A PIPELINE FOR CLASSIFYING CLOSE FAMILY RELATIONSHIPS WITH DENSE SNP DATA AND PUTATIVE PEDIGREE INFORMATION

by

Zhen Zeng

B.E., Harbin Institute of Technology, China, 2007

M.S., Shanghai Jiao Tong University, China, 2010

M.S. University of Pittsburgh, 2013

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Zhen Zeng

It was defended on

May 6th, 2015

and approved by

Dissertation Advisor: Eleanor Feingold, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Daniel E. Weeks, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Committee Member: George C. Tseng, Sc.D., Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Wei Chen, Ph.D., Assistant Professor, Division of Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC

A PIPELINE FOR CLASSIFYING CLOSE FAMILY RELATIONSHIPS WITH DENSE SNP DATA AND PUTATIVE PEDIGREE INFORMATION

Zhen Zeng, Ph.D.
University of Pittsburgh, 2015

ABSTRACT

When genome-wide association studies (GWAS) or sequencing studies are performed on family-based datasets, the genotype data can be used to check the structure of putative pedigrees. Even in datasets of putatively unrelated people, close relationships can often be detected using dense single-nucleotide polymorphism/variant (SNP/SNV) data.

A number of methods for finding relationships using dense genetic data exist, but they all have certain limitations, including that they typically use average genetic sharing, which is only a subset of the available information. We present a set of approaches for classifying relationships in GWAS datasets or whole genome sequencing datasets. We first propose an empirical method for detecting identity-by-descent segments in close relative pairs using unphased dense SNP data and demonstrate how that information can assist in building a relationship classifier. We then develop a strategy to take advantage of putative pedigree information to enhance classification accuracy. Our methods are tested and illustrated with two SNP array datasets from two distinct populations. With these new techniques, we propose classification pipelines for checking and identifying pair-wise relationships in datasets containing a large number of small pedigrees.

We also explore the performance of the pipeline on a whole exome sequencing dataset. Although the classifier based on SNP array data does not perform well on exome sequencing data, it can in principle be modified using new algorithm parameters and training data in order to achieve better performance.

Finally, we develop a method to reconstruct pedigrees from pair-wise relationship information. Our method can reconstruct core pedigrees with high accuracy and pair-wise relationship inferences can be further improved during this process.

Detecting close family relationships and reconstructing pedigrees are important in both population-based and family-based studies. Providing precise pedigrees and hidden relatedness information helps increase the accuracy and power of various genetic analyses and avoids false positive associations, making these studies more efficient in identifying the genetic basis of diseases. This is a crucial step on the path to developing better treatments and interventions and improving public health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation is dedicated to my family: my parents, wife and son. It is not possible without the unselfish support from them.

I give the sincerest thank to my advisor, Dr. Eleanor Feingold. She led me to the world of statistical genetics. Her wisdom encouraged me and inspired my work. I am also fortunate to have Dr. Daniel E. Weeks, Dr. Wei Chen and Dr. George C. Tseng in my committee. This work will lose much luster without their priceless advices and help.

Our work was funded by R01-HD038979. We sincerely thank Dr. Mary L. Marazita for providing the COHRA and GENEVA datasets. The development of these datasets was funded by grants R01-DE014899, R01-DE-016148, and U01-DE018903.

1.0 INTRODUCTION

1.1 OVERVIEW

When genome-wide association studies (GWAS) or sequencing studies are performed on family-based datasets, the genetic marker data can be used to check the structure of putative pedigrees. Even in datasets of putatively unrelated people, such as a large number of people from a local geographical region (e.g., sequenced by a regional healthcare system), close relationships can often be detected using dense single-nucleotide polymorphism/variant (SNP/SNV, we use SNP for short) data.

It is important to test putative relationships and also test for unexpected relationships in genetic studies. Validity of linkage analyses depends on accurate pedigree structures. Hidden relatedness may cause genomic inflation and affect ancestry inferences in population-based studies (Patterson, Price et al. 2006). Presence of hidden close relationships may also lead to false associations, especially in the analysis of rare variants. In health-system datasets without pedigree information, pedigrees can be inferred, adding power to genetic analyses.

An important step in relationship inference is to identify regions of identity-by-descent (IBD) between pairs of individuals. Shared IBD segments between individuals can be inferred by identical-by-state (IBS) of streaks of SNPs using dense marker data. Existing methods for detecting IBD segments have certain limitations: some are very time-consuming, some require

extra assumptions, such as independence between SNPs, and some need correctly phased genotypes as input. To strike a balance between computational time, accuracy, and extra requirements on the data, we developed an empirical method for detecting IBD segments and indentifying recombination events in close relative pairs. We explored combinations of different strategies (large sliding window vs. small fixed window; reference panel vs. no reference panel; windows based on physical distance vs. those based on a fixed number of SNPs) and developed a new algorithm that is computationally efficient and does not require knowledge of putative relationships.

A common limitation of existing relationship inference methods is that most of them use average genetic sharing, which is only a subset of the available information. By using our algorithm, spatial genetic sharing information can be extracted. We built a relationship classifier with information on both average and spatial genetic sharing for detecting relationships in datasets of putatively unrelated individuals or for checking relationships in datasets containing a large number of small pedigrees.

Putative pedigrees are typically generated based on subject interviews. These pedigree structures often contain errors, but most of the information is still correct and useful. Putative relationships based on the putative pedigree structures can be used as prior knowledge to adjust for relationship classification (Ray and Weeks 2008). Ambiguous relationships falling on the boundaries of two or more categories might be pushed to the right category by taking into account their putative relationships. We therefore developed a strategy to take advantage of putative pedigree information to enhance classification accuracy.

Finally, we propose a method to reconstruct pedigrees with pair-wise relationship predictions. By reconstructing pedigrees, not only can we further improve the accuracy of pair-wise relationship inference, but also generate the necessary input for various genetic analyses.

This dissertation consists of 5 chapters. This first chapter is an overview and provides the general background as well as a literature review. Chapter 2 is a research paper covering the most essential parts of our work. It has minor overlap with chapter 1, but those parts were retained in order to make the chapter more readable as a standalone article. Chapter 3 and chapter 4 present two extensions of the work described in chapter 2. The last chapter provides a summary and discusses some possible future work.

1.2 DENSE SNP DATA

With the development of high-throughput genotyping technologies, the cost of dense SNP data has become more and more affordable for large-scale samples. We define dense as having at least 100,000 successfully genotyped SNPs. Two main technologies to generate dense SNP data are microarrays and massive parallel sequencing.

1.2.1 SNP array

Microarray technology allows thousands to millions of molecules to be measured simultaneously on a small chip for a biological sample. A microarray is referred to as SNP array when it is designed for SNP genotyping. The genetic polymorphisms on a SNP array are preselected and

optimized to produce very accurate genotypes. As a result, with array technology, all genotyped individuals share a fixed set of SNPs. The selected SNPs are usually polymorphic in most populations and the SNP numbers for different arrays range from 100,000 to 5,000,000.

1.2.2 Next-generation sequencing

As automated Sanger sequencing is recognized as the first generation of sequencing technology, massively parallel sequencing technologies are referred to as the next-generation sequencing (Metzker 2010).

Two types of next-generation sequencing can potentially generate genome-wide dense SNP data: one is whole genome sequencing and the other is whole exome sequencing. As denoted by their names, whole genome sequencing produces sequence data covering the whole genome, and whole exome sequencing only targets exon sequences.

Sequencing data are essentially a collection of reads randomly distributed across the targeted area. Ideally such data should contain all the genetic variants in the targeted area, both common and rare. However, because of the randomness, there are some issues specific to sequencing data in practice. First, not all individuals have exactly the same set of SNPs being called from deep-sequencing data, so missingness is a difficult problem to deal with. Not only is this the case when the coverage is low or moderate, but is also true even for deep coverage. Secondly, the quality of genotypes called from sequencing data is often lower than that from SNP array data. Heterozygous genotypes are particularly inaccurate when the number of mapped reads is too few. Fortunately, when the sequencing depth is high enough, it always captures as many common SNPs without missingness and with almost as high quality as the SNP array.

For exome sequencing, the coverage is usually very deep. However, it is arguably not a genome-wide genotyping method, because there are gaps between exons. Although exons are quite dense and cover most of chromosomal regions, they are not distributed evenly. Therefore, we investigated if our methods are applicable to whole exome sequencing data and the results are shown in Chapter 3.

1.3 BASICS OF PEDIGREE INFERENCE AND A LITERATURE REVIEW ON EXISTING METHODS

The problem of pedigree testing and construction can be divided into three steps: segmental IBD detection, relationship inference, and pedigree inference. We review the background for each of them separately.

1.3.1 IBD segment detection

The first step in pedigree inference is to estimate IBD sharing between pairs of individuals. In principle, spatial information on genetic sharing can assist in determining relationships. For example, IBD states along chromosomes can be described as Markov processes with transition rates λ and 2λ for grandparent-grandchild and half-siblings. The process for avuncular relationship is non-Markov, but the transition rate is known to be $5/2\lambda$ (λ can be interpreted as the unit of genetic length) (Feingold 1993). In other words, the expected sojourn length between IBD state transition spots is different for different relationships. Accordingly, the observed times

of IBD state transition, which is a summary of the spatial IBD information, can help classify relationships.

Existing popular methods for detecting IBD segments include PLINK (Purcell, Neale et al. 2007), BEAGLE (Browning and Browning 2011), PARENTE/PARENTE2 (Rodriguez, Bercovici et al. 2014) and GERMLINE (Gusev, Lowe et al. 2009).

The IBD segment detecting algorithm embedded in PLINK is based on a hidden Markov model (HMM), which uses the observed IBS and overall genetic sharing between the pair. PLINK does not model the dependency among SNPs and requires that the input SNPs are in linkage equilibrium, so dense SNP data must be pruned before running PLINK.

The fastIBD algorithm implemented in BEAGLE also employs HMM. It takes into account the dependency among SNPs to simultaneously phase genotypes and detect shared IBD. Despite “fast” is in the name of this method, it requires quite intense computation, and is roughly one magnitude slower than other methods.

PARENTE and PARENTE2 employ a variant of a likelihood ratio test for the presence of IBD in a sliding window. While PARENTE assumes independent SNPs and requires SNP pruning, PARENTE2 relaxes this by accounting for linkage disequilibrium. However, both PARENTE and PARENTE2 need a training dataset. While PARENTE adopts a training set of unrelated pairs, PARENTE2 requires phased training data to empirically estimate haplotype frequencies.

GERMLINE (Genetic Error-tolerant Regional Matching with Linear-time Extension) reduces the quadratic-time for the number of individuals to linear. It starts with phased genotype data and looks for the matches of haplotypes among individuals and expands them to long segmental sharing. However, as a trade-off it requires phased genotypes as input.

In general, most of the existing methods are likelihood-based. All of them have certain limitations: some of them require phasing the genotypes, some are very time-consuming, some are based on strong assumptions such as independence between genetic markers, and some require extra information such as training data.

1.3.2 Pair-wise relationship inference

The second step in pedigree inference is inference about relationships between pairs of relatives. Existing methods for relationship inference can be grouped into the following three classes.

Methods for sparse genetic markers: Before the era of GWAS and whole-genome sequencing studies, the popular methods for inferring relationships models IBD states using likelihood-based methods, assuming independent markers. Examples include PREST (McPeek and Sun 2000) and RELPAIR (Epstein, Duren et al. 2000). These methods usually infer relationships based on hypothesis testing. Because they are based on a hypothesis-testing paradigm, they have built-in assumptions about what relationships are more likely.

Methods for dense SNP data without using spatial genetic sharing information: When dense SNP data became available, many new relationship inference tools were developed to use genome-wide data, such as PLINK (Purcell, Neale et al. 2007), KING (Manichaikul, Mychaleckyj et al. 2010) and REAP (Thornton, Tang et al. 2012). PLINK makes a strong assumption of a homogeneous population. KING and REAP are robust in the presence of population structure, while REAP also works for population admixture. These methods focus on estimating measures of average genetic sharing, such as kinship coefficients and probabilities of IBD sharing between each pair of individuals. Based on estimates of these quantities, using

predefined inference criteria, different relationships can be separated. But many very close relatives, such as grandparent-grandchild, half-siblings and avuncular pairs in the second-degree relationship category, share the same expected values for these quantities (Table 1.1).

Table 1.1 Expected IBD scores of different relationships

Relationship Category	P(IBD=0)	P(IBD=1)	P(IBD=2)
Monozygotic twins	0	0	1
Full siblings	1/4	1/2	1/4
Parent-offspring	0	1	0
Unrelated	1	0	0
First cousins	3/4	1/4	0
Grandparent-child	1/2	1/2	0
Half siblings	1/2	1/2	0
Avuncular pairs	1/2	1/2	0

Methods for dense SNP data using spatial genetic sharing information: More recent methods have begun to incorporate spatial information on genetic sharing to help increase the inference accuracy. Stevens et al (Stevens, Heckenberg et al. 2011) developed a method to calculate better estimates of average IBD sharing based on observed IBS within chromosomal windows. Then combining with the average IBS sharing, it empirically infers the degrees of relationships. GRAB (genetic relationship by averaged blocks) (Li, Glusman et al. 2014) employs a similar approach to segment the genome into blocks to obtain average IBD sharing, but uses a classification tree to infer relationship degrees. Although both of these methods consider spatial information on genetic sharing, they do not directly use it for relationship classification but instead use it for better estimates of average IBD sharing. ERSA (estimation of recent shared ancestry) (Huff, Witherspoon et al. 2011) constructs a likelihood ratio test for any

relatedness utilizing the number and genetic lengths of IBD segments shared between two individuals. They provide a maximum-likelihood estimate for the degree of relationship if significant relatedness is found between the two individuals. ERSA 2.0 (Li, Glusman et al. 2014) achieves better performance for whole-genome sequence data by masking several irregular genomic regions that exhibit excess spurious IBD in sequencing data, including certain centromeric regions, unmappable heterochromatic regions of the genome, and regions of long-range linkage disequilibrium. ERSA and ERSA 2.0 appear to be the only relationship inference tools that actually take advantage of spatial genetic sharing. With the aid of spatial genetic sharing information, improved accuracy has been shown for these methods. However, they still focus on separating degrees of relationships and none of them can effectively separate second-degree relatives, i.e., grandparent-grandchild, half-siblings, and avuncular pairs, from each other.

1.3.3 Pedigree reconstruction

Pedigree reconstruction with genetic data has a long history, but until the recent availability of whole-genome genotype data, it was not powerful enough to be practical. Several state-of-the-art methods using whole-genome data have been developed recently, including MLP-ILP (Maximum Likelihood Pedigree reconstruction using Integer Linear Programming) (Cussens, Bartlett et al. 2013), COP (Constructing Outbred Pedigrees) and CIP (Constructing Inbred Pedigrees) (Kirkpatrick, Li et al. 2011), IPED (Inheritance Path-based Pedigree Reconstruction) (He, Wang et al. 2013) and IPED2, PREPARE (Pedigree Reconstruction of Extant populations using Partitioning of Relatives) (Shem-Tov and Halperin 2014), and PRIMUS (Pedigree Reconstruction and Identification of the Maximally Unrelated Set) (Staples, Qiao et al. 2014).

These methods work under different assumptions and are suitable for different real-life problems.

Methods taking genotype data as input: MLP-ILP (Cussens, Bartlett et al. 2013) treats pedigrees as Bayesian networks and uses integer linear programming to search for the pedigree graph that maximizes the likelihood of the Bayesian network given the observed genotypes assuming all genetic markers are independent and the founder genotypes are in Hardy-Weinberg equilibrium. The advantage of this method is that it can handle inbred pedigrees and pedigrees with half-siblings. The downsides are that it only works for sparse and independently segregating genetic markers, and requires all the connecting individuals in pedigrees have fully observed genotype data.

Methods taking IBD segments between each pair of individuals as input: This is the most popular school of methods so far. These methods attempt to reconstruct pedigrees from an "extant population", which is defined as the latest generation in the population. They reconstruct pedigrees generation-by-generation, starting from the extant population, by grouping siblings among extant individuals and determining the parents of sibling group recursively. Of note, the extant population is assumed to be from the same generation, and only extant individuals have genotype data. Therefore, these methods are essentially reconstructing pedigrees by relating individuals with dummy parents and the dummy parents of dummy parents and so on. COP/CIP (Kirkpatrick, Li et al. 2011) first derives the expectation and variance of shared segment length between pairs of extant individuals that are related by different generations, analytically for outbred pedigrees (COP) and by simulation for inbred pedigrees (CIP). Then it calculates the observed average of shared length for each pair of individuals from the input data. With these quantities, tests are constructed for inferring sibling relationships in different generations.

Finally, COP/CIP applies a Max-Clique algorithm to identify sibling groups in each generation. This method does not consider pedigrees involving half-siblings, and the computation time for inbred pedigrees is exponential in the number of individuals. IPED (He, Wang et al. 2013) borrows the idea of COP/CIP and accelerated the computation for inbred pedigrees by dynamic programming. IPED still cannot deal with half-siblings in pedigrees. IPED2 extends the method to handle half-siblings. PREPARE (Shem-Tov and Halperin 2014) considers partitioning the relatives into maternal and paternal relatives and improves the accuracy of reconstructed pedigrees. It also can deal with half-siblings in pedigrees. An important shared issue with these methods is that they assume all the genotyped individuals are in the same generation, i.e., the extant population, which is rarely satisfied in reality. Also, performance of these methods relies on the accuracy of IBD segment inputs. Sometimes, inferring IBD segments is not trivial itself.

