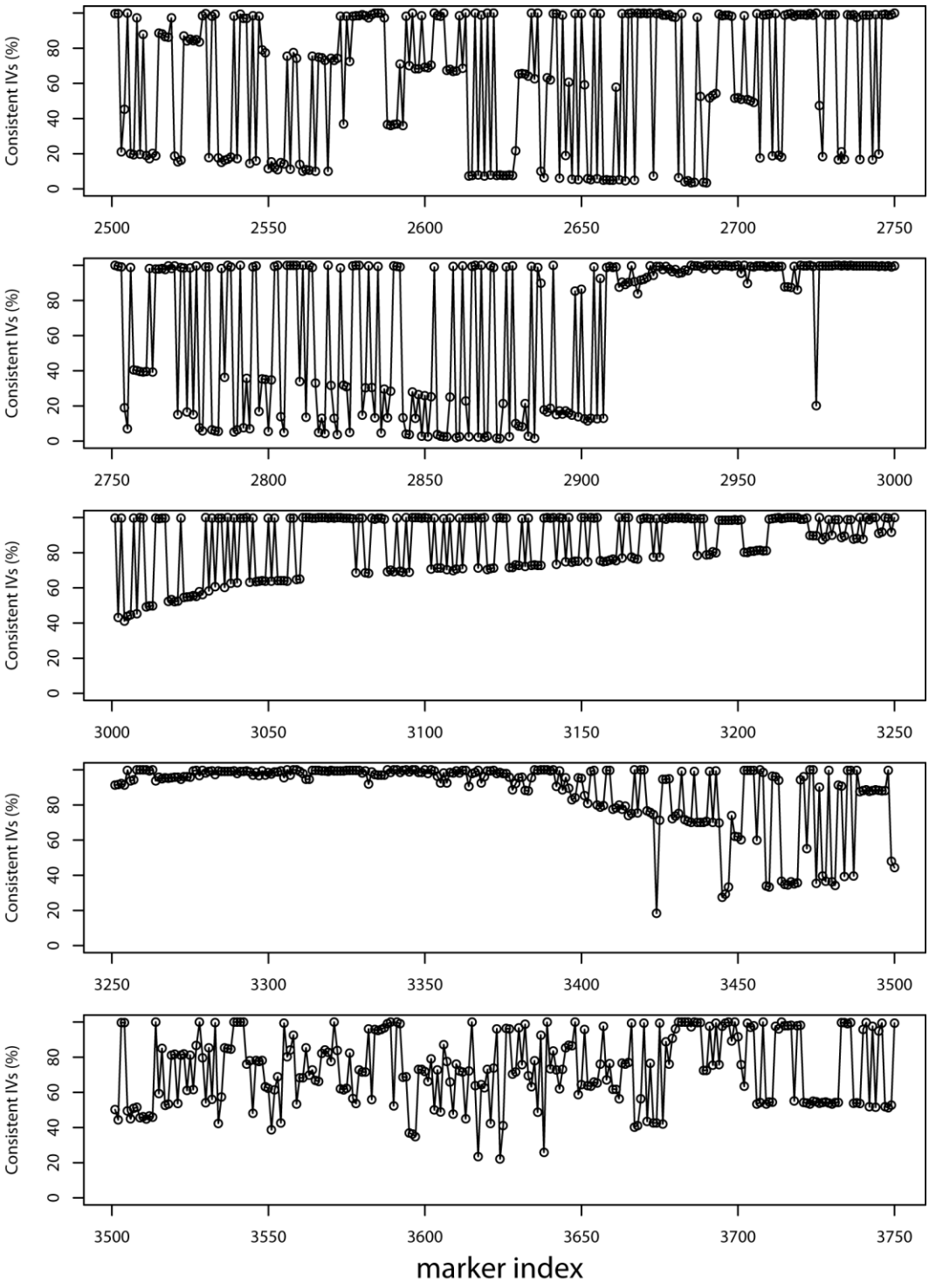
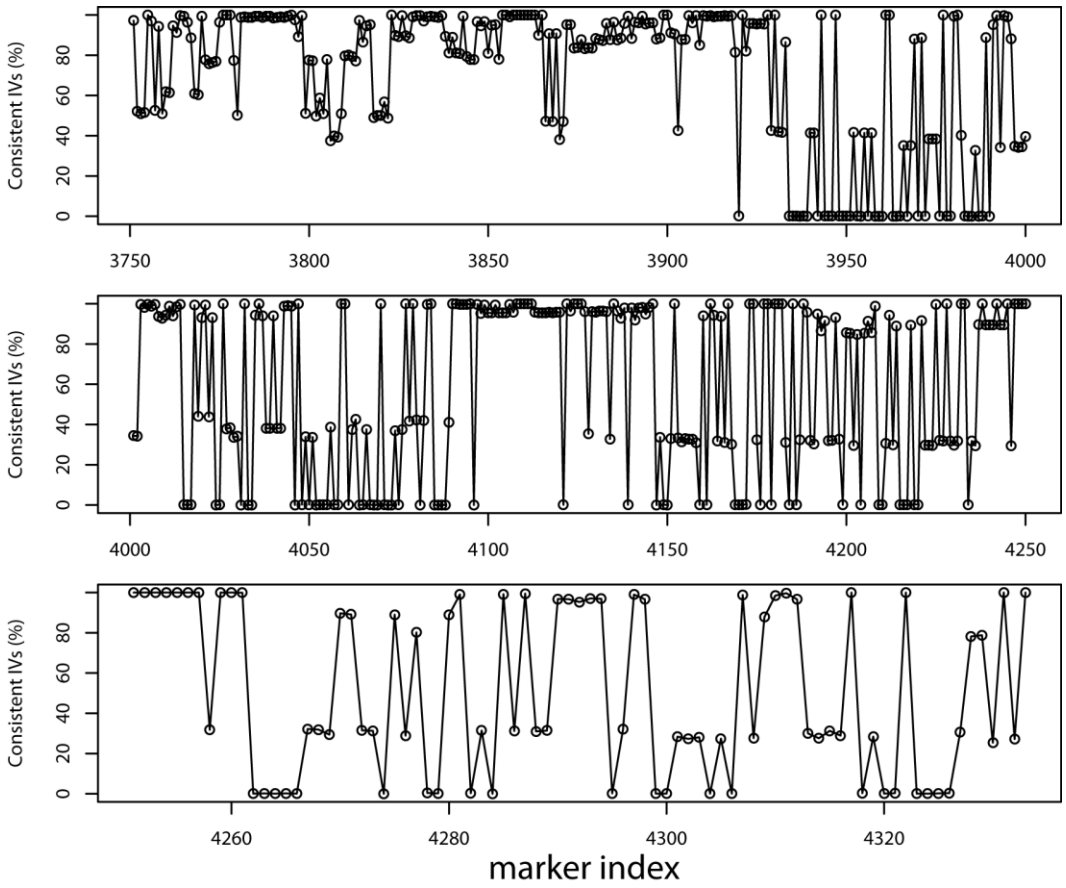



Figure 4.14 Plot of the percentage of consistent IVs by SNP marker index to detect MC genotyping errors in the real pedigree data.


Consistent IVs (%)
0
20
40
60
80
1250
1300
1350
1400
1450
1500
1500
1550
1600
1650
1700
1750
1750
1800
1850
1900
1950
2000
2000
2050
2100
2150
2200
2250
2250
2300
2350
2400
2450
2500
marker index


Consistent IVs (%)
0
20
40
60
80
2500
2550
2600
2650
2700
2750
2800
2850
2900
2950
3000
3050
3100
3150
3200
3250
3300
3350
3400
3450
3500
3550
3600
3650
3700
3750
marker index




The main reason why the percent consistency fluctuated greatly between nearby SNP markers is because different markers may contain different information about the chromosomal descent. It is important to recognize that most of these markers likely do not contain genotyping errors. Thus, a more likely explanation is that since SNP markers may have different genotype configurations and none of them are fully informative about the true descent pattern, genotypes at each marker may be consistent with a different set of IVs (Figure 4.15). Hence, the percentage of consistent IVs may still vary between test positions in a region without recombination.

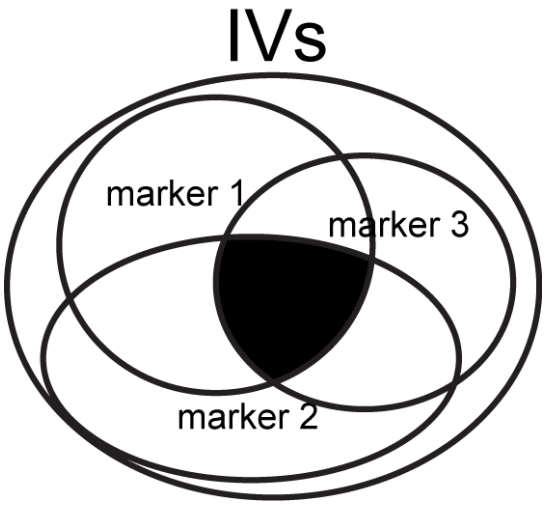



Figure 4.15 Schematic diagram of the set of sampled IVs that are consistent with each marker, as illustrated by inner ellipses. The black box represents the intersection of the sets of IVs that are consistent with all 3 markers.

As future direction, we can consider using dense SNP markers within a small neighborhood to jointly improve the existing inference of IVs. Existing inference of IVs uses gl_auto to sample IVs at framework positions and then uses the procedure described in Section 2.6 to sample IVs at a dense position. The inference of IVs from such set of IVs can be further refined by using dense SNP markers that are very close together. When a sampled IV is well-inferred, genotypes of nearby dense markers should all be consistent with this IV unless there are genotyping errors or unless this neighborhood contains recombination. Thus, joint-consistency between dense markers and a sampled IV can further assure that this sampled IV is well-inferred.

### 4.6.2 Visualizing MC Genotyping Errors Using Joint-Inconsistency Test

The joint-consistency approach can be used to detect MC genotyping errors. The rationale for this approach is that the use of multiple neighboring markers at each test position

can help select the best IVs for testing. This approach modifies the percent consistency approach (marginal approach) as described in Section 4.3. As in the marginal approach, the joint-consistent approach samples IVs at positions corresponding to the dense markers using IVs at the framework positions. Then, at each test position, each sampled IV is checked for joint-consistency with multiple neighboring markers but excluding the marker at the test position. A sampled IV is scored as consistent if it is consistent with the genotypes of all dense markers in this neighborhood and inconsistent otherwise. After repeating the joint-consistency check for each sampled IVs, the approach summarizes the percentage of joint-consistent IVs.

The joint-consistency approach flags for signatures of genotyping error (Figure 4.16). Unlike the marginal approach that flags for low percentage of consistent IVs, the joint-consistent approach searches for a signature pattern. This approach assumes that only one error exists within a neighborhood of markers. If this genotyping error at a dense marker is detectable by well-inferred IVs, then any test positions that include this marker as a neighboring marker will have a low percentage of joint-consistent IVs. However, the percentage of joint-consistent IVs should be considerably higher at the test position of the bad marker than at the neighboring test positions, since the joint-consistency at the bad marker position excludes performing consistency check with the bad marker. Thus, the presence of a genotyping error will lead to a pattern illustrated in Figure 4.16.

