RELATIONSHIP INFERENCE AND ANCESTRAL GENOME RECONSTRUCTION USING IDENTICAL BY DESCENT SHARING AMONG RELATIVES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ying Qiao

December 2021

RELATIONSHIP INFERENCE AND ANCESTRAL GENOME RECONSTRUCTION USING IDENTICAL BY DESCENT SHARING AMONG RELATIVES

Ying Qiao, Ph.D.

Cornell University 2021

The fraction of relatives in large genetic databases is continuously increasing, making it more vital to uncover the relationships among samples for further downstream analysis. Relatedness inference is a fundamental step for genetic association studies, population genetics, and genealogy. For close relatives, it is even possible to infer specific relationship types. However, due to the randomness of recombination and inheritance, different pairs of relatives, even in different degrees of relatedness, can have similar amounts of identical by descent (IBD) sharing. This makes relatedness inference difficult and especially causes ambiguities for inferring relationship types of the same degree. We first present an analysis that explores the possibility of improving the accuracy of relatedness inference by adding IBD segment numbers. We investigated the importance of IBD segment numbers between the pair of relatives via both a theoretical information theory analysis and a machine learning classification approach. Our study showed that the IBD segment number adds information for the relatedness inference in general, but the improvement of accuracy is weakened by the IBD detection error in practice. Next, we describe CREST, an accurate and fast method to identify specific relationship types of close relatives using multiway IBD sharing. More specifically, for a given second degree relative pair, we lever-

aged their mutual relatives to determine their relationship types—grandparent grandchild (GP), avuncular (AV), and half siblings (HS). CREST achieved high sensitivities when tested in both simulated and real dataset with sufficient mutual relatives. CREST also identifies the directionality for parent child (PC), AV, and GP pairs and has the potential to be extended to identify more distant relatives. Lastly, with the aid of IBD segments from relatives, we developed HAPI-RECAP to reconstruct parental genome from a set of genotyped siblings and their relatives using a combination of family-based phasing and IBD sharing. For families with eight or more children, HAPI-RECAP can reconstruct most of genotypes for two parents, with Comparable error rates to direct genotyping, using only genotype data from children. For smaller families with four to seven children, HAPI-RECAP is able to reconstruct large portion of two parents genome with the IBD segments of relatives as the reference.

BIOGRAPHICAL SKETCH

Ying Qiao received her B.S. from Zhejiang University in 2015, specializing in Mathematics & Applied Mathematics. Fascinated by the beauty and the power of math, she developed strong interests in mathematical modeling and its broad applications. In 2014, she participated in a program at UC Davis and was able to assist with research in the lab of Dr. Sharon Aviran. This opportunity opened her door to biostatistics and computational biology, and eventually led her to pursue a Ph.D. in this area. Ying started the doctoral program in Computational Biology at Cornell University and joined the lab of Dr. Amy Williams. In this lab, she worked on developing computational methods that characterize relatives in large genetic datasets. Ying was able to present her research in several international conferences. In 2019, she also interned at 23andMe, which gave her a better understanding of the meaning and applications of her research.

This document is dedicated to my family, friends, and everyone who supported and believed in me. You have my sincere appreciation.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the members of my special committee, my labmates and collaborators, my family, and my friends for all of your help, support, and encouragement throughout this process.

To Amy, thank you for the years of mentorship, your invaluable guidance, and constant support throughout the whole journey. You consistently rekindled my enthusiasm for research and shaped my scientific rigor. You are always there for me when I need help or encouragement, whether it is for my research or my personal growth. Thank you!

To Jason, thank you for your support and assistance! I could not have made it this far without your encouragement. Even when I didn't believe in myself, your faith in me lifted me up and changed my path. Thank you!

To Kilian, thank you for leading me into the machine learning world! Your instructions have opened the door and implanted in me the mindset of understanding the underlying principles of models. Thank you for enriching my graduate experience!

To my labmates and collaborators, Jens Sannerud, Jesse Smith, Daniel N. Seidman, and Siddharth Avadhanam, thank you for being great collaborators and supporting me throughout the years. I enjoyed working with you and discussing both science and nonsense in the lab!

To my parents, thank you for your unfailing faith in me over the years. Your irrational confidence in me has nurtured a great deal of optimism and courage in me. Your unconditional support and love have always been and will continue

to be my source of strength. I realize you do not know anything about my research and probably will never read this, but I am confident you will be quite proud of it!

To Afrah Shafquat and Manisha Munasinghe, I am fortunate to have known you since I arrived here and to be able to go through the entire process with you. Thank you for always being there when I needed help and for inspiring me by being great yourself! To Sue Bishop, thank you for all of your timely and valuable help over the years. Your support has made my graduate studies much easier.

Thank all of my friends who have supported and encouraged me throughout the years. As this would take more than one page to list all of your names, I will refrain from doing so, but my gratitude remains the same. Thank you for your support!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The Importance of Relatedness and Relationship Inference

Recently more and more large scale genetic datasets exist because of the ever-decreasing cost, allowing for novel genetic discoveries and studies. Related to the recent outburst in genetic study sample sizes, a higher proportion of individuals in a data set have at least one close relative, making relatives detection necessary. An example is the UK Biobank[7], where approximately 30% of genotyped individuals have a third degree (e.g., a first cousin) relative or closer in the cohorts. Some direct-to-consumer (DTC) genetic testing companies even maintain datasets with sample size in the milions and much higher fractions of relatives due to the less random samples[23]. As a result, both the needs and ability to identify genetic relatives continues to grow, necessitating the new analysis and approaches to characterize the enriched relatives in large scale datesets.

Relatedness and relationship inference has broad applications in genetic analyses. It is a fundamental component in studies that directly make use of genetic relatives, including pedigree reconstruction[32,58], pedigree-based linkage analysis for disease and trait mapping[62], heritability estimation[6,17,71,73], forensic genetics[33,35,66], and genetic genealogy[36]. In addition, to avoid bias in many population genetic models, relatives must be accounted for or removed. In particular, the genome-wide association studies (GWAS) traditionally needs to exclude cryptic relatedness to avoid spurious signals or biased effect sizes[64], since it assumes

that samples are unrelated and the violation can alter the genome-wide distribution per-SNP p-values[1]. Thus, successful and precise genetic studies hinges on the abilities to detect and account for relatives.

Recent efforts to account for close relatives in genetic models indicates that the higher resolution and more accurate identification of relatives is in demand. Generally, pruning out close relatives to avoid modeling violations[65] will dramatically reduce sample sizes in large datasets[7,56]. Instead, with relationship information among relatives, studies have shown that relatives in different relationship types might vary in their shared environmental effects, even if they have the same expected kinship, and thus the estimation of their heritabilities can be inflated[71,73]. In addition, given accurate relatedness and relationship inference, those relatives enables pedigree reconstruction, which empower association and disease mapping and genetic genealogy in DTC genetic testing companies[32]. However, the current approaches to inferring pedigrees from genetic data[22,37,58] has the limitations due to ambiguities in the relatives' pedigree relationships. These potential applications and discoveries have attracted attention to the question of how to accurately identify relationships in large genetic studies.

1.1.1 Identical by Descent Segments

Identical by descent (IBD) segments are inherited by two or more individuals from a common ancestor without recombination. IBD segments between the pair of relatives give insight into their shared ancestry and the proportion of genome that is IBD sharing reveals how distantly related they are. For instance, the parent-child will share one half of their genome since the parent always

randomly transmits half of his/her genome to the children. As the generation increases, the expected proportion of the genome that the common ancestor and the decedent share IBD will decrease exponentially. Thus, the IBD information between relatives plays an essential role to infer their relationships.

More specifically, there are three types of IBD status at each locus: IBD0, IBD1, and IBD2. If a pair of relatives shares two copies at a locus, with each inherited from the same common ancestor, that locus is denoted as IBD2 sharing. Similarly, if the pair shares one or zero haplotype copy at the locus, the locus is denoted as IBD1 or IBD0 correspondingly. Considering the IBD0, IBD1, and IBD2 all together, it is useful to calculate the kinship coefficients (Defined in 1.1.2) and infer relatedness. In particular, they can be used to distinguish specific relationship types under certain circumstances. For example, the parent-child pair is expected to have IBD1 sharing at everywhere in the genome; while the full siblings are expected to share $\frac{1}{2}$ of the genome IBD1, $\frac{1}{4}$ of the genome IBD0, and another $\frac{1}{4}$ of the genome IBD2. To look at each IBD status separately, it is easy to distinguish parent-child and full siblings, even if they have the same expected kinship coefficients.

1.1.2 The Degree of Relatedness and Relationship Types

The degree of relatedness represents the level of shared genome between relatives, with each degree corresponding to an expected genome-wide proportion of IBD sharing. The kinship coefficient ϕ is used to quantify this proportion. Assume that a pair of individuals, denoted as i and j, has the proportion $k_{ij}^{(1)}$ and $k_{ij}^{(2)}$ of their genomes that i and j share IBD1 and IBD2, respectively. Their kinship coefficient can be calculated as $\phi_{ij} = \frac{k_{ij}^{(1)}}{4} + \frac{k_{ij}^{(2)}}{2}$, meaning the probabil-

ity that two randomly selected alleles, each from individual i and j, at a locus are IBD. By definition, the kinship coefficient is the half of the proportion of the genome that individual i and j share IBD. Technically, the closer a pair of relatives, the lower the degree of relatedness and the greater the kinship coefficient. For example, the parent-child and full siblings are in first degree and they are expected to have kinship coefficients of $\frac{1}{4}$. However, the kinship coefficient can vary even within each degree since the randomness of segregation and recombination, which are determined during meiosis, will influence the amount of DNA that relatives share.

The estimated kinship coefficient can be used to infer the degree of relatedness. The KING paper[41] recommended the ranges of kinship coefficient values for each degree: for a given degree n, the range of acceptable kinship coefficient is $[2^{-n-\frac{3}{2}}, 2^{-n-\frac{1}{2}}]$. Note that the expected kinship coefficient is 2^{-n-1}, which will exponentially decrease as the degree of relatedness increase. In this way, we can map estimated kinship coefficient to the degree of relatedness between relatives. This approach has been widely used to infer relatedness. Notably, there are different relationship types with the same expected kinship coefficient, so they are classified as the same degree. As shown in Table 1.1, a few common relationship types, such as grand-parent, avuncular, and half-siblings, in the same degree also have the same expected proportion of the genome that is IBD0, IBD1, and IBD2. This indicates that these statistics of IBD are not enough to distinguish different relationship types within the same degree.

Degree	Relationships	IBD0	IBD1	IBD2	ϕ
1	Parent-child	0	1	0	$\frac{1}{2^2}$
	Full siblings	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2^2}$
2	Half-sibling	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2^3}$
	Avuncular	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2^3}$
	Grandparent	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2^3}$
	Double-cousins	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2^3}$
3	First cousins	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{2^4}$
	Great-grandparent	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{2^4}$

Table 1.1: Different relationship types in each degree of relatedness; expected proportions of the genome that is IBD0, IBD1, and IBD2; and expected kinship coefficient ϕ. This list only include several common relationship types for the reference.

1.1.3 Difficulties of Relatedness and Relationship Inference

Although kinship coefficient plays a crucial role for relatedness inference, the distributions of kinship coefficient for different degrees tend to overlap, especially as the degree increases. Since the randomness of Mendelian inheritance and recombination cause the variance of kinship coefficient, as the number of meiosis increases, the coefficient of variation in kinship coefficient also become larger. Thus, for higher degree, it is more possible to observe the overlapping in distributions, and the ranges to map kinship coefficient are typically not enough to cover the variations. This fact complicates the inference for higher degrees and limits the abilities of current approaches to accurately infer relatedness beyond third degree[49].

In the case of relationship inference within the same degree, the first degree rel-

atives are simple to distinguish—parent-child and full siblings have different proportion of the genome that is IBD0/IBD1/IBD2 (See Table 1.1). However, distinguishing relatives only one degree more distant, including grandparent-grandchild (GP), avuncular (AV), and half-sibling (HS) pairs becomes quite challenging. The genome-wide IBD statistics won't be effective enough to make an inference. Some studies found that the IBD segment number can be used to distinguish GP and AV[23], however, the IBD segment number distribution of HS still causes heavy overlap. The similarities of IBD segment distribution among these different relationship types introduce ambiguities.

In addition, the existence of background IBD segments and detection errors makes it more complicated in practice. Since any two individuals in a finite population are related, i.e., they must have a common ancestor at some point in the past. The different population structures might introduce different levels of background relatedness in the samples and therefore, cause the shift of estimated kinship coefficient distribution. On the other hand, high quality IBD detection is fundamental and detection errors could also prevent the improvement of relatedness and relationship inference.

1.2 Information Theory and Machine Learning Approaches

Relatedness and relationship inference can be treated as the classification problem, allowing for approaches using machine learning techniques and information theory analysis. Given each degree or each relationship types as a class, the difficulties of classification lie in that distributions of widely used features, i.e., kinship coefficients, proportions of the genome that is IBD0, IBD1, and IBD2, overlap among different classes. The key is to discover informative features

that can capture differences of classes.

Although many relatedness inference methods that only utilize kinship coefficients[14,41,49,53], the number of IBD segments also draws attentions of some approaches, such as ERSA[28]. However, a recent evaluation[49] found that its performance does not always exceed that of other methods that rely solely on kinship coefficients. To thoroughly investigate whether segment numbers provide additional information for relatedness inference beyond that provided by kinship coefficients, feature selection based on information theory brings an innovative view for evaluating additional features. More specifically, a commonly used measure—mutual information (MI)—can quantify the dependency between various features and the class variable (here, the degree of relatedness), as well as the dependency among the features themselves. This method has the advantage of not assuming a linear relationship between the features and can be calculated for both discrete and continuous variables[2]. This approach has been widely used in machine learning and data mining in fields as diverse as bioinformatics and pattern recognition[26,39,46] and can capture inherent links between variables from an information theory perspective in a way that classification cannot.

Machine learning approach for relatedness and relationship inference provide flexibility in terms of feature sets, models, and even datasets. First, there are machine learning models that does not make explicit assumptions about feature distributions, which can be hard to justify or violated easily due to population structures or IBD detection errors. In addition, features can be discrete or continuous values, and do not require to have no biological meanings. Second, since

machine learning models learn from training data, this approach can implicitly account for background relatedness and detection errors, given enough and proper training data. Instead, traditional approaches need to post-process data or adjust models to account for other factors based on specific datasets. Third, estimated probabilities or likelihoods for each class can also be generated using the machine learning approach. Afterwards, such as in pedigree reconstruction, this information is extremely valuable.

1.3 Ancestral Genome Reconstruction

Ancestor reconstruction has been always an interesting yet challenging subject. A recent study by Jagadeesan et al.[31] shows the possibility of accurate reconstruction and potential downstream analysis. This study reconstructed 38% of the maternal genome of Hans Jonatan (HJ), a man born in 1784 to an African mother and a European father. The fact that African gene flow to Iceland is very rare allows for using local ancestry inference results from 182 of HJ's genotyped descendants to reconstruct HJ's maternal genome. Analysis of the reconstructed genome indicated that HJ's mother was likely from the region spanned by Benin, Nigeria and Cameroon.

Besides this, other applications exist, including in the area of genome-wide association studies (GWAS). For example, Kong et al.[38] utilized descendants of an ungenotyped deceased lung cancer patient to impute his haplotypes. They recovered 1,001 SNPs in his two phased haplotypes and showed that the imputed region harbored variants associated with lung cancer. Such reconstructions make it possible to involve ungenotyped samples with phenotype information in GWAS, which can enable improved power and is especially meaning-

ful when the case is less attainable like in rare diseases.

These examples offer a hint into the future potential of reconstructing ancestral genomes. However, the large numbers of genotyped descendants and complete pedigrees required by previous methods are often unavailable in practice. Moreover, the prior information about the African ancestry is a very special case, thus that approach has limitations to reconstruct other distant ancestors. In contrast, it is easier to collect genome segments of more recent ancestors from a smaller number of descendants. Phenotype information for recent ancestors may also be more easily attainable.

To reconstruct genome of parents from children is the fundamental step, when accurate pedigree is not available. Children inherit two chromosome copies, one from each parent, with both formed via recombination. So only half of the genomes of each parent is transmitted to each child. When there are multiple children available, due to independent segregation and randomized recombination, different child will inherit different regions of genome from both parents. By expectation, n siblings will inherit a proportion of $1 - (1/2)^n$ of both parents' genomes. As n increases, it is possible to reconstruct a large portion of parental genomes from a set of genotyped children. The reconstruction of ungenotyped parents has the potential applications in GWAS, relatedness inference, and other downstream analysis.

In this thesis, we explore new approaches and methods to utilize IBD information to improve relatedness and relationship inference, as well as the possibility to reconstruct ancestral genome from descendants and relatives with the

aid of IBD. In Chapter 2, we conduct both a mutual information analysis and a machine learning classification to evaluate the importance of IBD segment numbers for relatedness inference. In Chapter 3, we present CREST (Classification of RElationShip Types), a novel approach to distinguish pedigree relationships of close relatives via multy-way IBD sharing. Chapter 4 introduces our method, HAPI-RECAP (REConstruct Ancestral geontyPes), to reconstruct parental genome from children and their close relatives.

CHAPTER 2

EVALUATING THE UTILITY OF IDENTITY-BY-DESCENT SEGMENT NUMBERS FOR RELATEDNESS INFERENCE VIA INFORMATION THEORY AND CLASSIFICATION

2.1 Abstract

Despite decades of methods development for classifying relatives in genetic studies, pairwise relatedness methods' recalls are above 90% only for first through third degree relatives. The top-performing approaches, which leverage identity-by-descent (IBD) segments, often use only kinship coefficients, while others, including ERSA, use the number of segments relatives share. To quantify the potential for using segment numbers in relatedness inference, we leveraged information theory measures to analyze exact (i.e., produced by a simulator) IBD segments from simulated relatives. Over a range of settings, we found that the mutual information between the relatives' degree of relatedness and a tuple of their kinship coefficient and segment number is on average 4.6% larger than between the degree and the kinship coefficient alone. We further evaluated IBD segment number utility by building a Bayes classifier to predict first through sixth degree relationships using different feature sets. When trained and tested with exact segments, the inclusion of segment numbers improves the recall by between 0.0028 and 0.030 for second through sixth degree relatives. However, the recalls improve by less than 0.018 per degree when using inferred segments, suggesting limitations due to IBD detection accuracy. Lastly, we compared our Bayes classifier that includes segment numbers with ERSA and IBIS and found comparable results, with the Bayes classifier and ERSA slightly outperforming

each other across different degrees. Overall, this study shows that IBD segment numbers can improve relatedness inference but that errors from current SNP array-based detection methods yield dampened signals in practice.