Methods taking pair-wise relationship predictions as input: PRIMUS (Staples, Qiao et al. 2014) is the most recent pedigree reconstruction method. It uses pair-wise relationship from six categories as the building blocks: parental, full-sibling, second-degree, third-degree, distant and unrelated relationships to reconstruct all possible pedigrees using relationship-likelihood vectors of all pair-wise relationships subject to several a priori restrictions. PRIMUS outperforms other current methods in most realistic settings, but it also has a few drawbacks, such as not using spatial information on genetic sharing (resulting in low accuracy for second-degree relationship predictions when the proportion of missing individuals is high) and being very slow for large pedigrees with a large number of missing individuals.

1.4 OUR METHODS

In summary, many existing methods share some strategies with the pipeline we propose. However, none combines the strategies we feel are best in a single pipeline. And our pipeline includes novel techniques and features. Those are:

- A fast empirical IBD algorithm
- Use of spatial IBD information in relationship classification
- Classification approach rather than hypothesis-testing
- An option to use putative pedigree information or not
- Reconstruction of pedigrees
- Fast, convenient, and imposing few assumptions

2.0 A PIPELINE FOR CLASSIFYING RELATIONSHIPS USING DENSE SNP/SNV DATA AND PUTATIVE PEDIGREE INFORMATION

This paper has been submitted to Genetic Epidemiology and is currently in the first round of revision. The co-authors include Zhen Zeng, Daniel E Weeks, Wei Chen, Nandita Mukhopadhyay and Eleanor Feingold. Eleanor Feingold proposed the initial ideas. I developed the ideas into detailed methods. Daniel E Weeks and Wei Chen provided advice and provided references to other relevant work. I implemented the methods, carried out simulations, and applied the methods to real datasets. Nandita Mukhopadhyay provided the analysis results of her recombination detection algorithm as a means for comparing our IBD segment detection algorithms. I drafted the first version of the paper. All the co-authors reviewed and contributed to the current version of this paper.

2.1 ABSTRACT

When genome-wide association studies (GWAS) or sequencing studies are performed on family-based datasets, the genotype data can be used to check the structure of putative pedigrees. Even in datasets of putatively unrelated people, close relationships can often be detected using dense single-nucleotide polymorphism/variant (SNP/SNV) data. A number of methods for finding

relationships using dense genetic data exist, but they all have certain limitations, including that they typically use average genetic sharing, which is only a subset of the available information. Here we present a set of approaches for classifying relationships in GWAS datasets or large-scale sequencing datasets. We first propose an empirical method for detecting identity-by-descent segments in close relative pairs using un-phased dense SNP data and demonstrate how that information can assist in building a relationship classifier. We then develop a strategy to take advantage of putative pedigree information to enhance classification accuracy. Our methods are tested and illustrated with two datasets from two distinct populations. Finally, we propose classification pipelines for checking and identifying relationships in datasets containing a large number of small pedigrees.

2.2 INTRODUCTION

In genetic studies it is important to test putative relationships and also test for unexpected relationships. Validity of linkage analyses depends on accurate pedigree structure. Hidden relatedness may cause genomic inflation and affect ancestry inferences in population-based studies (Patterson, Price et al. 2006). Presence of hidden close relationships may also lead to false associations, especially in the analysis of rare variants. Also, inferring relationship pairs is useful in genealogical studies and forensics.

A number of methods are available for testing relationships based on likelihoods and hypothesis testing, such as PREST (McPeek and Sun 2000) and RELPAIR (Epstein, Duren et al. 2000). These methods usually require sparse and uncorrelated genetic markers. Most of the

existing relationship inference tools for dense single-nucleotide polymorphism/variant (SNP/SNV, we will use SNP for short in the rest of the paper) data as from genome-wide association studies (GWAS) or sequencing studies, such as PLINK (Purcell, Neale et al. 2007), need a strong assumption of a homogeneous population. More recent additions, KING (Manichaikul, Mychaleckyj et al. 2010) and REAP (Thornton, Tang et al. 2012), are robust in the presence of population structure and admixture. However, although these methods are powerful for detecting first-degree, second-degree, and third-degree relationships, none of them can effectively separate second-degree relatives, i.e., grandparent-grandchild, half-siblings, and avuncular pairs, from each other. This is due to the fact that existing algorithms focus on estimating measures of average genetic sharing, such as kinship coefficients and probabilities of identity-by-descent (IBD) sharing, and the above mentioned second-degree relatives share the same expected values for these quantities.

Average genetic sharing is only part of the information available in genomic data. In principle, grandparent-grandchild, half-siblings, and avuncular pairs are separable if spatial information on genetic sharing is also considered. IBD states along chromosomes can be described as Markov processes with transition rates λ and 2λ for grandparent-grandchild and half-siblings. For avuncular relationship, the process is non-Markov, but the transition rate is known to be $5/2\lambda$ (λ can be interpreted as the unit of genetic length) (Feingold 1993). In other words, the expected sojourn length in different IBD states, a summary of the spatial IBD information, is different for different relationships. The observed times of transition can therefore help classify relationships. Several existing algorithms for detecting segmental sharing of IBD, such as PLINK (Purcell, Neale et al. 2007), fastIBD (Browning and Browning 2011), GERMLINE (Gusev, Lowe et al. 2009), PARENTE and PARENTE2 (Rodriguez, Bercovici et

al. 2014), can be used to generate such summary statistics of spatial information, but they all have certain limitations: PLINK and PARENTE do not model SNP dependency and require SNP pruning; fastIBD has to phase genotypes and call IBD segments simultaneously; GERMLINE needs correctly-phased genotype data as input; and PARENTE2 requires a phased training dataset.

Another important piece of relationship information is the putative pedigree that is typically generated based on subject interviews. These pedigree structures often contain errors, but most of the information is still correct and useful. Putative relationships based on the assumed pedigree structures could be used as prior knowledge to adjust for relationship classification (Ray and Weeks 2008). Ambiguous relationships falling on the boundaries of two or more categories based on IBD information might be therefore pushed to the right category by taking into account their putative relationships. Furthermore, the recombination rate in paternal meiosis (i.e., spermatogenesis) is known to be much lower than in maternal meiosis (i.e., oogenesis). The genetic length of the female autosomal genome is estimated to be 1.65 times that of the male (Kong, Gudbjartsson et al. 2002). Thus, expected IBD transition rates differ even within the same relationship category, depending on maternal meiosis or paternal meiosis, which could be inferred from sexes of intervening relatives. Therefore, sexes of pertinent relatives, if available, could be useful for further improving classification accuracy. So far, to our knowledge no existing method takes advantage of putative relationships and sexes of meiosis.

In this paper we propose a set of approaches for classifying relationship types in GWAS datasets or large-scale sequencing datasets. We first present a new empirical algorithm for finding regions of IBD in closely-related individuals using un-phased dense SNP data. A summary of IBD spatial information, observed recombination number (N), is generated. We then

demonstrate how that information can be used in principle to distinguish relationships. We also build a classifier and develop novel approaches taking advantage of information from putative pedigree structures. All the methods are tested and illustrated with two different datasets. Finally, we propose classification pipelines for checking and identifying relationships aimed at datasets containing a large number of small pedigrees. Computational tools for implementing our methods are provided.

2.3 METHODS

2.3.1 Datasets

Our methods were applied to two datasets from two distinct populations. One dataset consists of a homogeneous US sample (non-Latino whites) from the Center for Oral Health Research in Appalachia (COHRA) Project (dbGaP accession number phs000095.v3.p1). The other consists of a non-homogeneous Guatemalan sample from the Gene-Environment Association Studies (GENEVA) Guatemala Dental Caries Project (dbGaP accession number phs000440.v1.p1). Both datasets were genotyped using Illumina Human610-Quadv1_B BeadChip (Illumina, Inc., San Diego, CA, USA), and were cleaned to have genotyping rate per individual larger than 0.9, genotyping rate per SNP larger than 0.9, minor allele frequency larger than 0.01, and Hardy–Weinberg equilibrium test p-value larger than 10^{-4}. Approximately 540,000 autosomal SNPs were included in each dataset.

Table 2.1 Sample sizes and means of observed recombination number (N) by relationship category for the two training datasets

Relationship category	Guatemala		US	
	Training sample size	N* mean (sd)	Training sample size	N* mean (sd)
PO	100	0.2 (0.6)	100	0.4 (0.8)
GG	72	43.8 (10.1)	46	35.3 (9.5)
GGp	15	33.5 (9.6)	12	27.4 (8.6)
GGm	57	46.5 (8.5)	34	38.0 (8.3)
HS	60	76.4 (12.9)	100	68.6 (14.8)
HSp	1	65.0 (NA)	18	45.0 (5.3)
HSm	59	76.5 (12.9)	82	73.8 (10.5)
AV	100	82.3 (9.9)	100	75.9 (9.0)
FC	39	65.3 (10.9)	91	59.6 (13.6)
UN	100	7.5 (6.1)	100	1.5 (1.3)

*Not adjusted by the mean of UN pairs. PO: parent-offspring, GG: grandparent-grandchild, HS: half-siblings, AV: avuncular pair, FC: first-cousins, UN: unrelated pair, GGp and HSp: paternal-meiosis GG and HS, GGm and HSm: maternal-meiosis GG and HS.

Both datasets contain abundant close relationships. The pedigree files have been previously cleaned manually by experts. We selected a number of pairs of individuals with confident relationships for the following categories as the gold standard to train two separate classification models, one for each population: monozygotic twins (MZ), full-siblings (FS), parent-offspring (PO), grandparent-grandchild (GG), half-siblings (HS), avuncular pair (AV),

first-cousins (FC), and unrelated pair (UN). Duplicate GG pairs (a grandchild and each of the grandparents are duplicated pairs) and any problematic relationships identified during the current analyses were removed from the training data. The training data sizes by relationship categories are summarized in Table 2.1.

2.3.2 Algorithms for inferring IBD segments

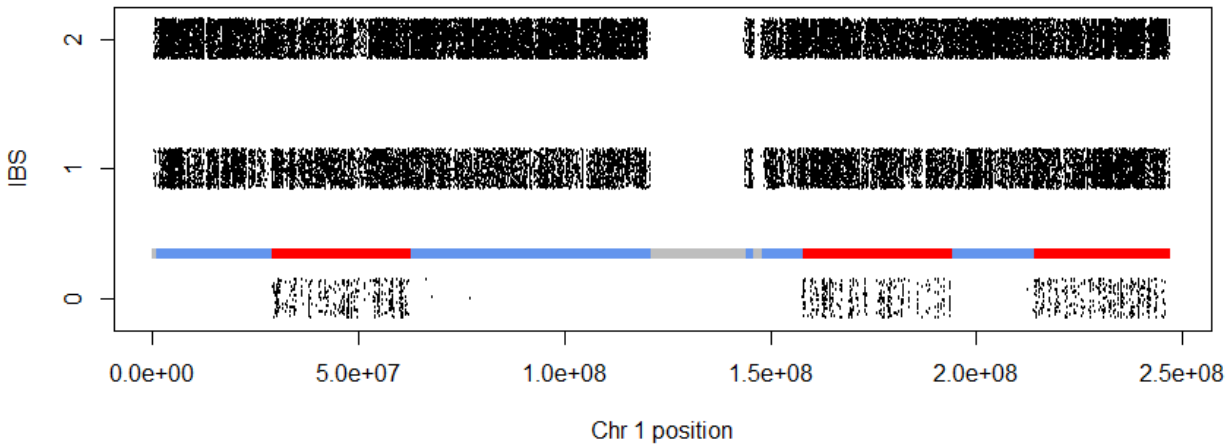


Figure 2.1 An example illustrating the IBD segments identified by our algorithm on chromosome 1 for a pair of individuals

Each dot represents a SNP. Red bars indicate IBD=0 segments. Blue bars indicate IBD=1 segments. Gray bars indicate uncertain regions where SNPs are too sparse. In this example, the observed recombination number (N) is 5.

Our algorithm is based on rules relating IBD and identity-by-state (IBS): assuming no genotyping error, in IBD=0 regions, the IBS state could be 0, 1 or 2; in IBD=1 regions, the IBS state could be either 1 or 2; in IBD=2 regions, the IBS state can only be 2. Under these assumptions, large IBD segments can be identified simply by eye (Figure 2.1). The algorithm

essentially automates this visual inspection process. The intuition behind the algorithm is to call the IBD state in small chromosomal segments and then fill any low-information gaps and filter out small IBD segments. We investigated eight variations on this algorithm, considering different ways to define chromosomal segments, whether to use a sliding window, and whether to use a reference panel consisting of UN pairs for calling segmental IBD states. After careful comparison (see Results), the final algorithm defines chromosomal segments with a fixed number of SNPs, and neither uses a sliding window nor a reference panel (algorithm 5 in Table 2.2). Steps for inferring IBD segments for a unilineal pair of individuals are shown in Figure 2.2 and described as follows.

Step 1: Divide each of the 22 autosomes into chromosomal segments each containing 200 SNPs. Count the number of SNPs with IBS=0 and IBS=1 within each segment, denoted as n_0 and n_1.

Step 2: Compute the P-value for IBD=1 in each chromosomal segment by 1-B($X<n_0$; $p=0.0001$, $n=n_0+n_1$), where B(•) is the CDF of binomial distribution and $p=Pr(IBS=0|IBD=1)$ is the probability of IBS=0 given IBD=1 resulting from genotyping errors.

Step 3: Call IBD states in each chromosomal segment. Uncertain small gaps are then filled according to their flanking IBD status (Figure 2.2). For regions where SNPs are particularly sparse, such as centromeres, the IBD states are labeled as unknown.

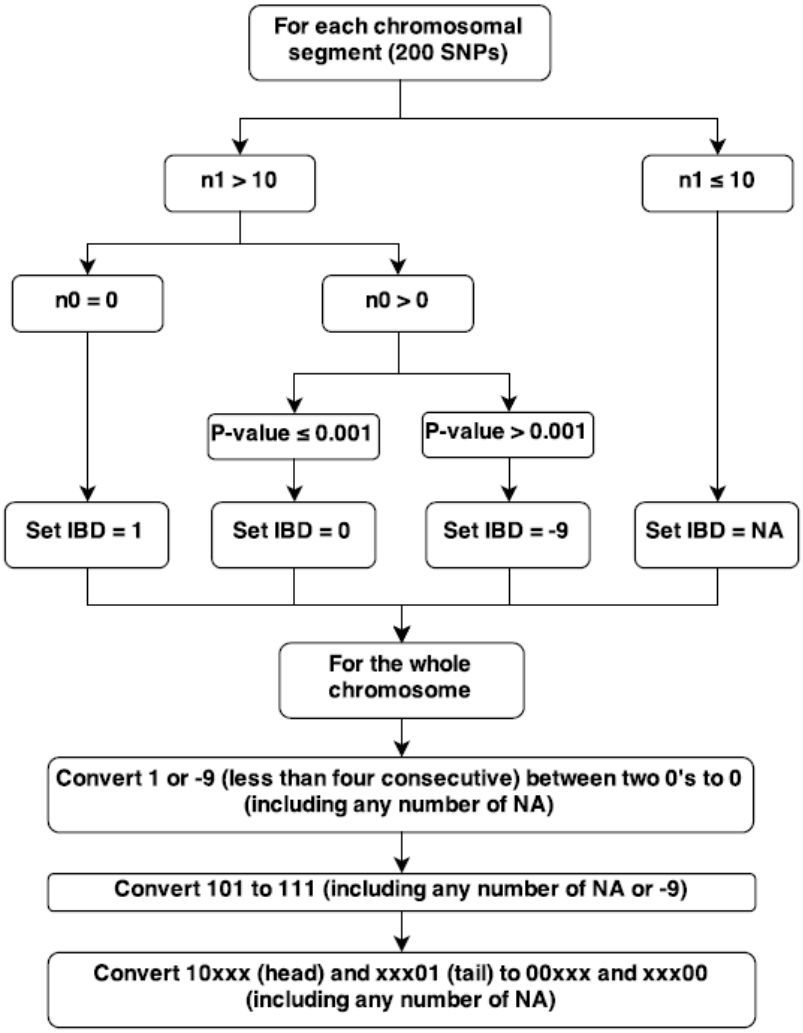


Figure 2.2 Algorithm flowchart

The algorithm can be described as two steps: first call IBD state within each chromosomal segment, and then fill the low information or uncertain gaps across the whole chromosome.

Our algorithm should work for SNP array data or sequence data as long as the data are genome-wide and markers are evenly distributed across genome. There are two tuning parameters in our method: the genotyping error parameter *p* and the number of SNPs in each

chromosomal segment. These can be chosen to accommodate the data features, such as SNP density and frequencies.

Distinct from most of the current IBD inference methods, our algorithm is not likelihood-based. It does not require SNPs to be independent, and does not need to estimate average IBD sharing in the study population or need the genotypes to be phased. The algorithm is designed to tackle the scenarios where IBD states are either 0 or 1 (i.e., unilineal relatives), but MZ and FS can easily be identified and separated in advance using conventional methods (e.g. PLINK (Purcell, Neale et al. 2007)).

2.3.3 Quantifying the accuracy of the algorithms by simulation and comparison

To choose the best algorithm from different combinations of strategies and evaluate the accuracy of the final algorithm, the eight proposed algorithms (Table 2.2) were evaluated in several ways.

Our first comparison was based on simulated data. We used the third generation Rutgers Combined Linkage-Physical Map of The Human Genome (Matise, Chen et al. 2007) to estimate the genetic position for each SNP. Assuming recombination is a Poisson process on chromosomes, we generated random variables from an exponential distribution with mean 100 as the distances between recombination events (i.e., length of IBD segments). Artificially synthesized IBD data were simulated by joining the IBD=0 and IBD=1 segments sampled from 30 UN pairs and 30 PO pairs randomly selected from the US dataset. Synthesized IBD data on chromosome 1 for 1,000 artificial pairs were simulated. Each of the eight proposed algorithms was used to infer IBD for the simulated data and the results were compared with the truth to estimate the false negative and false positive rates.