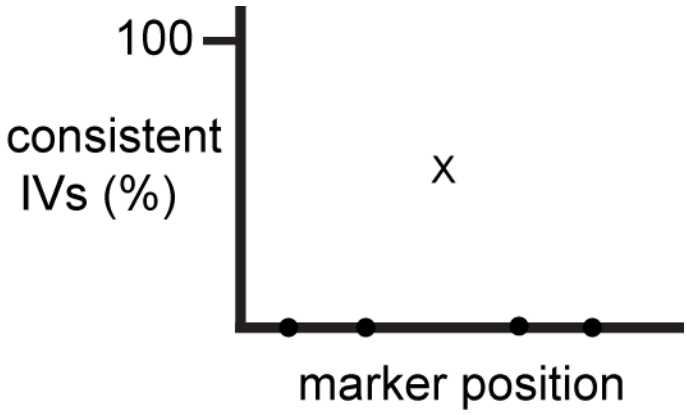



Figure 4.16 Pattern that resembles a signature of error

The joint-consistency approach requires us to define what a neighborhood is. This approach assumes that loci that are close together share the same underlying descent pattern. Here, I arbitrarily define a neighborhood to be up to 5 closest markers on each side within 0.20 cM from a test position. As the approach test different positions, it slides the neighborhood of markers using a "moving-window" approach. 2

### 4.6.3 Analysis

I applied the joint-consistency approach on the real pedigree data (Figure 4.12). The diagnostic plot had three interesting features. First, the joint-consistency approach flagged substantially fewer markers by using the previously described signatures to detect errors. For instance, while marker 145 and marker 1006 showed clear signatures of error (Figure 4.17), none of the 22 markers between marker 4260 and the last marker flagged by the marginal consistency approach (Figure 4.14) was also flagged by the joint-consistency approach (Figure 4.17). Even though there was a marker flagged by the joint-consistency approach, this marker was not flagged by the marginal consistency approach. Second, a large patch of test positions had very

low percentage of joint-consistent IVs. Almost all test positions between marker 3940 and the last marker had 0 percent consistency in the joint-consistency approach (Figure 4.17), whereas in the same region the percent consistency fluctuated greatly between high and low values in the marginal consistency approach (Figure 4.14). Third, the joint-consistency plot had much smoother trends than the marginal consistency plot (Figure 4.14). A good example is the test statistics between marker 1500 and marker 1700. Whereas the test statistics constantly fluctuated up and down in the marginal consistency plot, the test statistics were mostly on an increasing trend in the joint-consistency plot. A discussion about such trends is presented in Section 4.6.4.

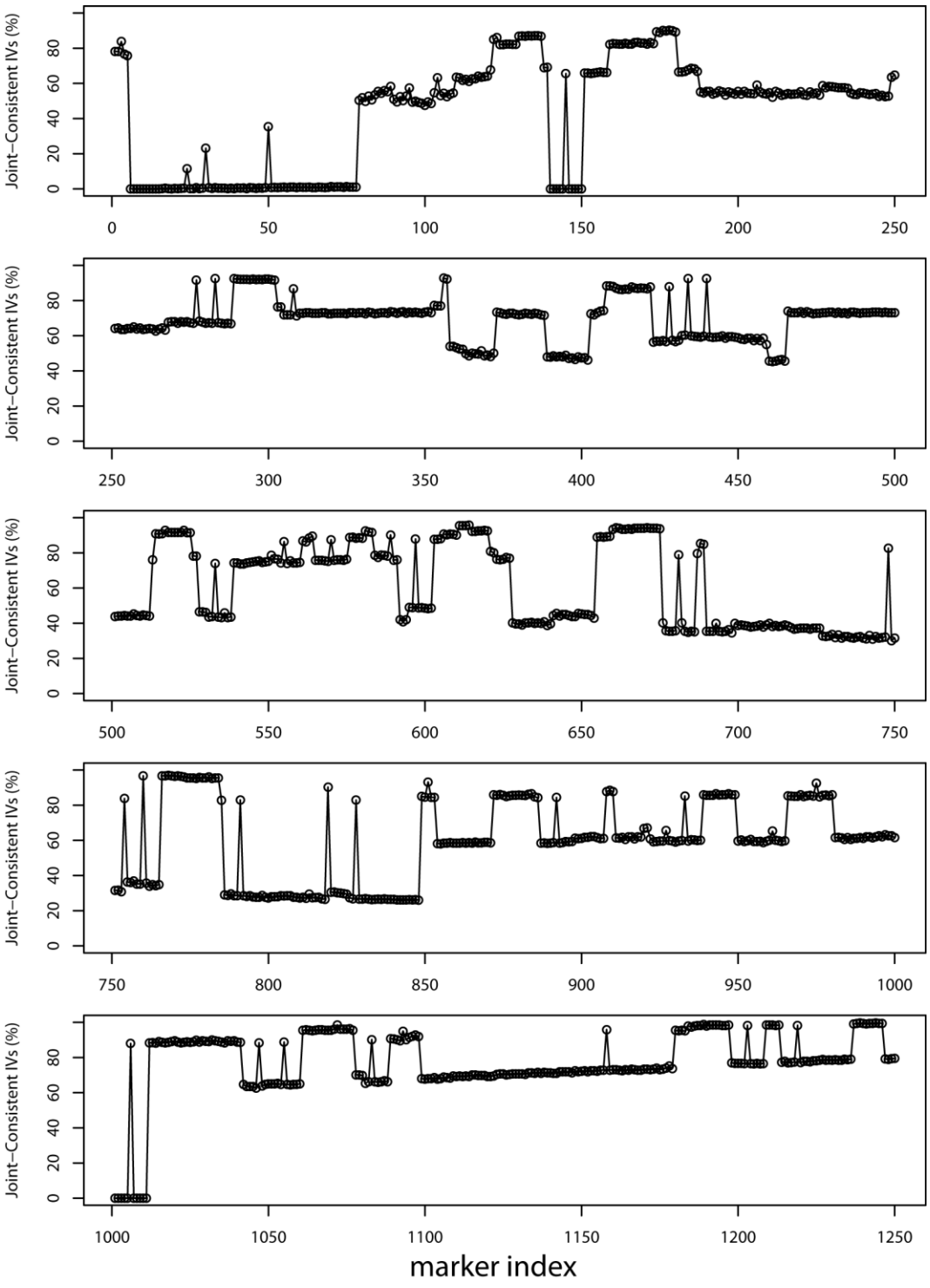
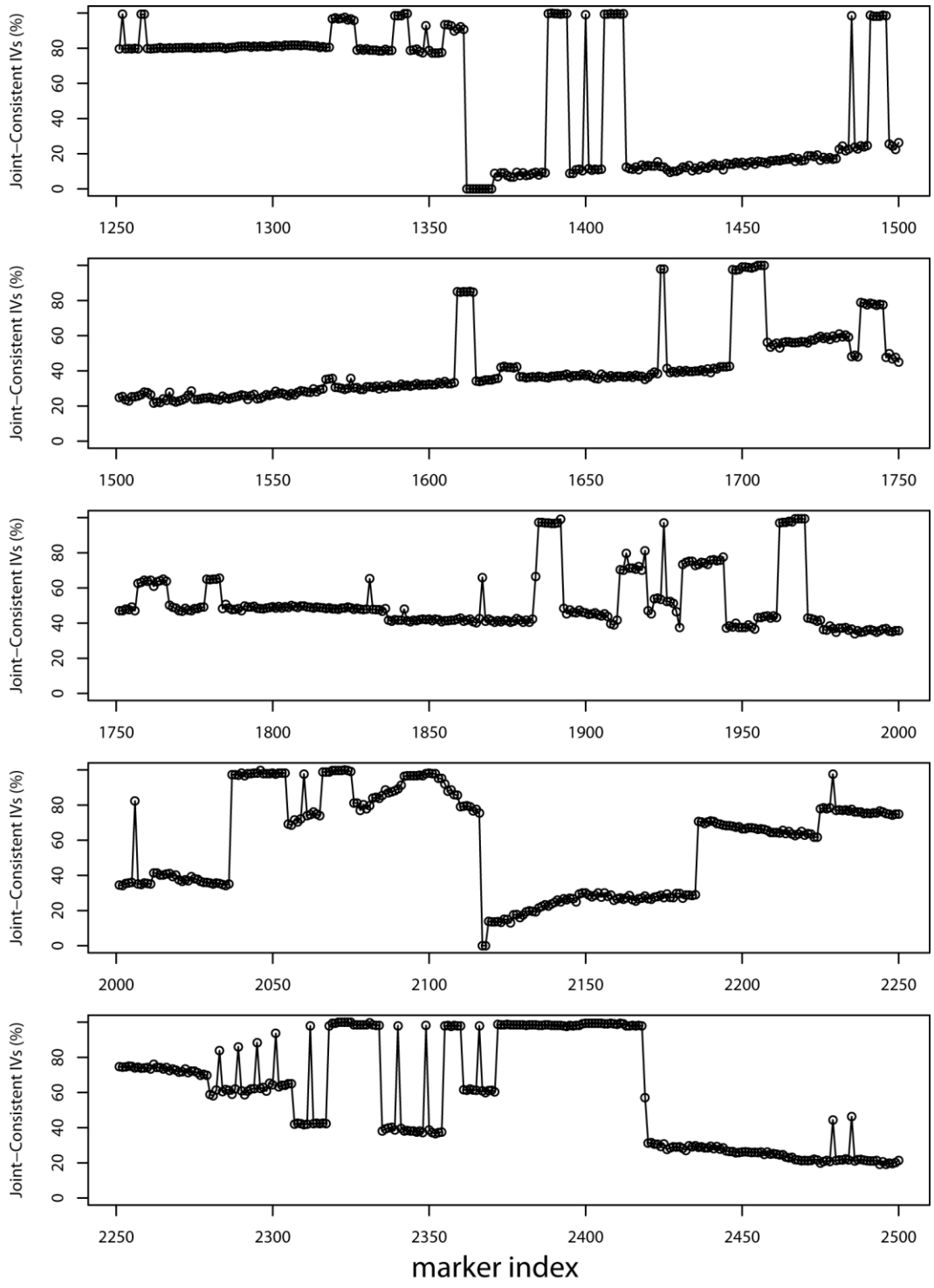
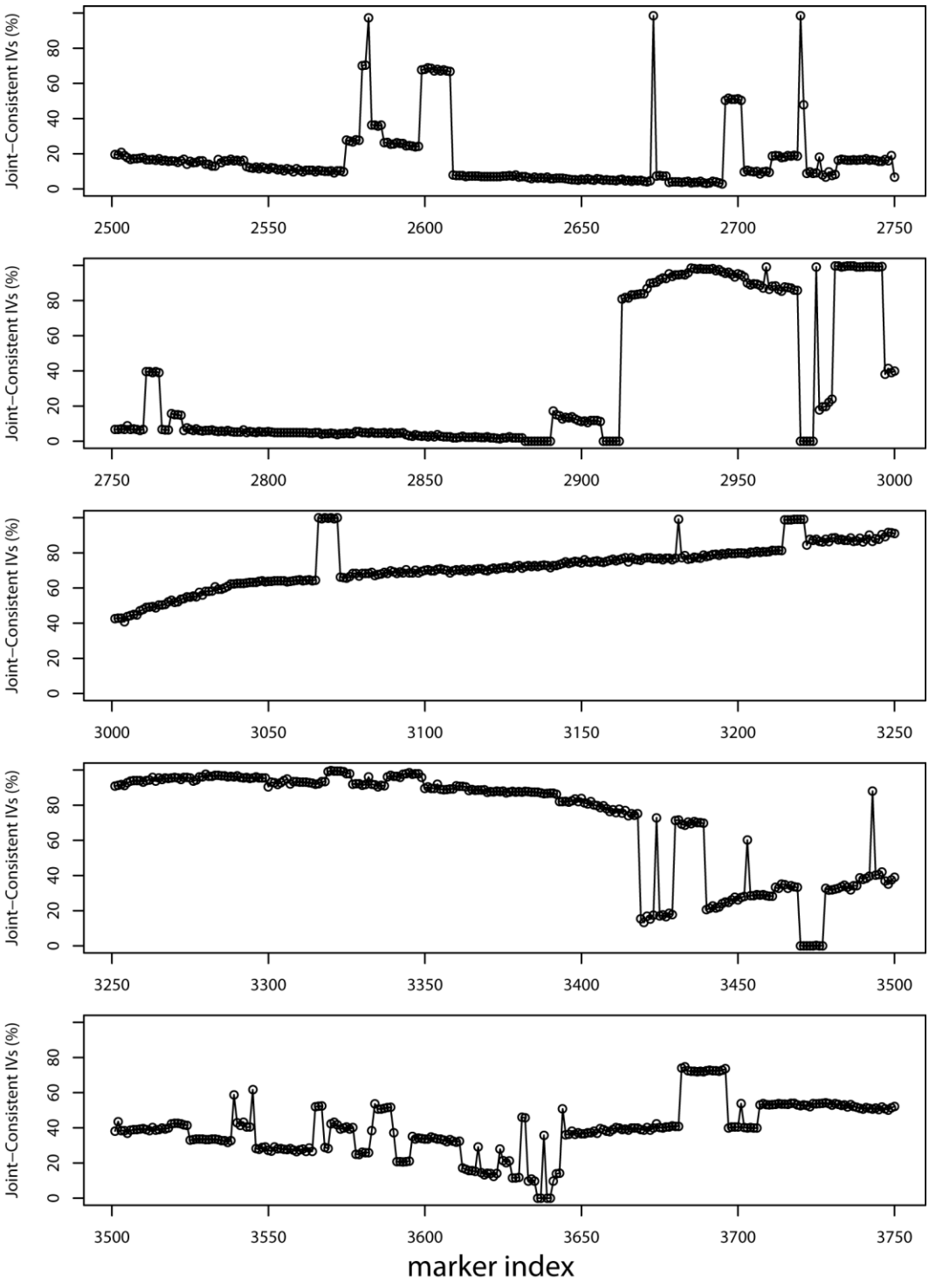
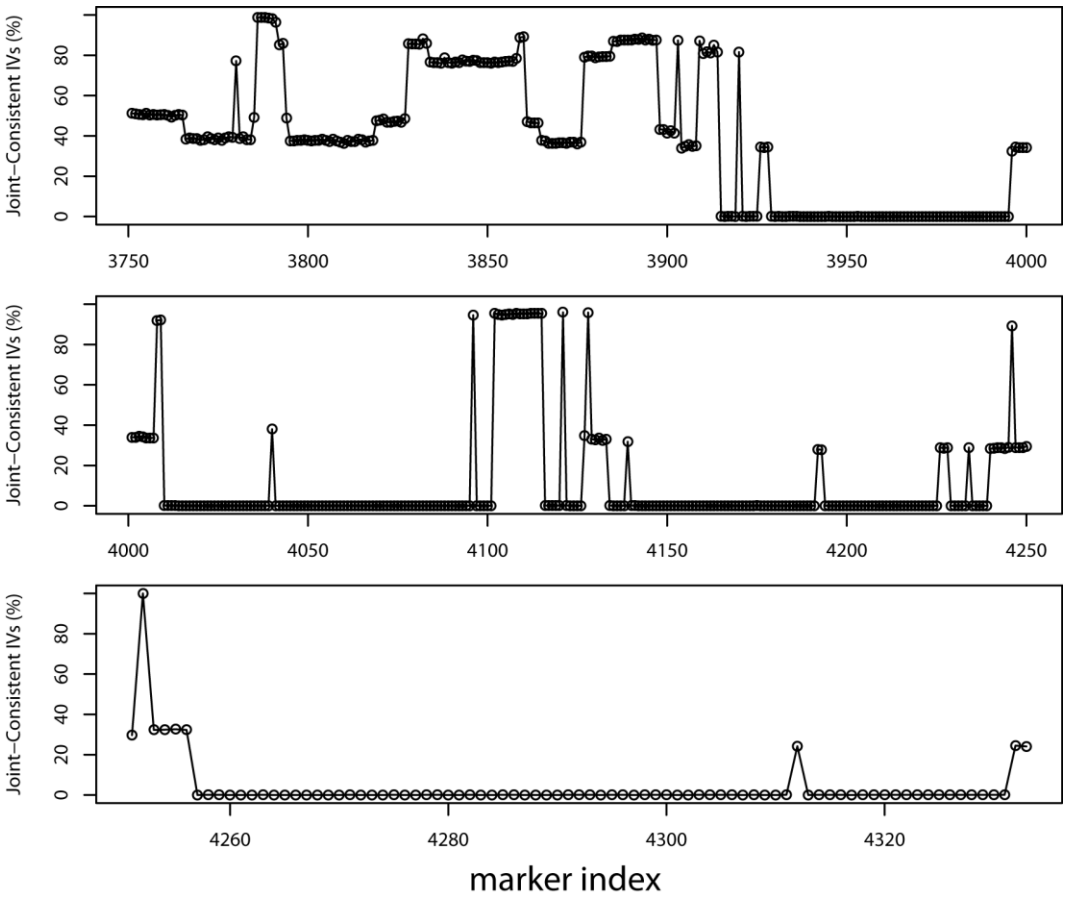