2.2 Introduction

Relatedness inference in genetic data often plays a fundamental role in enabling more accurate genetic analyses—both in studies that directly leverage relatives and those that prune them to avoid modeling violations. The need and opportunity to identify genetic relatives continues to increase as the scale of genetic datasets increase[7,23]. One notable example is the UK Biobank wherein roughly 30% of genotyped individuals have a third degree (e.g., first cousin) or closer relative in the study[7]. Applications that make use of genetic relatives are numerous and varied and include pedigree reconstruction[32,58], pedigree-based linkage analysis for disease and trait mapping[45], heritability estimation[71,73], forensic genetics[67], and genetic genealogy[55]—a popular tool among direct-to-consumer genetic testing customers. On the other hand, traditional genome-wide association study tests and many population genetic models assume that the study samples are unrelated, and, as such, must exclude inferred relatives to avoid spurious signals or inaccurate parameter estimates[64]. All these applications motivate a thorough analysis of the approaches used for relatedness inference to determine which of the various features the methods should leverage.

Many relatedness inference methods only utilize kinship coefficients[14,41,49,53], while some such as ERSA leverage the number of identity-by-descent (IBD) segments between a pair[28]. To date, the question of whether segment numbers provide information for relatedness inference beyond that of kinship coef-

ficients has not been carefully explored. A recent evaluation of 12 pairwise relatedness inference methods using real relatives highlighted three top performing approaches: ERSA and two IBD detection algorithms (i.e., using kinship coefficients derived from their output)[49]. Although ERSA models the distribution of both the number and lengths of IBD segments, that evaluation found that it does not always outperform other methods that only use kinship coefficients. One possible reason is that estimated segment numbers from most phase-based IBD detection methods are inflated due to switch errors that typically break up segments[14,19,53]. Alternatively, these results may indicate that IBD segment numbers and lengths do not better capture relatives' degrees of relatedness than kinship coefficients.

To determine whether incorporating the number of IBD segments in a model with kinship coefficients (or coefficients of relatedness) improves relatedness inference, we first performed an information theory-based analysis. Feature selection based on information theory is widely used in machine learning and data mining in fields as diverse as bioinformatics and pattern recognition[26,39,46]. We applied a commonly used measure—mutual information (MI)—to quantify the dependency between various features and the class variable (here the degree of relatedness) and also the dependency among the features themselves. An advantage of this approach is that MI does not make an assumption of linearity between the features and can be calculated for both discrete and continuous variables[2]. In addition, the MI analysis results do not depend on the specific classifier used downstream and can capture the relationship between variables from an information theory perspective that is distinct from classification.

We also conducted a classification-based analysis to determine the importance of IBD proportions and segment numbers for inferring degrees of relatedness. For this purpose, we developed a Bayes classifier with mathematical underpinnings that parallel those of MI. Bayes classifiers are a form of generative learning that seek to minimize the probability of misclassification by estimating the probability of a given data point being from each class[13]. In this work, we assign a pair of relatives to the maximum posterior probability degree, in contrast to approaches that map estimated kinship coefficients to degrees of relatedness using *a priori* fixed ranges of kinship[41,49,53]. The latter ignores the effect of population structure on IBD signals—including background IBD segments[67]. These effects are important to model since they vary by population and can meaningfully influence relatedness classification. Furthermore, bias in the detection of IBD segments can shift the distributions of both IBD proportions and segment numbers. Such biases may especially impact classification of more distant relatives as they have smaller ranges of kinship values that correspond to a given degree. In light of these concerns, we estimate the probability of the features given the degree (i.e., the likelihood) using training data simulated using genotypes from the target population. This implicitly accounts for the influence of background IBD segments as well as any errors in IBD segment detection. Researchers with access to data from a given population can also apply this strategy by using the available samples as founders in simulated pedigrees[8].

Finally, we benchmarked the performance of our relatedness classifier together with ERSA and IBIS using simulated genotypes. Overall, we obtained comparable classification results for all the methods, indicating that the Bayes classifier is reliable and suggesting that our approach can be used in practice given appro-

priate training data resources. Notably, the Bayes classifier performs similarly to IBIS (which does not use segment numbers) demonstrating that, in practice, incorporating segment numbers provides very little improvement in classification rates.

All the analyses in this paper leverage IBD segments from simulated data, either exact segments produced by the simulator or segments inferred from simulated genetic data. In particular, we investigated (1) MI quantities based on exact segments, (2) classification rates using exact segments, and (3) classification rates from inferred segments. In this way, the MI analysis quantifies the theoretical information gain available by fully exploiting relatedness signals captured by exact segment numbers. Additionally, the classification analysis using exact segments reveals how much improvement in relatedness inference is possible by incorporating IBD segment numbers in the limit of perfect IBD detection. Finally, comparing the classification results using exact versus inferred segments enables us to localize the influence of IBD detection errors.

2.3 Results

We analyzed the potential for using coefficients of relatedness r (defined below) either alone or both r and n, the number of IBD segments a pair of relatives share, to infer the pair's degrees of relatedness D. To begin, we quantified the inherent dependency between the IBD segment features and D by analyzing MI between the features and D. MI is a quantification of the information obtained about one random variable through observing another; in this case, we analyzed the information gained about D through observing the variables r, n, or (r, n). We compared our analysis of MI quantities with the corresponding Bayes

classification results based on features r, n, and (r, n); the conclusions we form about the classification effectiveness of different feature sets are therefore based on both the MI and classification results.

Throughout, we refer to IBD regions that two individuals share on only one haplotype copy as IBD1, and those the individuals share IBD on both chromosomes as IBD2.

Mutual information analysis

We used thousands of relative pairs to estimate mutual information $I(\vec{F}\,;\, D)$ between different IBD features $\vec{F}$ and the degree of relatedness D of each pair (Methods, "Estimating mutual information"). Specifically, we compared MI values of $I(n\,;\, D)$, $I(r\,;\, D)$, and $I((r, n)\,;\, D)$ calculated using units of bits. Let $k_{ij}^{(1)}$ and $k_{ij}^{(2)}$ denote the proportion of their genomes that individuals i and j share IBD1 and IBD2, respectively—i.e., the sums of genetic lengths of all IBD1 or IBD2 segments divided by the total genetic length of the genome analyzed. We calculate r as twice the kinship coefficient or $r = \frac{k_{ij}^{(1)}}{2} + k_{ij}^{(2)}$.[49]

The first analysis uses exact IBD segments from pairs of individuals that each have one of 13 genetic relationships (Table 2.1). To reduce the influences of randomness, we replicated this analysis by performing 80 independent simulations. We also analyzed three different distributions of numbers of pairs per degree D: uniform, exponential, or a slow-exponential function where the number of pairs increases exponentially with degree for both the exponential and slow-exponential distributions (Figure A.1). The exponential function is potentially a more realistic distribution of relatives than the uniform, while the

slow-exponential is intermediate between the two.

Degree	Relationships
1	Full siblings
2	Half-siblings, Avuncular
3	First cousins, Half avuncular
4	First cousins once removed, Half first cousins
5	Second cousins, Half first cousins once removed
6	Second cousins once removed, Half second cousins
7	Third cousins, Half second cousin once removed

Table 2.1: Simulated relationship types in each degree of relatedness. Half relatives share only one common ancestor while other types have two common ancestors.

Figure 2.1(a) shows the average MI of the simulated pairs computed over all 80 runs (Methods, "Simulated data"). For each distribution shape, the MI between the multivariate feature (r, n) and univariate D is the greatest, followed by $I(r\,;\,D)$ and $I(n\,;\,D)$. To quantify the relative increase in MI when including both n and r, we used a normalized MI gain $G_N(x) \equiv \frac{I((r,n)\,;\,D) - I(x\,;\,D)}{I((r,n)\,;\,D)}$ where $x \in \{r, n\}$. The normalized MI gain $G_N(r)$ (the increase in information gained from using (r, n) beyond that of only using using r) is 0.030 for the uniform distribution, 0.040 for the slow-exponential, and 0.068 for the exponential. Greater MI indicates a stronger dependency between the features and D, and therefore classifying D based on features with greater MI should yield greater recall. At $G_N(r)$ of 0.068 for the exponential distribution, we expect that incorporating numbers of perfectly detected segments could meaningfully improve classification of degrees of relatedness compared to using r alone, especially for higher order degree pairs. In turn, the normalized gain over using segment number alone, $G_N(n)$, is 0.15 for the uniform distribution, 0.14 for the slow-exponential,

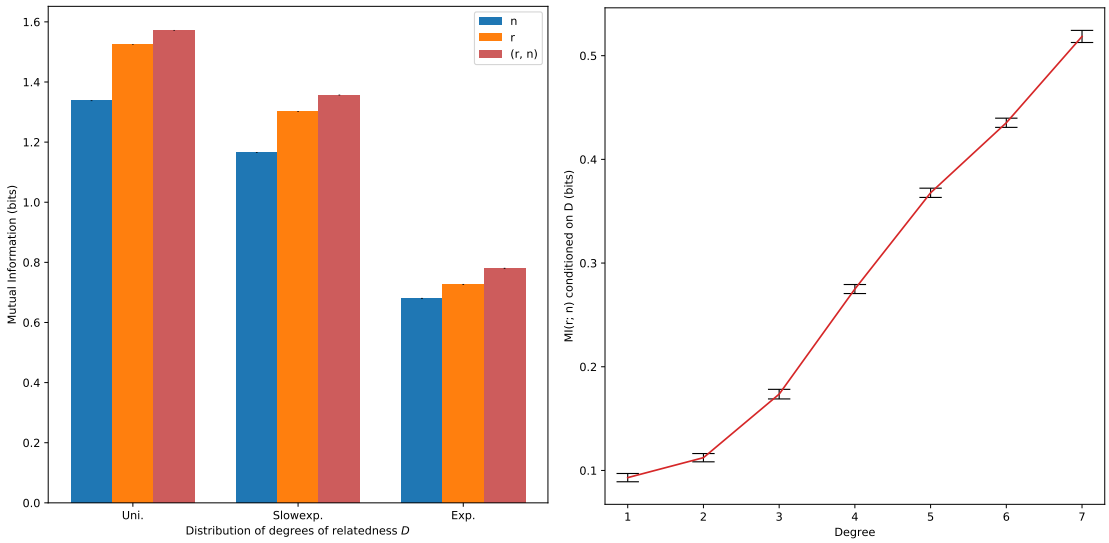

Figure 2.1: MI between relative pairs calculated using exact IBD segments. MI between (a) various IBD feature sets and D and (b) r and n conditioned on the relatives' degree of relatedness. All MI quantities are averaged over 80 independent runs, and the values in (b) are calculated using the uniform distribution with 33,000 pairs per degree. Error bars indicate one standard error and are barely visible in (a) (all of order 10^{-3}).

and 0.13 for the exponential, demonstrating that use of r dramatically improves classification rates compared to only using n, regardless of the distribution of D.

Across all three feature sets, the MI is maximal for the uniform D distribution and decreases as the distribution becomes more exponential. By construction, the exponential distributions have a higher proportion of high-degree relative pairs compared to the uniform distribution. Therefore, the IBD features from higher degree pairs share less information with D than lower degree pairs. This is consistent with observations from classification analyses that show that the recall of degree inference decreases as the degree increases[49].

To better understand how r and n relate to each other as well as to D, we calculated MI between these two features using the exact IBD segments and conditioned on the degree of relatedness (Figure 2.1(b)). The amount of shared information between features r and n monotonically increases with degree of relatedness, meaning that in higher degree pairs r and n have increased redundancy. Therefore, using both features has less benefit for classification in higher degrees. Nevertheless, both r and n individually become less informative about D with increasing degree, so any additional information can be useful.

Bayes classification and statistical tests of exact and inferred IBD segments

As MI quantities from exact IBD segments suggest the potential for sizeable improvements by using (r, n) to determine D, we sought to understand whether parallel results arise from explicit relatedness classification. To that end, we simulated another 210,000 pairs of relatives for training, this time producing genetic data for them using genotypes from UK Biobank unrelated samples as pedigree founders (Methods, "Simulation"). We detected IBD segments in these samples with IBIS and used the resulting r and n quantities to train Bayes classifiers. For comparison, we further trained a separate set of classifiers using the exact IBD segments from the same simulated pairs (Methods, "Bayesian classification"). Using Bayes classification allowed us to incorporate our prior knowledge of the distribution of D to better determine the pairs' degrees, and also more closely mirrors the mathematical basis of MI. For both the inferred and exact statistics, we generated a set of three classifiers, one trained only on the coefficient of relatedness r, one on the IBD segment number n, and a third on the vector (r, n).

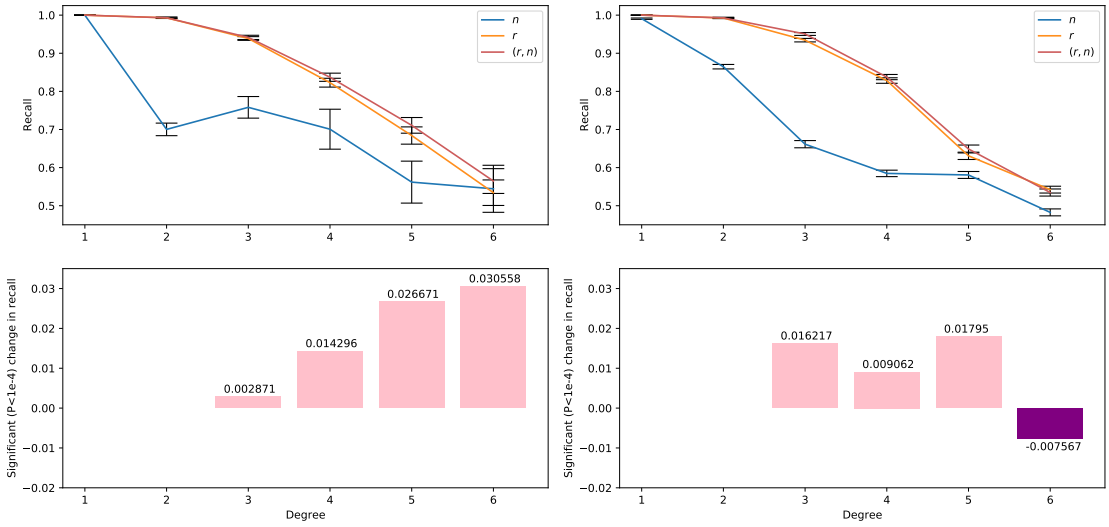

Figure 2.2: Recalls of Bayes classifiers for first through sixth degree relatives. Results are from classifiers trained on (a) exact and (b) inferred segments with features n, r, or (r, n). The recalls for both (a) and (b) are calculated using the uniform distribution of 3,000 pairs per degree and averaged over 80 independent runs. For each degree, the lower subplot shows the corresponding significant ($P < 10^{-4}$) change in recall between classifiers (r, n) and r (positive values have greater recall in the (r, n) classifier). Significant increases and decreases per degree are shown in pink and purple, respectively. Error bars indicate one standard error.

We tested both the exact and inferred segment classifiers on 80 independent simulated datasets containing 3,000 simulated relative pairs per degree, again inferring segments with IBIS. (Genetic data for testing pairs was produced identically to the training pairs, as noted above.)

Figure 2.2 shows the recalls of these classifiers as a function of degree and also shows the recall differences between classifiers trained on (r, n) and r. We also show the proportions and types of misclassifications in the inferred and exact datasets in Figure A.2 and A.3. Almost all misclassified pairs are inferred as an adjacent degree of relatedness compared to the truth (i.e., one degree closer or

more distant). Note that we do not report accuracy results for seventh degree relatives as these pairs act as an "unrelated" class that provide bounds on sixth degree relatedness classification.

Overall, recalls for all three classifiers decrease monotonically as a function of the degree of relatedness. For first and second degree pairs, the classifiers trained on r and (r, n) both have nearly perfect recall values of over 0.99. For higher degree pairs from third through sixth degree, the recalls of the r and (r, n)-trained classifiers fall from over 0.93 (third degree) to below 0.55. This is consistent with previous observations from real relatives[49], and aligns well with our results based on MI: The features of higher degree pairs share less information with D, meaning that the IBD signals of higher degree pairs tell the classifier less about their true D (see misclassification rates in Figure A.2 and A.3. The classifier trained on n alone performs poorly in all but degree one: For second degree relatives, the classifier trained on inferred segments has a recall of only 0.86, and in third through sixth degree relatives its recall is 0.06 to 0.27 units lower than those of the classifier trained on r. The results for the classifier trained on exact segments are qualitatively similar to those of the inferred-segment classifier.

In general, when using both exact and inferred IBD segments, the classifiers trained on (r, n) outperform those trained on r for every degree. One exception is in the inferred IBD segments for sixth degree pairs, where the classifier trained on r has a recall of 0.54 while the classifier trained on (r, n) has a recall of 0.53. This decrease in recall is counter-intuitive because the (r, n) classifier is trained on a strictly larger feature set and so has more information than the r

classifier. In addition to general stochasticity introduced by segment detection for these distant relatives, it may be that this decrease is caused by the distributions of segment numbers inferred by IBIS (Figures A.4 and A.5): IBIS does not detect segments smaller than 7 cM and so the distribution of numbers of detected segments for fifth and sixth degree pairs have lower means and are more similar to each other.

We ran two-sided independent sample t-tests on the recalls from the (r, n) and r classifiers trained on the inferred IBD segments. Except for the first degree relatives, in which all three classifiers have recalls of nearly 1.0, and the second degree pairs, in which the two classifiers containing r have above 0.99 recall, the differences in recall between the (r, n) and r classifiers are significant ($P < 10^{-7}$) but small in magnitude. These differences range from -0.00756 to 0.0179 in third through sixth degree pairs. In turn, for the classifiers trained on exact IBD segments, the (r, n) classifier has significantly greater recall than the r classifier in third through sixth degree relatives ($P < 10^{-4}$). The improvement in recall ranges from 0.0029 to 0.031, suggesting that better IBD segment inference would meaningfully benefit classification with (r, n) (Figure 2.2).

Comparison with IBIS and ERSA

To put these results in the context of existing methods, we compared our Bayes classifier with IBIS's built-in relative classifier and with ERSA, another method that models relatedness using IBD segment number (as well as with segment length). This analysis uses for testing another independent set of 3,000 pairs per degree, again simulated from UK Biobank individuals. Our Bayes classifier remained trained on the same 210,000 pairs as above.