Besides simulation, we also evaluated the accuracy of algorithms by quantifying the concordance between duplicated GG pairs that share exactly the same IBD patterns but in opposite phase and by comparison with a recently developed recombination detection method based on known relationships (unpublished).

2.3.4 Calculating the observed recombination number

The observed recombination number (N) for a pair of individuals was defined as the total number of alternations between different IBD states across 22 autosomes after editing out the chromosomal regions with unknown IBD state.

2.3.5 Estimating IBD scores

Our classifiers were based on N and IBD scores. PLINK was used to estimate IBD scores k_0, k_1 and k_2 (probabilities of sharing zero, one and two IBD alleles) in the US sample. Due to the presence of population stratification and admixture in the Guatemalan sample, an ancestrally informative marker pruning technique (Morrison 2013) was applied to generate correct IBD estimates for the Guatemalan dataset.

2.3.6 SVM classification and cross-validation

A support vector machine (SVM) was used to build the classifiers. Unadjusted classifiers without putative pedigree information were based on k_0 and N only. The k_1 score was not considered as a feature because for unilineal relatives, k_1 and k_0 are collinear. To adjust for systematic difference

in N among populations due to different population background relatedness, we subtract from N the mean of UN pairs in each population, and set to 0 if it becomes negative after adjustment.

To incorporate putative relationship information, a feature-weighted SVM was adopted. Indicator variables were created to specify the relationship category to which each pair belongs (0=no; 1=yes). The number of indicator variables matched that of relationship categories. The indicators were then included as additional features together with k_0 and N in adjusted SVM models. Both k_0 and N were scaled in the adjusted classifiers, but not the indicators. Instead, a tuning parameter *s* was introduced to weight the indicators. Let x_i^T be the feature vector for data point *i* after scaling k_0 and N, x_i^{*T} be the weighted feature vector, $I_1, \ldots, I_n$ be the indicators, and n be the number of indicators

$$x_i^{*T} = x_i^T P$$

where $x_i^T = [k_0, N, I_1, \ldots, I_n]$, and $P = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & s & \\ & & & \ddots \\ & & & & s \end{bmatrix}_{(n+2)\times(n+2)}$

We used a radial basis kernel function in the SVM, with parameters selected using a grid search. Other kernel functions were explored but none of them achieved better performance. 1,000 iterations of 5-fold cross-validation were carried out for assessing model performance. The software "libsvm" (implemented in the R package "e1071") was used to realize the SVM models (Chang and Lin 2011).

2.4 RESULTS

2.4.1 Comparison of different strategies for the IBD detection algorithm

Table 2.2 shows metrics of accuracy and computational time for the eight combinations of algorithm strategies. Algorithm 5, which uses fixed number of SNPs to define chromosomal segments and does not use a sliding window or a reference panel, is among the best for all metrics and is faster than others. Thus, algorithm 5 was chosen as our final algorithm. We investigated algorithm 5's performance by examining all the IBD segments omitted or mistakenly identified in the simulation. 150 IBD segments were false negatives and 21 were false positives. We found all the false positive segments were in the same region and were from a single pair sampled repeatedly in the simulation. The distribution of genetic length of false negative segments indicates the omitted segments are usually quite short (Figure 2.3). This is natural because our algorithm filters out small uncertain regions. The filtering caused more false negatives than false positives and therefore introduced a small bias, which can be seen from the mean differences between inferred N and the truth in the simulation (Table 2.2). However, the bias is reasonably small. Also, it should be noted that the false positives were either due to genotyping errors within IBD=1 regions or due to population background relatedness in IBD=0 regions, which should be prevented aggressively, while false negatives are small IBD segments usually due to close double-recombination. In reality, the double-recombination interference results in fewer small IBD segments compared to the simulation with Poisson process, so our algorithm should have even less bias for real data.

Table 2.2 Comparison of different proposed algorithm strategies

Algorithm	Strategies			Discordance between 7 duplicated pairs		Discordance with a recent relationship-aware method on 30 pairs		Simulation results on chromosome 1 of 1,000 pairs (2,784 simulated recombination events)		Computational time for 53 pairs (in seconds)
	Partitioning chromosomes by	Call IBD with sliding window	Use of reference panel	l_1 norm of differences in N over pairs	l_1 norm of differences in N over chromosomes and pairs	l_1 norm of differences in N over pairs	l_1 norm of differences in N over chromosomes and pairs	l_1 norm of differences between N and the truth	Mean of truth minus mean of N	
1	physical distance	No	No	21	41	42	58	310	0.248	244
2	physical distance	No	Yes	14	40	46	62	455	0.309	337
3	physical distance	Yes	No	21	59	90	116	374	0.202	419
4	physical distance	Yes	Yes	22	62	72	104	452	0.148	610
5	number of SNPs	No	No	12	34	33	45	282	0.208	93
6	number of SNPs	No	Yes	13	45	38	52	417	0.353	182
7	number of SNPs	Yes	No	7	31	77	101	326	0.278	253
8	number of SNPs	Yes	Yes	25	69	106	122	454	0.380	446

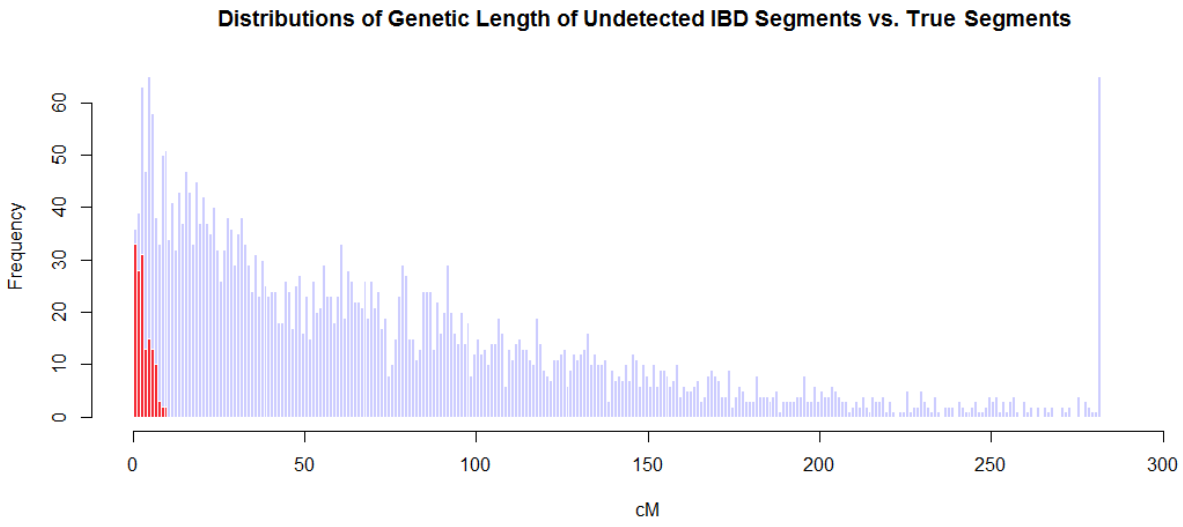


Figure 2.3 Genetic lengths of 150 false negative IBD segments (red) and 3,784 true segments (blue)

2.4.2 Classifying relationships using N and k_0

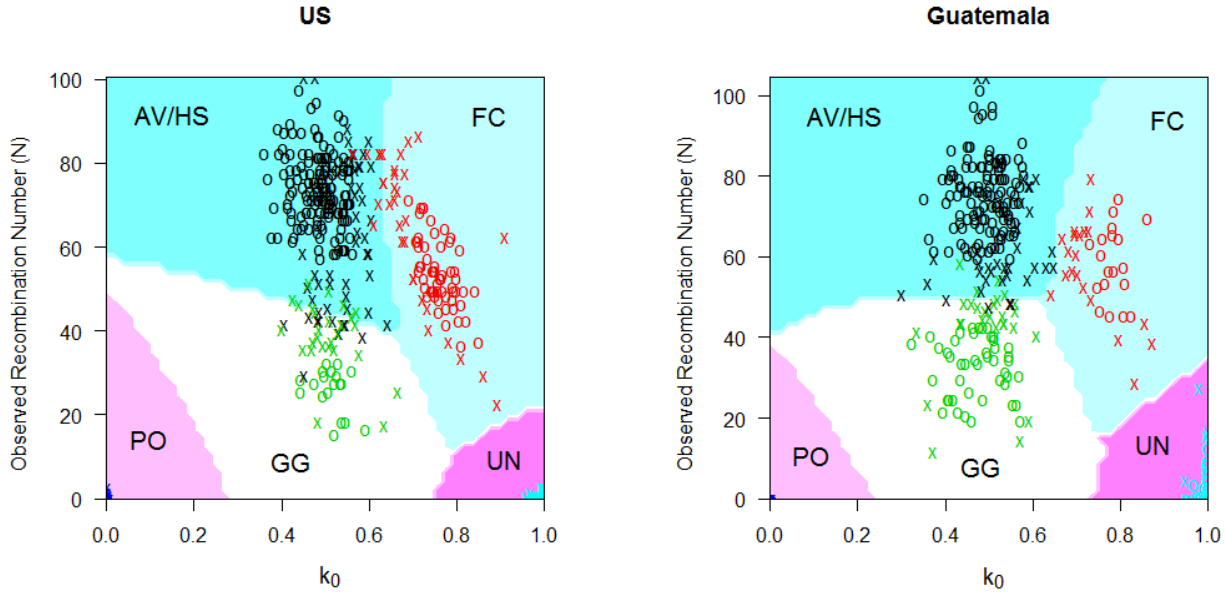


Figure 2.4 SVM classifiers with features N and k_0 for the US and Guatemalan populations

Colored areas illustrate different relationship categories. Training data are plotted with circles. Cross symbols indicate the support vectors. AV and HS are grouped as one category.

Table 2.3 Prediction accuracy (in percentage) and associated 95% confidence interval for the US and Guatemalan datasets based on 1,000 5-fold cross-validation

US

Predicted	True Relationship				
	AV/HS	FC	GG	PO	UN
AV/HS	93.8 (92.5, 95.0)	5.8 (4.4, 7.7)	15.5 (10.9, 21.7)	0	0
FC	0	94.2 (92.3, 95.6)	0	0	0
GG	6.2 (5.0, 7.5)	0	84.5 (78.3, 89.1)	0	0
PO	0	0	0	100 (100, 100)	0
UN	0	0	0	0	100 (100, 100)

Guatemala

Predicted	True Relationship				
	AV/HS	FC	GG	PO	UN
AV/HS	95.5 (94.4, 96.3)	0.4 (0, 2.6)	4.8 (2.8, 5.6)	0	0
FC	1.2 (0.6, 1.9)	99.6 (97.4, 100)	0	0	0
GG	3.3 (3.1, 3.8)	0	95.2 (94.4, 97.2)	0	0
PO	0	0	0	100 (100, 100)	0
UN	0	0	0	0	100 (100, 100)

Table 2.4 Results of cross-population prediction between the US and Guatemalan datasets

US predicts Guatemala

Predicted	True Relationship				
	AV/HS	FC	GG	PO	UN
AV/HS	157	1	14	0	0
FC	2	38	0	0	1
GG	1	0	58	0	0
PO	0	0	0	100	0
UN	0	0	0	0	99
Accuracy	98.1%	97.4%	80.6%	100%	99%

Guatemala predicts US

Predicted	True Relationship				
	AV/HS	FC	GG	PO	UN
AV/HS	182	14	1	0	0
FC	0	76	0	0	0
GG	18	0	45	0	0
PO	0	0	0	100	0
UN	0	1	0	0	100
Accuracy	91%	83.5%	97.8%	100%	100%

N and k_0 were used as two features to train the classifiers. Figure 2.4 shows the scatter plots of N and k_0 for the US and Guatemala training data and visualizes the two classifiers. AV and HS cannot be distinguished in most cases, so these two relationships were pooled together and treated as one category (see Discussion). Cross-validation results are shown in Table 2.3. The prediction accuracy was greater than 90% for all the relationship categories, except for GG (84.5% in the US sample). Cross-population prediction results, i.e., using the classifier built in one population to predict the training data in the other population, were also satisfactory, with the accuracy for all relationship categories better than 80% (Table 2.4). The adjustment for N with the mean of UN pairs is crucial. Essentially, the two populations have very different background relatedness. The Guatemalan sample has inflated N compared to the US sample. In other words, excessive shared IBD segments were observed between Guatemalan unrelated individuals, probably due to background distant relatedness in the population. Therefore, in practice, obtaining the mean N from a set of UN pairs to adjust for N is a useful extra step to enhance the robustness of our classifiers.

2.4.3 Incorporating putative relationships

The use of putative relationship information is a double-edged sword: when the information is correct, it improves the classification; otherwise, the classification may be misled and may give worse results. Therefore, how to weight the prior pedigree information is crucial. A reasonable value of the tuning parameter *s* should be selected to take advantage of correct information while retaining the ability to recover from misleading wrong prior information.

To assess the improvement in prediction accuracy, the relationship-indicator-adjusted classification results using correct relationship indicators were compared with the unadjusted

ones. Classification accuracy of each relationship category was estimated by 1,000 time 5-fold cross-validation. To assess the recovery rate for different types of misspecification in putative pedigree information, relationships were intentionally misspecified and the modified data were predicted with the adjusted classifier. Recovery is defined as a prediction escaping the misspecified relationship: the predicted category could be the true category or any other category, even an incorrect one. Whenever a prediction differs from its presumed one, it will be classified again using the unadjusted two-feature classifier without the putative relationship indicators. The rationale is that if the putative relationships are specified correctly, better classification accuracy will be achieved; if the putative relationships are wrong, there is a good chance to be recovered and reclassified by the unadjusted classifier.

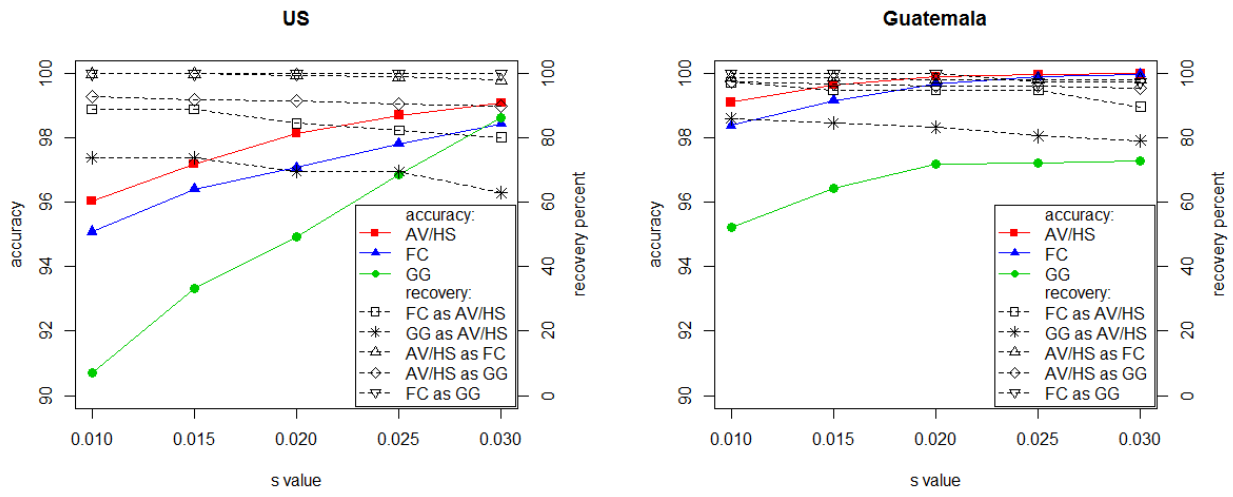


Figure 2.5 Classification accuracy and recovery percentage as functions of *s* value in the US and Guatemalan populations

The red, blue and green solid curves represent the classification accuracy based on 1,000 5-fold cross-validation for relationship categories AV/HS, FC and GG respectively, when including relationship indicators as features and the putative relationships are correct. Dashed curves represent the percentage of

pairs being recovered when the putative relationships are misspecified as shown. Classification accuracy for relationship categories not shown and recovery percentage of types of misspecification not shown are 100% across different *s* values.

Figure 2.5 shows the improvement of classification accuracy and the decrease of recovery rates as *s* value increases in the two samples. To balance the gain and loss, a value of 0.025 for *s* is recommended because the improvement is substantial (prediction accuracy > 96% for all relationship categories in both samples), while the recovery rates of all types of misspecification are above 80% except for GG being misspecified as AV/HS in the US samples. However, this type of misspecification is presumably rare in most cases.

In practice, different *s* values can be selected by users depending on how much they would like to trust the putative pedigree structures. If the prior information is not reliable, a smaller *s* is recommended so that the prior information contributes less to the prediction. In contrast, if an investigator has good reasons to trust the collected pedigree data, a larger *s* is proper and the prior information would be weighted more to enhance the prediction. However, in any case, we do not recommend using an *s* beyond the scope of 0.01 and 0.03.

2.4.4 Considering sex information of meiosis for GG

The GG category was divided into two subgroups, paternal-meiosis GG (GGp) and maternal-meiosis GG (GGm) by sex of the intervening parent. The training of SVM classifiers and the adjustment using putative relationships were the same as before. Better prediction accuracy was not observed in either dataset (data not shown).