Figure 4.17 Plot of the percentage of joint-consistent IVs by SNP marker index to detect MC genotyping errors in the real pedigree data.


Joint−Consistent IVs (%)
0
20
40
60
80
1250
1300
1350
1400
1450
1500
Joint−Consistent IVs (%)
0
20
40
60
80
1500
1550
1600
1650
1700
1750
Joint−Consistent IVs (%)
0
20
40
60
80
1750
1800
1850
1900
1950
2000
Joint−Consistent IVs (%)
0
20
40
60
80
2000
2050
2100
2150
2200
2250
Joint−Consistent IVs (%)
0
20
40
60
80
2250
2300
2350
2400
2450
2500
marker index


Joint−Consistent IVs (%)
0
20
40
60
80
2500
2550
2600
2650
2700
2750
2750
2800
2850
2900
2950
3000
3000
3050
3100
3150
3200
3250
3250
3300
3350
3400
3450
3500
3500
3550
3600
3650
3700
3750
marker index




### 4.6.4 Discussion

Compared to the marginal consistency plot, the joint-consistency plot had much smoother trends. This is an expected result because the use of multiple neighboring markers allows adjacent test positions to use similar information to refine the inference of IVs. In the moving windows approach, discrepancy between adjacent test positions results mainly from the use of a slightly different set of markers because neighborhood of markers to include are different at different test positions. The percentages of consistent IVs at nearby positions do not fluctuate much, which implies that the influence from the excluded markers at the boundary of adjacent test positions is often small. However, two adjacent test positions can still have very different joint-consistent percentages when one test position uses an informative marker that is

inconsistent with many sampled IVs and the other test position does not use this marker. The reduction of fluctuation is a nice property because we can now determine how well the IVs are sampled along the chromosome by evaluating the percentage of joint-consistency in a general region.

The main advantage of this joint-consistency approach is that it can decrease the false detection rate. The joint use of dense markers helps us to identify regions where we cannot sample IVs well. The use of multiple markers, some of which are informative, allows us to visualize regions where we cannot sample IVs well by detecting test positions with low percentages of joint-consistent IVs. At such regions, we cannot tell whether markers indeed have errors. Thus, to avoid false positives, we may not want to flag markers in those regions.

The joint-consistency approach, however, is potentially more conservative than the marginal-consistency approach. We call IVs well-inferred when the sampled IVs are consistent with all markers in the neighborhood. A concern is that the requirement for consistency with multiple markers is too stringent. This is because even if the sampled IVs are not entirely accurate in the entire pedigree, these sampled IVs may still be adequate for error detection. What ultimately matters is that the relevant part of the IVs that are needed to detect genotyping errors is sampled sufficiently well.

The number of markers in a neighborhood affects the joint-consistency approach. While the main advantage for using more markers in the neighborhood is to refine the set of sampled IVs, there are two notable disadvantages. First, the use of more markers in the neighborhood increases computational burden, since the computational time increases linearly with the number of markers in the neighborhood. Second, the joint-consistency approach assumes that adjacent test positions only contain one bad marker. The presence of multiple errors occurring at nearby

test positions can affect the detection of errors in this region, since the signature of error would be wiped out by a continuous patch of low joint-consistent percentages caused by errors at multiple nearby markers.

The decision of whether to use the marginal-consistency or joint-consistency depends on how stringent the user wants the test to be. If higher sensitivity is desired, the marginal-consistency approach should be used because it flags errors even when we have suspicion that inconsistency between observed genotypes and sampled IVs is because the sampled IVs are not similar to the true IV. On the other hand, if higher specificity is desired, the joint-consistent approach should be used because it would minimize calling genotyping errors in regions of a chromosome where IVs of the entire descent of a pedigree cannot be sampled well.

## 4.7 Concluding Remark

It is important to minimize the chance of false findings when we conduct genetics studies. Because of the amount of time, money, and effort we invest in these studies, we should never neglect proper checks to improve the quality of genotype data before we carry out any major analyses. My practical approach and implemented program to detect genotyping errors enables researchers to efficiently perform this crucial quality check of dense genotypes in pedigrees. My approach provides a line of defense against false findings. However, when we have indication of the existence of genotyping errors, it may be worthwhile to re-determine the genotypes in the laboratory.