In general, all three methods performed comparably. The accuracy of the Bayes classifier closely tracks that of IBIS, which may be because the Bayes method takes IBIS segments as input. At the two extremes of relatedness we considered, all three methods have similar recalls for first degree and sixth degree relatives with differences smaller than 0.01. The Bayes classifier has nearly identical recall to IBIS in second and third degree pairs (the differences are bounded above by 0.004), whereas ERSA's recalls for these degrees are 0.06 and 0.02 units smaller, respectively. (An analysis with real relatives also found that ERSA's second degree classification rates are reduced compared to other approaches[49].) For fourth degree relatives, ERSA has a recall 0.01 units higher than the Bayes classifier, and 0.035 units higher than IBIS. ERSA also outperformed the Bayes classifier and IBIS on fifth degree pairs by 0.067 and 0.054 units, respectively. ERSA's improved performance compared to the other two methods may be because of its use of $\geq$ 2.5 cM segments (instead of $\geq$ 7 cM segments from IBIS). Consistent with this, simulated fourth and fifth degree relatives have a non-trivial proportion of 3–7 cM segments (Figure A.6)—suggesting that these undetected IBD segments may lead to more erroneous calculations by IBIS and the Bayes classifier. Another factor benefiting ERSA is its population model that accounts for background relatedness, which may help it in this and other datasets. Additionally, we used perfectly phased data as input to GERMLINE[20], and we supplied the resulting segment calls to ERSA (Methods, "Simulated data"). Notably, ERSA's higher recalls for fourth and fifth degree pairs are close to the range of the Bayes classifier's recalls using exact IBD segments (in fact, ERSA outperforms the exact Bayes classifier in these degrees by 0.012 and 0.0031, respectively). Finally, considering run time, the Bayes classifier is efficient, taking on average 1 minute 40 seconds to analyze the test data and 7 seconds to train

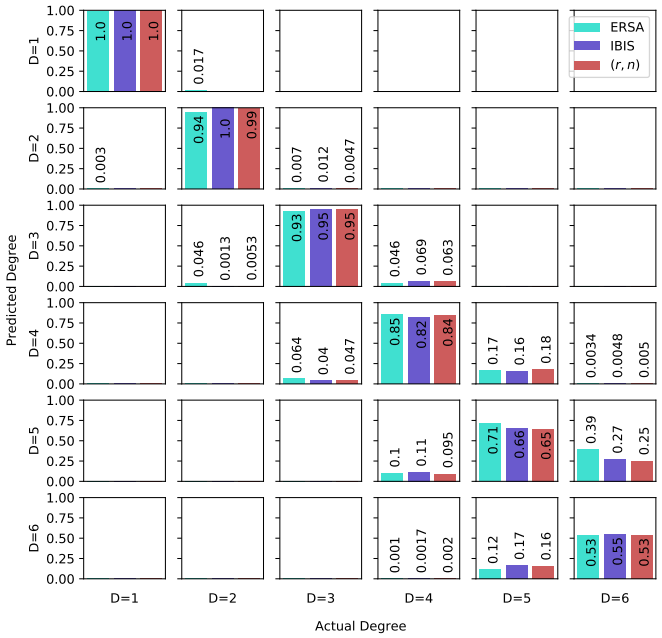

Figure 2.3: Confusion matrix of recalls of ERSA, IBIS, and our Bayes classifier trained on (r, n).

on the 210,000 training pairs. In contrast, ERSA takes more than 3.5 CPU days to classify the testing pairs.

2.4 Methods

Mutual information discrete definition and binning approaches

MI is difficult to calculate for continuously valued variables without a known distribution and whose distribution must therefore be estimated from finite data. Furthermore, estimating the MI between one continuous and one discrete random variable is in general non-trivial and multiple approaches exist for this estimation, such as nearest-neighbor[51] and binning methods. To enable our MI calculations (such as $I(r\,;\,D)$), we used a procedure that bins data points of r

and avoids biased MI estimates in our finite but large sample size. In computing MI, we treated the binned feature vector $\vec{F}$ (where $\vec{F}$ has the possibility of being one dimensional when representing r or n) and the degree of relatedness D as two discrete random variables with realizations f and $d \in [1, 7]$, respectively. If we know the probability mass functions (pmfs) of the discrete random variables X and Y with realizations x and y, we can calculate MI using its definition as

$$I(X\,;\,Y) = \sum_{x}\sum_{y} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}, \tag{2.1}$$

where $p_{X,Y}$ is the joint pmf of X and Y and p_X, p_Y are the marginal pmfs of X and Y respectively.

Binning a continuous variable in order to use Equation (2.1) introduces the difficulty of picking the right bin size. It has been shown[51] that MI is sensitive to bin size and that its stability with respect to this variable is dependent on the sample distribution. Our distributions and sample sizes of r yielded a large range of bin sizes that have stable MI estimates (see the flat regions of each curve in Figure A.7). Because the fraction G_N is normalized by MI, its correct calculation relies on the unbiasedness of the various MI quantities that form it. At bin sizes smaller than 150 pairs per bin (ppb), both the means and standard deviations (Figure A.8) of our MI quantities increased rapidly. Given this, in our calculations of MI, we binned r at 150 ppb, where our binning converts a continuous value of r to its nearest bin-value in 150 evenly spaced numbers from $[\min(\vec{r}), \max(\vec{r})]$. Here and below $\vec{r}$ represents all sampled training and testing data points r.

Estimating mutual information

Calculating Equation (2.1) without access to the entire spaces $\mathcal{X}$ and $\mathcal{Y}$—i.e., estimating MI from sampled data—is contingent on the estimation of marginal and joint probabilities p_X, p_Y, and $p_{X,Y}$. We used a simple counting approach to calculate each probability, assigning $\hat{p}_F(f) = \frac{1}{N}\sum_i^N \mathbb{1}(\text{bin}(F_i) = \text{bin}(f))$. Here $\vec{F}$ is the vector of realized data points representing all N sampled values for the desired feature r, n, or (r, n); $\text{bin}(x)$ denotes the function that takes a continuous realization to its binned value; and $\mathbb{1}(X = Y)$ is the indicator function. By binning r to 150 ppb as noted in the previous subsection, we were able to use this discrete maxium likelihood estimator (MLE) approach for calculating every desired pmf and obtain stable results in MI.

We performed calculations of MI on the exact IBD segment data restricted to three distribution shapes: A uniform distribution, a "slow-exponential" distribution, and an exponential distribution (see Figure A.1). We accounted for different distributions of D in the calculations of $I(\vec{F}\,;\, D)$ by decomposing the joint pmf relating $\vec{F}$ and D as $p_F(f, d) = p_F(f|d)p_D(d)$, and also decomposing the marginal pmf on $\vec{F}$ (with the law of total probability $p_F(f) = \sum_{d'} p_F(f|d')p_D(d')$). Equation (2.1) is then expressed as

$$I(\vec{F}\,;\, D) = \sum_f \sum_d p_F(f|d)p_D(d) \log \frac{p_F(f|d)}{\sum_{d'} p_F(f|d')p_D(d')} \tag{2.2}$$

by canceling the $p_D(d)$ terms in the numerator and denominator. $p_F(f|d)$ is the pmf of realizations of feature $\vec{F}$ in a given degree, and $p_D(d)$ is the distribution shape (from Figure A.1). This approach removes noise associated with calculating the pmfs $\hat{p}_{F,D}$ for different distribution shapes, which stems in part from random factors in finite sample sizes (including smaller numbers of pairs in the

non-uniform distributions). In particular, because the probabilities of f conditioned on any given degree of relatedness d are identical across each distribution shape, we estimated $p_F(\vec{F}|d)$ once only from the uniform distribution data. Note too that the differences in MI due to the D distribution are entirely accounted for in the probabilities $p_D(d)$, and these are exactly calculable given the equation for each distribution.

Probability density estimation of features

In the context of the Bayes classifier, we estimated the probability of a feature realization f conditioned on the training data $\vec{T}^d$ in degree d according to the degree-wise count as

$$\hat{p}(f|d,\vec{T}) = \frac{1}{N_d^T}\sum_{i=1}^{N_d^T}\mathbb{1}(\mathrm{bin}(T_i^d) = \mathrm{bin}(f)), \tag{2.3}$$

T_i^d being a particular realization of the training data $\vec{T}^d$ with total count N_d^T. However, we only had access to the frequencies of realizations f that occur at least once in the training data, so Equation (2.3) is only calculable for these values. The total training data $\vec{T}$ and testing data $\vec{\tau}$ are of dimension equal to their respective number of data points N^T or N^τ. To generate posteriors $p(D|\tau_i)$ for realizations in $\vec{\tau}$ at values where there are no training data points in $\mathrm{bin}(\tau_i)$, we linearly interpolated the values given by Equation (2.3) within the convex hull (see Figure A.9) specified by the bounds of the training data. (Strictly speaking, these posteriors are then incorrect pmfs with mass greater than 1—however, in practice this is only relevant for a vanishingly small number of points.) We used the `scipy` packages `interp1d` and `griddata` for the linear interpolations in one-dimension (when $\vec{F}$ is either r or n) and two-dimensions (when $\vec{F}$ is (r,n)), respectively.

In the case that $\vec{F}$ is (r, n), the two-dimensional linear interpolation of $\hat{p}(f|\vec{T})$ values are only well-defined inside of the two-dimensional convex hulls of the training data. Therefore, we could not assign posteriors to realizations of the testing data that lay outside the bounds of the training data. For these data points (labeled in Figure A.9 as "Unscored under (r, n)"), we assigned probability values according to the one-dimensional interpolation of $p(f|\vec{T}^d)$ values with $\vec{F} = r$. The one-dimensional interpolations for $\vec{F} = r$ (or $\vec{F} = n$) only remained undefined when they occurred outside the interval of training values $[\min(\vec{r}), \max(\vec{r})]$ (or $[\min(\vec{n}), \max(\vec{n})]$), in which case they remained unclassified in our analysis. In the inferred segment data, there was only a maximum of one point per degree that remained unclassified.

Bayes classification

Our classifiers use the posterior probabilities $p(D|\vec{F}, \vec{T}) = \frac{p(\vec{F}|D,\vec{T})p(D)}{p(\vec{F}|\vec{T})}$ for the single and multivariate features $\vec{F}$ to infer D in the testing data. The priors $p(D)$ are the known shape of the degree distribution (Figure A.1), and we generated the probability of our data $p(\vec{F}|\vec{T})$ as the sum across degrees according to the law of total probability $\sum_d p(\vec{F}|d, \vec{T})p(d)$. We calculated likelihoods $p(\vec{F}|D, \vec{T})$ according to the estimator in Equation (2.3). To classify a testing pair τ_i to a certain degree, we calculated $\log p(D|\tau_i)$ for each degree and classified the pair as the maximum a posteriori degree:

$$D_i^P = \underset{d' \in \mathcal{D}}{\operatorname{argmax}} \log p(D = d'|\tau_i), \tag{2.4}$$

where D_i^P is the predicted degree, and $\mathcal{D}$ is the set of possible degrees $\{1, \ldots, 7\}$. The recall of a particular classifier for degree d is $\frac{1}{N_d^\tau} \sum_i^{N_d^\tau} \mathbb{1}(D_i^P = d)$.

The classifier takes input IBIS segments and calculates the IBD proportion and segment numbers for all pairs of individuals with at least one detected IBD segment. It classifies any pair with $r < 2^{-15/2}$ (a common lower bound for seventh degree classification[41,49]) as unrelated, and, for all other pairs, predicts their degree using Equation 2.4.

Simulated data

For the exact IBD segment data, we used Ped-sim[8] to simulate 231,000 relative pairs of 13 relationship types from seven degrees of relatedness (Table 2.1) (replicated 80 times for the MI analysis and once for the classification-based analysis) and leveraged the IBD segments this tool prints. Thus these segments are free of error and we refer to them throughout as *exact*. We used both sex-specific genetic maps[3] and crossover interference modeling[27] for these simulations.

For each degree, we simulated an equal number of pairs from each of two relationship types. The one exception is first degree relatives where we only considered full sibling pairs since parent-child pairs always have $r = 0.5$ and are trivial to identify. We doubled the number of full sibling pairs (to the total number assigned from the distribution shape) so that the first degree relatives included the full number of pairs. We calculated the IBD proportion by adding the lengths of all outputted IBD segments and dividing by the total length of the sex averaged genetic map—halving the length of IBD1 segments (see the equation for r in Results). We calculated the segment number by counting the number of outputted IBD segments from either Ped-sim (exact) or IBIS (inferred, as described next).

To simulate relatives with genetic data, we used autosomal genotypes from participants in the UK Biobank[7] as founders in Ped-sim runs. We used the phased data distributed by the UK Biobank[7] and, before simulating, filtered the samples to include the white British ancestry subjects. To filter out close relatives, we first performed SNP quality control filtering on the UK Biobank unphased genotypes (filtering SNPs with minor allele frequency less than 2%, missing data rate greater than 1%, and retaining only SNPs used for phasing in the original analysis[7]). Next we ran IBIS v1.20.8 on the filtered genotypes with the `-maxDist 0.12` option and with IBD2 segment detection enabled. This provided kinship coefficients that we then input to PRIMUS[58], running it with `--no_PR` (which corresponds to not reconstructing pedigrees: executing only IMUS[57]) and `--rel_threshold 0.022` to filter out relatives with a kinship coefficient greater than 0.022 (i.e., retaining only pairs no more closely related than fifth degree[41]). We ran Ped-sim as described above (using sex-specific genetic maps and crossover interference modeling) and otherwise used default options (including genotyping error and missing data rates of 10^{-3}). Finally, we used IBIS v1.20.7 (enabling IBD2 detection with `-2`) to detect IBD segments between these simulated relatives.

Running ERSA

To get relatedness estimates from ERSA[28], we first ran GERMLINE[20] v1.5.1 with `-err_het=1` and `-err_hom=2` (the options recommended by the ERSA authors) on the simulated Ped-sim haplotypes. That is, we provided ERSA perfectly phased data output by the simulator. We then ran ERSA with default options on the resulting GERMLINE segments.

Runtimes

We ran both ERSA and our Bayes classifier on a machine with an AMD EPYC 7702 2.0 GHz processor and 1 TB of RAM. We supplied 16GB to ERSA and 8GB to our Bayes classifier. Both methods are single threaded.

2.5 Discussion

In this paper, we sought to examine how much incorporating the number of IBD segments together with the coefficient of relatedness of a relative pair improves degree of relatedness inference. We thus provided both a theoretical MI analysis using simulated exact IBD segments and a machine learning-based classification analysis using exact and inferred segments. The results using exact segments show that including IBD segment numbers can non-trivially enhance related inference quality, especially for distant relatives. However, the results using inferred segments reveal that IBD detection errors—including false negatives for segments shorter than 7 cM—meaningfully limit this improvement. Indeed, the performance of our machine learning classifier is almost indistinguishable from IBIS (Figure 2.3), which does not use segment numbers. With the potential development of more accurate IBD detection tools in the future—including for whole genome sequencing data—use of IBD segment numbers in relatedness inference models may be worth considering.

We introduced a machine learning-based classifier and demonstrated that it has comparable accuracy to two state-of-the-art methods and is computationally efficient. Because we fit the classifier to population-specific training data (instead of using fixed kinship thresholds for each degree[49,53]), it implicitly accounts for

background IBD sharing and erroneous IBD signals. This approach differs from model-based methods such as ERSA in that it makes no assumptions about the distributions of IBD segment lengths or numbers with respect to relatedness degrees. Those assumptions can be violated in populations with small effective size or a historical founder effect[28]. Our trials of this machine learning method suggest that even without data for large numbers of (labeled) real relatives, simulating relatives enables this data-driven approach to relatedness inference. Additionally, both the machine learning classifier and the MI analyses can be easily extended to include other IBD features such as the minimum or maximum IBD segment length between a pair.

An important factor in attempting to utilize IBD segment numbers is their accurate detection. Switch errors profoundly influence segment number estimates when using phase-based IBD detectors[14,19,44,53]. Our use of IBIS segments in our classifier was motivated by IBIS's ability to call IBD segments in unphased data—one of only a few methods to do so[14]—which is key to avoiding biased segment number estimates. ERSA takes inferred IBD segments from the phase-based IBD detector GERMLINE. To exclude the possibility of phasing errors impacting ERSA's performance, the phased data we provided GERMLINE was that generated by the simulator, thus being perfect up to the limit of the haplotypes input to Ped-sim. In particular, these haplotypes do not contain switch errors in IBD segments between the simulated relatives. It is possible that ERSA's superior performance in classifying fifth degree relatives is enhanced by its segment detection in these data.

In general, our analyses are consistent with prior work showing that relatedness

inference can achieve high recall for up to third degree relatives. However, two recent papers have focused on distinguishing relationship types of the same degree, especially three types of second degree relatives[47,70]. In this setting, IBD segment numbers can provide useful information, such as for distinguishing avuncular from grandparent-grandchild pairs[23]. Still, for degree of relatedness inference, even when using exact IBD segments, the classification recalls for distant relatives—i.e., those beyond fourth degree—are limited (Figure 2.2(a)). This suggests that pairwise IBD information might not be sufficient to reliably infer distant relatives, regardless of segment quality. Approaches that leverage multi-way IBD signals to infer more distant relatives can achieve considerably higher accuracy than those of pairwise methods[50,59]. Even so, these multi-way methods are built on pairwise classifiers, so understanding and improving pairwise relatedness classification remains an important fundamental problem for relatedness inference.

Acknowledgements

We thank Debbie Kennett for conversations about genetic genealogy. Funding for this work was provided by NIH grant R35 GM133805. Computing was performed on a cluster administered by the Biotechnology Resource Center at Cornell University. This research has been conducted using the UK Biobank Resource under Application Number 19947.

Data Availability

Data for exact segments are generated using the open source Ped-sim simulator and it is possible to generate data with the same expected summary statistics

given the pedigree definition (def) files from this study. Genetic data were simulated using Ped-sim based on input UK Biobank haplotypes. The latter are available to qualified researchers from the UK Biobank. Code used to calculate mutual information and perform Bayes classification is available from the authors upon request.

CHAPTER 3

DISTINGUISHING PEDIGREE RELATIONSHIPS VIA MULTI-WAY IDENTITY BY DESCENT SHARING

3.1 Abstract

The proportion of samples with one or more close relatives in a genetic dataset increases rapidly with sample size, necessitating relatedness modeling and enabling pedigree-based analyses. Despite this, relatives are generally unreported and current inference methods typically detect only the degree of relatedness of sample pairs and not pedigree relationships. We developed CREST, an accurate and fast method that identifies the pedigree relationships of close relatives. CREST utilizes identical by descent (IBD) segments shared between a pair of samples and their mutual relatives, leveraging the fact that sharing rates among these individuals differ across pedigree configurations. In simulated data, CREST correctly classifies 91.5-100% of grandparent-grandchild (GP) pairs, 80.0-97.5% of avuncular (AV) pairs, and 75.5-98.5% of half-siblings (HS) pairs compared to PADRE's rates of 38.5-76.0% of GP, 60.5-92.0% of AV, 73.0-95.0% of HS pairs. Turning to the real 20,032 sample Generation Scotland (GS) dataset, CREST identified seven pedigrees with incorrect relationship types or maternal/paternal parent sexes, five of which we confirmed as mistakes, and two with uncertain relationships. After correcting these, CREST correctly determines relationship types for 93.5% of GP, 97.7% of AV, and 92.2% of HS pairs that have sufficient mutual relative data; and it completes this analysis in 2.8 hours including IBD detection in eight threads.