2.4.5 Pipelines for classifying relationships with and without prior pedigree information

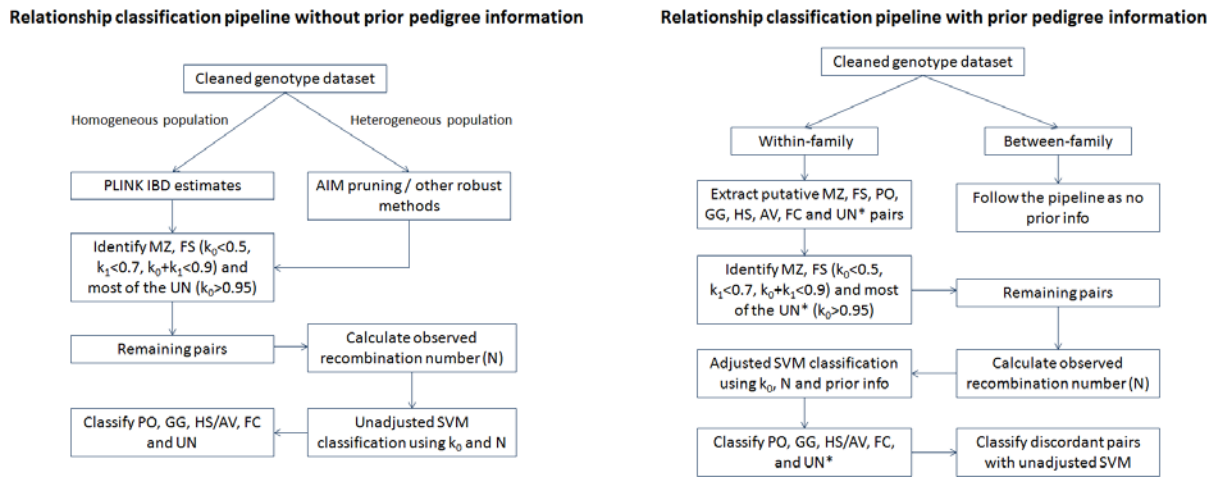


Figure 2.6 Classification pipelines with or without prior pedigree information

UN*: within-family founder pairs. AIM: ancestrally informative marker

Combining our methods with existing approaches, we propose general pipelines for relationship classification (Figure 2.6). In general, the pipelines can be summarized into four stages.

Stage 1: Clean genotype data and get the list of individual pairs for testing. When no prior pedigree information is available or detecting between-family relationships is of interest, all pair-wise relationships should be examined. When checking within-family putative relationships is of interest, all close relationships of the eight categories should be extracted. An R tool is provided for extracting putative relationships according to putative pedigree structures.

Stage 2: Calculate IBD estimates. Separate MZ, FS and most of the UN pairs according to k_0 and k_1. A conservative cutoff for UN is k_0>0.95. MZ or FS pairs can be arbitrarily defined as satisfying k_0<0.5, k_1<0.7, and k_0+k_1<0.9. IBD estimates can be generated by either PLINK (Purcell, Neale et al. 2007) for a homogeneous population or other robust methods (Manichaikul, Mychaleckyj et al. 2010; Morrison 2013) for a non-homogeneous population. The remaining pairs are left for SVM classification.

Stage 3: Obtain N for all the remaining pairs and adjust N with the mean estimated from a number of UN pairs. Our R package automatically treats all pairs with $k_0>0.95$ as UN to calculate the offset value.

Stage 4: Carry out the classification with the SVM classifier (using either one of the two provided or a user-defined population-matched classifier). One could adjust the classifiers using prior pedigree information. If putative relationships are used, those predictions disagreeing with corresponding putative relationships should be reclassified by unadjusted classifiers. Based on the final classification results, corrections can be made to the pedigree file.

2.5 DISCUSSION

We developed a new algorithm for IBD segment detection to utilize spatial information on genetic sharing between individuals to facilitate relationship classification. Our classification models can take advantage of putative relationships as prior information to enhance classification accuracy. Based on these new schemes, detailed pipelines for relationship classification are proposed, for checking within-pedigree putative relationships or detecting unknown relationships in population-based studies including many small families.

We demonstrated our methods with two real datasets, one from the US population and the other from the Guatemalan population. Systematic differences between the two populations were observed. Basically, inflated observed recombination number N's were observed for all relationship categories in the Guatemalan sample (Table 2.1). The inflation indicates the presence of shared chromosome segments in the population, presumably due to distant population background relatedness. In practice, if possible, we encourage investigators to collect user-defined training data and build population-specific classifiers. However, as long as a

number of UN pairs can be obtained and their mean is used to adjust N, the cross-population prediction accuracy is quite satisfactory, as seen in our US and Guatemalan samples.

We were unable to show classification improvements by considering sex of meiosis. Limited by training data sizes, we only attempted separate-sex classification for GG pairs. Even so, the training sample sizes of maternal-meiosis GG (GGm) and paternal-meiosis GG (GGp) were quite small. In theory, sexes of meiosis could be considered for several other relationships. Table 2.5 lists all the relationships that can be divided into subtypes by sexes of their pertinent relatives and be modeled in the same way as GG. Thus, more advanced classifiers could be built accordingly if there were enough data. So, despite our failure to show improvement, sex information still has potential to enhance the classification performance, and is worth further investigation in the future. It should be noted that sex information is also putative in practice, since it is obtained from putative pedigree information. Effects of sex misspecification should also be investigated.

Two issues regarding the classification accuracy should be noted. One is the composition of the data to be tested. In our classifiers, some categories contain subtypes, such as GG (comprises GGm and GGp) and AV/HS (comprise AV and HS), and the results are combined, ignoring subtypes. When the prediction accuracy differs among subtypes, composition of test data would influence the prediction accuracy of a category. For example, GGp is inherently classified better than GGm (since the N of GGm is closer to AV/HS). The more paternal GG in the test data, the better the classification results of GG will be. The other issue is the number of instances of each category in the training datasets. Since parameter selection is based on the overall prediction accuracy, small categories will automatically sacrifice for larger ones, i.e., the classifier will be trained to be more accurate for classes with higher frequency in the dataset.

Because we do not have balanced training data sizes for all relationship categories, our classifiers may have a preference for the categories with larger training data size. This issue can be solved easily when more training data are available. We suggest using confirmed relationships as additional training data to build better classifiers when possible.

Because HS and AV have the same expected k_0 and similar expected N (2λ for HS and 2.5λ for AV), our methods are not able to distinguish them. It has been shown with simulated segmental IBD sharing (Hill and White 2013) that if one simultaneously takes into account the likelihood on the observed numbers, positions and lengths of shared IBD segments, correct relationships for HS and AV could be assigned with a probability of 0.83. This provides an upper bound of classification accuracy under the assumptions that all these quantities are measured perfectly and their distributions are known. In reality, the measures are approximate and we do not know the true distributions, so HS and AV are difficult to distinguish in practice.

Our methods were implemented in R. In terms of computational efficiency, the most time-consuming step is calculating N. It took 441 seconds system time to compute N for the US sample (546 pairs) and 378 seconds for the Guatemalan sample (488 pairs) with two quad-core 2.93 GHz CPUs and 24 GB of memory. Basically, the computing time increases linearly with the number of pairs to be tested. Also, the time required to read in the genotype data is not trivial when the dataset is very large. Data size is proportional to both the number of individuals and the number of SNPs. For computational efficiency, we recommend eliminating irrelevant individuals from the genotype file in the data cleaning step and transforming the data to a better format before processing it with R. Example code is given for transforming the data with PLINK and shell commands and can be found in the documentation of our R package. In addition, we recommend removing most of the confidently unrelated pairs with a conservative k_0 score to

reduce the number of pairs to be tested as suggested in the pipelines (Figure 2.6). Our algorithm can be easily parallelized by both chromosomes and individual pairs to deal with extremely large datasets.

Table 2.5 Relationships for which the sex of pertinent relatives can be used to create subcategories

Relationship	Number of meioses involved	Number of meioses pertinent to expected N	Number of pertinent relatives	Description of pertinent relatives	Possible sexes of pertinent relatives	Relationship subcategories
GG	1	1	1	Parent of the grandchild relating to the grandparent	Male	Paternal GG
					Female	Maternal GG
HS	2	2	1	Common parent	Male	Paternal HS
					Female	Maternal HS
AV	5	1	1	Parent of the nephew/niece relating to the uncle/aunt	Male	Paternal AV
					Female	Maternal AV
FC	6	2	2	Two siblings as the parents relating the cousins	Male and male	Paternal FC
					Female and female	Maternal FC
					Male and female	Mixed FC

Our IBD transition detecting algorithm is developed for both whole-genome SNP array data and sequence data. However, it is unclear whether it will work for whole-exome sequence data, which lie in between whole-genome and targeted sequencing. We will need to examine the algorithm performance on such data.

Our relationship classification pipelines focus on generating accurate pair-wise relationships. It is also important to reconstruct the pedigrees with individual relationship pairs. Of note, some relationships may conflict with each other during pedigree reconstruction, which implies classification errors. It might be of interest to consider modeling relationship

classification and pedigree construction together so that such errors can be avoided while further improving the relationship classification accuracy. A recent pedigree constructing tool has made use of such a notion but it treated all second-degree relationships as one category (Staples, Qiao et al. 2014). By dividing second-degree relationships with our methods, more accurate pedigrees might be reconstructed.

We implemented the putative relationship extraction tool, IBD transition detecting algorithm, and relationship classifiers in an R package (available through http://relcla.sourceforge.net/).

3.0 APPLYING RELATIONSHIP CLASSIFIERS TO WHOLE EXOME SEQUENCING DATA

3.1 MOTIVATION

Our proposed IBD segment detection algorithm works for dense SNP datasets as long as the data are genome-wide and the markers are relatively evenly distributed across the genome. In addition to SNP array data, whole genome deep sequencing data also can be handled by our algorithm. However, due to cost, most sequencing studies are currently whole exome rather than whole genome. We therefore want to assess the extensibility of our algorithm for whole exome sequencing data.

A few practical issues should be noted when applying our algorithm to whole exome sequencing data. In general, exons are not equally spaced in the genome. Also, while arrays are designed to include a fixed set of SNPs and all individuals have the same set of SNPs being genotyped except for sporadic missingness, sequencing usually produces much more missing data even when the coverage is quite deep. Lastly, the genotyping accuracy of sequencing data is one or two magnitudes lower than that of array data.

To evaluate the actual performance, we tested our algorithm on real data. We investigated the distribution of markers for whole exome sequencing data. Different missingness filters and minor allele frequency filters were used to explore their impact on the signal/noise ratio.

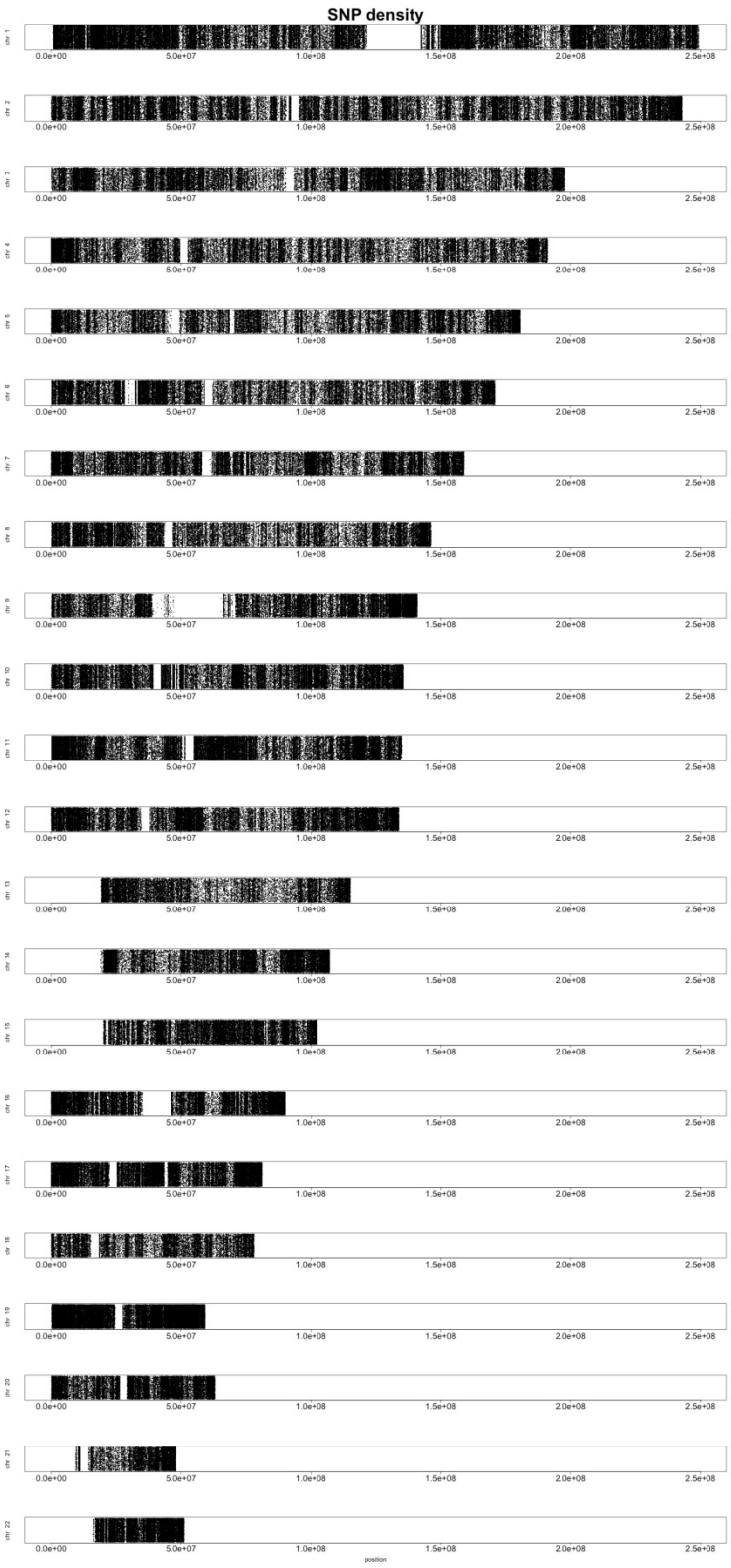
3.2 DATASET

A family dataset of age-related macular degeneration was provided by the National Eye Institute. Whole exome sequence data were generated on Illumina HiSeq 2500 platform for 82 subjects. The average sequencing depth was more than 50X, which means the expected number of reads are more than 50 for any given locus across the exome. Genotype calling was performed with GATK (McKenna, Hanna et al. 2010). The majority of the study subjects are of European ancestry. 18 subjects were removed (4 due to inbreeding, 7 due to non-European ancestry and 7 because of unknown ancestry or unknown pedigree information), resulting a total of 64 individuals belonging to 16 families in the final analysis.