# Chapter 5

# EFFICIENT SELECTION OF SUBJECTS FOR SEQUENCING

When we have a limited budget, we should design our experiment in a cost effective way. Because sequencing is still expensive and we may not have the resources to sequence all subjects in a pedigree, it is ideal to answer the same scientific question by only sequencing a small number of subjects. To facilitate such a design, I developed a genotype imputation approach to impute missing genotypes (Chapter 3). I showed that the approach can achieve high accuracy when alleles are called using deterministic constraints from IVs (Section 3.4). However, the number of alleles that the approach can call deterministically depends greatly on which subjects are sequenced. Thus, it is important to develop an approach to prioritize subjects for sequencing. The purpose of this chapter is to present some ideas that I will investigate in the future.

## 5.1 General Considerations

There are a few aspects we should consider in the selection of subjects for sequencing. First, selection should depend on the purpose of the study and our scientific hypothesis. If we want to identify rare risk variants, we should select subjects who can increase our chance of seeing these rare variants. Second, selection should depend on prior available information. For instance, do we expect risk variants to be segregating within a region we already identified to have evidence for linkage, or do we expect the risk variants to be anywhere on the chromosome? If we can narrow down our search space to a specific region, we should select subjects that

optimize our chance to extract the most information for a local region. In addition, subject selection should depend on the information that we know about phenotypes of interest. For instance, if a disease is segregating within a family, we may prefer to ascertain some subjects who are affected and who are connected to the central pedigree. Third, subject selection may need to consider the specific method of analysis that we propose to use. For instance, if the method of analysis uses only subjects with phenotype data, we should only consider choosing among subjects who have phenotype data.

## 5.2 Thoughts for Selection of Subjects

Here, we consider a specific scenario: (1) our goal is to detect risk variants, (2) we only have a budget to genotype a fixed number of subjects, (3) we have identified a region that most likely contains the rare variant, (4) we have existing genotype data that we can use to infer IVs, and (5) we plan to sequence a few subjects and then impute missing genotypes.

We assume that our goal is to select a set of subjects that maximizes the total number of alleles that we can reliably impute in the pedigree. Under the scenario above, inferred IVs can guide subject selection. The idea is to select subjects who offer the most incremental value.

### 5.2.1 Statistical Framework for Subject Selection

I introduce a statistical framework for subject selection that uses coverage as a metric. Coverage is the expected fraction of the copies of alleles in a variant that are determined after pedigree-based genotype imputation conditional on an IV. As an expected value, this measure naturally accounts for whether the observed alleles in the pedigree can be assigned

unambiguously to the FGLs, which is a necessary condition to imputing alleles deterministically using GIGI. Thus, coverage reflects the fraction of alleles that are either well-imputed or sequenced. When all subjects are not genotyped, coverage has a value of 0. On the other hand, when all subjects are genotyped for this variant or if all alleles in the pedigree can be imputed deterministically, coverage has a value of 1. By using coverage, we can compare among designs for subject selection.

Coverage can be computed by using exact calculation. The calculation first involves translating IV into disjoint identity-by-descent-graphs, as denoted by $ibdg_i$, for $i = 1,2,\dots,I$. Define $N$ to be the number of subjects in the pedigree, so $2N$ is the total number of alleles in the pedigree at a locus. In each $ibdg_i$, there is a probability $p_i$ that the observed alleles are deterministically assigned to $ibdg_i$ and a probability $q_i$ that the observed alleles are not deterministically assigned to $ibdg_i$. If the observed alleles are deterministically assigned to $ibdg_i$, a total of $A_i$ alleles in the pedigree can be determined unambiguously. If alleles are not deterministically assigned to $ibdg_i$, a total of $B_i$ alleles can be determined unambiguously. Coverage is expressed as

$$\text{Coverage} = \frac{1}{2N}\sum_i (A_i p_i + B_i q_i) \qquad (1),$$

$A_i$ and $B_i$ in Equation (1) are simple to calculate. $A_i$ is equal to the total number of copies of FGLs that are shared by FGLs in $ibdg_i$. This is because if alleles are deterministically assigned to $ibdg_i$, all other alleles which share the same FGLs as the observed alleles in the $ibdg_i$ would be determined. $B_i$ is 2 times the number of subjects from $ibdg_i$ + number of subjects who share both alleles IBD with any people from $ibdg_i$. This is because if alleles are not deterministically assigned to $ibdg_i$, only the observed genotypes or genotypes from

unobserved subjects who share both alleles identical-by-descent with at least an observed subject can be determined unambiguously.

For di-allelic variant such as the majority of SNPs, it is simple to calculate $q_i$ and $p_i$ in Equation (1). $q_i$ is equal to the probability that alleles from $ibdg_i$ display a pattern of alternating allelic types. For instance, if an $IBDG_1$ is a linear graph $1-3-5$, $q_i = P_A P_a P_A + P_a P_A P_a = P_A P_a$, where $P_A$ denotes the population allele frequency of the major allele and $P_a$ is the population allele frequency of the minor allele. $p_i = 1 - q_i$.

### 5.2.2 Estimating Coverage

While coverage is a conceptual quantity defined for an arbitrary IV, for practical use, we extend the concept of coverage to local and genome-wide coverage. First, local coverage is the estimated coverage at a locus of interest. Using the approach described in Chapter 2, we can sample IVs at a locus of interest by using framework markers. To estimate local coverage, coverage is called and averaged across a set of inferred IVs. When prior information about a candidate chromosomal region is available, the use of local coverage can be used as the metric to optimize genotype imputation in a local region. Second, genome-wide coverage is the estimated expected coverage at a random locus in the genome. Genome-wide coverage can be approximated by averaging coverage calculated across a large set of randomly sampled IVs. When the study is interested in optimizing genotype imputation without knowledge of a candidate region, the use of the genome-wide coverage may be a more appropriate measure.

### 5.2.3 Joint-Prioritized Selection Method

I developed a "joint-prioritized" subject selection method to sequentially select $m$ subjects from $n$ potential subjects. The method consists of a few steps. First, the method iterates through the entire list of subjects and ranks choices using the estimated coverage as a metric. It retains a ranked list of top $c$ (e.g. $c = 5$) choices that has the highest estimated coverage and discards the unselected choices. These top choices are called templates for the next step. Second, the method selects the second subject using each template one-by-one. Using each selected subject from the template, the method adds a new subject that is not in the template and computes the estimated coverage. Thus, a total of $c(n-1)$ coverage scores are calculated. Third, the method retains $c$ unique combinations of selected subjects with the highest coverage. These top $c$ selections now become the new templates for the next step. Fourth, step two and three are repeated until $m$ subjects are selected. After $m$ selection steps, the final templates become the final choices.

The "joint-prioritized" subject selection method offers a few theoretical properties. First, this method is superior to the stepwise-forward selection because it explores many more combinations of selection choices. However, as a base case when $c = 1$, the method reduces to the stepwise-forward selection. Second, when $c > 1$, the method maintains flexibility in who to select to achieve higher chance of finding better designs after all selection steps. Unlike forward-stepwise selection, the joint-prioritized selection method does not select a list of subjects with certainty after each step but instead continues to refine selection on top of templates. This scheme is more enticing than a selection scheme that makes permanent choices after each step. Third, the method builds on multiple first selections. This scheme is potentially beneficial because different starting choices may have high influence on subsequent choices. Fourth, the

method keeps computation low by focusing only on high potential building blocks. The premise is that only top templates are likely to be high quality building blocks. Fifth, this scheme enables efficient computation. The number of calculation is only $c$ times more than forward-stepwise selection, where $c$ is a fixed number much smaller than $n$. Hence, computation is still on the order of O(nmc), where c is the number of templates to keep, n is the total number of subjects to select from, and m is the number of subjects to select, which requires just a few times more computation than stepwise-forward selection, of which computation complexity is on the order O(nm).

## 5.3 Thoughts for Selecting the Minimum Number of Subjects for Imputing Rare Alleles under Specific Scenario

I now look at this question from the perspective of trying to sequence the minimum number of subjects to impute the rare risk allele. For consistency between the recessive and dominant mode of disease, the risk allele will be labeled by the lower case letter "a". As a starting point for later investigation, I use a set of assumptions.

Assumptions:

(1) We can infer the descent pattern perfectly using already collected markers. For instance, we already collected very informative markers for linkage analyses.
(2) A single disease variant causes the disease in every affected subject in the pedigree
(3) Mendel's Law of Segregation

*Recessive Disease*

If the disease is recessive, as defined by true homozygosity in the same base pair variant on both chromosomes, we only need to sequence as few as 1 affected subject. The pair of risk alleles that this affected subject has is trivially phased deterministically to a pair of FGL on the IBD-graph (Figure 5.1). Thus, we can impute the risk alleles on any subjects who share any FGLs with this affected individual. On the other hand, if the descent pattern suggests that more than 2 founder chromosomes contain the same risk allele, then we have to sequence other affected subjects to obtain the other distinct copies of the risk allele (Figure 5.2).

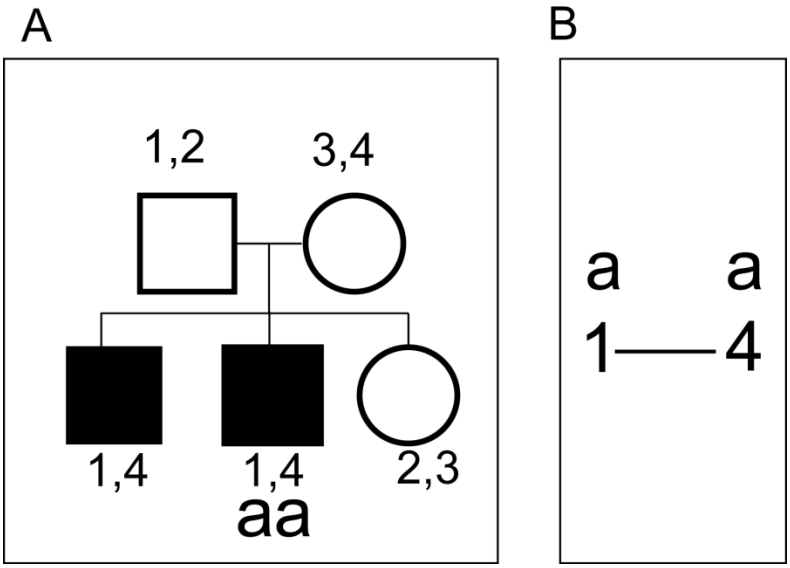



Figure 5.1 In this recessive disease example, we only need to sequence 1 subject to impute all copies of the risk alleles in all carriers. FGL 1 and 4 have the "a" (risk) alleles. A) FGLs are labeled by numbers and genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A).