3.2 Introduction

Modern scale genetic datasets contain tens to hundreds of thousands of individuals, sample sizes within which numerous close relatives exist[7,56]. Characterizing relatives within such datasets is essential to avoid spurious signals and to improve power in genetic association studies[34,64,74], but standard models considering only kinship estimates while ignoring the potential for different relationship types to vary in their shared environmental effects and therefore their heritabilities[71,73]. Moreover, while population genetic studies typically filter close relatives to avoid modeling violations[65], such an approach will dramatically reduce sample sizes in large datasets[7,56]. One way to enable analyses of more study samples is to directly model the transmission of shared haplotypes—i.e., identical by descent (IBD) segments[63]—using the pedigree structure of each set of relatives, but this requires accurate determination of those pedigrees. And although several approaches exist for inferring pedigrees from genetic data[22,37,58], ambiguities in the samples' true pedigree relationships limit the utility of these methods.

Identifying pedigree relationships is simple for first degree relatives[41]—parent-child (PC) and full sibling pairs—yet distinguishing relatives only one degree more distant, including grandparent-grandchild (GP), avuncular (AV), and half-sibling (HS) pairs remains a challenge. Most methods infer only the degree of relatedness of a pair using either the number and length of pairwise IBD segments[23,28], or the proportion of their genome a pair shares IBD[41,49]. However, an existing method that leverages these pairwise signals provides limited ability to discriminate among second degree relationships[16]. Still, IBD segment number distributions overlap little between GP and AV types[23], and segment position

may be possible to leverage to infer relationship types using only pairwise IBD segments[25]. Turning to multi-way IBD approaches, a recent method detects aunts/uncles of siblings[50], but it requires at least two siblings to work and can only identify their aunts/uncles.

We developed CREST (Classification of RElationShip Types), a novel approach for inferring pedigree relationships that leverages multi-way IBD sharing. CREST utilizes multi-way IBD sharing to differentiate relationship types, relying on the fact that a pair of close relatives is expected to share IBD regions with their mutual relatives at different rates depending on the pair's relationship. For example, consider a mutual relative that is the parent of the genetically older member of a second degree relative pair. Because each meiosis leads to the transmission of half a parent's DNA, a grandchild will, in expectation, inherit 1/4 of the regions shared IBD between the grandparent and the mutual relative—i.e., the parent of that grandparent. In the case of AV pairs, since two full siblings have equal IBD sharing with their parent, the child of one sibling—the niece/nephew of the other—is expected to share 1/2 as many sites IBD with her/his grandparent as the aunt/uncle does. Lastly, two half-siblings have equal IBD sharing with their common parent. These same sharing rates—the genetically younger sample in GP, AV, and HS pairs sharing fractions of 1/4, 1/2, and 1 compared to the older sample, respectively—arise for many other types of mutual relatives, enabling the classification of relationship types. Thus, we derived IBD sharing quantities based on this idea and trained kernel density estimation models (KDEs) to classify these three types of second degree relatives in CREST.

This approach of leveraging IBD sharing with mutual relatives not only determines the pedigree relationship types of second degree relatives, it also identifies the directionality of the relationship—that is, which sample is genetically older (e.g., which is the grandparent or aunt/uncle). In particular, the sample with higher levels of IBD sharing with mutual relatives is most likely to be from an earlier generation. (Other pedigree inference methods similarly identify this information using kinship coefficients[37,58].) CREST applies this logic to GP and AV pairs and to PC relatives to detect which sample is the parent. When available, age information unambiguously implies the genetically older sample for direct descendants (PC and GP relationships), but can fail for AV pairs since a niece/nephew may be (temporally) older than an aunt/uncle.

We used a combination of simulated and real pedigree data to evaluate CREST, the latter from the Generation Scotland[43,54] (GS) cohort. The GS data consist of 20,032 samples recruited as part of families, and include 848 GP, 6,599 AV, and 381 HS pairs. We also compared CREST's results in simulated data to those of PADRE[59], a composite likelihood method that infers pedigree structures for two sets of close relatives when members of the sets are also related to each other. PADRE makes use of the relationship between the two sets to choose the PRIMUS pedigree that maximizes its composite likelihood and, in the process, implicitly infers the pedigree relationship of the second degree pairs.

In addition to classifying second degree relatives, the CREST approach may be extended to infer more distant relationship types. For example, when using simulated pedigrees that include a pair of third degree relatives and two first cousins of the genetically older sample, CREST can also distinguish third de-

gree relatives with high accuracy, thus highlighting the potential for expanding CREST's target relationships as datasets further grow in size.

3.3 Methods

CREST takes inferred IBD segments from a set of samples as input and applies a multi-way IBD sharing analysis to classify pedigree relationships among pairs. The multi-way IBD segment analysis calculates ratios from the IBD regions that a target pair of close relatives and their mutual relatives share, as described below. The algorithm then uses KDEs we trained on ratios from simulated relative sets to infer the pair's relationship type. CREST is open source and freely available (Web Resources).

We used IBIS[53], an approach that operates on unphased genotype data, to infer both IBD segments and degrees of relatedness. While CREST can use good quality IBD segments inferred by any method, IBIS produces IBD segments that are largely free of internal gaps[53], with the trade-off that by default it identifies ≥ 7 cM segments. Furthermore, our experimental results indicate that use of these long segments suffices for discriminating between second degree relationship types. Still, the use of shorter, gap-free IBD segments has the potential to increase the quality of CREST's inference further.

Throughout, we refer to IBD regions that two or more samples share on only one haplotype copy as IBD1 segments, and those the individuals share IBD on both chromosomes as IBD2 regions. Correspondingly, IBD0 regions are those where the given samples do not share an IBD segment.

Multi-way IBD sharing ratios

CREST utilizes the IBD regions shared between a pair of close relatives x_1 and x_2 and one or more of their mutual relatives to distinguish their relationship. The expected IBD rates we adopt are based on the assumption that each mutual relative y is related to both x_1 and x_2 only through the most recent common ancestor(s) (MRCA(s)) of x_1 and x_2. CREST further assumes that there is only one lineage from the MRCA(s) to both x_1 and x_2, thus excluding cases of close inbreeding. Under these assumptions, all IBD segments shared between y and one or both of x_1 and x_2 must have been transmitted by this/these MRCA(s) through one lineage. For example, if x_1 is the grandparent of x_2, we take their MRCA to be x_1 itself, and if y is the half-sibling of x_1, y is related to both x_1 and x_2 only through x_1 (via the common parent of x_1 and y), so the assumptions hold. However, if y is the half-sibling of the grandchild x_2, y is related to x_2 through their common parent, and not only through x_1, in conflict with the assumption. In fact, mutual relatives that are descendants of either x_1 or x_2 violate the assumption in many cases. To exclude direct descendants of x_1 and x_2, we only analyze mutual relatives that are third degree (e.g., a first cousin) or more distant relatives of both x_1 and x_2. Because most genetic datasets only span two or three generations, this strategy should generally prevent analyses involving descendant mutual relatives.

The intuition behind the approach CREST uses is that x_1 and x_2 will have different relative amounts of IBD sharing with a given mutual relative y depending on their relationship. We use two ratios to quantify the IBD sharing rates:

$$R_i = \frac{length(IBD(x_1, x_2, y))}{length(IBD(x_i, y))}, i \in \{1, 2\}.$$

Here $IBD(s_1, s_2, \ldots, s_n)$ denotes the set of IBD regions that all samples

$s_1, s_2, \ldots, s_n$ share, i.e., the intersection of the IBD segments each of the $\binom{n}{2}$ pairs share. The $length$ function sums the genetic length (i.e., Morgan [M] length) of a set of IBD segments, accounting for the diploid status of each segment. That is, for a given set of IBD segments I,

$$length(I) = \sum_{i \in I} \begin{cases} \frac{1}{2}\ell(i) & \text{if } i \text{ is IBD1} \\ \ell(i) & \text{if } i \text{ is IBD2,} \end{cases}$$

where $\ell(i)$ denotes the (M) genetic length of an IBD segment i, here from a sex averaged genetic map. The numerators are the same in both ratios and give the genetic length of IBD regions shared jointly by all three samples. The denominators are the length of IBD segments shared by x_1 and y in R_1, and by x_2 and y in R_2.

These ratios differ according to the relationship type of the second degree relatives. Specifically, for a GP pair, if x_1 is the grandparent of x_2, the numerator $length(IBD(x_1, x_2, y)) = length(IBD(x_2, y))$ since x_2 will inherit a subset of the IBD segments x_1 shares with y (Figure 3.1A). Additionally, $E[length(IBD(x_2, y))] = \frac{1}{4} \cdot length(IBD(x_1, y))$ since x_2 is two meioses away from x_1 and each meiosis leads to the transmission of an average of one-half of the IBD segment length any pair of relatives shares. Thus, $E[R_1] = \frac{1}{4}$ and $E[R_2] = 1$. Similarly, each member of a HS pair independently inherits one-half of the genome of their common parent $\tilde{p}$, so the probability that they both inherit a given IBD region that $\tilde{p}$ and y share is $\left(\frac{1}{2}\right)^2$ (Figure 3.1A). Therefore the expected numerator is $\frac{1}{4} \cdot length(IBD(\tilde{p}, y))$, and the expected denominator is $\frac{1}{2} \cdot length(IBD(\tilde{p}, y))$ for both R_1 and R_2, so $E[R_1] = E[R_2] = \frac{1}{2}$. In the case of an AV pair, the aunt/uncle inherits half the genome of her/his parent $\tilde{g}$—the grandparent of the niece/nephew—that is related to y. And, as in the

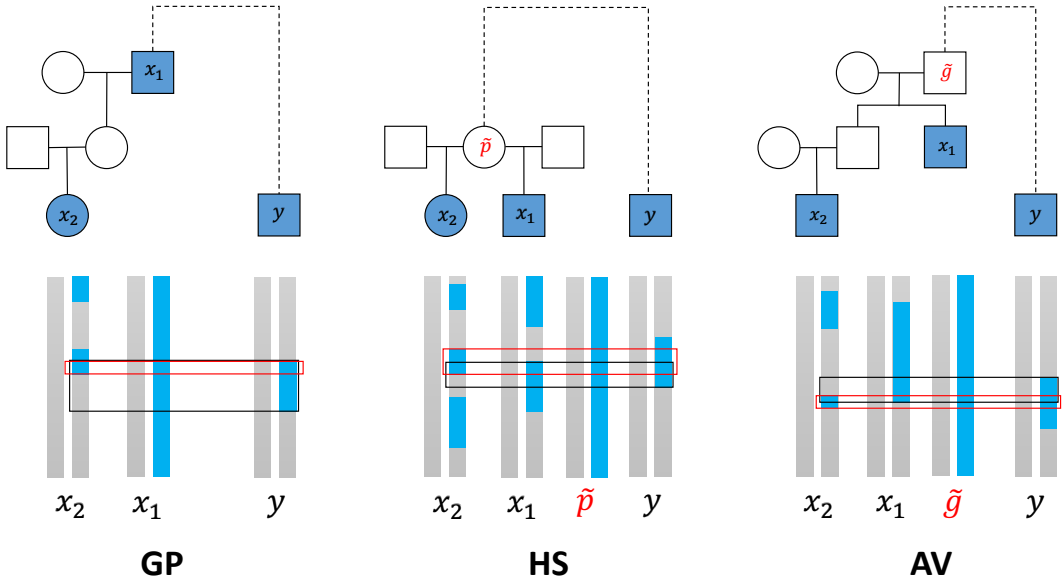

Figure 3.1: Example IBD sharing between the three types of second degree relatives and one of their mutual relatives. Samples with filled shapes are those for which data are available and include the close relative pair x_1 and x_2 and their mutual relative y. The dashed line connecting an MRCA of x_1 and x_2 to y indicates that the pedigree structure between that MRCA and y need not be known. Sexes here are arbitrary and the pedigree relationship type inference works identically for all sample sexes. Haplotypes for the genotyped individuals appear below each pedigree plot as blue or grey vertical bars, with haplotypes for ungenotyped common ancestors of the HS and AV pairs that are related to y also shown. The blue regions are either one haplotype of an MRCA of x_1 and x_2 or IBD segments other individuals share with this haplotype. (Grey portions of the vertical bars are not IBD with the blue haplotype in the MRCA and do not enter the analysis.) The black boxes outline the regions shared IBD between x_1 and y, and the red boxes outline the regions x_2 and y share IBD.

GP case, the niece/nephew is expected to inherit one-quarter of the genome of $\tilde{g}$ (Figure 3.1C). Therefore the expected numerator is $\frac{1}{2} \cdot \frac{1}{4} \cdot length(IBD(\tilde{g}, y))$, the expected denominator of R_1 is $\frac{1}{2} \cdot length(IBD(\tilde{g}, y))$, and that of R_2 is $\frac{1}{4} \cdot length(IBD(\tilde{g}, y))$, resulting in $E[R_1] = \frac{1}{4}$ and $E[R_2] = \frac{1}{2}$.

In practice, the above ratios vary around their expectations. This variability arises from three sources: errors in IBD segment detection, the variance in IBD

sharing between the close relative pair (i.e., depending on the outcome of the small number of meioses that separate them), and the variance in the meioses that separate y from the MRCA(s) of x_1 and x_2. This latter variance increases for greater meiotic distance. More specifically, mutual relatives y with a large meiotic separation share on average a comparatively small fraction of their genome IBD with the MRCA(s) of x_1 and x_2, and they have a higher coefficient of variation for this sharing rate than closer relatives[24], leading to higher variance in the ratios. Therefore, the more closely related y is to the MRCA(s) of x_1 and x_2, the more precise the ratios will be.

In large samples, data for multiple mutual relatives can be common, and considering only a single y will typically provide less information than combining data from multiple samples. In particular, combining IBD regions from multiple mutual relatives will often capture a larger fraction of the IBD regions that the MRCA(s) of x_1 and x_2 transmitted to the pair. Our approach to incorporating multiple mutual relatives into the ratios is to take the union over these samples of their three- and two-way IBD sharing regions. This effectively reconstructs the IBD sharing pattern of one or more ungenotyped sample[50] that is more closely related to x_1 and x_2 than any single y, thereby reducing the variance of the calculated ratios (Figure 3.2). The ratios are:

$$R_i = \frac{length(\bigcup_{y_j} IBD(x_1, x_2, y_j))}{length(\bigcup_{y_j} IBD(x_i, y_j))}, i \in \{1, 2\},$$

where y_j ranges over the mutual relatives that are available in the dataset and satisfy CREST's assumptions.

Ideally, the union operation in the above would be defined on two possible haplotypes of each x_i such that, if different relatives y_m and y_n share IBD segments to

a given x_i on different haplotypes and in the same region, the segments would be merged into an IBD2 segment. For example, as shown in Figure B.1, a grandparent can share overlapping IBD regions with a maternal relative and a paternal relative on different haplotypes. Merging these into a single IBD1 segment would yield biased ratios—reducing the grandparent's IBD sharing length by 1/2 at this location. A challenge in addressing this is that IBIS and some other IBD detectors do not report which haplotype a segment resides on. Thus we extended CREST to determine when a set of shared IBD regions belong to the same or different haplotypes. This procedure utilizes the fact that if either sample x_i has overlapping IBD regions on the same haplotype with any two relatives y_m and y_n, these regions should also be IBD between y_m and y_n. That is, regions x_i shares IBD1 to these relatives should have three-way IBD sharing such that $IBD(x_i, y_m) \cap IBD(x_i, y_n) \subseteq IBD(y_m, y_n)$. On the other hand, if y_m and y_n share IBD segments to the same region on different haplotypes of x_i, the corresponding haplotypes of y_m and y_n will not, in general, be IBD in that region. Thus, in regions where y_m and y_n are IBD0, CREST treats x_i as being IBD2 to the set of mutual relatives (which is equivalent to the $IBD^{(011)}$ concept implemented in DRUID[50]). Note that this approach does not detect all instances of IBD2 sharing: it is possible for y_m and y_n to be IBD1 to each other on one of their haplotypes while sharing their other haplotypes to each of x_i's two haplotypes. Therefore, this method is an approximation that does not consider this latter case since we lack information to distinguish which haplotypes the samples share.

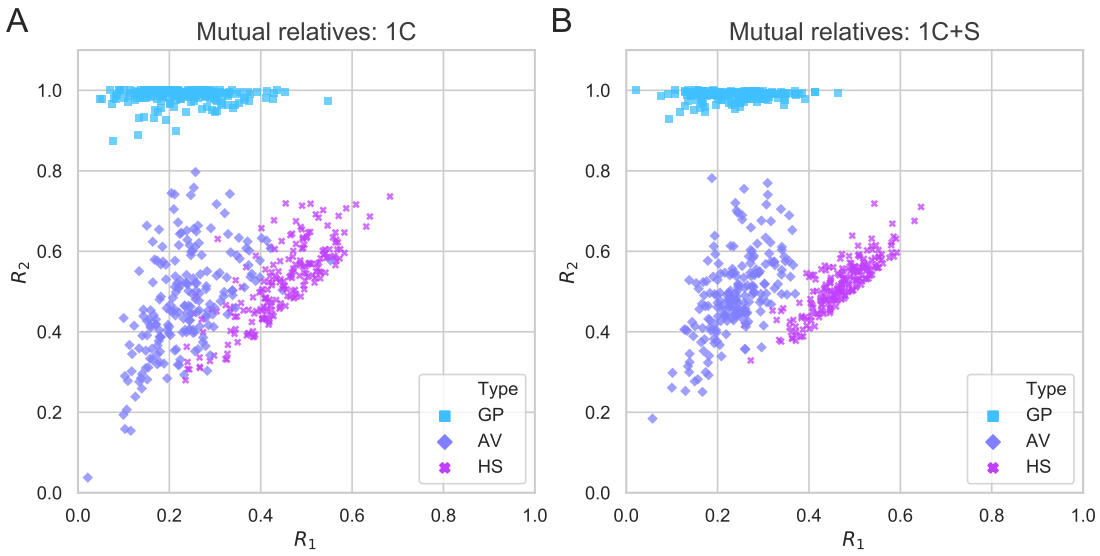

Figure 3.2: The R_1 and R_2 ratios cluster more tightly when using multiple mutual relatives. Ratios R_1 and R_2 from 200 simulated pairs of each relationship type, calculated using (A) one first cousin (1C) of the genetically older sample, and (B) combining one first cousin and his/her sibling (1C+S). Here we swap labels if needed so that $R_1 \leq R_2$.