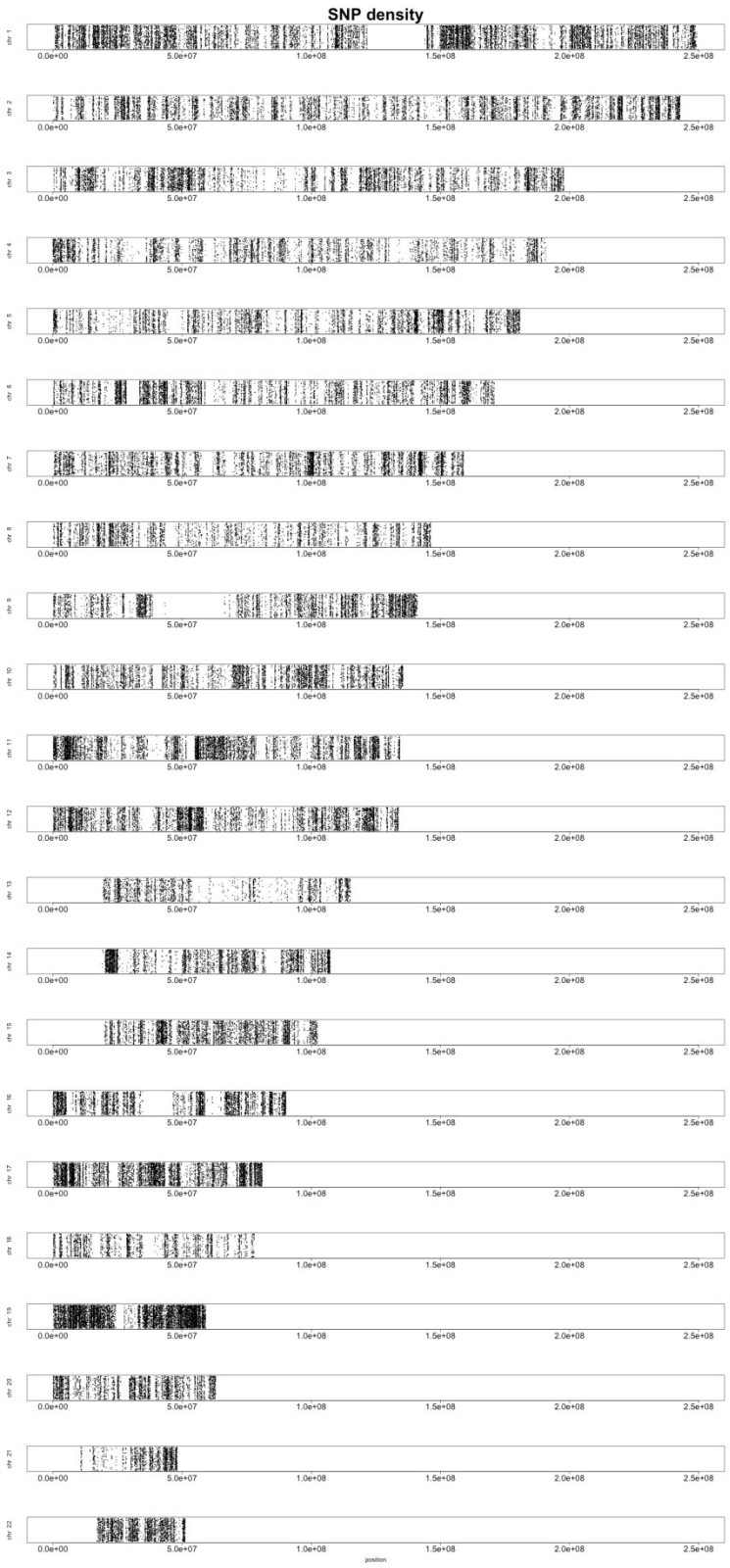
3.3 RESULTS

3.3.1 Distribution of SNPs

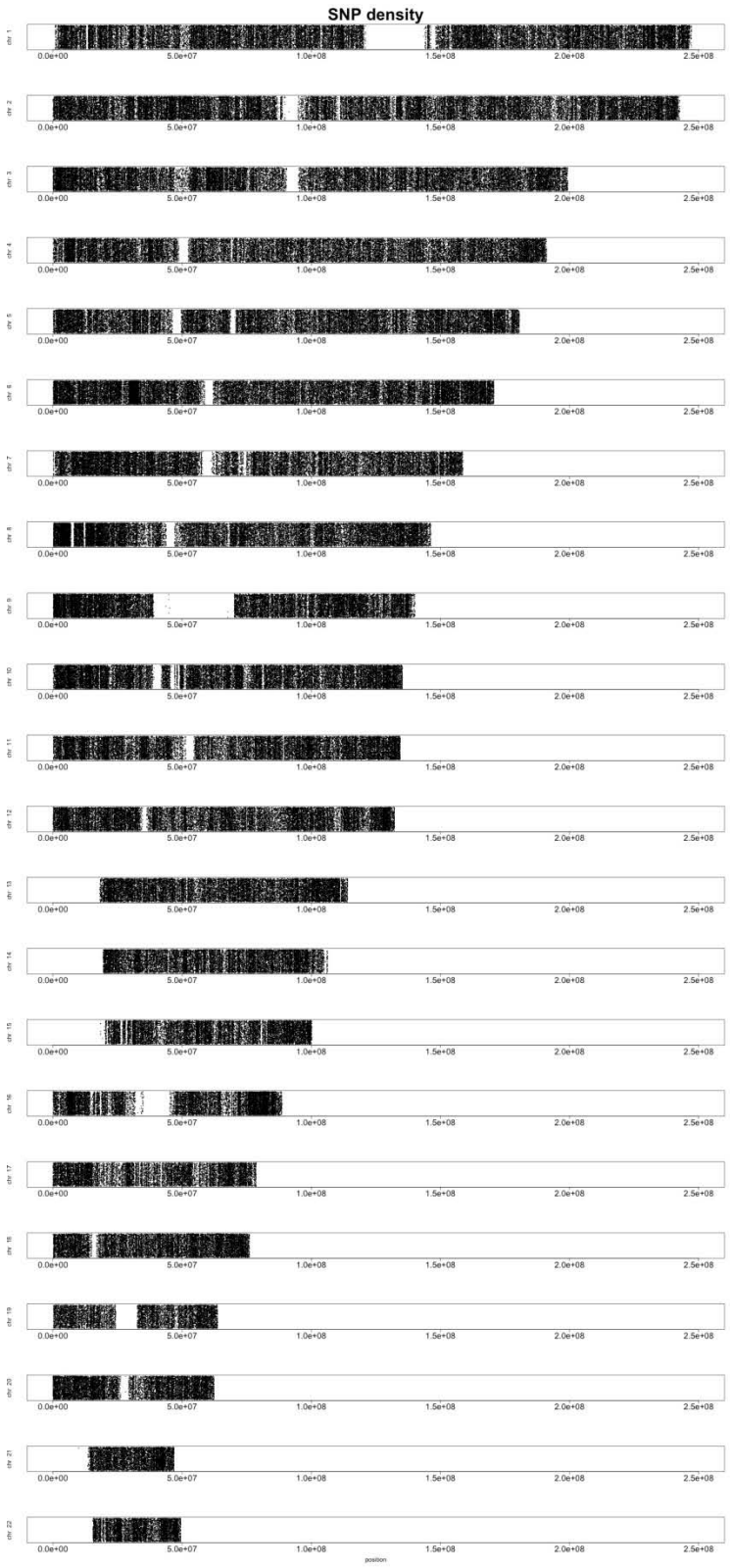
Figure 3.1a illustrates the density of SNPs across the genome for this whole exome sequencing dataset before applying any filtering. We can see that although the density is not uniform, most chromosomal regions are covered. However, after applying SNP genotype non-missing rate (Nmiss) and SNP minor allele frequency (MAF) filters to the data, SNPs become too sparse to cover all the areas (Figure 3.1b). In contrast, the SNP density for array data is fairly uniform across all the chromosomal regions (Figure 3.1c).



a. Whole exome sequencing data without filtering



b. Whole exome sequencing data with filters (Nmiss=1.0 and MAF>0.01)



c. SNP array data (the US dataset from Chapter 2)

Figure 3.1 Density of SNPs across the 22 autosomes

3.3.2 IBD score estimates

Since the sample is homogeneous, the pair-wise IBD score estimates for the 64 individuals were generated using PLINK (Purcell, Neale et al. 2007). Different combinations of SNP filters were tried to obtain the best IBD estimates in terms of the clear separation of different relationship degrees in the IBD plot, and Nmiss=1.0 and MAF>0.01 were selected. According to the IBD plot (Figure 3.2), FS (k_0<0.5, k_1<0.7, and k_0+k_1<0.9) and MZ (k_0<0.1 and k_1<0.1) relationships can be easily identified, as well as most of the UN pairs (k_0>0.95).

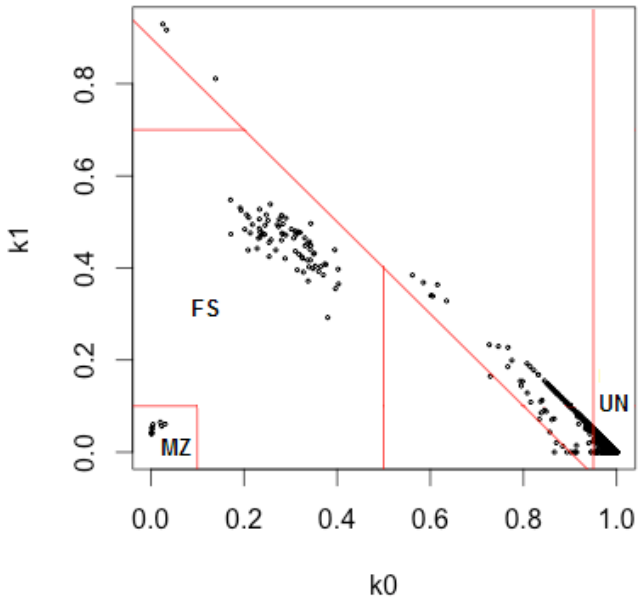
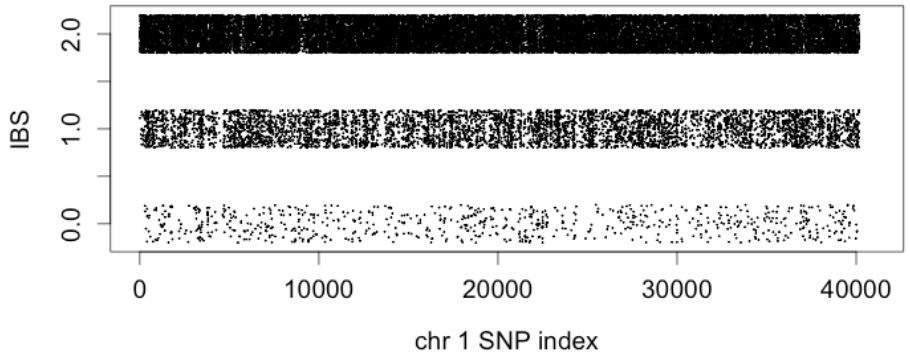


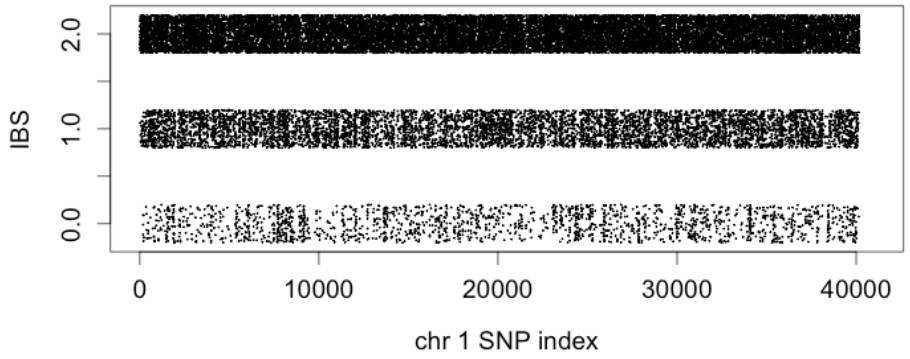
Figure 3.2 IBD plot for the 64 individuals

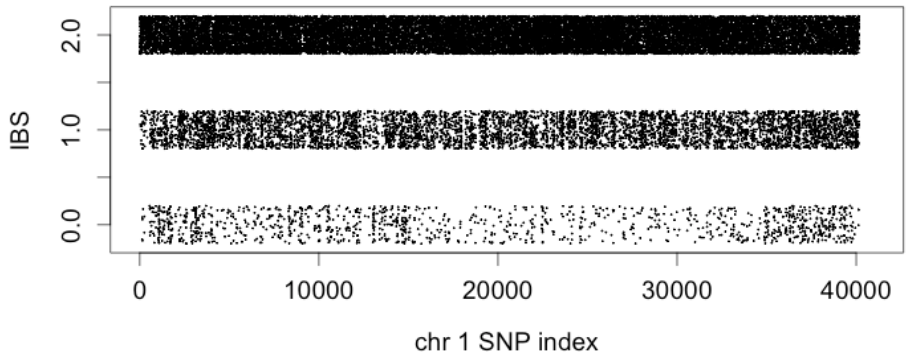
k_0: P(IBD=0), k_1: P(IBD=1). Red lines show the boundaries defining FS, MZ, and most of UN.

3.3.3 The effect of SNP/SNV filters on the signal/noise ratio

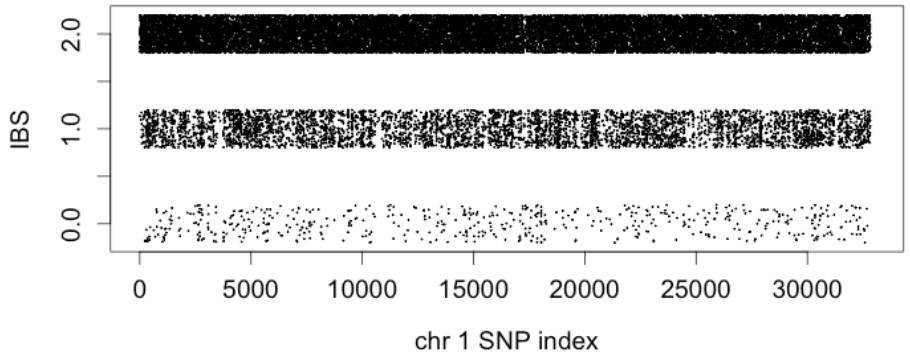
We again explored the effect of different combinations of non-missing rate filters and minor allele frequency filters on the genotype error rate, the number of informative SNPs, and the signal/noise ratio. The results are shown in Figure 3.3 and Table 3.1. For PO pairs in non-inbred families, IBD statuses are always 1 for all SNPs, so IBS can only be 1 or 2. Therefore, SNPs with IBS equal 0 indicate genotyping errors. We can use the ratio between SNPs with IBS 0 and SNPs with IBS 1 to measure this error rate, which corresponds to the noise. On the other hand, for UN pairs, IBD statuses are always 0, and the ratio between SNPs with IBS 0 and SNPs with IBS 1 can be deemed as the signal for IBD=0 status. The higher the ratio is, the greater the signal is. In this way, we defined the signal/noise ratio for IBD=0 status as a metric to select the best SNP filters, together with the number of informative SNPs. We hope to achieve larger signal/noise ratios so that segments with IBD=0 status can be distinguished more easily from segments IBD=1 status. Figure 3.3 illustrates the signal and noise for different SNP filters using the IBS plots of chromosome 1 for examples of different relationship pairs: PO, UN, and AV.

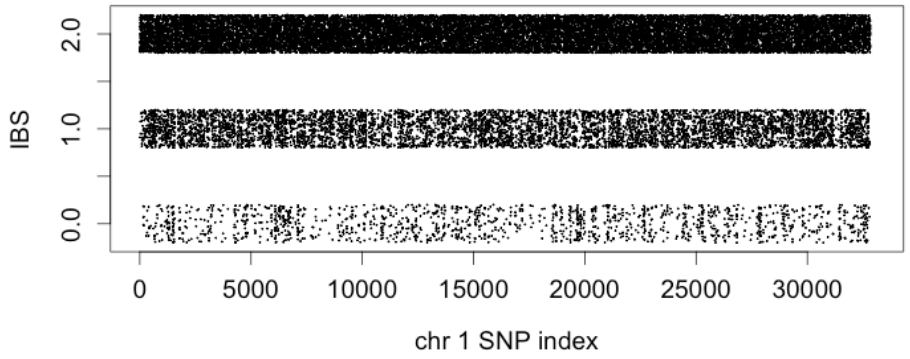


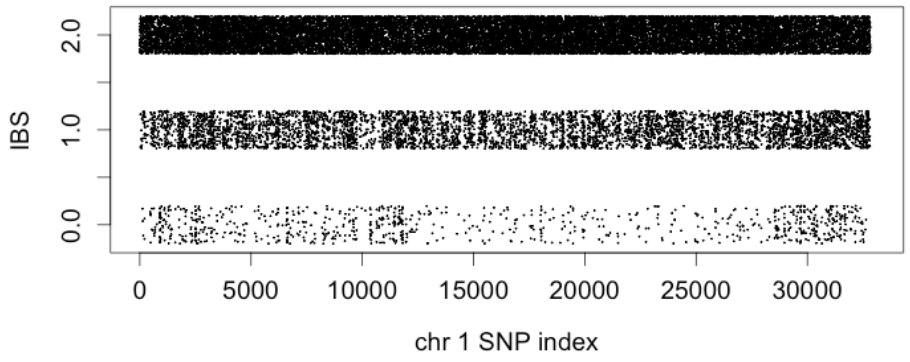




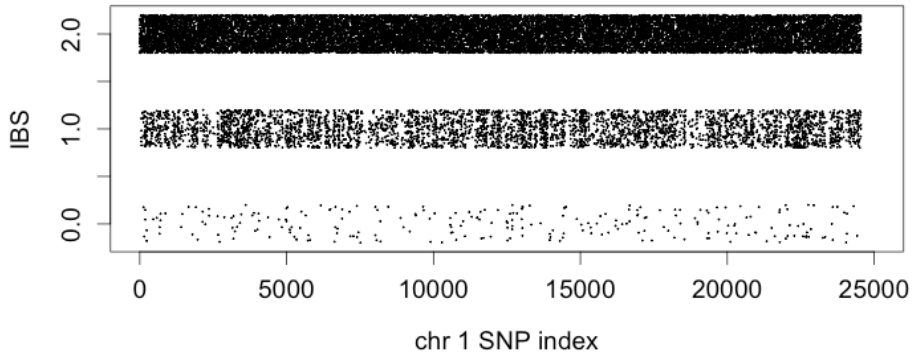
a. Nmiss>0.5 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)