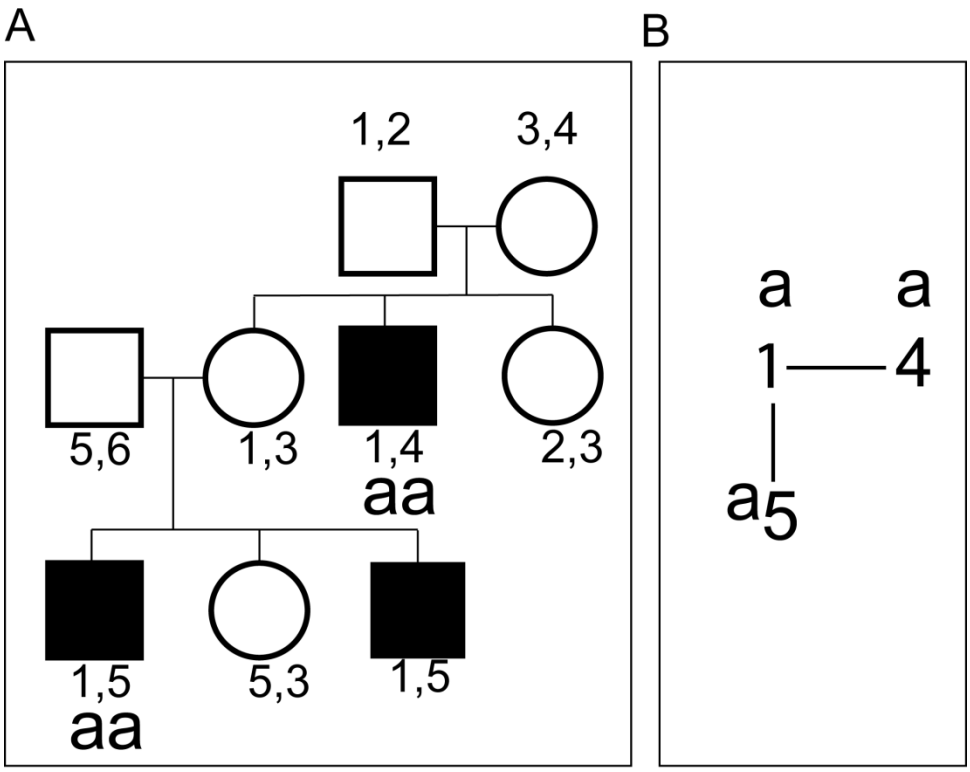



Figure 5.2 Recessive disease with multiple affected subjects. In this case, we need to sequence 2 subjects to impute all copies of the risk alleles in all carriers. FGL 1, 4, and 5 have the "a" (risk) alleles. A) FGLs are labeled by numbers and genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A).

*Dominant Disease*

If we have a dominant disease, we need to sequence a minimum of 2 subjects. Similar to the case of recessive disease, the first person to sequence should be one of the affected subjects. Sequencing this person ensures that we observe a copy of the disease variant. If this subject is homozygous for the risk allele, we end up with the scenario above. On the other hand, if this subject has 1 risk allele and 1 normal allele, then we need to sequence one more subject for the purpose of unambiguously assigning the observed alleles to FGLs , which I will refer to as

phasing alleles to FGLs.

The second subject to sequence should be a relative of this first person who is IBD for exactly 1 allele with the first person and homozygous for either the risk or normal allele at the risk locus. The reason for sequencing the second subject is to ensure that the risk allele can be phased deterministically to a FGL on the IBD-graph (Figure 5.3). In this example, once the mother in the first generation and the affected child are sequenced, we can impute the risk allele in all remaining affected subjects. After determining the phase, we can assign the phased risk allele to all individuals who share the same FGL. If we make one further assumption that P(disease | at least 1 risk allele) = 1, then the requirement for the second person is guaranteed if we sequence either the unaffected parent or the unaffected child of the affected subject we already sequenced. More generally, if the penetrance of the disease is high, sequencing the unaffected child or the unaffected parent of the affected subject are both sensible options.

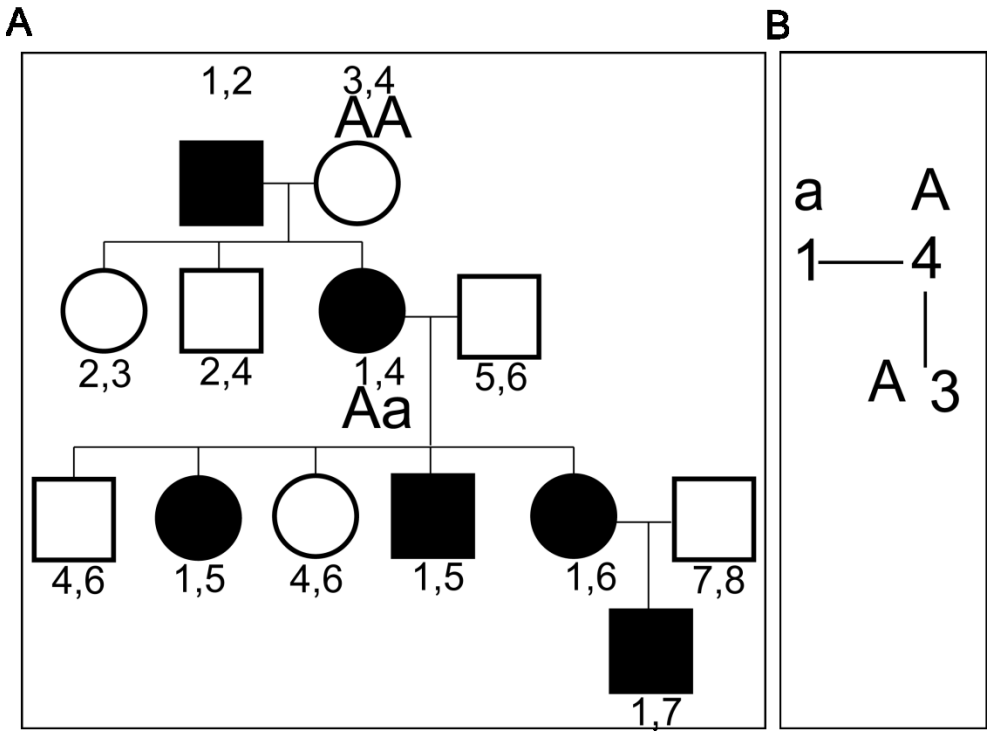



Figure 5.3 In this dominant disease example, we only need to sequence 2 subjects. Any subjects who have FGL=1 inherit the "a" (risk) allele. A) FGLs are labeled by numbers and genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A).

The assumption that P(disease | at least 1 risk allele) = 1 is often not realistic. In complex disease, a risk allele increases the risk of the disease but does not always lead to the disease. If we have a rare disease variant in the population, the normal married-in spouse of the subject who carries the risk allele likely is homozygous for the normal allele. Hence, the unaffected married-in parent can be used to phase the rare risk allele. Thus, sequencing the married-in is a sensible option (Figure 5.3).

Sequencing both parents, one of whom is affected, will not enable phasing (Figure 5.4). Even though all alleles are observed when both parents are sequenced in a nuclear family, we can not know whether the affected parent passes down the risk or normal allele to each offspring. Likewise, sequencing the affected parent and the affected child (Figure 5.5) or sequencing two affected sibs (not shown) does not enable phasing if both of them are heterozygous for the risk alleles.

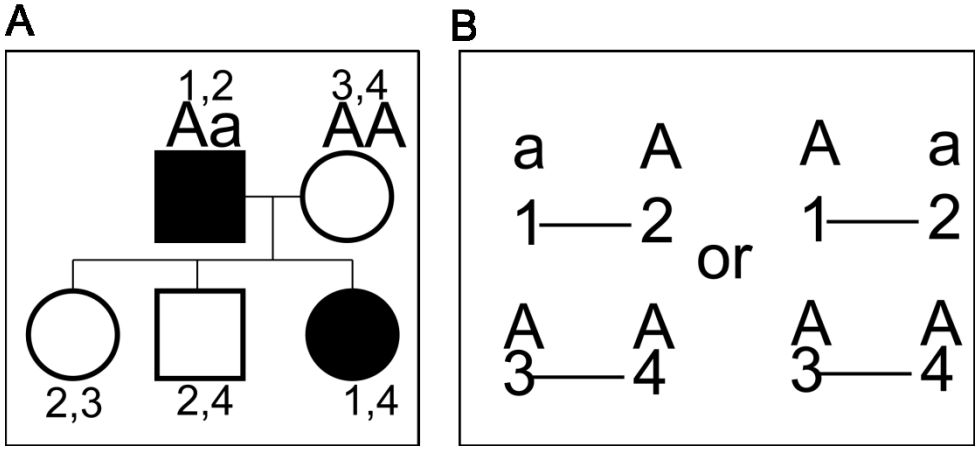



Figure 5.4 An example showing that the phase is unresolved when sequencing only the parents. A) FGLs are labeled by numbers and genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A).

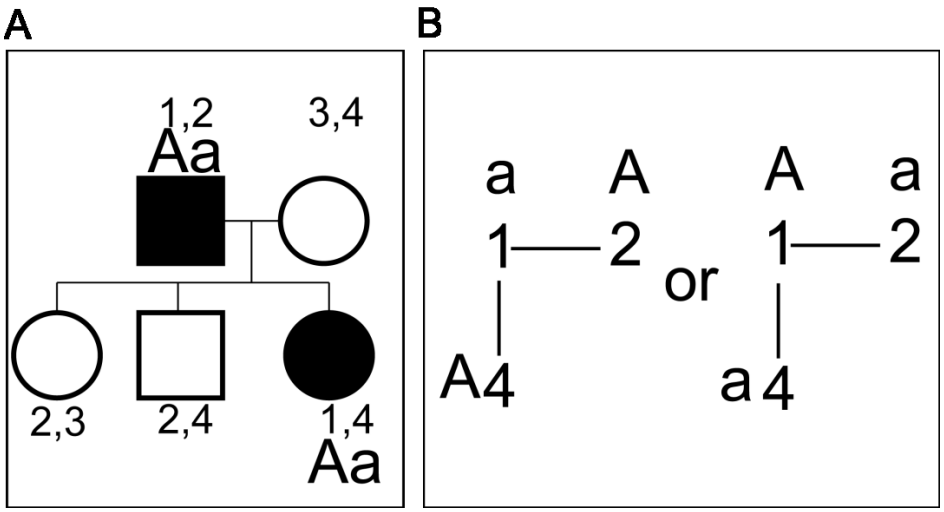



Figure 5.5 An example showing that the phase is unresolved when sequencing two connected subjects with heterozygous genotypes. This illustrates that sequencing only affected subjects would not enable us to unambiguously assign the risk allele to the FGL. A) FGLs are labeled by numbers and genotypes are labeled by letters. B) IBD-graph is constructed by using information from (A).

In real situation, typing more affected subjects than the minimum of two subjects may be needed. First, we can be more confident that an identified rare allele is not a genotyping error if multiple affected subjects are observed for this rare allele. Multiple observation of a rare allele significantly decreases the probability of identifying a false rare allele. Second, the assumption that IVs are well inferred may be false. For instance, genotyping errors can negatively affect the inference of IVs, as was noted in Chapter 4. Since my approach to impute genotypes and decide who to sequence depends largely on the quality of the inferred IVs, we may need to sequence additional subjects also if there is uncertainty in the inferred IVs.