Classifying relationship types using kernel density estimation models

CREST adopts KDEs to classify the three second degree relationship types using the ratios R_1 and R_2 as features. To train and evaluate the KDEs, for each such relationship type, we first simulated genotype data for a range of pedigree structures that include various mutual relatives, and we derived R_1 and R_2 ratios from the IBD segments that IBIS[53] detects in the simulated genotypes (see "Simulations" for details). Because the R_1 and R_2 values are ordered, and since we only seek to classify the relationship types (with directionality considered separately), CREST exchanges the order of the two ratios if needed such that $R_1 \leq R_2$. This shrinks the space the features range over, increasing precision. We then trained separate KDEs for each relationship type and used

five-fold cross validation to select both their optimal bandwidth (from 10^{-2} to $10^{-1/2}$) and kernel function from among the 'Gaussian', 'Linear', and 'Exponential' forms.

As noted earlier, the closer the mutual relatives are to the target pair, the less variance the ratios will tend to have, yielding more reliable classification. Therefore, to build models that account for this, we incorporate another feature that is associated with the variance: what we term the *genome coverage rate*, C, of the pair for a given set of mutual relatives. We define this as $C = \max\left(\frac{1}{L}length(\bigcup_{y_j} IBD(x_1, y_j)), \frac{1}{L}length(\bigcup_{y_j} IBD(x_2, y_j))\right)$, where L is the total (M) genetic length of the genome. Thus, it is the larger of either the IBD sharing rate between x_1 and the mutual relatives or that of x_2. This genome coverage rate is anti-correlated with the variance in the ratios (Figure B.2) since it is related to how much of the genome of x_1 and x_2's MRCA(s) is/are covered by IBD segments in the mutual relatives.

To incorporate genome coverage into our models, we built KDEs stratified by C, one for each of several bins. When $C < 0.2$, the bins span intervals of size 0.025, and we use only one bin for $C \geq 0.2$ because the variances of R_1 and R_2 appear more constant above this threshold (Figure B.2). CREST does not attempt to classify pairs with a $C < 0.025$ since distinguishing relationships is difficult with such a low signal. For a given genome coverage bin, we trained KDEs using five-fold cross validation as noted above for each bin separately.

To classify a pair's relationship type, CREST calculates the posterior probability of each type. It outputs these probabilities, calculated as $\Pr(T \mid$

$R_1, R_2, C) = \Pr(R_1, R_2 \mid T, C) \cdot \Pr(T) / (\sum_{T'} \Pr(R_1, R_2 \mid T', C) \cdot \Pr(T'))$, where $T \in \{\text{GP}, \text{AV}, \text{HS}\}$ is the type, and $\Pr(T)$ is the prior probability of the given type, which defaults to $\frac{1}{3}$ for all T, but can be specified by the user. $\Pr(R_1, R_2|T, C)$ is the likelihood of R_1 and R_2 for a given relationship T from the KDE applicable to the given genome coverage value C. As CREST reports all these probabilities, users can choose to use the maximum a posteriori relationship type or to incorporate the probabilities into downstream analyses. In Results, we use the maximum a posteriori type unless otherwise specified. When $C < 0.025$ (including when no mutual relatives are available) or $R_1 = R_2 = 0$ (i.e., $length(\bigcup_{y_j} IBD(x_1, x_2, y_j)) = 0$, so there is no detected multi-way IBD sharing to the mutual relatives), CREST does not infer the relationship but outputs the prior probabilities.

Inferring the directionality of the relationship

CREST leverages the ratios R_1 and R_2 to determine the directionality of the relationships. More specifically, CREST identifies which sample is the grandparent, aunt/uncle, and parent in GP, AV, and PC pairs, respectively, by comparing these ratios. In principle, the genetically older sample in the pair should inherit more DNA from the MRCA(s) than the younger sample. Thus, the union of pairwise IBD sharing over mutual relatives for the genetically older sample is expected to be greater than that of the younger sample. This pairwise IBD sharing quantity is in the denominator of the ratios, so CREST uses $D = \log_2 \frac{R_2}{R_1} = \log_2 \frac{length(\bigcup_{y_j} IBD(x_1, y_j))}{length(\bigcup_{y_j} IBD(x_2, y_j))}$ to determine the directionality. For instance, if x_1 is genetically older, then D is more likely to be positive. We trained KDE models with D values from simulated GP, AV, and PC pairs and CREST uses these to calculate the probability of the relationship directionality.

Simulations

To train and test CREST's relationship type inference, we used Ped-sim[8] to simulate a range of pedigree structures that include one GP, AV, or HS pair and one or more of their mutual relatives (Figure B.3). In all cases we used sex-specific genetic maps[3] and crossover interference[9] modeling in these simulations, and a collection of European descent samples[12] as the input phased data. (The latter were previously phased using Beagle[4], and filtered so that no pair is more closely related than fifth degree[50].)

The simulated data we used for training include mutual relatives that vary from first cousins to second cousins of the genetically older sample in the second degree pair. We simulated enough samples to obtain 1,000 pedigrees within each KDE genome coverage bin. As the coverage rate varies for a given pedigree structure, we simulated 1,000 pedigrees for each relationship type and pedigree structure class in five batches of 200 pedigrees each. We then mapped these to the corresponding genome coverage bin based on the IBD segments IBIS inferred, and we randomly downsampled to obtain 1,000 pedigrees per bin. The pedigrees include nine different combinations of mutual relatives that have the following relationships to the genetically older sample: one first cousin; one first cousin and his/her sibling; two first cousins that also are first cousins to each other (i.e., non-sibling first cousins); three first cousins that are first cousins to each other; one first cousin and his/her niece/nephew; one first cousin once removed and his/her sibling; one first cousin once removed and his/her niece/nephew; one second cousin; one second cousin and his/her sibling. Thus, we include third degree relatives (first cousins) and as far as seventh degree relatives (second cousins twice removed of a grandchild) for train-

ing.

To compare CREST with PADRE[59], we also simulated seven different pedigree structures that include the second degree pair and mutual relatives consisting of (again with respect to genetically older sample): one first cousin and his/her sibling (1C+S); one first cousin and his/her child (1C+C); one first cousin and his/her niece/nephew (1C+N); one first cousin once removed and his/her sibling (1C1R+S); one first cousin once removed and his/her child (1C1R+C); one first cousin once removed and his/her niece/nephew (1C1R+N); and one second cousin and his/her sibling (2C+S). We tested both methods using 200 replicate pedigrees of each structure for all three types of second degree relatives.

We further evaluated CREST's inference sensitivity and specificity across genome coverage bins. For this analysis, we simulated 200 copies for each relationship type of the same nine pedigree structures we used for training (above). We then mapped these to genome coverage bins and randomly downsampled to obtain 200 copies per bin. To generate calibration curves, we performed another five batches of simulations of the same nine pedigree structures and analyzed 1,000 pairs for each bin following random downsampling.

Parameters used to run each method

To collect IBD segments for the relationship type of CREST, we first ran IBIS v1.19.1 with default parameters on the simulated data. Since PADRE requires results from ERSA[28] and PRIMUS[58] as inputs, we ran them separately on the simulated data. To run PRIMUS (v1.9.0), we first used the `--no_IMUS` and

`--no_PR` options, which corresponds to only running PLINK[10] (v1.90b2k) to calculate relatedness estimates. We then filtered the output file from PLINK to only include pairs from the same pedigree. Next we ran PRIMUS on this file to reconstruct pedigrees, allowing it to search for up to second degree relatives using the `--degree_rel_cutoff 2` option (all simulation pedigrees it applies to include only first and second degree relatives). Meanwhile, ERSA needs inferred IBD segments from GERMLINE[20] as input, while GERMLINE works on phased data, so we ran Eagle[40] v2.4 to phase the simulated unphased genotypes.

Each of the Ped-sim simulation runs for the PADRE comparison generated data for 200 pedigrees for all three relationship types, and each pedigree includes data from four samples, for a total of 2,400 samples output by one Ped-sim run. After running Eagle on these 2,400 samples separately for all seven of the pedigree structure types used to compare CREST and PADRE, we ran GERMLINE v1.5.1 with the options `-err_het 2 -err_hom 1 -min_m 1 -bits 64` as specified in the ERSA paper. Then we ran ERSA v2.1 with default settings on the GERMLINE output for each dataset.

After all these steps, we ran PADRE v1.0. We found that PADRE initially crashed in some tests, with the source of the crashes being some of the pedigrees PRIMUS inferred, so we removed the pedigrees that cause the crashes from consideration by PADRE (as in another PADRE analysis[50]). This avoids calling these tests as PADRE failures, thereby improving its performance.

In a separate test, to exclude the possible effects of phasing quality on PADRE's

results, we simulated replicates of the same pedigree structures and used the true haplotypes produced by the Ped-sim `--keep_phase` option, keeping the subsequent analysis steps the same as described above.

Runtimes on the simulated data are from the same server configuration as in the real data tests (below).

Real data processing

To test CREST's relationship type inference on the GS dataset, we ran IBIS v1.20 using `-maxDist 0.116131` and otherwise with default parameters. The `-maxDist` option sets the maximum genetic distance between SNPs and can reduce false positive segment calls[53]. Following this, we used CREST to analyze the second degree relatives that IBIS inferred and excluded potential double cousins or other pairs that potentially violate CREST's assumptions by requiring the IBD2 sharing fraction between these pairs to be less than 0.02 (a default CREST option). We also restricted CREST's analysis to mutual relatives that are third to sixth degree relatives of both members of the target pairs since IBIS has been validated on relatives up to sixth degree[53]. (Note that we used all mutual relatives for the analyses of simulated data.)

Some GS samples are part of multiple target pairs—for example, one grandparent can have several grandchildren resulting in several GP pairs—and we averaged the classification results across those pairs for each relationship type. The reason for this is that each sample shares the same IBD segments with his/her relatives regardless of which pair CREST analyzes it in, so the ratios of pairs with overlapping members are correlated. Thus we averaged the sensitivity

and specificity of all pairs that have the same genetically older sample in GP and AV pairs, and also averaged the results for HS pairs with the same common parent (Figure B.4). For instance, for a grandparent with four grandchildren, each pair contributes a count of 1/4 towards the sensitivity and specificity metrics. We calculated the averages within relationship types, so a given sample can be both a grandchild and a half-sibling, with results from the two types considered independently.

The runtimes we report are from servers with four Xeon E5 4620 2.20 GHz processors, and we ran IBIS with eight threads on the real data. (CREST is not multithreaded.)

3.4 Results

To evaluate CREST's ability to distinguish among second degree relationship types, we first compared its performance with that of PADRE using simulated pedigrees. We also used simulated data to characterize CREST's performance across variable genome coverage rates; its ability to infer directionality for PC, AV, and GP pairs; and its potential to classify third degree relationship types. To validate CREST in real samples, we ran it on the GS dataset and compared its inferred second degree relationship types with those of the reported relationships.

Classifying second degree relationship types using CREST and PADRE

We tested CREST and PADRE using simulated data from seven different types of pedigrees. These pedigrees include the target second degree pair and two of their mutual relatives, and we define them by the relationship of the mutual relatives to the genetically older target sample: 1C+S, 1C+C, 1C+N, 1C1R+S, 1C1R+C, 1C1R+N, and 2C+S (Methods). PADRE was designed to infer degrees of relatedness but can be used to classify relationship types of close relatives when given data from their more distant relatives[50]. In fact, its accuracies for inferring second degree relationship types are higher than those previously reported from RELPAIR[16], a close relationship type classifier (below). More specifically, PADRE assigns the degrees of relatedness that maximize the composite likelihood between two sets of close relatives, this likelihood being the product of (a) the PRIMUS-inferred pedigree likelihoods[59] for each close relative set and (b) the pairwise relatedness likelihoods[28] between members of different sets. We read off the second degree relationship type of the target pair from the corresponding maximum composite likelihood PRIMUS pedigree. PRUMIS pedigrees must contain at least two closely related samples to work, and PADRE analyzes a pair of related PRIMUS pedigrees. Thus, all the simulated pedigrees we used to compare PADRE and CREST include the target second degree pair and two mutual relatives that are first or second degree relatives of each other. However, we note that CREST works even with only one mutual relative of the target pair.

We ran both CREST and PADRE on 200 replicates of each of the pedigree structures. As noted in Methods, PADRE crashed for some tests, and we applied

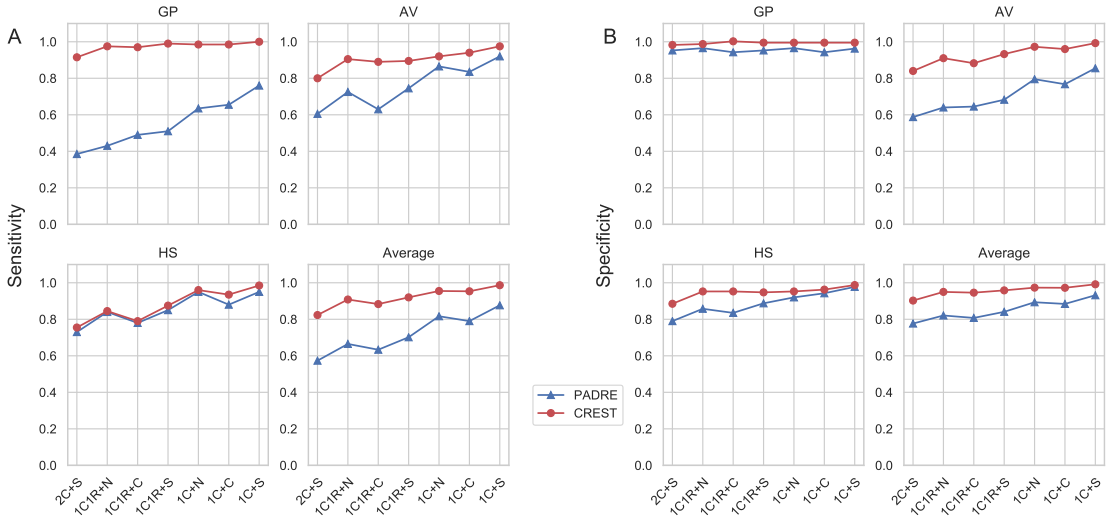

Figure 3.3: Performance of CREST and PADRE for second degree relationship type classification. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types in simulated data, along with the average of these rates across the three relationships. The x-axis indicates the mutual relative types included in the analysis (abbreviations in Methods), with each target relationship type and mutual relative combination including data from 200 target pairs.

a previously used fix[50] that enabled it to analyze most of these cases, but it continued to crash for 2.10% of the pedigree structures. In turn, for 0.830% of pedigrees, CREST did not infer a type due to: IBIS not inferring the target pair as second degree relatives, $C < 0.025$, or $R_1 = R_2 = 0$ (Methods). To account for the effects of these pairs, we show classification results both with and without the unclassified pairs.

Figure 3.3 plots the sensitivity and specificity from all 200 pedigrees for the seven types of pedigree structures. (If a tool did not classify a target pair, we scored it as having a sensitivity of 0 and a specificity of 0.) CREST's overall sensitivity (Figure 3.3A) ranges from 0.915-1.00 for GP, 0.800-0.975 in AV, and 0.755-0.985 in HS pairs across the seven types of mutual relatives. In contrast, PADRE's overall sensitivity is 0.385-0.760 for GP, 0.605-0.920 in AV, and

0.730-0.950 in HS pairs. This corresponds to an increase in sensitivity of 0.110-0.250 in CREST across all mutual relative types, averaged over the three target relationships (Figure 3.3A). Turning to specificity (Figure 3.3B), CREST's performance rates are 0.978-0.995 in GP, 0.885-0.993 in AV, and 0.903-0.988 in HS, while PADRE's rates are 0.943-0.965 in GP, 0.589-0.855 in AV, and 0.790-0.978 in HS. Averaged over the three relationship types, CREST's specificity is 0.060-0.130 higher (Figure 3.3B). When only considering the subset of pairs that both PADRE and CREST classify (97.1% of pairs), PADRE's average sensitivity and specificity over all relationship and pedigree types increase, respectively, by 0.016 and 0.019 (Figure B.5). CREST's comparative performance remains similar, as its sensitivity and specificity are 0.101-0.250 and 0.051-0.125 higher on average, respectively.

To determine whether phasing quality adversely impacts PADRE's results, we compared CREST and PADRE on another 200 replicates of the same pedigree structures but used perfectly phased haplotypes output by the simulator. This step should not affect CREST's performance since IBIS ignores phase information. Use of these optimal haplotypes improves PADRE's sensitivity by 0.039 on average, and most especially improves its sensitivity for GP pairs, by a range of 0.105-0.330 (Figures B.6, B.7). Nevertheless, CREST's average sensitivity is still 0.107-0.203 higher in these data, and its specificity is 0.059-0.116 greater, averaged over the three relationship types.

In general, for the types of mutual relatives we tested, both CREST and PADRE perform well at classifying HS pairs, while CREST has higher sensitivity for AV and GP pairs. PADRE's high performance in HS pairs may be because the

mutual relatives are equally close to the target samples for this relationship type. Alternatively, previous work indicated that PADRE may be biased against GP relationship classification and in favor of HS[50]. Along these lines, the confusion matrices show that PADRE misclassified more GP pairs as AV when given more distant mutual relatives (Figures B.8, B.9). In turn, CREST tends to mix HS and AV classifications, and is better at identifying GP pairs.

Considering the runtime of these analyses, the IBD detector IBIS ran on the 2,400 samples simulated for each of the seven types of mutual relative classes in an average of 11.2 CPU minutes (single threaded), and CREST completed its classification in another 1.75 minutes on average. On the other hand, the pre-processing steps for running PADRE require that the samples be phased, have IBD detected (with GERMLINE), and be analyzed using both PRIMUS and ERSA. Phasing using Eagle and ERSA together take more than two CPU days to finish processing data from one of the mutual relative type simulations.

The performance of CREST under variable genome coverage rates

As discussed in Methods and depicted in Figure 3.3, classification using close mutual relatives has better performance than using more distant relatives. To ensure that CREST's KDE distributions more accurately represent the true relationship probabilities for a given target pair and their mutual relatives, we trained stratified KDEs based on the genome coverage rate C of a set of mutual relatives (Methods).

Figures 3.4A and 3.4B show the sensitivity and specificity of CREST in simulated data across the same bins of genome coverage rates on which we trained separate KDEs. As expected, the sensitivity and specificity both increase as the coverage grows. For coverage rates between 0.125-0.15, or roughly that expected when using one first cousin, CREST's sensitivity and specificity are both 1.00 for GP, 0.983 and 0.957 for AV, and 0.913 and 0.992 in HS pairs, respectively. Even when C is in the lowest bin of 0.025-0.05, CREST still achieves sensitivities and specificities, respectively, of 0.928 and 0.985 for GP, 0.819 and 0.789 in AV, and 0.650 and 0.924 in HS pairs. Notably, the inference of GP pairs generally has quite high sensitivity and specificity regardless of the genome coverage rate. This is likely because, if x_i is the grandchild, in theory $R_i = 1$, with no variance from the meioses that separate x_i from the grandparent, but only due to false positive and/or false negative IBD segments.