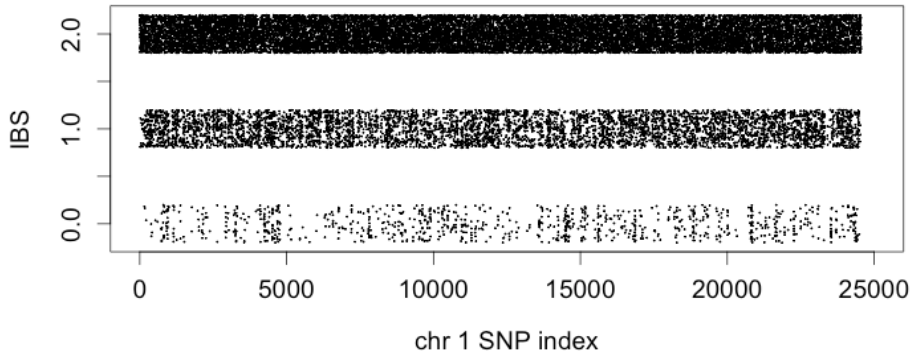


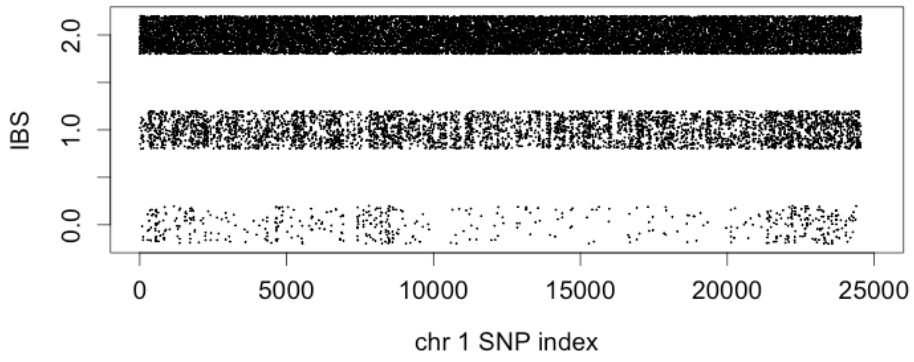




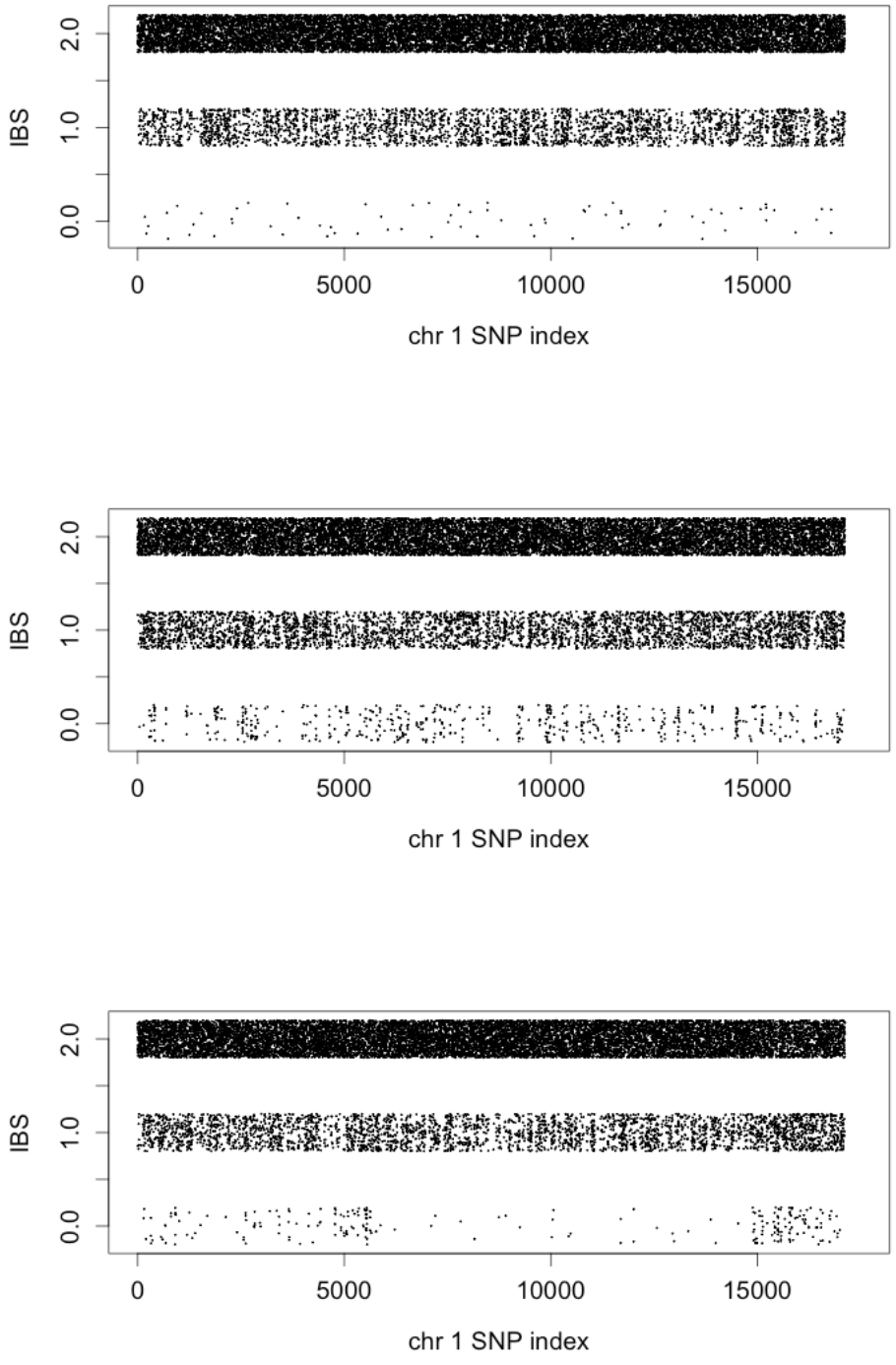
b. Nmiss>0.7 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)



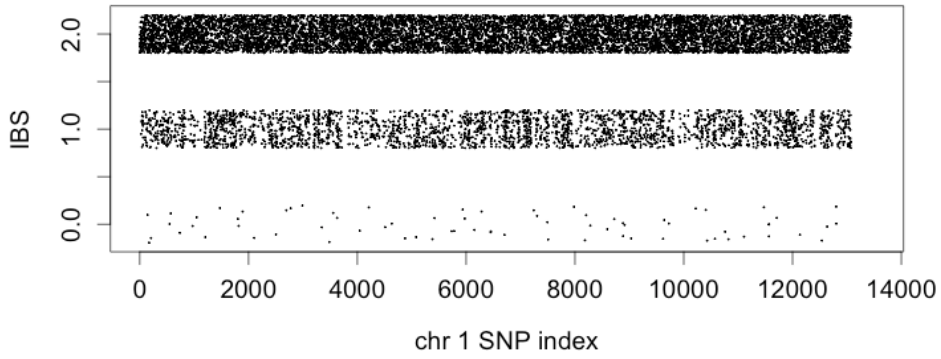


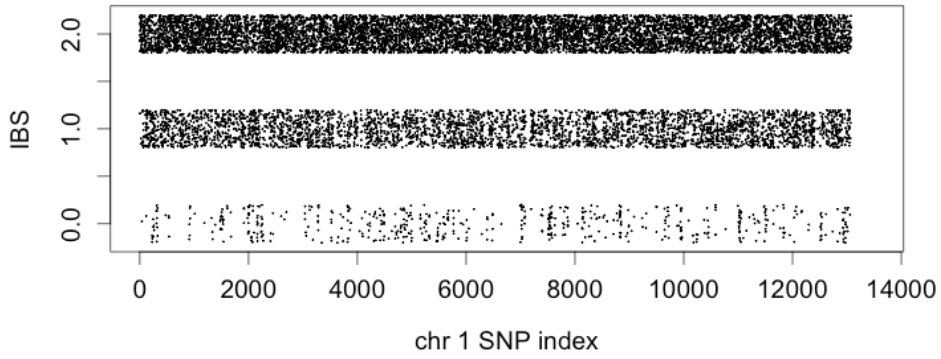


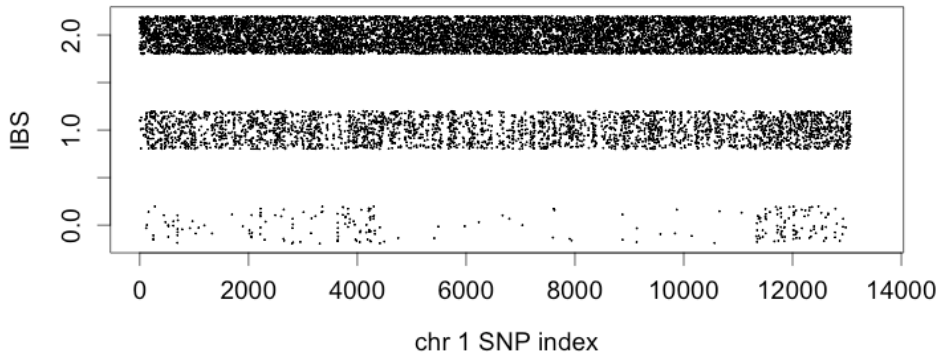
c. Nmiss>0.9 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)



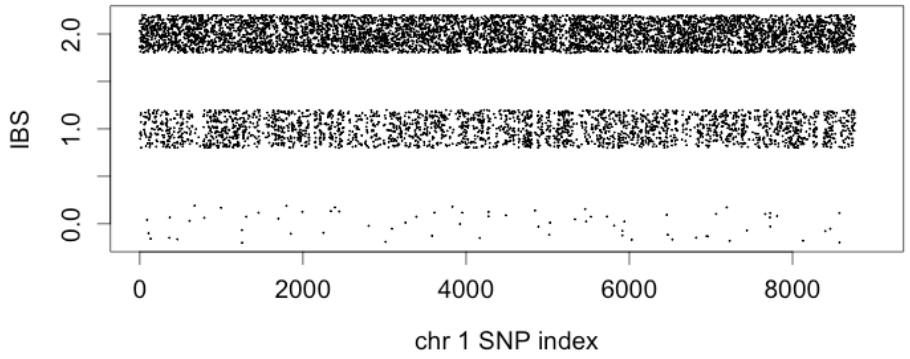
d. Nmiss=1.0 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)

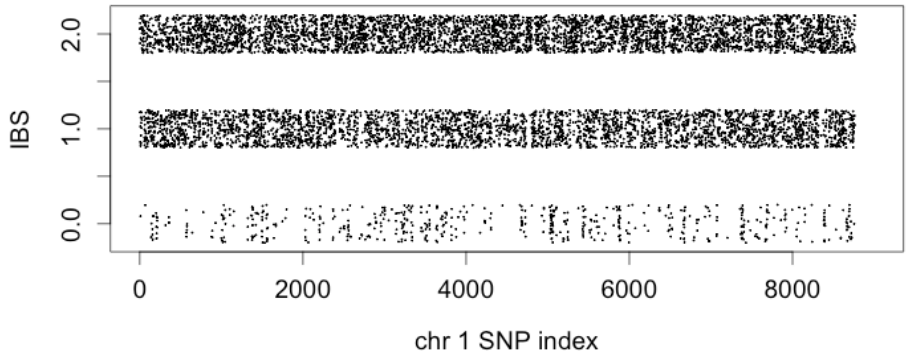


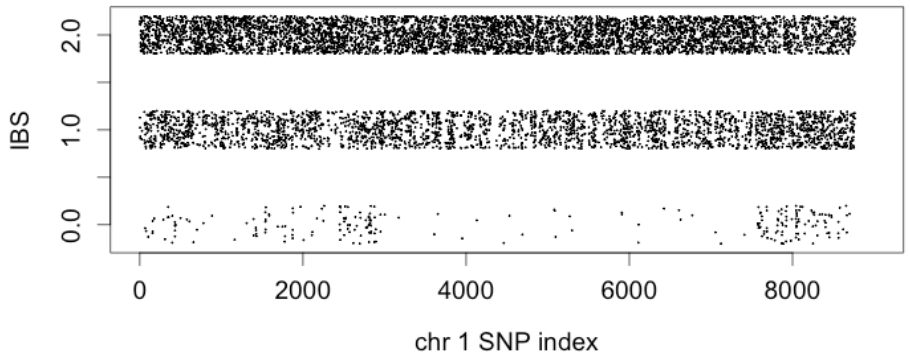




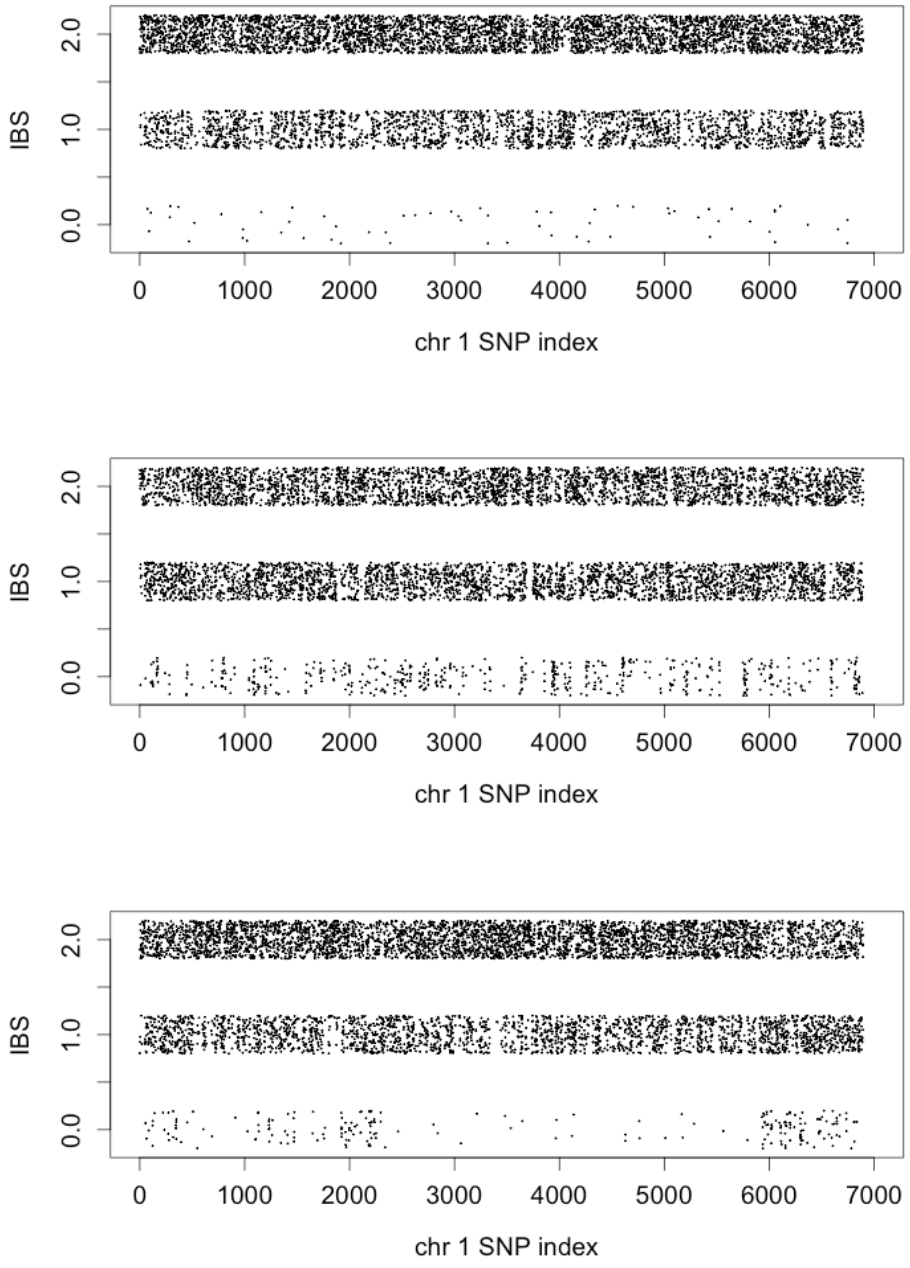
e. Nmiss=1.0 and MAF>0.01 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)







f. Nmiss=1.0 and MAF>0.05 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)



g. Nmiss=1.0 and MAF>0.1 (**upper**: PO pair; **middle**: UN pair; **bottom**: AV pair)

Figure 3.3 IBS plots of chromosome 1 for PO, UN, and AV pairs after applying different SNP filters

SNPs in the plots were sorted by position. IBS were plotted against SNP index rather than physical position.

Table 3.1 Number of informative SNPs and signal/noise ratio for different SNP filters

Filters	Total Number of SNPs	UN pair			PO pair			Average Signal/Noise Ratio
		Number of SNPs with IBS=1	Number of SNPs with IBS=0	IBS=0 /IBS=1	Number of SNPs with IBS=1	Number of SNPs with IBS=0	IBS=0 /IBS=1	
Nmiss>0.5	422,089	69,956	19,955	0.285	56,229	8,083	0.144	1.984
Nmiss>0.7	340,159	66,530	15,673	0.236	53,526	5,830	0.109	2.163
Nmiss>0.9	246,706	56,399	10,046	0.178	43,026	2,759	0.064	2.778
Nmiss=1.0	168,472	39,166	5,572	0.142	30,461	736	0.024	5.888
Nmiss=1.0 & MAF>0.01	129,362	38,684	5,572	0.144	29,717	736	0.025	5.816
Nmiss=1.0 & MAF>0.05	87,312	35,177	5,526	0.157	27,397	690	0.025	6.237
Nmiss=1.0 & MAF>0.1	68,929	30,684	5,367	0.175	24,910	633	0.025	6.883

From the IBS plots and the fraction IBS=0/IBS=1 for PO pairs we can see that the genotype errors are much more frequent compared with those of array data, for which the fraction IBS=0/IBS=1 is less than 0.001. While missing rate is strongly associated with genotyping quality, minor allele frequency does not affect genotyping quality and signal/noise ratio dramatically.

The best IBS=0/IBS=1 for PO pairs was achieved using the most stringent missingness filter, i.e., excluding SNPs with any missing values. On the other hand, excluding rare variants can result in better signal/noise ratio, but we also lose more informative markers. To balance both number of informative SNPs and signal/noise ratio, we used Nmiss=1.0 as the only filter.

3.3.4 Relationship classification results

The fact that the noise for exome sequencing data is much higher than that for array data makes it harder to determine IBD status transitions. As a result, the IBD segments are no longer easily

determined by visual inspection, as illustrated in the IBS plots for an AV pair (Figure 3.3). By modifying parameters in our IBD segment detection algorithm (setting the genotyping error parameter *p* at 0.01 and the number of SNPs per segment at 300), we can still run the algorithm, but the observed recombination number N is severely underestimated compared to the estimates from SNP array data. In theory, this underestimation is not necessarily a problem, since the goal of the procedure is not to estimate the recombination but rather to distinguish relative types.

We classified the pair-wise relationships for 270 pairs with k_0 estimate and N using the US training data. The results are shown in Table 3.2. Based on these results, the relationship classification accuracy is not satisfactory, so our relationship classifiers built for SNP array data cannot be readily used for whole exome sequencing data. The unsatisfactory relationship calls are primarily due to the underestimated N. There are two major reasons for the underestimated N. The first is the high level of noise due to more genotype errors making our algorithm "insensitive" to IBD status transitions. The second is the unevenly distributed SNPs. Exons are clustered in some chromosomal regions and are too few in some other regions. After applying filters, SNPs become too sparse to provide accurate IBD segment information for some regions, making it impossible to detect any IBD status transitions within those areas.

Table 3.2 Relationship classification results for 270 pairs in the whole exome sequencing dataset

Relationship Category	Mean of N in Training Data (SNP Array)	Mean of N* in Testing Data (Exome Sequencing)	Number of True Pairs	Predicted
PO	0.4	1.9	3	3 PO
AV/HS	72.3	28.0	5	5 GG
FC	59.6	20.9	30	20 FC + 6 UN + 3 GG + 1 AV/HS
UN	1.5	1.9	232	232 UN

*Adjusted by the mean of confident UN pairs

3.4 DISCUSSION

In general, based on our exploratory test results, we conclude that the relationship classifiers based on array data cannot be directly used for whole exome sequencing data. This does not depend on the sequencing coverage, considering the sequencing coverage for the test dataset is more than 50X, but is due to issues embedded in exome sequencing data. However, it does not mean our methods and pipeline are totally useless. If we have more data to build specific relationship classifiers for whole exome sequencing data, we may alleviate the effect of the bias in N estimates. It is also worthwhile to try better genotype calling methods to enhance genotype quality. The currently genotypes are called individually, if we called samples jointly using the information on linkage disequilibrium, missingness can be reduced and genotype quality may be improved.