If the goal is to maximize the number of total imputed alleles in multiple loci across a

region, then sequencing the affected child and married-in parent may not always be the best design. For instance, consider a pedigree where the inner parent has multiple siblings. If we sequence the inner parent instead of the married-in parent, we can likely impute many more total alleles in the siblings of the inner parent. This is because the inner parent shares many alleles with the siblings.

# Chapter 6

# DISCUSSION

I developed an approach to impute genotypes and detect genotyping errors that can be used for dense variants on large pedigrees. My genotype imputation approach allows us to infer missing dense genotypes using dense genotypes collected on a few subjects with existing sparse genotypes. This is likely a useful tool in the immediate future because dense genotyping is still expensive and because it is not always possible to generate dense genotypes on all subjects. My genotyping error detection approach facilitates quality checking of genetic data by detecting hidden errors not identified by standard Mendelian Inconsistency checks. Both approaches are computationally efficient and are suitable for use with dense variants. Also, the fact that my approaches can be used on large pedigrees is especially important because large pedigrees have important roles in genetics. My approaches are implemented in computer programs and are available to the genetics community.

## 6.1 Thoughts for Future Directions

The framework for inferring IVs is central to many applications of pedigree-based genetic analysis. In the past, this framework has allowed important applications such as finding evidence for linkage between chromosomal regions and phenotypes [Almasy and Blangero 1998; Kruglyak, et al. 1996], identifying causal variants through fine-mapping methods [Almasy and

Blangero 2004], and inferring relatedness among individuals [Epstein, et al. 2000]. Here I used inferred IVs to impute genotypes (Chapter 3) and to detect Mendelian consistent genotyping errors (Chapter 4). This framework also allows me to tackle new problems.

Aside from the topic of effective selection of subjects for sequencing, my research also generates other potential future directions. These topics include genotype calling in sequencing and potential for low coverage sequencing in pedigrees, haplotype determination in dense markers and precise determination of recombination, multiple imputation in genotype imputation, benefit of using imputed genotypes, and detection of *de novo* mutations. We further discuss these topics below.

### 6.1.1 Calling Genotypes and Potential for Low Coverage Sequencing

Next-generation sequencing is becoming an important tool for discovering functional variants in both population-based and family-based studies. In this type of data, the probability of making a correct genotyping call increases as a function of sequencing depth. It is important to have enough sequencing reads to achieve the desired call accuracy. Within related individuals in a family, there may be redundant information due to the sharing of segments of chromosomes between relatives. We can use this level of information to improve genotype calling. Hence, it may be possible to reduce sequencing depth to obtain a desired average calling accuracy.

One extension would be to use our framework for inferring IVs to call genotypes for next-generation sequencing data in pedigrees. Using MCMC samplers to realize IVs, the approach will be computationally feasible for pedigree of any size. This topic is a natural extension of genotype imputation. Loosely speaking, this approach to call genotypes is genotype imputation with an additional layer of complexity that the observed genotypes are not

deterministic. In genotype imputation, the input data consists of observed genotypes with the assumption that they are correct. The pair of alleles in the genotype of an observed individual is assumed to be observed. In this model for joint-calling and imputation, we would not assume that both alleles are sequenced. We would calculate the probability of genotype configurations for subjects who are sequenced or not sequenced, conditional on sequencing reads of subjects who are observed. This calculation will likely to be similar to the calculation of missing genotype presented in Chapter 3, except that the probability of genotype is also calculated for each observed subject because it is possible that not both alleles in each genotype are ascertained. The probability that both alleles are typed increases with more reads at the position of the variant. We can then call genotypes probabilistically on subjects who are sequenced and impute genotypes on those who are not sequenced.

### 6.1.2 Phasing Haplotypes of Dense Markers

Our framework for inferring IVs also allows us to develop a method to phase dense haplotypes in pedigrees. This method allows us to statistically phase haplotypes of tightly-linked markers in observed individuals and can then potentially impute haplotypes on individuals unobserved for genotypes. The ability to phase genotypes depends largely on the structure of the pedigree and the observed pattern for genotypes. When I impute genotypes, I often see that many observed alleles are deterministically matched to an inferred set of IVs. Thus, I can use inferred IVs to perform long range phasing of haplotypes. This proposed scheme to phase haplotypes of dense markers by first using framework markers to infer IVs will be an extension to the genotype imputation approach that I developed in Chapter 3.

I propose an outline of a potential approach to perform long-range phasing of dense

genotypes that can handle large pedigrees. These dense genotypes can include framework genotypes used to infer IVs. Existing approach such as Simwalk [Sobel and Lange 1996] is not designed for phasing haplotypes of dense markers on large pedigrees because (1) the use of dense markers may lead to mixing issues when using a MCMC sampler to perform computation, (2) time of computation is impractical when the number of markers to phase is large on large pedigrees, and (3) the use of dense markers may violate the LD assumption. Violation of the LD assumption might affect the reliability of inferred haplotypes [Schaid, et al. 2002]. This potential work aims to avoid these problems.

Outline:

(1) Use sparse framework markers to infer IVs at dense marker positions

(2) Identify the set of alleles from observed subjects that are deterministically assigned to a sampled IVs. Repeat this for each marker of interest. Repeat this exercise for each set of jointly-sampled IVs. The deterministic information provides the backbone of phasing.

    a. Uncertainty in phase exists at many levels: the set of phased alleles may be different across IVs, subjects, and markers

        i. Not all alleles in a marker can be assigned deterministically to the FGLs of an IV. If we are interested in calculating the probability of the haplotypes of these markers later, we need to incorporate them in the next step

        ii. We need to account for the possibility of label switching in FGLs when we phase haplotypes across IVs.

    b. This deterministic phasing information must be stored in a condensed way. We

can use a template to store phased haplotypes. We record departures from this haplotype template as we process through IVs and record the number of those haplotypes as we phase different IVs. We can use a tree structure to dynamically grow and shrink the template of phased haplotypes. Storing the information efficiently is a challenge.

(3) Summarize uncertainty in phases. This is another challenge.

Results from Chapter 4 suggests that even though the framework markers used to infer IVs are sparse, the accuracy of imputing genotypes is high when calls are made using the deterministic constraints from inferred IVs. We have to evaluate if IVs inferred by sparse framework markers are also adequately well-inferred for phasing. For instance, one concern is that the phasing of dense genotypes may require us to have a finer resolution of where recombination breakpoints occur. It might be necessary to develop an improved method to infer IVs that can incorporate additional information from dense genotypes.

Hence, another related topic is whether there is a practical way to integrate dense genotypes that are not in the framework panel into inferring IVs. Results from Chapter 4 suggest that dense markers offer deterministic constraints that can be utilized for inferring IVs. To infer IVs on large pedigrees, we have resorted to using MCMC-based samplers. Under this approach, we cannot add dense markers because of the MCMC mixing problems and the assumption that markers are in linkage equilibrium. In addition, computational burden increases with more markers. Even if dense markers cannot be practically integrated as usual into the sampling of IVs, we can potentially devise a sequential approach to integrate deterministic information from dense markers to refine the inference of IVs after we realize IVs using sparse framework

markers.

### 6.1.3 Evaluating the Benefit of using Imputed Genotypes for Analysis

As mentioned in Chapter 3, we can use imputed genotypes for analysis. The use of imputed genotypes potentially facilitates cost-effective analysis, but we have not yet performed an in-depth evaluation of how useful imputed genotypes are. We can perform simulation to evaluate how much power we can gain from using imputed genotypes to identify rare causal variants. Such investigation can also enable us to have a better understanding of what summary of imputed results we should use for downstream analyses. For example, in a test for association, preliminary study suggested that the use of imputed genotypes, as summarized by dose of the minor allele as the fixed effect in a linear mixed model, could potentially increase the ability to detect causal alleles [Marchani, et al. 2013].

### 6.1.4 Multiple Imputation in Genotype Imputation

It is straightforward to implement an option for multiple imputation to my existing genotype imputation approach. Currently, a probability distribution of each genotype is estimated marginally for each unobserved individual in the pedigree. For practical purpose, a user may decide to impute each genotype by choosing the most likely configuration or by using a confidence threshold. However, calling genotype once does not take into account of the uncertainty in imputation. Multiple imputation of genotypes allows us to capture uncertainty in imputed results, so downstream statistical analyses may become more valid than if analyses are based only on single imputation calls. Moreover, imputed genotypes, together with observed genotypes, are not guaranteed to be Mendelian consistent because the current method call

genotypes marginally across IVs.

### 6.1.5 Integrating Information from Linkage Disequilibrium to Impute Genotypes

My approach to impute genotypes may benefit from integrating information from LD. In GIGI, we impute alleles using IVs and observed genotypes marginally without using information from LD. On the other hand, population-based imputation approaches use LD to impute genotypes. Since information from both LD and IVs are useful, combining the two sources of information may improve imputation. My results in Chapter 3 suggest that BEAGLE excels in imputing common variants while GIGI excels in imputing rare variants.

My approach is to combine calls from GIGI and BEAGLE. As a first attempt, I suggested an algorithm to combine the results of GIGI and BEAGLE [Marchani, et al. 2013]. First, I called alleles using GIGI by using a very high call threshold ($t_1$=0.99 and $t_2$=0.995). Second, I called additional alleles using results from BEAGLE with the default threshold of $t_1$=0.8 and $t_2$=0.9 only if the calls from BEAGLE are not in conflict with those of GIGI. Results from Chapter 3 suggested that imputation using GIGI is accurate when calls rely on the use of information from pedigree structure and inferred IBD graphs. The use of high call threshold forced most calls in GIGI to be made using such information. Results suggested that combining calls from GIGI and BEAGLE improves call rate and accuracy for common variants over the use of only GIGI or BEAGLE [Marchani, et al. 2013].

Alternatively, we can use an approach that conditions on each inferred IVs. After sampling IVs, we combine imputed genotypes from BEAGLE and only let BEAGLE to override imputed calls if the imputed genotypes are compatible with those from GIGI. We can then average those calls across IVs.

### 6.1.6 Detection of De Novo Mutations

My current approach to detect genotyping error assumes the absence of *de novo* mutations. In reality, *de novo* mutations exist. Both *de novo* mutations and genotyping errors can lead to inconsistency between observed genotypes and inferred IVs and are not distinguishable in my current error detection approach. However, when a *de novo* mutation occurs in an observed individual and in an observed descendant who inherits the same mutation, it may be possible to distinguish between the two possibilities. For instance, if there is an unexpected change in allele in an individual and also in at least one of the individual's descendent, we may be more willing to believe that having a *de novo* mutation has occurred instead of two independent genotyping errors. It may be possible to extend the error model that I currently use in my error detection approach. This modification would involve the use of a *de novo* mutation rate and need to allow for mutant allele to be matched to a single FGL as the original allele. Evaluation of this potential extension through simulation will elucidate the strength and limit of this approach. In particular, evaluation using different pedigree structures will enable us to understand what pedigree structures are needed to accurately distinguish genotyping errors from *de novo* mutations. While this section focuses on *de novo* mutations, an extension to my error detection framework may also allow us to detect other structural variants such as deletions.