The results above consider only the highest posterior probability relationship as the type that CREST infers, but this probability is informative about CREST's confidence and can be used in applications of the method. Figure B.10 depicts calibration curves for each relationship type in each genome coverage bin. In general, CREST gives reasonably well-calibrated probabilities across bins, though there are some biases evident for HS and AV pairs for lower coverage values. GP probabilities are well calibrated regardless of the coverage, while the probabilities for AV and HS are well-calibrated for coverage rates larger than 0.125. For lower coverage rates, the probabilities are still informative, especially for values near 0 or 1.

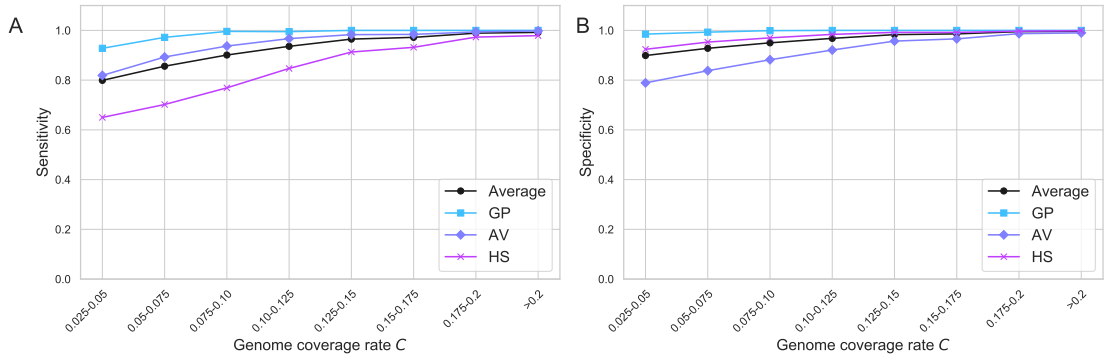

Figure 3.4: CREST performance on simulated relatives. (A) The sensitivity and (B) specificity within genome coverage rate (C) bins for GP, AV, and HS pairs, and the average across these three types.

Detecting the directionality of relationships

To test CREST's ability to detect the directionality of relationships, we used the same simulated pedigree structures as in the above genome coverage analysis, but instead of analyzing HS pairs, we took their common parent and one of the half-siblings to serve as PC pairs. We applied the KDE classifier to infer which sample is the grandparent, aunt/uncle, or parent in 200 pairs for each genome coverage bin. As shown in Figure B.11, averaged over all pairs with $C > 0.025$, or roughly using one fifth degree or more closely related mutual relative, CREST achieved sensitivity of 1.00 in determining the directionality of GP pairs, 0.99 for AV, and 1.00 for PC pairs. Moreover, the probabilities from this test are nearly perfectly calibrated (Figure B.12).

CREST has the potential to infer third degree relationship types

In principle, the CREST approach need not be limited to second degree relationships, as a similar logic applies to more distant relatives. To analyze the

potential for CREST to distinguish third degree relatives, we tested its ability to classify four third degree relationship types: great-grandparent (GGP), grand-avuncular (GAV), half-avuncular (HAV), and first cousin (1C). Assuming that x_1 is the genetically older sample, for a GGP pair, $E[R_1] = \frac{1}{8}$ and $E[R_2] = 1$; for a GAV pair, $E[R_1] = \frac{1}{8}$ and $E[R_2] = \frac{1}{2}$; for a HAV pair, $E[R_1] = \frac{1}{4}$ and $E[R_2] = \frac{1}{2}$; and for a 1C pair, $E[R_1] = E[R_2] = \frac{1}{4}$.

To train and test this extension of CREST, we simulated 1,000 pedigrees for each of the third degree relative types, with each pedigree including two first cousins of the genetically older sample as mutual relatives. After calculating R_1 and R_2, we trained KDEs using 800 pairs and five-fold cross validation for each type. We then tested on the remaining 200 pairs, and found that the inference accuracy is high, with sensitivities of 0.990 for GGP, 0.940 for GAV, 0.925 for HAV, 0.975 for 1C pairs (Figure B.13). Furthermore, the classification probabilities are well calibrated (Figure B.14). Thus, CREST has potential utility to distinguish relationship types even for third degree pairs given sufficient mutual relative data.

Validation in Generation Scotland data

In order to test our model in real data, we used CREST to classify second degree relationships in the GS samples, which are enriched in close relatives and include reported pedigree structures. Analyzing these data required 2.8 hours to run IBIS using eight threads, 2.7 CPU minutes to infer relationship types. We considered those pairs that IBIS detects as second degree relatives and who have at least one sufficiently related mutual relative for performing relationship type inference (Methods).

When analyzing CREST's performance for inferring relationship types, we found a few relative pairs it confidently infers as having a conflicting type, so we inspected the pairs using sample ages and IBD sharing to other relatives. For two pairs, CREST shows strong evidence that they are AV instead of HS and GP as reported (inferred probability of 1.00 with $C = 0.162$ and $C = 0.121$, respectively). For the pair labeled as GP, we found that the (ungenotyped) intermediate parent is listed as five years younger than his labeled father, indicating that this pair cannot be GP and supportive of the AV type. The other pair was labeled as paternal HS, but, denoting the individuals as A and B, we found that individual A has IBD sharing with B's maternal relatives (A is a fourth degree relative of B's maternal first cousin), and, in turn, B does not share IBD segments with A's maternal aunt. This indicates that they cannot be either paternal or maternal HS. In addition, B is 24 years older than A, supporting CREST's prediction of an AV relationship. A third case concerns a set of labeled maternal HS pairs, where we found that purported paternal first cousins of some of these samples are in fact their niece and nephew. We confirmed this by calculating an $IBD^{(011)}$ rate of 129 cM; this is a signal DRUID uses to detect aunts and uncles of two or more siblings, with a threshold of 50 cM reliably discriminating aunts and uncles[50]. However, after correcting this part of the pedigree, we noticed other inconsistent degrees of relatedness among the relatives, and the true relationship of the labeled HS pairs is difficult to determine. We therefore excluded this entire pedigree (which contains only 1 reported HS pair after averaging) from our analysis. After relabeling the HS and GP pairs as AV and removing the noted pedigree, the relationship type analysis includes 233 GP, 2,616 AV, and 344 HS pairs.

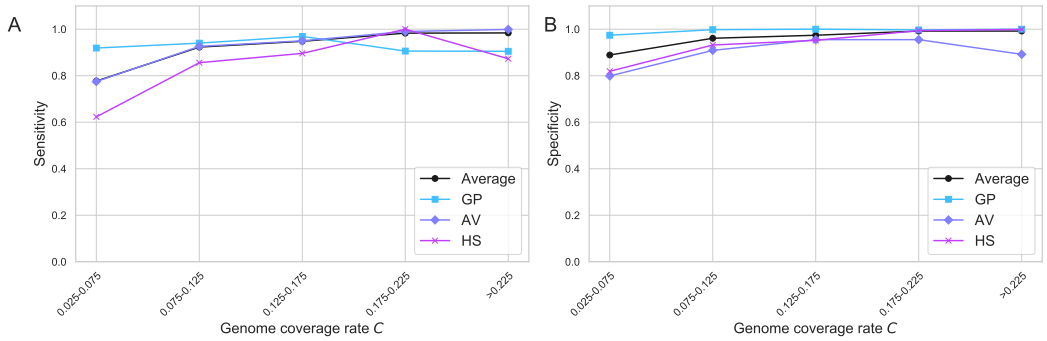

Figure 3.5: CREST performance on the Generation Scotland data. (A) The sensitivity and (B) specificity of relationship type classification for GP, AV, and HS pairs, and the average across these three types in the GS dataset. These plots use a genome coverage rate (C) bin size of 0.05 because several bins have a small number of HS and GP pairs with a bin size of 0.025 (minimum of 7 for HS and 7 for GP using 0.025 vs. 14 and 16 here).

Figures 3.5A and 3.5B plot CREST's relationship type inference sensitivity and specificity in the GS data across different genome coverage rates C. As expected, both the sensitivity and specificity tend to increase with C. Overall, for $C > 0.125$, CREST's sensitivity is relatively high at 0.935 for GP, 0.977 for AV, and 0.922 for HS pairs. Similarly, the specificity is high in this coverage range, with values of 0.999 for GP, 0.937 for AV, and 0.979 for HS pairs. However, relative to the next lower coverage bin, the sensitivity of GP pairs drops when $C > 0.175$, and that of HS pairs drops for the $C > 0.225$ bin. For the GP pairs, these last two bins include only 1.5 and 2 misclassified pairs (after averaging), and for HS pairs, the last bin has 1.75 misclassified pairs. These misclassifications are due to CREST using mutual relatives that either: (a) include another grandchild of the grandparent that IBIS infers as a third degree relative of the grandparent, or (b) violate CREST's MRCA assumptions but only occur with three or more generations of sample collection (e.g., a great-grandchild or descendants of a HS member's full sibling). We note that GS's recruitment provides more of the latter category of relatives than is typical for population-based studies[7],

so fewer assumption violations may occur in population samples. Still, extending CREST to detect mutual relatives that violate its MRCA assumptions is the subject of future work.

3.5 Discussion

Pedigrees have wide ranging utility throughout genetics, with the modeling of transmitted haplotypes among relatives and/or the use of their IBD sharing fractions being central to both linkage analysis and recent heritability estimation procedures[71,73]. Family data are also needed to identify *de novo* recombinations[3,9,21] and mutations[48,52], and to enable family-based phasing and imputation, the gold-standard means of addressing these problems[5].

Given these applications, several methods exist for pedigree reconstruction and for confirming or disproving reported pedigree relationships[16,22,37,58,59,61]. However, differentiating among the relationships that map to a given degree of relatedness has remained challenging. Pairwise relatedness measures, the standard signal for detecting relatives until recently[49], have limited information to enable the classification of relationship types[16].

We developed CREST, an approach that infers both pedigree relationships and directionality. CREST assumes that mutual relatives connect to both members of a target pair only through one or more MRCA(s) of the target pair. To enforce this assumption, which is most readily violated by descendants of the MRCA(s), CREST does not analyze first and second degree relatives of the target pair. However, such close relatives carry IBD segments that span a large frac-

tion of a target sample's genome—i.e., they have high coverage rates—and so have the potential to be very informative for relationship type inference. On the other hand, in the GS dataset, some relatives that violate the MRCA assumption are more distantly related than first or second degree, and CREST's use of these samples lowered its performance in the high coverage rate bins (Figures 3.5A and 3.5B). We view the proper utilization of such samples as a subject of interest for future work.

At present, CREST does not require age information even though the difference in age of the target pair is also informative for distinguishing among relationship types. However, the age difference distribution in the GS data reveals large overlapping ranges between HS and AV pairs, and between AV and GP pairs (Figure B.15). Still, straightforward extensions of CREST may benefit from use of ages when they are available.

Here we applied CREST to simulated and real relatives using IBD segments detected with IBIS. In both forms of data, so long as the mutual relatives do not violate CREST's assumptions, the method appears relatively insensitive to errors in the IBD segments. Nevertheless, the quality of IBIS and other IBD detectors depend on several factors, including SNP density. Therefore, users must be careful to ensure that the detected IBD segment quality does not adversely impact CREST. One way to accomplish this is to simulate relatives with properties such as marker density and population membership similar to the target samples and tune the IBD detector's parameters accordingly to ensure that CREST's performance matches the user's goals.

While this paper was under review, PONDEROSA[70]—a method for pedigree reconstruction and second degree relationship type inference in endogamous populations—was released. PONDERSA uses highly reliable phased IBD segments to make inference, leveraging both segment numbers and whether the segments reside on only one haplotype in order to distinguish among types. These signals are distinct from those that CREST uses, and PONDEROSA is therefore complementary to CREST. Indeed, depending on haplotype phase quality and the availability of mutual relatives, one approach may shed light on a pairs' type when the other method falls short.

As direct-to-consumer genetic testing companies provide customers with estimated relationships among samples, CREST has several uses. Most apparently, it can enable these companies to report specific relationship types, including which parent an individual is related through for some relationships. Additionally, while the mutual relatives of a target pair inform the pedigree structure between the pair, providing this pedigree structure to the method DRUID can enable more exact detection of the distance between those close relatives and their more distant mutual relatives[50]. Thus, an iterative procedure is possible, with mutual relatives of unknown relationship to a set of close relatives enabling the detection of the latter pairs' relationship types, and the resulting pedigrees enabling more precise characterization of their distance to the mutual relatives.

Lastly, a key factor influencing CREST's performance is the genome coverage rate of the available mutual relatives. In general, more closely related pairs will have a higher genome coverage. Consequently, with ever increasing sample

sizes—and therefore datasets with greater numbers of relatives, including close relatives—CREST's inference of relationship types will have greater reliability going forward.

Acknowledgements

We thank Archie Campbell for help in evaluating the relationships of Generation Scotland individuals, Reka Nagy for deciphering the original Generation Scotland pedigree information, Daniel Seidman for support in using IBIS, and Giulio Genovese for his observations about inferring parent-child directionality. We also thank Shai Carmi for discussions regarding time-dependent Poisson rates and for pointing out that non-IBD regions in half-sibling and grandparent-grandchild pairs have analogous properties to the interior of IBD segments. Funding for this work was provided by NIH grant R35 GM133805, an Alfred P. Sloan Research Fellowship, and a seed grant from Nancy and Peter Meinig to A.L.W. J.S. was partially supported by National Institutes of Health grants T32 GM007617. C.H. is supported by an MRC University Unit Programme Grant MC_UU_00007/10 (QTL in Health and Disease). Computing was performed on a cluster administered by the Biotechnology Resource Center at Cornell University. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z). This study makes use of data generated by the Wellcome Trust

Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

Web Resources

CREST, https://github.com/williamslab/crest

Data and Code Availability

The code generated during this study is available at https://github.com/williamslab/crest.

Genotype data for Generation Scotland subjects are available through an application from https://www.ed.-ac.uk/generation-scotland/.

The simulated genotype data supporting the current study have not been deposited in a public repository because they are easy to reproduce from Ped-sim and give no more relevant information but are available from the corresponding author to request.

CHAPTER 4

RECONSTRUCTING THE GENOTYPES OF PARENTS FROM SIBLINGS AND OTHER RELATIVES

4.1 Abstract

The opportunity exists to reconstruct partial genomes of the ancestors of genotyped individuals by identifying identical by descent (IBD) segments shared among genotyped relatives, noting that these segments must have been transmitted through ungenotyped ancestral individuals. Inferring genotype data for these ancestors has the potential to empower genome-wide association studies (GWAS) by adding ungenotyped samples, improve relationship inference, and allow phenotypic estimates in ancestors, among other applications. We propose a novel approach to infer the genotypes of ancestors using a combination of family-based phasing and IBD sharing. Specifically, we develop HAPI-RECAP to infer genotypes of parents from a set of genotyped children and their relatives using an extension of the original HAPI method. This extension first enables HAPI to jointly phase multiple genotyped siblings even without data from parents. By combining this phase information with inferred IBD among the siblings and one or more genotyped relatives, we are able to further resolve ambiguous phase and assign the haplotype segments to the two parents. Validated with San Antonio Mexican American Family Studies (SAMAFS) data, HAPI-RECAP reconstructed 80.3% to 100.0% of genotypes for one missing parent, when the other parent and three or more children are available; 67.8% to 72.6% of genotypes for two parents, using genotypes of four to seven siblings and their relatives; above 94.0% for two parents in families with eight or more

children. The error rates of reconstructed genotypes are near or below 10^{-3}, which are comparable to genotyping error rates.

4.2 Introduction

The genomes of individuals can be considered as a mosaic of segments inherited from different ancestors. Regions of shared mosaic patterns among two or more relatives—so called identical by descent (IBD) segments—provide genetic information for ancestors that must have transmitted these segments. The process of locating these IBD segments and identifying which ancestor the segments descend from would enable inference of partial genomes of ungenotyped ancestors. Such an approach has the potential to not only resolve the geographic origins of those ancestors' genetic segments[31], but also to improve relatedness inference[32,50], and increase power in genome-wide association studies (GWAS) by adding data from ungenotyped individuals[30,38,72]. This is especially meaningful when the case is less attainable like in rare diseases. In addition, reconstructed parents genome can be used to study direct and indirect genetic effects[72]. With high quality reconstructed ancestral genomes, it is also possible to infer some traits of ancestors.

Reconstructing ancestral genomes has drawn increasing attention given the wide applications, however, previous methods usually require the complete and accurate pedigree or population information[11,18,60]. Recently Jagadeesan et al.[31] reported the reconstruction of 38% of the maternal genome of Hans Jonatan, a man born in 1784, relying on data from 182 of his genotyped descendants and a deep genealogically-derived pedigree from Iceland. This example demonstrates the potential for inferring the regional origins of ancestors

by reconstructing their genetic segments. However, this is a special case since African DNA was not common in Iceland in the 18th century and this work was empowered by the reliable identification of African segments. Moreover, such large numbers of genotyped descendants and complete pedigrees required by previous methods are often unavailable in practice.

In comparison to methods that focus on the reconstruction of a single distant ancestors of many individuals, it is possible to reconstruct considerably larger fractions of the genomes of more recent ancestors, even with data from a smaller number of individuals. Health information may also be more easily attainable for recent ancestors, making them good candidates for downstream analyses that rely on reconstructed genomes. For example, Kong et al.[38] imputed two phased haplotypes composed of 1,001 SNPs in an ungenotyped deceased lung cancer patient and showed that the imputed region harbored variants associated with lung cancer. Thus, imputation of ancestral genotypes in GWAS has been done previously, and use of ungenotyped samples with phenotype data holds great promise. Moreover, large genetic datasets contain a significant number of close relatives and provide a good opportunity to retrieve genetic information of their ancestors. Notably, parent-child pairs and full siblings are common relationships that can be identified very accurately even without pedigree information. While each child inherits only half of the genomes of each parent, independent segregation and recombination are randomized such that n siblings will inherit on average a proportion of $1 - \frac{1}{2^n}$ of both parents' genomes. Thus, the opportunity exists to reconstruct partial genomes of parents from a set of genotyped children.

We developed a novel approach to reconstruct the genotypes of parents using a combination of family-based phasing of a set of siblings and IBD sharing to other close relatives. The problem of inferring the genome of a parent includes two main components: first, phasing a set of siblings using a family-based phasing approach. This step provides rough haplotype data for the parents, up to ambiguities in parental assignment of genotypes. This is mainly caused by multiple possibilities of phasing given children' genotypes. The second component is to assign inferred haplotype segments to correct parents by leveraging IBD information between genotyped siblings and relatives of one or more parents. For the first component, we use an extension of HAPI, a method for inferring minimum recombinant haplotypes in nuclear families[69]. When genotype data for one or both parents is missing, the new version of HAPI enables joint phasing of siblings and infers haplotype segments of the missing parents using only the children's genotype data. This joint phasing has low error and, when given data for 8 or more siblings, often provides chromosome-scale haplotypes for the parents. For more moderate numbers of siblings, the phasing results in a number of multi-megabase long segments where which parent they belong to needs to be resolved. To resolve the parental origin and further reconstruct genotypes for the second part, we leverage IBD segments shared between the children and other relatives. The basis of this inference is that, if we assume that a relative is related only to children through one parent, the IBD shared between children and their relative should only come from that one parent. Thus, we treat the IBD regions as a reference to distinguish which parent the genome segments descend from. When an IBD region occurs, we can infer not only the haplotypes of the parent those segments were transmitted by but those that belong to other parent in the region spanned by the IBD segment as well. Taken all

steps together, we developed the new tool, HAPI-RECAP (REConstruct Ancestral genotyPes), to reconstruct genotypes of parents.