Interestingly, it has been shown recently that analyzing multiple samples together considering their family information can further improve the genotype call accuracy in sequencing studies (Chen, Li et al. 2013). However, actual relationships must be validated with high quality genotype data. This is a paradox in practice. Therefore, it is important for investigators to decide which information is more reliable and how it might be used to improve the information on the other side.

4.0 PEDIGREE RECONSTRUCTION WITH PAIR-WISE RELATIONSHIP INFERENCES

4.1 MOTIVATION

Pair-wise relationships do not provide complete pedigree information, but they provide a basis for reconstructing more complete pedigrees. During the process of pedigree reconstruction, some relationship pairs may be found that are contradictory with others, implying incorrect pair-wise inferences. By reconstructing pedigrees, those contradictory relationships may be corrected, which in turn improves the accuracy of pair-wise relationship inference. Also, reconstructing pedigrees is necessary for various pedigree-based analyses, such as linkage analysis, family-based association studies, checking Mendelian errors, and estimating heritability. So, our ultimate goal is to develop a tool to reconstruct pedigrees with the pair-wise relationship predictions, similar to PRIMUS (Staples, Qiao et al. 2014).

The major difference between our method and PRIMUS is that we focus on reconstructing many small pedigrees fast and accurately. Therefore, all the relationships that we consider are close ones. Also, we use finer relationship categories as the building blocks: the separation of GG from AV/HS may let our method achieve better performance.

Since we focus on either identifying very close relationships in population-based studies or correcting pedigrees in family-based studies with many small pedigrees, we assume the study

sample only contains families with three generations at most. Also, our method allows half-siblings, but we assume the multiple partners do not occur within the first generation (the oldest). In addition, we assume there are no recent inbred relationships or bilineal relationships other than MZ and FS in the families. As a result, there are seven relationship categories under consideration: MZ, FS, GG, AV/HS, PO, FC, and UN, which are all within the scope of our pair-wise relationship classifiers.

When all the connecting individuals in a pedigree are genotyped, the pedigree can easily be reconstructed using only the first-degree relationships, i.e., all MZ, FS and PO relationships. Because the inferences for these relationships are very accurate, the information from other pair-wise inferences can be simply ignored. However, problems arise when some connecting individuals in pedigrees are missing, in which case pedigree structures cannot be determined only with those confident relationships. We must then take into account other pair-wise relationships that are not as certain. Sometimes, pedigrees can be found to fit in all the inferred pair-wise relationships, but sometimes conflicts exist among some inferred relationships and there is no pedigree consistent with all the pair-wise relationship inferences.

Ideally, we want to identify the most likely pedigrees among all the possibilities. A way to construct the likelihood of pedigrees given the predicted (observed) pair-wise relationships is described as follows. The likelihood of pedigrees given the observed pair-wise relationships can be expressed as a product of a series of conditional probabilities. Assuming the probabilities of all possible pedigrees are equal, we have:

$$\Pr\{\text{pedigree}|\text{observed relationships}\} = \frac{\Pr\{\text{observed relationships}|\text{pedigree}\} * \Pr\{\text{pedigree}\}}{\Pr\{\text{observed relationships}\}}$$

$$\propto \Pr\{\text{observed relationships}|\text{pedigree}\} = \Pr\{\text{observed relationships}|\text{true relationships}\}(*)$$

$$= \prod_i \Pr\{\text{observed relationship}_i|\text{true relationships}\}\ (**)$$

$$= \prod_i \Pr\{\text{observed relationship}_i|\text{true relationship}_i\}$$

(*) Given a pedigree is equal to given all pair-wise relationships, since they provide the same information.

(**) Given all the true relationships, each observation of relationship is mutually independent.

Note that the "observed relationships" are not "putative relationships from putative pedigree information", but are "the relationships predicted by the classifier".

Therefore, maximizing $\Pr\{\text{pedigree}|\text{observed relationships}\}$ is the same as maximizing $\prod_i \Pr\{\text{observed relationship}_i|\text{true relationship}_i\}$ over all possible pedigrees.

After building a relationship classifier, we can estimate the classification accuracy for each relationship category and the error rate for each type of misclassification using the training data by cross-validation. These quantities can be interpreted as the probabilities of each relationship inference given the truth. Therefore, with these estimates we can directly calculate $\prod_i \Pr\{\text{observed relationship}_i|\text{true relationship}_i\}$ for any given pedigree.

Then, the problem becomes to enumerate all the possible underlying pedigrees and search for the pedigrees with maximum likelihood. However, enumerating all possible underlying pedigrees is NP-hard. More importantly, no simple algorithm is available for automating the enumeration. Although the search space may be reduced by the aforementioned assumptions and some relationship inferences are accurate enough to rule out certain kinds of misclassification, the problem still cannot be easily solved.

We therefore developed a compromise method to reconstruct “core pedigrees” that are extremely accurate and capture as many pair-wise relationships as possible with the assistance of individual age and sex information as well as constraints imposed by the assumptions. Improved relationship inferences for all relative pairs are provided as a part of our algorithm. Although some conflicts may still exist, all the pair-wise relationships captured by the core pedigrees are consistent. These “core pedigrees” can then be used as an initial point to manually reconstruct full pedigrees combining other sources of information.

4.2 METHODS

4.2.1 Identify families

Since we start with pair-wise relationship inferences, when putative pedigree information is not available, the first task is to identify individuals from the same family. Basically, all pair-wise relationships can be treated as an undirected graph with individuals being the vertices and relationships between relatives being the edges. Here we only focus on searching for individuals from the same families rather than the actual pedigree structure, so there is no need to consider the directions of relationships, e.g., who is the parent and who is the child in a PO pair. Then the problem is equivalent to finding out all the tree structures within the huge undirected graph, since individuals in each tree belong to a family. An algorithm for this purpose was developed using the idea of breadth-first search (BFS). Our algorithm starts with randomly picking an individual, and then searches for all the relatives of this individual. If there are any relatives, it keeps searching for the relatives of the relatives from the rest of the graph until no new relatives can be

found, i.e., the current tree is complete. When a tree is complete, it stores all its individuals, deletes the corresponding vertices from the graph, and starts over again until all the tree structures are identified in the graph. An important assumption of our method is that all the inferences of UN are correct, i.e., both specificity and sensitivity are 100%. This is the case in our training datasets, but may be unrealistic in reality. The time complexity of our algorithm is between O(n) and O(n^2), where n is the number of individuals.

4.2.2 Steps for reconstructing core pedigree for each family

After isolating all families, we collect the pair-wise relationships except for UN among individuals within each family. We use the following steps to reconstruct the core pedigree for each family.

Step 1: Build an initial pedigree with confident pair-wise relationships MZ, FS, and PO (if no such relationships exist, no pedigree will be constructed). With age and sex information of pertinent individuals, a unique pedigree can be determined, i.e., the parents and children can be figured out for PO pairs using ages, and fathers and mothers can be determined using sexes. Dummy individuals will be added when necessary. If any conflicts occur among relationships at this step, manual inspections are recommended.

Step 2: Extract other relative pairs implied by the initial pedigree, i.e., GG, AV, HS, and FC. If any conflicts occur between the extracted relative pairs and the previously inferred pair-wise relationships, treat the extracted ones as correct, because the extracted ones are based on confident relationships.

Step 3: Expand the initial pedigree based on a couple of constraints. One constraint is that if an individual is confirmed to have a grandchild in the initial pedigree, any pair-wise relationship inferences of GG, AV/HS, and FC involving this individual must be actually GG, otherwise the assumptions about

pedigrees must be violated. Also, the relationships among grandchildren of a grandparent must be FS, HS or FC, and any GG's should be actually HS, since FS and FC cannot be misspecified as GG but HS may (table 4.1). Then, based on the relationships among grandchildren, dummy intermediate individuals are added to the initial pedigree when necessary to result in consistent pertinent relative pairs (Figure 4.1).

Step 4: Repeat step 2. For the relationships not implied by the expanded pedigree, just report the previous inferred pair-wise relationships. Note that conflicts may exist among these relationships.

Step 5: Output the expanded pedigree as the final one (which is what we call core pedigree) and all the updated pair-wise relationships.

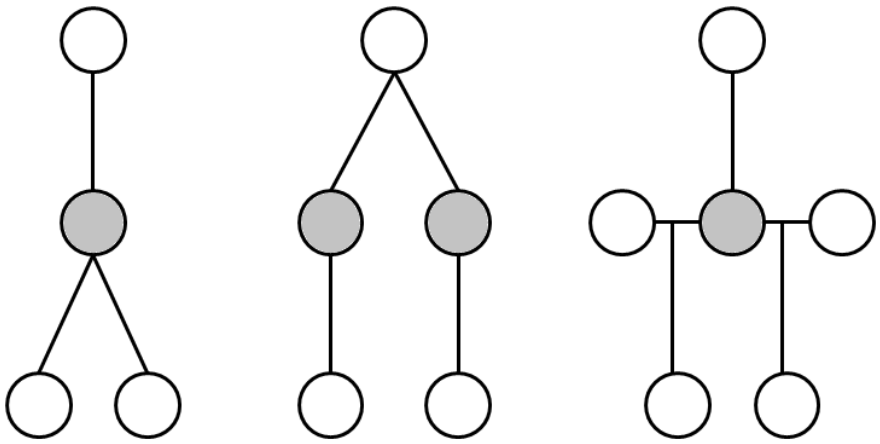


Figure 4.1 Three cases of adding dummy intermediate individuals between a grandparent and multiple grandchildren

Left: when two grandchildren are FS; middle: when two grandchildren are FC; right: when two grandchildren are HS. Shaded circles indicate added dummy individuals. Sexes are not differentiated here.

4.2.3 An example of reconstructing core pedigree

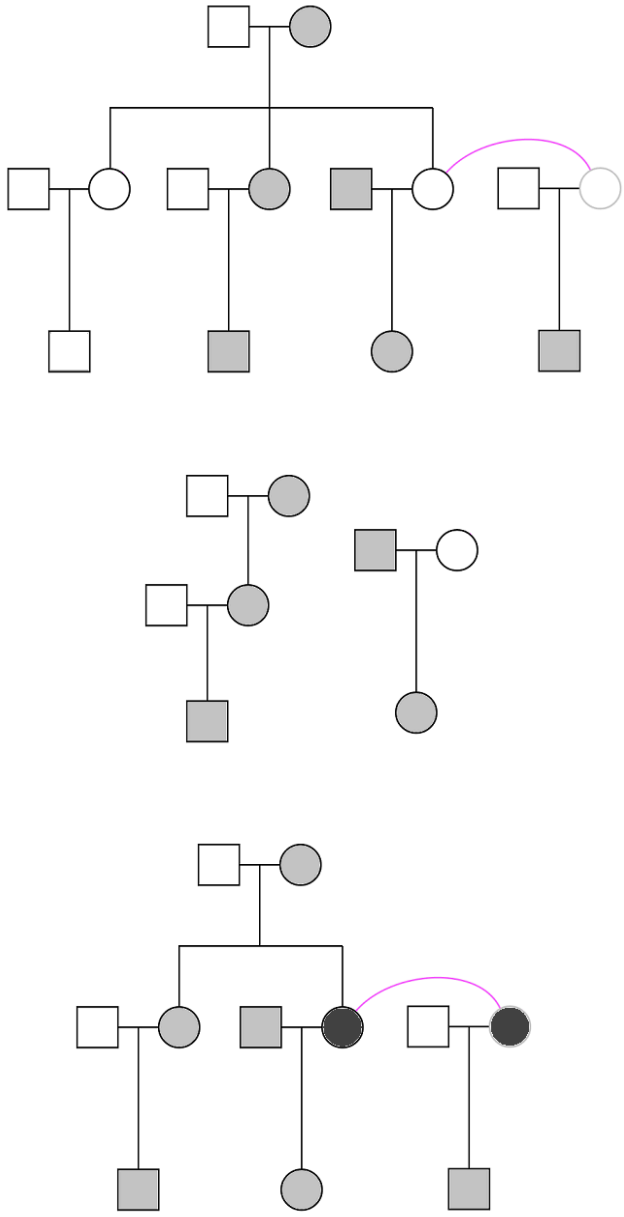


Figure 4.2 An example of key steps in core pedigree reconstruction

Upper: the underlying full pedigree. Only shaded individuals are observed; **Middle**: the initial pedigree built upon PO pairs; **Bottom**: the final core pedigree after adding the dummy intermediate individual (filled) based on the relationships among grandchildren.

Figure 4.2 illustrates how a core pedigree is reconstructed using a concrete example. In the example, 6 out of 13 individuals are observed (genotyped). During Step 1, an initial pedigree is determined by PO relationships. Note that initial pedigrees may not have one single tree structure, as shown in the middle panel of figure 4.2, and so may core pedigrees. During Step 3, an important dummy intermediate individual is added based on the relationships among grandchildren (in this case two FCs and one AV/HS), resulting in a larger pedigree, i.e., the final core pedigree. The correctness of core pedigree relies on the pair-wise relationship inference among grandchildren. For example, if the AV/HS is mistakenly inferred as FC, then the dummy intermediate individual will be added to pedigree incorrectly, causing an erroneous core pedigree. Also, depending on number and importance of the unobserved individuals, the core pedigree may not recover the whole pedigree, as shown in the bottom panel of figure 4.2. However, as long as the reconstructed core pedigree is a partial one of the true pedigree, we regard it as accurate.

4.2.4 Simulation

We tested our method with simulation. A fabricated pedigree of 14 subjects was used as the building block of the simulation (Figure 4.3). This fabricated pedigree satisfied all our assumptions and contained all kinds of relationships of interest. The ages and sexes of the pedigree members are also assumed to be known. 5,000 pedigrees were generated, all of which are the same. Subjects within each pedigree were randomly selected to be missing (ungenotyped), so the pedigrees have different pattern of missingness. Four different levels of missingness were considered: 0% (14 genoptyped subjects per pedigree), 21.4% (11 genotyped subjects per pedigree), 42.9% (8 genotyped subjects per pedigree), and 51.7% (6 genotyped

subjects per pedigree). Pair-wise relationship inferences were simulated based on the error rates estimated with the US training dataset in Chapter 2. Essentially, the true relationships were specified as different categories with probabilities shown in Table 4.1. After introducing missingness to pedigrees and simulating the pair-wise relationship inferences, we applied our method to identify families, reconstruct core pedigree for each family, and generate improved pair-wise relationship inferences.

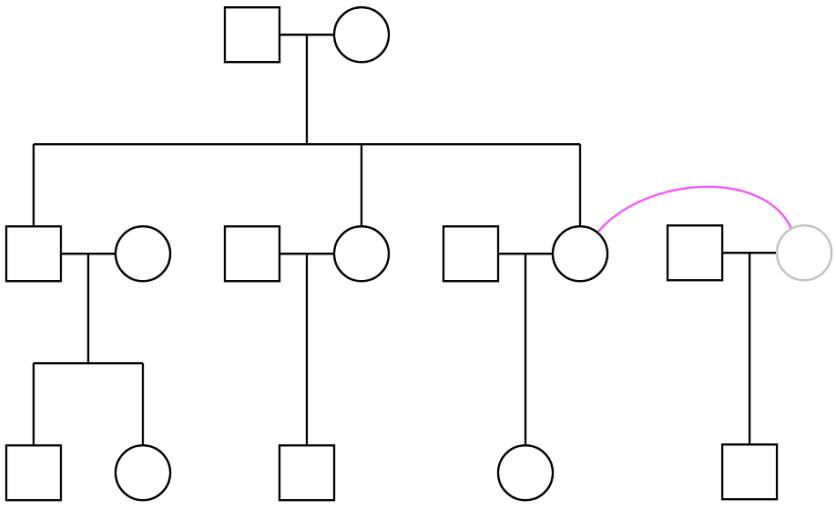


Figure 4.3 The fabricated pedigree for simulation

Table 4.1 Pair-wise relationship inference error rates estimated by cross-validation in the US sample

Truth	Inferred Relationship (%)		
	AV/HS	FC	GG
AV/HS	93.8	0	6.2
FC	5.8	94.2	0
GG	15.5	0	84.5

Note: 100% for MZ, FS, PO, and UN.

4.3 RESULTS

The simulated pair-wise relationship inference results were compared with the truth before and after reconstructing core pedigrees. Table 4.2 summarizes the results by relationship category for different levels of missingness in pedigrees. The prediction accuracy of each relationship category was also shown in Figure 4.4. Note that two new categories, AV and HS, which were not separable in previous pair-wise relationship classification, were presented because they can be inferred using the reconstructed core pedigrees.

Table 4.2 Relationship inference accuracy based on 5,000 simulated pedigrees with and without reconstructing core pedigrees by individual missingness

a. 0% missing (all the 14 subjects observed in each pedigree)

True Relationship Category	Prediction Ignoring Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	46,863 (93.7%)	0	0	3,137 (6.3%)	0	93.7%*
HS	4,688 (93.8%)	0	0	312 (6.2%)	0	93.8%**
GG	7,822 (15.6%)	0	0	42,178 (84.4%)	0	84.4%
FC	2,395 (6.0%)	0	0	0	37,605 (94.0%)	94.0%
True Relationship Category	Prediction Considering Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	0	50,000 (100%)	0	0	0	100%*
HS	0	0	5,000 (100%)	0	0	100%**
GG	0	0	0	50,000 (100%)	0	100%
FC	0	0	0	0	40,000 (100%)	100%

*The proportion of AV predicted as AV or AV/HS. **The proportion of HS predicted as HS or AV/HS.