The use of pedigrees is an important design used in genetics. As genetics is the study of heredity, it is natural to use pedigrees to find answers to questions that are related to inheritance. In my research, I demonstrated that the use of statistical thinking can improve the efficiency and validity of pedigree-based studies. As we continue to face new challenges in genetics studies, statistical thinking will continue to provide guidance for future research.

## Bibliography

Abecasis GR, Cherny SS, Cardon LR. 2001. The impact of genotyping error on family-based analysis of quantitative traits. European Journal of Human Genetics 9(2):130-134.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 30(1):97-101.

Akey JM, Zhang K, Xiong MM, Doris P, Jin L. 2001. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. American Journal of Human Genetics 68(6):1447-1456.

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics 62(5):1198-1211.

Almasy L, Blangero J. 2004. Exploring positional candidate genes: linkage conditional on measured genotype. Behavior Genetics 34(2):173-177.

Amberger J, Bocchini C, Hamosh A. 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM (R)). Human Mutation 32(5):564-567.

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occuring in statistical analysis of probabilistic functions of markov chains. Annals of Mathematical Statistics 41(1):164-&.

Becker T, Valentonyte R, Croucher PJP, Strauch K, Schreiber S, Hampe J, Knapp M. 2006. Identification of probable genotyping errors by consideration of haplotypes. European Journal of Human Genetics 14(4):450-458.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics 40(6):695-701.

Boehnke M. 1994. Limits of resolution of genetic-linkage studies - implications for the positional cloning of human-disease genes. American Journal of Human Genetics 55(2):379-390.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic-linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32(3):314-331.

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. 2003. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. American Journal of Human Genetics 73(3):612-626.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics 84(2):210-223.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81(5):1084-1097.

Buetow KH. 1991. Influence of aberrant observations on high-resolution linkage analysis outcomes. American Journal of Human Genetics 49(5):985-994.

Burdick JT, Chen WM, Abecasis GR, Cheung VG. 2006. In silico method for inferring genotypes in pedigrees. Nature Genetics 38(9):1002-1004.

Chang YPC, Kim JDO, Schwander K, Rao DC, Miller MB, Weder AB, Cooper RS, Schork NJ, Province MA, Morrison AC and others. 2006. The impact of data quality on the

identification of complex disease genes: experience from the Family Blood Pressure Program. European Journal of Human Genetics 14(4):469-477.
Chen WM, Abecasis GÇR. 2006. Estimating the power of variance component linkage analysis in large pedigrees. Genetic epidemiology 30(6):471-484.
Cherny SS, Abecasis GR, Cookson WOC, Sham PC, Cardon LR. 2001. The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. Genetic Epidemiology 21:S117-S122.
Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics 11(6):415-425.
Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare Alleles contribute to low plasma levels of HDL cholesterol. Science 305(5685):869-872.
Collins FS, Guyer MS, Chakravarti A. 1997. Variations on a theme: Cataloging human DNA sequence variation. Science 278(5343):1580-1581.
Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams Ja, Goddard ME. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189(September):317-327.
Daw EW, Heath SC, Lu Y. 2005. Single-nucleotide polymorphism versus microsatellite markers in a combined linkage and segregation analysis of a quantitative trait. Bmc Genetics 6.
Delaneau O, Marchini J, Zagury J-F. 2011. A linear complexity phasing method for thousands of genomes. Nature Methods(decembeR).
Douglas JA, Boehnke M, Lange K. 2000. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. American Journal of Human Genetics 66(4):1287-1297.
Douglas JA, Skol AD, Boehnke M. 2002. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. American Journal of Human Genetics 70(2):487-495.
Elston RC, Stewart J. 1971. General model for genetic analysis of pedigree data. Human Heredity 21(6):523–542.
Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. American Journal of Human Genetics 67(5):1219-1231.
Fishelson M, Geiger D. 2004. Optimizing exact genetic linkage computations. Journal of Computational Biology 11(2-3):263-275.
Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, Ch'ang LY, Huang W, Liu B, Shen Y and others. 2003. The International HapMap Project. Nature 426(6968):789-796.
Goldstein DR, Zhao HY, Speed TP. 1997. The effects of genotyping errors and interference on estimation of genetic distance. Human Heredity 47(2):86-100.
Gordon D, Heath SC, Liu X, Ott J. 2001. A transmission disequilibrium test that allows for genotyping errors in the analysis of single nucleotide polymorphism data. American Journal of Human Genetics 69(4):507-507.
Gordon D, Heath SC, Ott J. 1999. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Human Heredity 49(2):65-70.
Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. 2011. Evolutionary evidence of the effect of rare variants on disease etiology. Clinical Genetics 79(3):199-206.

Hackett CA, Broadfoot LB. 2003. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90(1):33-38.

Haldane JBS. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. Journal of Genetics 8(4):299-309.

Heath SC. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics 61(3):748-760.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the Next Generation of Genome-Wide Association Studies. Plos Genetics 5(6):15.

Huang QQ, Shete S, Amos CI. 2004. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. American Journal of Human Genetics 75(6):1106-1112.

Idury RM, Elston RC. 1997. A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Human Heredity 47(4):197-202.

Kennedy J, Mandoiu I, Pasaniuc B. 2008. Genotype error detection using Hidden Markov Models of haplotype diversity. Journal of Computational Biology 15(9):1155-1171.

Kirk KM, Cardon LR. 2002. The impact of genotyping error on haplotype reconstruction and frequency estimation. European Journal of Human Genetics 10(10):616-622.

Knapp M, Becker T. 2004. Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). American Journal of Human Genetics 74(3):589-591.

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T and others. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nature Genetics 40(9):1068-1075.

Krithika S, Valladares-Salgado A, Peralta J, Escobedo-de La Pena J, Kumate-Rodriguez J, Cruz M, Parra EJ. 2012. Evaluation of the imputation performance of the program IMPUTE in an admixed sample from Mexico City using several model designs. BMC Med Genomics.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. American Journal of Human Genetics 58(6):1347-1347.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences of the United States of America 84(8):2363-7.

Lathrop GM, Hooper AB, Huntsman JW, Ward RH. 1983. Evaluating pedigree data.1. The estimation of pedigree error in the presence of marker mistyping. American Journal of Human Genetics 35(2):241-262.

Lathrop GM, Lalouel JM, Julier C, Ott J. 1984. Strategies for multilocus linkage analysis in humans. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 81(11):3443-3446.

Lebrec JJP, Putter H, Houwing-Duistermaat JJ, Van Houwelingen HC. 2008. Influence of genotyping error in linkage mapping for complex traits - an analytic study. Bmc Genetics 9.

Leigh SEA, Foster AH, Whittall RA, Hubbart CS, Humphries SE. 2008. Update and analysis of the University College London Low Density Lipoprotein receptor familial hypercholesterolemia database. Annals of Human Genetics 72:485-498.

Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong XY, Aponte JL, Mooser VE, Chissoe SL, Whittaker JC and others. 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. Plos One 6(9).

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annual review of genomics and human genetics 10:387-406.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34(8):816-834.

Little RJA, Rubin DB. 1987. Statistical analysis with missing data. New York: J. Wiley & Sons.

Lunetta KL, Faraone SV, Biederman J, Laird NM. 2000. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. American Journal of Human Genetics 66(2):605-614.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Investigation 118(5):1590-1605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Marchani EE, Cheung CYK, Glazner CG, Conomos MP, Lewis SM, Sverdlov S, Thornton T, Wijsman EM. 2013. Identity-by-Descent graphs offer a flexible framework for imputation and both linkage and association analyses. GAW18 Proceedings In press.

Markus B, Birk OS, Geiger D. 2011. Integration of SNP genotyping confidence scores in IBD inference. Bioinformatics 27(20):2880-2887.

Matise TC, Chen F, Chen WW, De la Vega FM, Hansen M, He CS, Hyland FCL, Kennedy GC, Kong XY, Murray SS and others. 2007. A second-generation combined linkage-physical map of the human genome. Genome Research 17(12):1783-1786.

Mitchell AA, Cutler DJ, Chakravarti A. 2003. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. American Journal of Human Genetics 72(3):598-610.

Mukhopadhyay N, Buxbaum SG, Weeks DE. 2004. Comparative study of multipoint methods for genotype error detection. Human Heredity 58(3-4):175-189.

Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ and others. 2010. Exome sequencing, ANGPTL3 mutations, and Familial Combined Hypolipidemia. New England Journal of Medicine 363(23):2220-2227.

O'Connell JR, Weeks DE. 1997. PedCheck: A program for identifying marker typing incompatibilities in linkage analysis. American Journal of Human Genetics 61(4):A288-A288.

Ott J, Kamatani Y, Lathrop M. 2011. Family-based designs for genome-wide association studies. Nature Reviews Genetics 12(7):465-474.

Rosenthal EA, Ronald J, Rothstein J, Rajagopalan R, Ranchalis J, Wolfbauer G, Albers JJ, Brunzell JD, Motulsky AG, Rieder MJ and others. 2011. Linkage and association of phospholipid transfer protein activity to LASS4. Journal of Lipid Research 52(10):1837-1846.

Sanna S, Li BS, Mulas A, Sidore C, Kang HM, Jackson AU, Piras MG, Usala G, Maninchedda G, Sassu A and others. 2011. Fine mapping of five loci associated with low-density

lipoprotein cholesterol detects variants that double the explained heritability. Plos Genetics.
Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. 2002. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. American Journal of Human Genetics 71(4):992-995.
Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78:629-644.
Sieberts SK, Wijsman EM, Thompson EA. 2002. Relationship inference from trios of individuals, in the presence of typing error. American Journal of Human Genetics 70(1):170-180.
Sobel E, Lange K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. American Journal of Human Genetics 58(6):1323-1323.
Sobel E, Papp JC, Lange K. 2002. Detection and integration of genotyping errors in statistical genetics. American Journal of Human Genetics 70(2):496-508.
Sobel E, Sengul H, Weeks DE. 2001. Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. Human Heredity 52(3):121-131.
Speed T, Waterman MS. 1996. Genetic mapping and DNA sequencing. Friedman A, Gulliver R, editors. United States of America: Springer-Verlag New York, Inc.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68(4):978-989.
Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO. 2009. Methodological issues in multistage Genome-Wide Association Studies. Statistical Science 24(4):414-429.
Thompson E. 2011. The structure of genetic linkage data: From LIPED to 1M SNPs. Human Heredity 71(2):86-96.
Thompson EA. 2000. Statistical inference from genetic data on pedigrees. United States of America: Institute of Mathematical Statistics.
Thompson EA, Heath SC. 1999. Estimation of conditional multilocus gene identity among relatives. Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology 33:95-113.
Tong L, Thompson E. 2008. Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. Human heredity 65(3):142-53.
Viterbi AJ. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Ieee Transactions on Information Theory 13(2):260-+.
Walters K. 2005. The effect of genotyping error in sib-pair genomewide linkage scans depends crucially upon the method of analysis. Journal of Human Genetics 50(7):329-337.
Wijsman EM, Brunzell JD, Jarvik GP, Austin MA, Motulsky AG, Deeb SS. 1998. Evidence against linkage of familial combined hyperlipidemia to the apolipoprotein AI-CIII-AIV gene complex. Arteriosclerosis Thrombosis and Vascular Biology 18(2):215-226.
Wijsman EM, Rothstein JH, Igo RP, Brunzell JD, Motulsky AG, Jarvik GP. 2010. Linkage and association analyses identify a candidate region for apoB level on chromosome 4q32.3 in FCHL families. Human Genetics 127(6):705-719.
Wijsman EM, Rothstein JH, Thompson EA. 2006. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical

approaches for genome scans on general pedigrees. American Journal of Human Genetics 79(5):846-858.
Wilcox MA, Pugh EW, Zhang HP, Zhong XY, Levinson DE, Kennedys GC, Wijsman EM. 2005. Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for genetic analysis workshop 14: Presentation groups 1, 2, and 3. Genetic Epidemiology 29:S7-S28.
Zhi DG, Wu JH, Liu NJ, Zhang K. 2012. Genotype calling from next-generation sequencing data using haplotype information of reads. Bioinformatics 28(7):938-946.
Zou GH, Pan DY, Zhao HG. 2003. Genotyping error detection through tightly linked markers. Genetics 164(3):1161-1173.