To test our approach, we applied this method on data from the San Antonio Mexican American Family Studies (SAMAFS)[15,29,42]. We restricted our tests to the families with both parents available as we would need real data to compare with. The validation includes three scenarios: using one parent and three or more children in 116 families; using four to seven children and their relatives in 64 families; using eight or more children in ten families. We reconstructed genotypes of missing parent(s) in each scenario and compared with the real data to evaluate how much genotypes HAPI-RECAP can accurately reconstruct (See Results). Varied by different scenarios, HAPI-RECAP is able to reconstruct large portion of parental genotypes with similar error rates to direct genotyping. As large-scale datasets lead to the recruitment of family data, our work holds promise to enable high quality reconstruction of parent genotypes, opening the door to further analyses using inferred genotypes from individuals not directly collected.

4.3 Results

To evaluate the reconstruction quality, we compared the reconstructed genotypes with the real sequence data for each site and investigated two main measurements: the reconstructed coverage and the error rate. The reconstructed coverage is defined as $\frac{N_2+0.5*N_1}{N_0} * 100\%$, here N_1 and N_2 are the number of inferred sites only on one copy and on both copies respectively. N_0 is the total number of sites for real sequence data that passed quality checks. The error rate quantifies the chance that one inferred site is incorrect and is defined as

$\frac{2*N_{2d}+N_{1d}}{2*N_2+N_1}$. N_{1d} and N_{2d} are the number of inferred sites that differ from real sequence on only one copy and on both copies, respectively. Note that it is possible to have two copies inferred but only one copy is identical to the real sequence.

Reconstructing genotypes for the missing parent with only one parent available

We analyzed 116 families with both parents from SAMAFS by reconstructing each parent's genotypes, assuming that one parent is missing, in each family. The number of genotyped children in these families varies from 3 to 12, with the average of 4.57. The reconstructed coverage, i.e., the number of the sites that are reconstructed divided by the total number or sites, for all 232 parents varies from 80.3% to 100%. As shown in Fig 4.1, with as few as three children in 42 families, HAPI-RECAP reconstructed 87.5% of one perent's genotypes on average, with the range from 80.3% to 91.7%. When there are more children available, the amount of reconstructed genotypes also increase. With eight or more children, HAPI-RECAP is able to reconstruct at least 98.8% of one parent's genotypes. When comparing reconstructed genotypes with the real data, the error rates is below 10^{-3} on average, with 36,338 differed among over 114 million SNPs.

Reconstructing genotypes for both parents in large families

In the case that both parents are unavailable, we first tested HAPI-RECAP in ten families with eight or more children. As shown in Table 4.1, HAPI-RECAP is able to reconstruct 94.2% of both parents' genotypes on average, with the

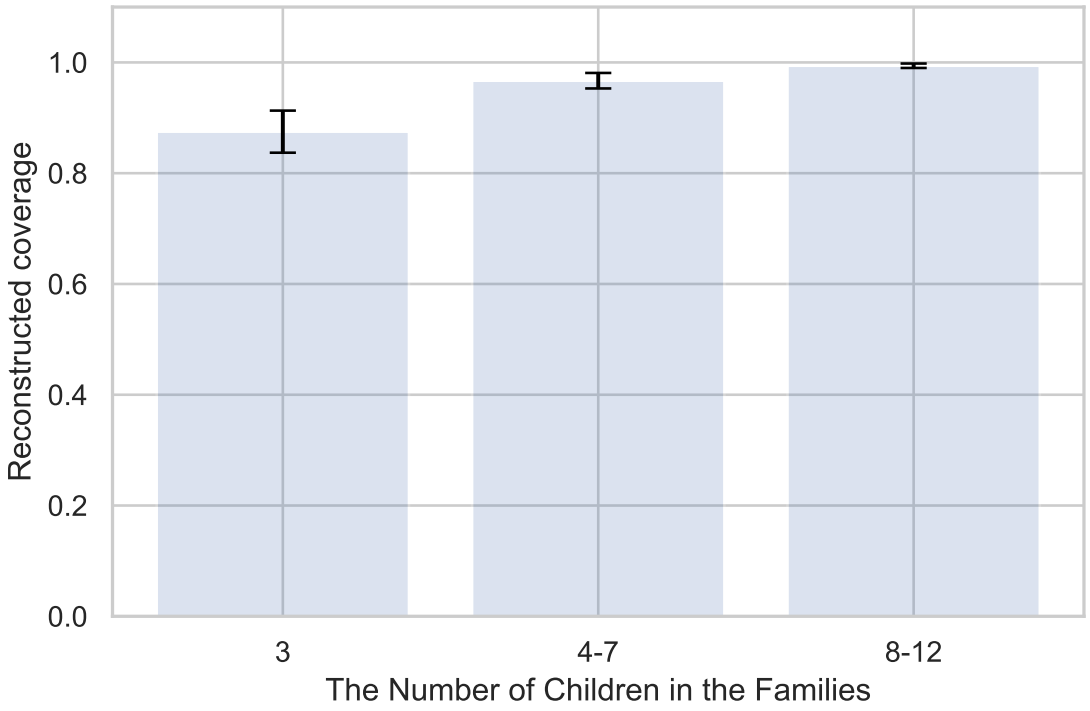

Figure 4.1: The Coverage of reconstructed genotypes when one parent is available Each bar is the average reconstructed coverage over the families with one parent plus three children, four to seven children, and eight or more children, respectively. The error bar is the standard deviation.

range from 94.0% to 95.2%, using only genotypes data of children. The error rate for the reconstruction is below 10^{-3}, which is comparable to the genotyping error rate. For the largest family in the SAMAFS dataset with 12 children, HAPI-RECAP reconstructed 95.2% of the two parents' phased haplotype with the error rate of 1.1×10^{-4} (Fig 4.2). Note that HAPI-RECAP does not infer the site where parents are both homozygosity, since all children would have the same genotypes at those sites and provide almost no information of parents. Considering the proportion of homozygosity sites, HAPI-RECAP can reconstruct almost parents' entire genome when the number of children is large.

Number of siblings	8	9	10	11	12
Coverage	94.0%	94.0%	94.3%	94.4%	95.2%
Error Rate (10^{-3})	1.0	1.0	0.26	0.18	0.11

Table 4.1: Average Coverage and error rates of reconstructed parental genotypes in large families.

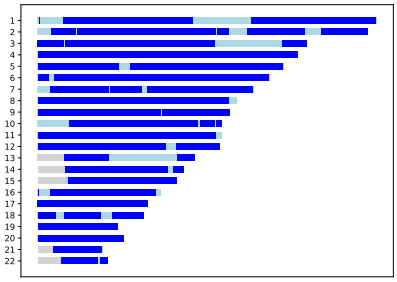

Figure 4.2: Reconstructed haplotypes of one parent in the 12 children family The blue segments represent reconstructed two copies in 22 chromosomes. The regions where only one copy is reconstructed are in light blue.

Reconstructing genotypes for both parents in small families

For families with four to seven children, HAPI-RECAP leverages the IBD segments between children and their relatives to resolve ambiguities. The IBD segments are collected from a group of inferred relatives, which are related to children only through one parent (see Methods). This analysis includes 27 four children families, 23 five children families, 9 six children families, and 5 seven children families, with detected second to sixth degree relatives of these children. We cluster these relatives to ensure that we use IBD segments from relatives only related to the children through one parent (see Methods). As shown in Fig 4.3, as expected, the increase of the number of available children help with the reconstruction. HAPI-RECAP reconstructed 67.8% to 72.6% of the genotypes for the parent which is related to the group of relatives, on average over different families with the same number of children, and 65.4% to 71.3% of the genotypes for the parent that is unrelated. We notice that there is a small disparity between related parent and unrelated parent, with higher reconstructed coverage for the parent that is related to the group relatives. That is related to the regions where

only one haplotype is inferred and could be considered belonging to the parent that is related to the group of relatives, if the IBD regions are consistent. The error rates of reconstructed genotypes for both parents range from 10^{-4} to 10^{-3} per site.

Note that the ability of HAPI-RECAP to reconstruct genotypes also depends on the availability of IBD segements. We thus evaluate how the amount of IBD segments will influence the reconstructed regions with four children families. To increase the number of four children families, we down-sampled the number children in five to seven families, resulting in 69 families (See Methods). As shown in Fig C.2, the IBD coverage, the total genetic length of IBD segments divided by the total length of genome, is highly associated with reconstructed coverage. While the error rates of reconstructed regions are still low and do not change much as the IBD coverage decreases.

When using IBD segments as the reference to reconstruct parental genotypes, we calculated the ratio of difference between similarities of haplotypes from two parents compared to IBD regions (See Methods). This ratio is also related to the reconstructed genotypes quality and can used as a threshold to achieve lower error rates. We showed the reconstructed genotypes coverage and error rates as the change of this ratio in Fig C.3. As the ratio increases, it is more strict when reconstructing genotypes, corresponding to lower coverage and error rates.

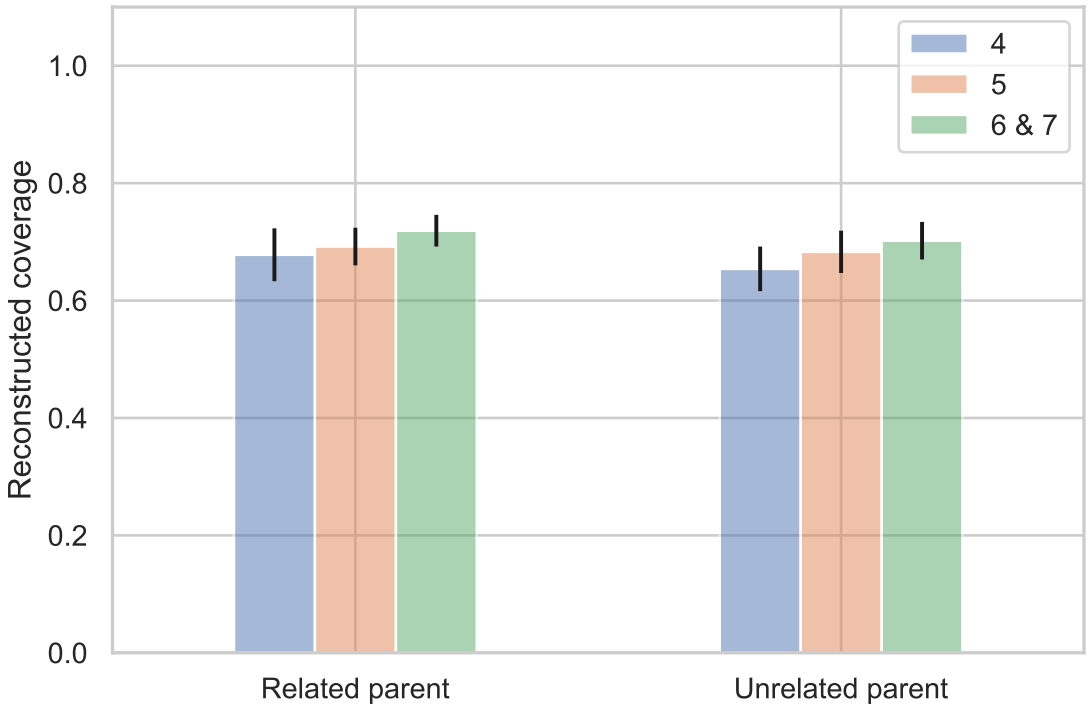

Figure 4.3: The Coverage of reconstructed genotypes for two parents Average reconstructed coverage for both parent that is related the relatives and the other unrelated parent, when there are four, five, and six or seven children in the families. The error bar is the standard deviation.

4.4 Methods

Data processing

We used SNP array data from the the SAMAFS[15,29,42], followed by the quality control filters carried out in the previous studies[68]. This process includes mapping the SNP array probe sequences to GRCh37 and more detailed steps. After these procedures, we also filtered out SNPs with more than 2% missingness, and samples with greater than 10% missing sites. Since the MZ twins share the same genotypes, we also remove six individuals to keep only one in each MZ twins. 2479 individuals at 521,184 SNPs remains in the dataset.

To get the IBD information of relatives, we ran IBIS[53] on all individuals to infer both IBD segments and relatedness. Yet HAPI-RECAP can accept high-quality IBD segments derived from any method. We ran IBIS v1.20 using default parameters and `-maxDist 0.116131` additionally. This `-maxDist` parameter determines the maximum genetic distance between SNPs, which can help to prevent false positive segment calls.[53].

In this analysis, we only consider families with three or more children and both parents available for the validation purpose. For the families with four to seven children, we selected relatives of siblings in each family that are inferred as second to sixth degree from IBIS. In the case where the relative lists are different for children in the same family, we considered the intersection of these relatives. To better evaluate smaller families, we created more four children families by down sampling the five to seven children families. For five and six children families, we randomly dropped one or two children to get four children families. For seven children families, we first split the children into two groups randomly: one has three and the other has four children. Then we randomly select one child from the four children group, combining with three children group to form a new four children families. This allows us to create two four children families for each seven children families, with only one children overlapping.

Joint phasing of siblings

We used an extension of HAPI[69] to jointly phase siblings and reconstruct the haplotype segments of missing parents. HAPI uses heavily optimized hidden Markov Model for phasing nuclear families and can infer minimum recombi-

nant phase in polynomial time. The extension of HAPI is more robust to genotyping errors and allows for missing data in the parents, and can jointly phases siblings and infer parents' haplotypes even without data from parents. It takes plink format data and outputs phased haplotype of siblings and reconstructed parental genotypes. In this analysis, we ran HAPI v1.92 under two scenarios: only one parent and three or more children are available; four or more children with no parent data. We used the `--no_err_max 1` option to restrict that maximum number of recombinations attributable to a single marker before it is called an error to be 1. It only takes 20 minutes for HAPI to analyze all of 116 families.

Resolving ambiguities with the relatives' IBD segments as reference

When there is one parent along with children available, the extension of HAPI is able to reconstruct the other parent's genotypes. In large families with eight or more children, the extension of HAPI can successfully generate chromosome-scale haplotypes for missing parents. In smaller families, however, the extension of HAPI would output a number of reconstructed segments, and which parent these segments belong to is unclear. We propose to use the IBD regions between children and one parent's relatives as reference to resolve the parental origin of these segments. Assume we also have one grandparent available, the grandparent must transmit some portion of his/her genome to children only through one parent, when there is no inbreeding between two parents. Thus, the IBD regions between this grandparent and children must be three-way IBD sharing with this parent as well. By comparing the relative's genotype with re-

constructed segments in these IBD regions, we can distinguish which segments belong to the parent that is related to the relative. This also implies that the other reconstructed segments in the same regions would belong to the parent that is unrelated to the relative. In practice, the grandparents are not always available, however, this idea can apply to any relatives as long as they are related to children only through one parent. Note that most second to sixth degree relatives, except double cousins, etc., would have IBD sharing residing only on one haplotype. As a result, the relative and one of the children would only have one haplotype that are identical. One problem is that IBIS and several other IBD detectors do not detect which haplotype these segments belong to. To be able to use relatives' genotypes as the reference, we only consider homozygous sites in relatives. In this way, we do not need to phase relatives and distinguish the haplotype that is IBD, and high density homozygous sites can still provide enough information to distinguish two parents.

Once the parental origins of the reconstructed segments have been determined, we can connect these segments to reconstruct chromosome-scale genotypes of parents. Technically, one parent should have the identical genotype with the relative in the IBD region, however, there are genotyping errors and since we only use homozygous sites of relatives, both parents can share similar genotypes to the relatives to some extent. To control the reconstruction quality, we introduce a measurement to quantify the similarity of genotypes from two parents to the IBD regions. For each segment that overlaps with IBD regions, we count the different sites between genotypes of each segment with those of relatives in the overlapping regions. In particular, we define a ratio of difference as: $r = \frac{|D(P_1,IBD)-D(P_2,IBD)|}{N(IBD)}$, here, $D(x, y)$ is the number of different sites between

genotypes of x and y, $N(IBD)$ is the total number of sites in the overlapping IBD regions. This ratio quantifies the relative difference of two parents compared to the IBD regions. The larger this ratio, the more similar that one parent is to relatives than the other. Thus, it is more confident to distinguish the parent that is related to relatives and the parent that is not. If this ratio is too small, the homozygous sites in IBD regions probably does not provide enough information to distinguish two parents and it can lead to misclassification. To use this ratio as a threshold, we can only rebuild the regions where there are enough differences so that we have lower error rates. As a result, there will be a trade off between the coverage and error rates of reconstructed genotypes. In the meanwhile, there are segments where reconstructed genotypes are similar enough between two parents, regardless of the IBD regions. In this case, the parental origins have little influence on the reconstruction results, therefore, we retrieve these regions with the ratio of reconstructed genotypes differences between two parents below 10^{-3} to increase the coverage.

Collecting IBD segments from relatives

Since we need the genotypes of relatives as the reference to resolve the origin of reconstructed segments, it is important that we only use relatives that related to the same parent. For each relative from second to sixth degree to children in the same family, the IBD regions between this relative and each child will come from the same parent, assuming the two parents are not related to each other and the relative isn't related to both. Therefore, we first collect the union of IBD regions between one relative and all children and consider this union of the IBD regions as the IBD sharing between this relative and this family. However, when there are multiple inferred relatives, these relatives can related to children from

either parent and it is incorrect to attribute the IBD regions to the same parent. A straight forward solution is to only use the closest relative to the family, since this would give us the largest length of IBD regions. To get more IBD regions than of the closest relative, we instead chose to cluster these relatives according to the IBD sharing among them. When there is no inbreeding for two parents, i.e., no IBD sharing between two parents, and no common relatives of the father and the mother, the relatives on different sides usually do not share IBD segments with each other (See Fig C.1). We detected the IBD segments among relatives of each family and clustered those relatives which share IBD segments with each other into the same group. That is, if two relatives are also related to each other, we assumed they are both related to the children through the same parent.