Table 4.2 Continued

b. 21.4% missing (11 subjects observed per pedigree)

True Relationship Category	Prediction Ignoring Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	28,448 (93.9%)	0	0	1,841 (6.1%)	0	93.9%*
HS	2,845 (93.6%)	0	0	194 (6.4%)	0	93.6%**
GG	4,757 (15.8%)	0	0	25,323 (84.2%)	0	84.2%
FC	1,393 (5.8%)	0	0	0	22,707 (94.2%)	94.2%
True Relationship Category	Prediction Considering Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	171 (0.56%)	30,109 (99.41%)	0	9 (0.03%)	0	99.97%*
HS	39 (1.28%)	0	2,996 (98.59%)	4 (0.13%)	0	99.87%**
GG	39 (0.13%)	0	0	30,041 (99.87%)	0	99.87%
FC	33 (0.14%)	0	6 (0.02%)	0	24,061 (99.84%)	99.84%

*The proportion of AV predicted as AV or AV/HS. **The proportion of HS predicted as HS or AV/HS.

c. 42.9% missing (8 subjects observed per pedigree)

True Relationship Category	Prediction Ignoring Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	14,450 (94.1%)	0	0	909 (5.9%)	0	94.1%*
HS	1,451 (94.1%)	0	0	91 (5.9%)	0	94.1%**
GG	2,320 (5.0%)	0	0	13,145 (85.0%)	0	85.0%
FC	708 (5.7%)	0	0	0	11,698 (94.3%)	94.3%
True Relationship Category	Prediction Considering Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	2,344 (15.3%)	12,882 (83.9%)	0	133 (0.9%)	0	99.1%*
HS	373 (24.2%)	0	1,144 (74.2%)	25 (1.6%)	0	98.4%**
GG	389 (2.5%)	0	0	15,076 (97.5%)	0	97.5%
FC	214 (1.7%)	0	22 (0.2%)	0	12,170 (98.1%)	98.1%

*The proportion of AV predicted as AV or AV/HS. **The proportion of HS predicted as HS or AV/HS.

Table 4.2 Continued

d. 57.1% missing (6 subjects observed per pedigree)

True Relationship Category	Prediction Ignoring Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	7,754 (93.9%)	0	0	504 (6.1%)	0	93.9%*
HS	774 (93.3%)	0	0	56 (6.7%)	0	93.3%**
GG	1,249 (15.2%)	0	0	6,966 (84.8%)	0	84.8%
FC	379 (5.7%)	0	0	0	6,295 (94.3%)	94.3%
True Relationship Category	Prediction Considering Underlying Pedigrees					
	AV/HS	AV	HS	GG	FC	Accuracy
AV	4,481 (54.3%)	3,542 (42.9%)	0	235 (2.8%)	0	97.2%*
HS	458 (55.2%)	0	337 (40.6%)	35 (4.2%)	0	95.8%**
GG	583 (7.1%)	0	0	7,632 (92.9%)	0	92.9%
FC	227 (3.4%)	0	9 (0.1%)	0	6,438 (96.5%)	96.5%

*The proportion of AV predicted as AV or AV/HS. **The proportion of HS predicted as HS or AV/HS.

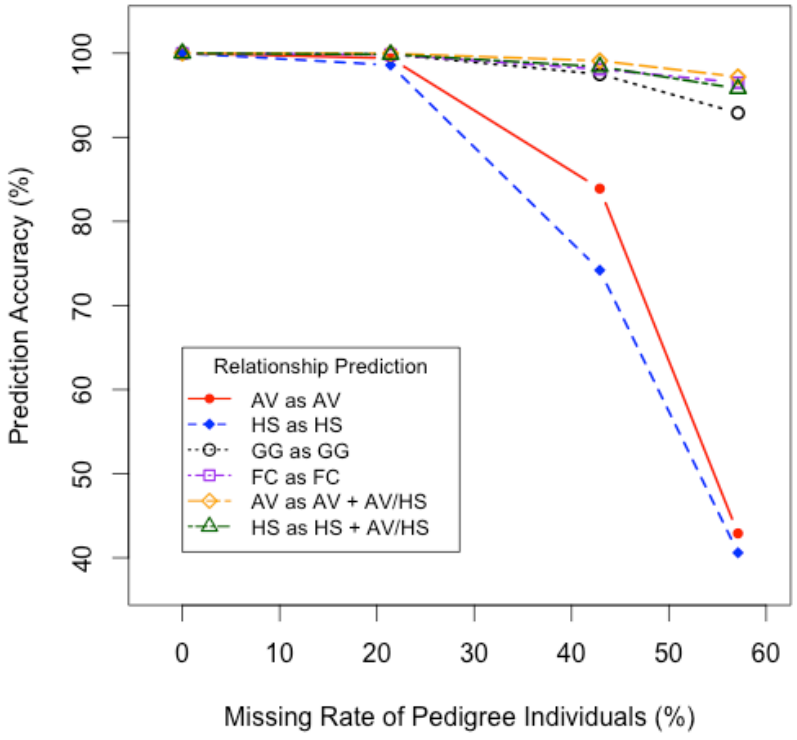


Figure 4.4 Prediction accuracy by relationship category as functions of individual missingness

When all the subjects in pedigrees were observed, relationships were inferred perfectly (Table 4.2a). With the increase in missingness in pedigrees, the prediction accuracy decreases for all relationship categories. However, even when more than half of the subjects were missing, the accuracy was still above 90% for all relationship categories, although the prediction accuracy of AV and HS to the exact relationships dropped dramatically. Most importantly, improvement was shown in all cases by reconstructing pedigrees for the second- and third-degree relationships.

Table 4.3 Core pedigree errors and coverage* for different pedigree individual missing rates

Observed Individuals per Pedigree (Among A Total of 14)	Missing Rate	Erroneous Core Pedigrees** (Among 5,000 Simulated Pedigrees)	Total Relative Pairs	Relative Pairs Captured by Core Pedigrees
14	0%	0	245,000	100%
11	21.4%	4	148,125	99.4%
8	42.9%	18	75,429	88.3%
6	57.1%	8	40,373	69.4%

*Defined as the percentage of relative pairs that are captured by core pedigrees. **Defined as implying any relative pairs not matching the truth.

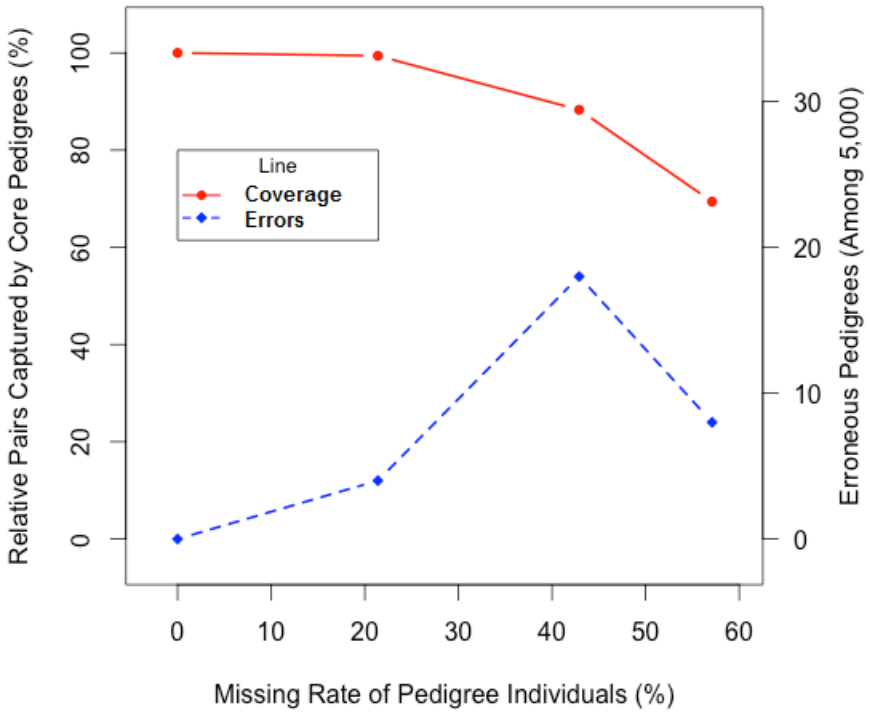


Figure 4.5 Coverage* and errors of reconstructed core pedigrees as functions of individual missingness**

*Defined as the percentage of relative pairs that are captured by core pedigrees. **Defined as implying any relative pairs not matching the truth.

Beside pair-wise relationships, we also investigated the qualities of reconstructed core pedigrees. A reconstructed pedigree was deemed as erroneous if it implies any relative pairs that are different from the truth. Among 5,000 simulated pedigrees at each level of individual missingness, only very few errors were detected (Table 4.3 and Figure 4.5), and the largest error rate was only 0.36%. Therefore, it is safe to conclude that our core pedigrees are very accurate.

We are also interested in the percentage of relative pairs that can be captured by core pedigrees, which we denote as the coverage. The more relative relationship pairs a core pedigree can capture, the more complete and useful it is. From Table 4.3 and Figure 4.5 we can see that

even if more than half individuals were missing, the reconstructed core pedigree on average is able to capture more than 60% of relative relationship pairs.

4.4 DISCUSSION

Instead of attempting to reconstruct maximum-likelihood pedigrees fitting all relative pairs, our method focuses on building "core pedigrees" that are highly accurate using the observed first-degree relatives, assisted by subject sex and age information and constraints imposed by assumptions. The core pedigrees can then be used as an initial point for manually reconstructing the whole pedigrees after incorporating more external information. A byproduct of pedigree reconstruction is improved pair-wise relationship inferences and depending on the level of individual missingness in pedigrees, a great number of AV and HS may become separable.

Our methods and results are based on a few assumptions, which should be reemphasized. Firstly, pedigree reconstruction is an extension of our pair-wise relationship classification pipeline. In general, we want to deal with the situations of a large number of small pedigrees, so only 7 relationship categories were considered: MZ, FS, PO, UN, GG, AV/HS, and FC. With these relationship categories, the possible underlying pedigrees must have three generations at most. Also, the pedigrees may not involve multiple spouse marriage in the first generation, and there should be no bilineal relatives other than MZ and FS. Secondly, we used the classification error rates from the US training data as the parameter to simulate pair-wise relationship inferences. MZ, FS, PO and UN pairs were assumed to be classified perfectly. The inference of MZ, FS and PO are indeed nearly perfect in reality, but it may be too optimistic for UN. If there are any mistakes about the UN inferences, no matter the cases of true UN being inferred as other

relationships or other true relative pairs being classified as UN, it will affect family identification at the very beginning. In this sense, our simulation results may be compromised in practice. Thirdly, our method requires the sex and age of subjects as input when constructing the core pedigrees, although in most studies such information should be available, this may not be always the case.

There are a few issues that should be noted in our simulation. First, we generated pair-wise relationship category labels instead of starting from simulating genome-wide marker data on the pedigree structures, and then inferring relationship labels from those data. The two procedures may have some differences. Also, the relationship labels were simulated independently, but in real data errors may be correlated. For example, a poorly genotyped individual may have less accurate relationships with everyone. Lastly, we randomly introduced missingness to pedigrees. However, the missing patterns of pedigrees in reality depend heavily on research aims and study populations. In general, order people are more likely to be missing. For some studies where only probands are collected, most individuals are not included and the probands are often from the same generation.

In population- or community-based studies, relative pairs are usually sporadic, which can be interpreted as having very high level of individual missing rate in pedigrees. In these cases, there may not be enough information to build meaningful core pedigrees. However, close pair-wise relationships are usually what we are interested in such studies anyway and pedigree reconstruction is more important for studies that are designed to be family-based.

5.0 SUMMARY AND FUTURE WORK

Our work provided a sound pipeline to classify close family relationships using un-phased dense genotype data and optional putative pedigree information. We demonstrated our pipeline on SNP array datasets and explored the extensibility to whole exome sequencing data. With pair-wise relationship inferences, we also developed a method to reconstruct accurate core pedigrees for further improving pair-wise relationship inferences and providing a basis for manual reconstruction of the whole pedigrees. When the assumptions hold, our methods were shown to be fast and accurate.

A few future directions may be considered to improve our work.

Firstly, the datasets used to build the relationship classifiers have limited number of training datasets and the training examples are not balanced among different relationship categories. If more training data are collected, better classifiers may be trained and classification error rates may be estimated more precisely. We have showed that the classifiers built with SNP array data cannot be directly used for whole exome sequencing data, so it is of interest to train classifiers specific for whole exome sequencing data.

Secondly, while our pedigree reconstruction method has been applied to simulated data to demonstrate its function and performance, it would be of interest to apply the method to real data before any pedigree cleaning, to compare the inferred pedigrees with the best pedigrees arrived

at by manual reconstruction. Also, it would be interesting to find a way to directly compare our method with PRIMUS.

Another possible improvement to our work is to consider more categories in the relationship classifiers, e.g., “great grandparent-grandchild” and “other relationships”, which stands for all other distant relationships not belonging to any categories. This may enable our methods to deal with datasets containing large pedigrees involving more complex relationships. With additional relationships the assumptions of reconstructed pedigrees can be relaxed.

In addition, instead of generating core pedigrees, a fully automated method is still desirable for reconstructing the maximum likelihood pedigrees, although this may be very challenging. A possible breakpoint is to dissect large pedigrees into small pieces and reconstruct the whole pedigrees in cascade. Even if it is hard to determine one single best pedigree, it will be informative to draw each part and indicate the relationships connecting them.

Finally, before any actual reconstruction process for maximum likelihood pedigrees taken place, a more fundamental and important problem should be solved, which is to determine when a maximum likelihood pedigree can be found uniquely, since sometimes there are situations where ties for the maximum likelihood pedigree structures exist.

BIBLIOGRAPHY

Browning, B. L. and S. R. Browning (2011). "A fast, powerful method for detecting identity by descent." Am J Hum Genet **88**(2): 173-182.

Chang, C. C. and C. J. Lin (2011). "LIBSVM: A Library for Support Vector Machines." Acm Transactions on Intelligent Systems and Technology **2**(3).

Chen, W., B. Li, et al. (2013). "Genotype calling and haplotyping in parent-offspring trios." Genome Res **23**(1): 142-151.

Cussens, J., M. Bartlett, et al. (2013). "Maximum likelihood pedigree reconstruction using integer linear programming." Genet Epidemiol **37**(1): 69-83.

Epstein, M. P., W. L. Duren, et al. (2000). "Improved inference of relationship for pairs of individuals." Am J Hum Genet **67**(5): 1219-1231.

Feingold, E. (1993). "Markov-Processes for Modeling and Analyzing a New Genetic-Mapping Method." Journal of Applied Probability **30**(4): 766-779.

Gusev, A., J. K. Lowe, et al. (2009). "Whole population, genome-wide mapping of hidden relatedness." Genome Res **19**(2): 318-326.

He, D., Z. Wang, et al. (2013). "IPED: inheritance path-based pedigree reconstruction algorithm using genotype data." J Comput Biol **20**(10): 780-791.

Hill, W. G. and I. M. S. White (2013). "Identification of Pedigree Relationship from Genome Sharing." G3-Genes Genomes Genetics **3**(9): 1553-1571.

Huff, C. D., D. J. Witherspoon, et al. (2011). "Maximum-likelihood estimation of recent shared ancestry (ERSA)." Genome Res **21**(5): 768-774.

Kirkpatrick, B., S. C. Li, et al. (2011). "Pedigree reconstruction using identity by descent." J Comput Biol **18**(11): 1481-1493.

Kong, A., D. F. Gudbjartsson, et al. (2002). "A high-resolution recombination map of the human genome." Nature Genetics **31**(3): 241-247.

Li, H., G. Glusman, et al. (2014). "Relationship estimation from whole-genome sequence data." PLoS Genet **10**(1): e1004144.

Li, H., G. Glusman, et al. (2014). "Accurate and robust prediction of genetic relationship from whole-genome sequences." PLoS One **9**(2): e85437.

Manichaikul, A., J. C. Mychaleckyj, et al. (2010). "Robust relationship inference in genome-wide association studies." Bioinformatics **26**(22): 2867-2873.

Matise, T. C., F. Chen, et al. (2007). "A second-generation combined linkage physical map of the human genome." Genome Res **17**(12): 1783-1786.

McKenna, A., M. Hanna, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

McPeek, M. S. and L. Sun (2000). "Statistical tests for detection of misspecified relationships by use of genome-screen data." Am J Hum Genet **66**(3): 1076-1094.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Morrison, J. (2013). "Characterization and correction of error in genome-wide IBD estimation for samples with population structure." Genet Epidemiol **37**(6): 635-641.

Patterson, N., A. L. Price, et al. (2006). "Population structure and eigenanalysis." PLoS Genet **2**(12): e190.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.

Ray, A. and D. E. Weeks (2008). "Relationship uncertainty linkage statistics (RULS): affected relative pair statistics that model relationship uncertainty." Genet Epidemiol **32**(4): 313-324.

Rodriguez, J. M., S. Bercovici, et al. (2014). "Parente2: A fast and accurate method for detecting identity by descent." Genome Res.

Shem-Tov, D. and E. Halperin (2014). "Historical pedigree reconstruction from extant populations using PArtitioning of RElatives (PREPARE)." PLoS Comput Biol **10**(6): e1003610.

Staples, J., D. Qiao, et al. (2014). "PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent." Am J Hum Genet **95**(5): 553-564.

Stevens, E. L., G. Heckenberg, et al. (2011). "Inference of relationships in population data using identity-by-descent and identity-by-state." PLoS Genet **7**(9): e1002287.

Thornton, T., H. Tang, et al. (2012). "Estimating kinship in admixed populations." Am J Hum Genet **91**(1): 122-138.