Appendix A

**Convergence property of the estimators:**

Under the assumption that dense marker $v$ is in linkage equilibrium with the framework markers, the estimator $\tilde{P}\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right) = \frac{\sum_{k=1}^{n} P\left(G_{pv}\middle|S_v^k, G_v^{ob}\right) P(G_v^{ob}|\, S_v^k)}{\sum_{k=1}^{n} P(G_v^{ob}|\, S_v^k)}$ converges to $P\left(S_v = s\middle|G_F^{ob}, G_v^{ob}\right)$ in probability as $n$ goes to infinity. The modified estimator $\hat{P}\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right) = \frac{1}{n^*}\sum_{k=1}^{n} P\left(G_{pv}\middle|S_v^k, G_v^{ob}\right)$, where $n^* = \sum_{k=1}^{n} I(P(G_v^{ob}|S_{vk}) > 0)$, converges to a quantity like $P\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right)$ that replaces the emission probability $P(G_v^{ob}|s)$ by the emission function $I(P(G_v^{ob}|s) > 0)$ at the position $v$. However, in test datasets, the estimates from the two estimators are often similar (data not shown).

Proof of convergence:

To see this, let $h(G_v^{ob}, s)$ be the *generic* emission function at the position $v$, conditional on the IV $s$ at the position $v$. Most commonly, $h(G_v^{ob}, s) = P(G_v^{ob}|s)$.

We show equality between $P(S_v = s|G_F^{ob}, G_v^{ob})$ and $\tilde{p} = \frac{P\left(S_v=s\middle|G_F^{ob}\right)h\left(G_v^{ob},s\right)}{\sum_w P\left(S_v=w\middle|G_F^{ob}\right)h\left(G_v^{ob},w\right)}$ when $h(G_v^{ob}, s) = P(G_v^{ob}|\, s)$. For brevity, we omit the inclusion of allele frequencies into the equation below. We assume that the dense marker $v$ is not in linkage disequilibrium with the framework markers, which are indexed from 1 to $M$.

Define $\alpha_j(s) = P(G_1^{ob}, \ldots, G_j^{ob}, S_j = s)$ and $\beta_j(s) = P(G_{j+1}^{ob}, \ldots, G_M^{ob}|S_j = s)$.

$$\tilde{p} = \frac{P(S_v=s|G_F^{ob})h(G_v^{ob},s)}{\sum_w P(S_v=w|G_F^{ob})h(G_v^{ob},w)}$$

$$= \frac{P(S_v=s,G_F^{ob})h(G_v^{ob},s)}{\sum_w P(S_v=w,G_F^{ob})h(G_v^{ob},w)}$$

$$= \frac{\sum_x \sum_y P(S_v=s|S_j=x,S_{j+1}=y,G_F^{ob})P(S_j=x,S_{j+1}=y,G_F^{ob})h(G_v^{ob},s)}{\sum_w \sum_x \sum_y P(S_v=w|S_j=x,S_{j+1}=y,G_F^{ob})P(S_j=x,S_{j+1}=y,G_F^{ob})h(G_v^{ob},w)}$$

$$= \frac{\sum_x \sum_y P(S_v=s|S_j=x,S_{j+1}=y)P(S_j=x,S_{j+1}=y,G_F^{ob})h(G_v^{ob},s)}{\sum_w \sum_x \sum_y P(S_v=w|S_j=x,S_{j+1}=y)P(S_j=x,S_{j+1}=y,G_F^{ob})h(G_v^{ob},w)}$$

$$= \frac{\sum_x \sum_y P(S_v=s|S_j=x,S_{j+1}=y)\alpha_j(x)P(S_{j+1}=y|S_j=x)P\left(G_{j+1}^{ob}\middle|S_{j+1}=y\right)\beta_{j+1}(y)h(G_v^{ob},s)}{\sum_w \sum_x \sum_y P(S_v=w|S_j=x,S_{j+1}=y)\alpha_j(x)P(S_{j+1}=y|S_j=x)P\left(G_{j+1}^{ob}\middle|S_{j+1}=y\right)\beta_{j+1}(y)h(G_v^{ob},w)}$$

$$= \frac{\sum_x \sum_y \alpha_j(x)P(S_v=s|S_j=x)h(G_v^{ob},s)P(S_{j+1}=y|S_v=s)P\left(G_{j+1}^{ob}\middle|S_{j+1}=y\right)\beta_{j+1}(y)}{\sum_w \sum_x \sum_y \alpha_j(x)P(S_v=w|S_j=x)h(G_v^{ob},w)P(S_{j+1}=y|S_v=w)P\left(G_{j+1}^{ob}\middle|S_{j+1}=y\right)\beta_{j+1}(y)}$$

$$= \frac{P(S_v=s,G_F^{ob},G_v^{ob})}{\sum_w P(S_v=w,G_F^{ob},G_v^{ob})} \text{ [when } h(G_v^{ob},s) = P(G_v^{ob}|s)]$$

$$= P(S_v = s|G_F^{ob}, G_v^{ob})$$

, which holds because

$$P\left(S_v = s\middle|S_j = x, S_{j+1} = y\right)P\left(S_{j+1} = y\middle|S_j = x\right) = P\left(S_v = s\middle|S_j = x\right)P\left(S_{j+1} = y\middle|S_v = s\right)$$

by the property of the Haldane Map function.

If $h(G_v^{ob},s) = P(G_v^{ob}|s)$, this equation becomes the usual calculation of $P(S_v = s|G_F^{ob}, G_v^{ob})$. This result tells us that the proper way to update the probability distribution of $S_v$ after adding genotypes of the dense marker $v$ is to reweight the top and bottom by the emission probability of the dense marker $v$. Alternatively, we can define

$$h(G_v^{ob},s) = P(G_v^{ob} is\ compatible\ with\ s)$$

$$= \begin{cases} 1\ if\ P(G_v^{ob}|s) > 0 \\ 0\ if\ P(G_v^{ob}|s) = 0 \end{cases}$$

$$= I(P(G_v^{ob}|s) > 0)$$

This emission function only uses the deterministic information of $G_v^{ob}$ and does not depend on the allele frequency of the dense marker $v$. Hence, the use of this emission function avoids the negative impact as a result of falsely assuming that the tightly linked markers are in linkage equilibrium with each other, which is an unrealistic assumption.

Now, we calculate $P\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right)$.

$$P\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right) = \sum_s P\left(G_{pv}\middle|S_v = s, G_F^{ob}, G_v^{ob}\right)P(S_v = s|G_F^{ob}, G_v^{ob})$$

$$= \sum_s P\left(G_{pv}\middle|S_v = s, G_v^{ob}\right)P(S_v = s|G_F^{ob}, G_v^{ob})$$

$$= \sum_s P\left(G_{pv}\middle|S_v = s, G_v^{ob}\right)\frac{P\left(S_v=s|G_F^{ob}\right)P\left(G_v^{ob}|s\right)}{\sum_w P(S_v=w|G_F^{ob})P\left(G_v^{ob}|s\right)}$$

A natural estimator of $P\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right)$ is to plug in $\hat{P}\left(S_v = s|G_F^{ob}\right)$ for $P\left(S_v = s\middle|G_F^{ob}\right)$.

$\hat{P}\left(S_v = s|G_F^{ob}\right) = \frac{1}{n}\sum_{k=1}^n I(S_{vk} = s\left|G_F^{ob}\right.)$ is an empirical estimator of $P\left(S_v = s|G_F^{ob}\right)$ using the realized MCMC samples $S_{v1}, \ldots, S_{vn}$. We propose the estimator

$$\tilde{P}\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right) = \sum_s P\left(G_{pv}\middle|S_v = s, G_v^{ob}\right)\frac{\hat{P}\left(S_v=s|G_F^{ob}\right)P\left(G_v^{ob}|s\right)}{\sum_w \hat{P}\left(S_v=w|G_F^{ob}\right)P\left(G_v^{ob}|s\right)}$$

$$= \frac{\sum_s P\left(G_{pv}|S_v=s,G_v^{ob}\right)\frac{1}{n}\sum_{k=1}^n I(S_v^k=s|G_F^{ob})P\left(G_v^{ob}|s\right)}{\sum_w \frac{1}{n}\sum_{k=1}^n I(S_v^k=w|G_F^{ob})P\left(G_v^{ob}|s\right)}$$

$$= \frac{\sum_{k=1}^n P\left(G_{pv}|S_v^k,G_v^{ob}\right)P\left(G_v^{ob}|S_v^k\right)}{\sum_{k=1}^n P\left(G_v^{ob}|S_v^k\right)}$$

$$\xrightarrow{p} P\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right)$$

If we replace $P(G_v^{ob}|s)$ by $I(P(G_v^{ob}|s) > 0)$ at dense marker $v$, the second estimator becomes

$$\hat{P}\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right) = \frac{\sum_{k=1}^n P\left(G_{pv}|S_{vk},G_v^{ob}\right)I(P(G_v^{ob}|\ S_{vk})>0)}{\sum_{k=1}^n I(P(G_v^{ob}|\ S_{vk})>0)}$$

$$= \frac{1}{n^*}\sum_{k=1}^n P\left(G_{pv}\middle|S_{vk}, G_v^{ob}\right)I(P(G_v^{ob}|\ S_{vk}) > 0)$$

In Chapter 4, I used $\hat{P}\left(G_{pv}\middle|G_F^{ob}, G_v^{ob}\right)$ as the estimator because this estimator does not require the assumption that the dense marker $v$ and framework markers are in linkage equilibrium.

## VITA

Charles Yin Kiu Cheung was born in Hong Kong. He grew up in Vancouver, British Columbia, Canada. He completed his Bachelor of Science degree at the University of British Columbia. In 2013, Charles earned his doctoral degree in Biostatistics at the University of Washington.