More specifically, we implemented a modified depth first search (DFS) algorithm by considering the relatives as the nodes in a graph. If two relatives share IBD segments, we add an edge to connec them in the graph. DFS then can be used to find the connected components where any two relatives are connected to each other by paths. This may result in more than two components, since some groups can come from the same side, but the relatives in each group should be related to the same parent. However, common relatives of both parents will exist in theory because two parents will coalesce to the common ancestor eventually. These common relatives will connected groups on two sides of parents and cause confusion. In practice, the common relatives usually appear only as distant relatives, so we add a few steps to reduce the possibility: (1) when we connect relatives, we only allow for up to second degree relationship. That is, two relatives will be connected with a path if they are second degree or closer

relatives. This limitation will reduce the probability of too distant relatives connecting all relatives into one group. This is a parameter that can be changed and depends on the sparsity of relatives in the datasets; (2) for each clustered group, if there are multiple relatives, we take the union of IBD regions between each relative and the family. During this process, we further check whether overlapping shared IBD regions belong to the same or different haplotypes in a similar to a previous study[47]. As stated in that paper, if two relatives come from two sides of parents, and they share overlapping IBD regions with children, the overlapping part should happen in two haplotypes. Otherwise, when two relatives are on the same side, the overlapping regions should be a three-way sharing among two relatives and the child. Thus, we exclude the relatives that do not have a three-way sharing of overlapping IBD regions with other relatives. With these steps, we reduce the cases where multiple relatives come from different sides of parents and increase the IBD regions that can be used as reference. Note that it is possible that some relatives might pass our check and introduce uncertainty when they are different sides but share IBD segments on their the other haplotypes. For example, they could be half-siblings of each parents, but they share another common parent. However, the cases are not very common and we did not find this scenario in our dataset.

Validation of reconstructed parental genotypes

To evaluate HAPI-RECAP, we compared the reconstructed genotypes of parents with the real data in terms of the coverage and error rates. When there is one parent available or there are four to seven children, we count the sites where genotypes are inferred incorrectly and the sites that are not inferred. For large families with eight or more children, we keep the longest consecutive region

with no ambiguities about parental origins for each chromosome, and the regions with the ratio of reconstructed genotypes differences between two parents below 10^{-3} as well. Then we compared reconstructed haplotypes with directly collected genotypes for the parents.

4.5 Discussion

HAPI-RECAP is a fast, effective method to reconstruct parental genotypes from a set of siblings and relatives. Different than previous approaches, it does not need pedigree or population information, and takes in unphased genotype data. In large scale genetic datasets, such as UK biobank, the complete and accurate pedigrees are usually not available, while the proportion of relatives, including both siblings and close or distant relatives, would make it possible to apply this approach. The direct to consumer (DTC) genetic testing companies, such as 23andMe, also have enriched relatives in the datasets and would be interested in potential applications of reconstructed ancestors. Recently, researchers show increasing interests to involve the information of relatives in genetic studies, such as GWAS. Some studies use statistical frameworks to estimate dosage of parents from children or other relatives[30,38]. Our approach provides the accurate and informative reconstruction for ungenotyped samples, with the promise for similar applications.

Our validation in the SAMAFS dataset shows that HAPI-RECAP is able to effectively reconstruct a large portion of parental genome, given informative genotyped children and relatives. It can detect Mendelian errors and other forms of errors, however, the quality control checks on genotypes of the children and relatives are necessary for reliable results. The core model of HAPI relies on each

site to infer recombination and inheritance vectors, thus, if the genotype data is not in good quality, such as, too much genotyping errors or missing data in the family, it might lead to low quality reconstructions. On the other hand, HAPI-RECAP is quite robust to the detection of IBD regions. It is challenging to infer highly accurate start and end positions for IBD regions or shorter segments[14,53], but HAPI-RECAP is not very sensitive to IBD detection quality.

We notice that there are a few common cases that HAPI-RECAP has trouble with. First, HAPI-RECAP skips the sites where all children are heterozygous. In this case, genotypes of children can not provide any information about whether parental genotypes is homozygous or heterozygous. For example, if all children are A/C at one site, the genotypes of two parents can be A/A and C/C, A/C and A/C, or other possible cases. Second, it is possible that some regions of parental genome does not get inherited by any child at all, especially in small families. For one parent, the probability that one haplptype does not transit to any child is $\frac{1}{2^n}$, n is the number of children. If there is only one child, then half of the genome will not be inherited by the child for each parent. In this case, HAPI-RECAP can not infer the missing parental genome from children. This is also a reason that HAPI-RECAP performs better with more children. One possible approach to solve these issues is to involve the population information and impute these sites and regions, but this will require reliable reference panels. We view this as a subject of interest for future work.

In this paper, we present how to reconstruct parental genotypes from children, and this approach has the potential to apply to a large scale of relatives. If reconstructed parents have enough siblings available, it is possible to further

infer grandparents and go up more generations in an iterative way. Another idea is to involve other relatives directly, such as half-siblings, if they can be inferred correctly. Reconstructing DNA from the shared parent of half-siblings is quite possible, and would be made even more effective if data for the non-shared parent of one or more of those half-siblings is available. Using other types of relatives solely or combined would provide alternative information or even more accurate reconstruction.

Acknowledgements

We thank the San Antonio Mexican American Family Study participants that made this analysis possible. We also thank Ethan Jewett for helpful discussions. Funding for this work was provided by NIH grant R35 GM133805. Computing was performed on a cluster administered by the Biotechnology Resource Center at Cornell University.

Data Availability

The SAMAFS sample data are available on dbGaP under accession numbers phs000847 and phs001215.

CHAPTER 5

SUMMARY AND CONCLUDING REMARKS

The number of relatives in genetic datasets increase dramatically as the sample sizes exploded in recent years, bringing in intensive possibilities and opportunities for new discoveries. On the one hand, relatives and pedigree information enable valuable genetic analyses and applications, such as linkage analysis, disease and association mapping, population genetics, genealogy, and even forensic genetics. The enrichment of relatives in large genetic datasets, including UK biobank, make it meaningful to analyze these relatives directly other than removing them. Genetic testing companies such as 23andMe and AncestryDNA have been always analyzing relatives for genealogy findings. The importance and availability of relatives necessitate the development of efficient and accurate methods. On the other hand, the various relatives in the datasets allow for new approaches and methods to utilize the information, not limiting to the traditional approaches that only focus on pair-wise information. Using information among multiple relatives can improve the accuracy and enable new studies.

In this thesis, we explored the new approaches to characterize relatives in large datasets, utilizing IBD sharing information. In Chapter 2, we considered the possibility to add a new feature, the IBD segment numbers, to improve relatedness inference. To evaluate whether the IBD segment number could enhance relatedness inference, we conducted both a information theory based feature importance analysis and a classification analysis using a Bayes classifier. The theoretical view of information theory analysis suggests that IBD segment num-

bers added information to relatedness inference, however, the detection errors make the classification improvement limited in practice. This study augmented the understandings of the dependency among the IBD segment number, kinship coefficient, and relatedness inference; it also showed and encouraged the new attempts to apply information theory and machine learning to similar questions. In Chapter 3, we presented a novel approach—CREST—to identify pedigree relationships of close relatives. CREST differs from other approach in that it utilizes the multi-way IBD sharing among relative pairs and their mutual relatives. We thus came up with new features and built stratified models to classify relationship types in a machine learning approach. CREST achieved the state of the art performance and outperformed PADRE in simulation data, and obtained over 92.0% sensitivities for classifying second degree relative types when tested with sufficient mutual relatives in the GS dataset. We also showed that CREST can identify the genetic older samples in PC, GP, and AV pairs with over 95.0% sensitivities with the aid of mutual relatives. The performance of CREST supports the possibility to use mutual relatives in the datasets for high resolution inference. In Chapter 4, we proposed a new method, HAPI-RECAP, to reconstruct missing parental genotypes from genotyped siblings and relatives. For this challenging problem, HAPI-RECAP jointly phased siblings to infer parental genotypes, and used IBD regions between relatives and children as the reference when the number of children is not large enough. We validated HAPI-RECAP in SAMAFS data under three scenarios: when there are three or more children and one parent; when there are four to seven children and their relatives; when there are eight or more children. The results show that HAPI-RECAP is able to reconstruct the large portion of parental genome under these three scenarios with the error rates comparable to the genotyping errors. This analysis gives

the promise to even infer unavailable genotypes by retrieving information from current samples and relatives.

All these studies consider how to better utilize the information among relatives to solve problems with practical applications, and show the possibilities to apply machine learning and other computational approaches. For the future work, it is possible to continue applying the frameworks or methods to similar questions, such as evaluating other meaning features, extending to infer more distant relationships, or reconstructing multi generational ancestors. On the other hand, the results of these analyses have meaningful down stream applications. CREST not only provides the classification results, but also generates the probability for each class, which can be used to reconstruct pedigree and further connect relatives in the datasets. In addition, when there are more relatives available in the datasets, these methods can benefit from more information and provide even more accurate results.

APPENDIX A

SUPPLEMENTARY FOR CHAPTER 2

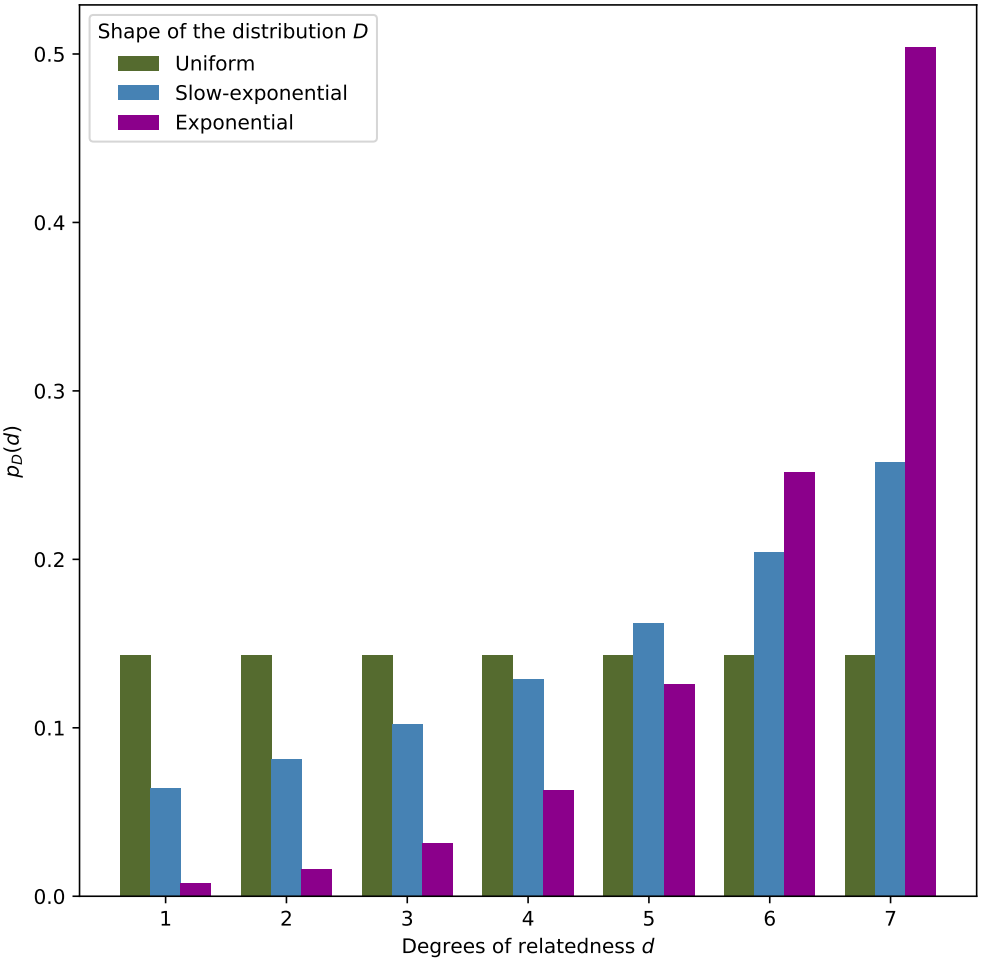

Figure A.1: Probability mass functions of different distribution shapes for D as a function of degree of relatedness d, where uniform=$1/7$, slow-exponential=$(1000/15541) \times 2^{(d-1)/3}$, and exponential=$(160/20320) \times 2^{d-1}$. Total pair counts for the testing data are 21,000 for uniform, 15,541 for slow-exponential, and 20,320 for the exponential.

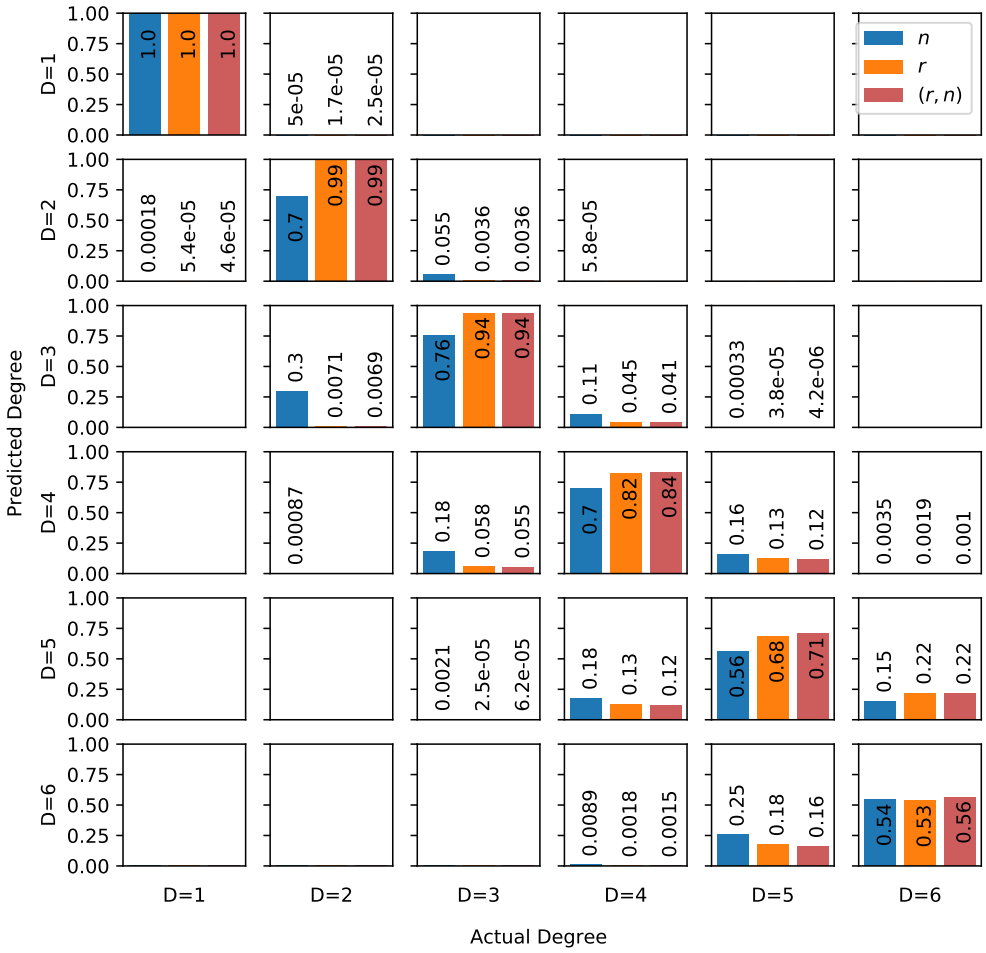

Figure A.2: Confusion matrix (with respect to degree of relatedness) of Bayes classifiers trained on exact segments with features n, r and (r, n) from the uniform distribution. Most misclassifications occur in diagonal-adjacent cells (off-by-one-degree misclassifications).

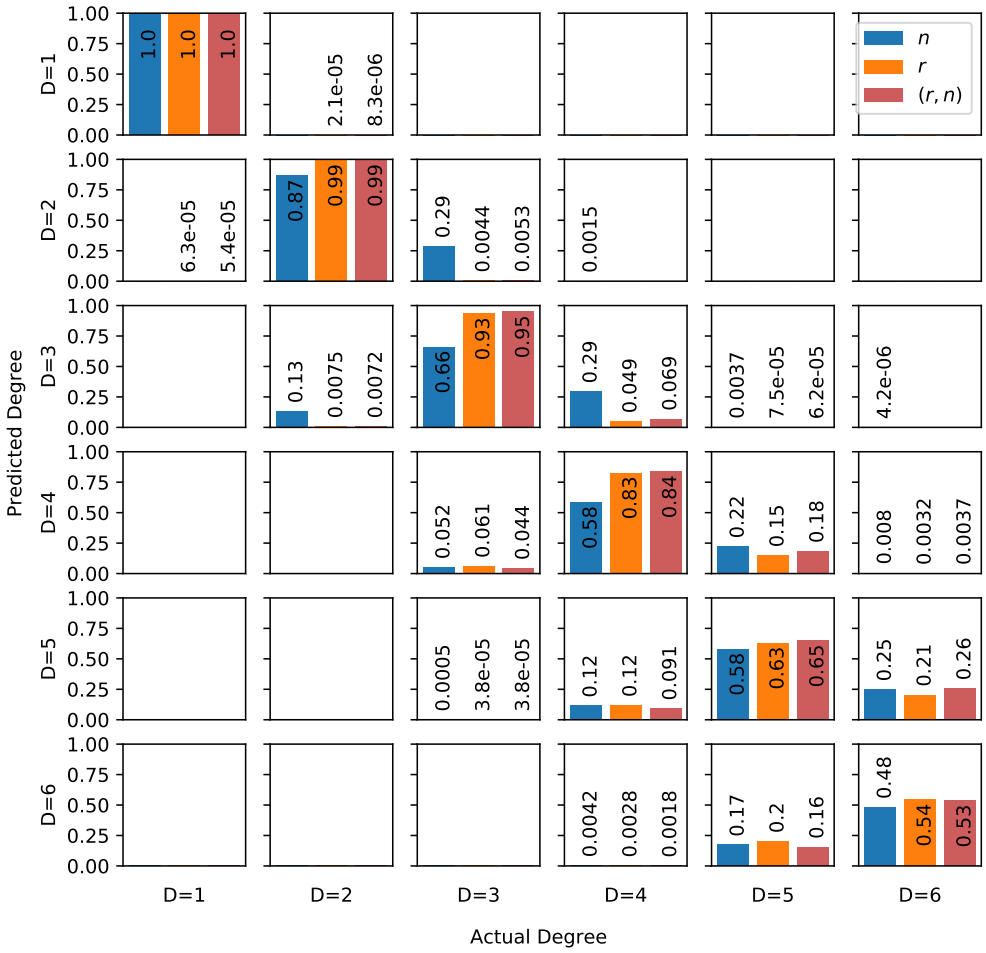

Figure A.3: Confusion matrix (with respect to degree of relatedness) of Bayes classifiers trained on inferred segments with features n, r and (r, n) from the uniform distribution. Most misclassifications occur in diagonal-adjacent cells (off-by-one-degree misclassifications).

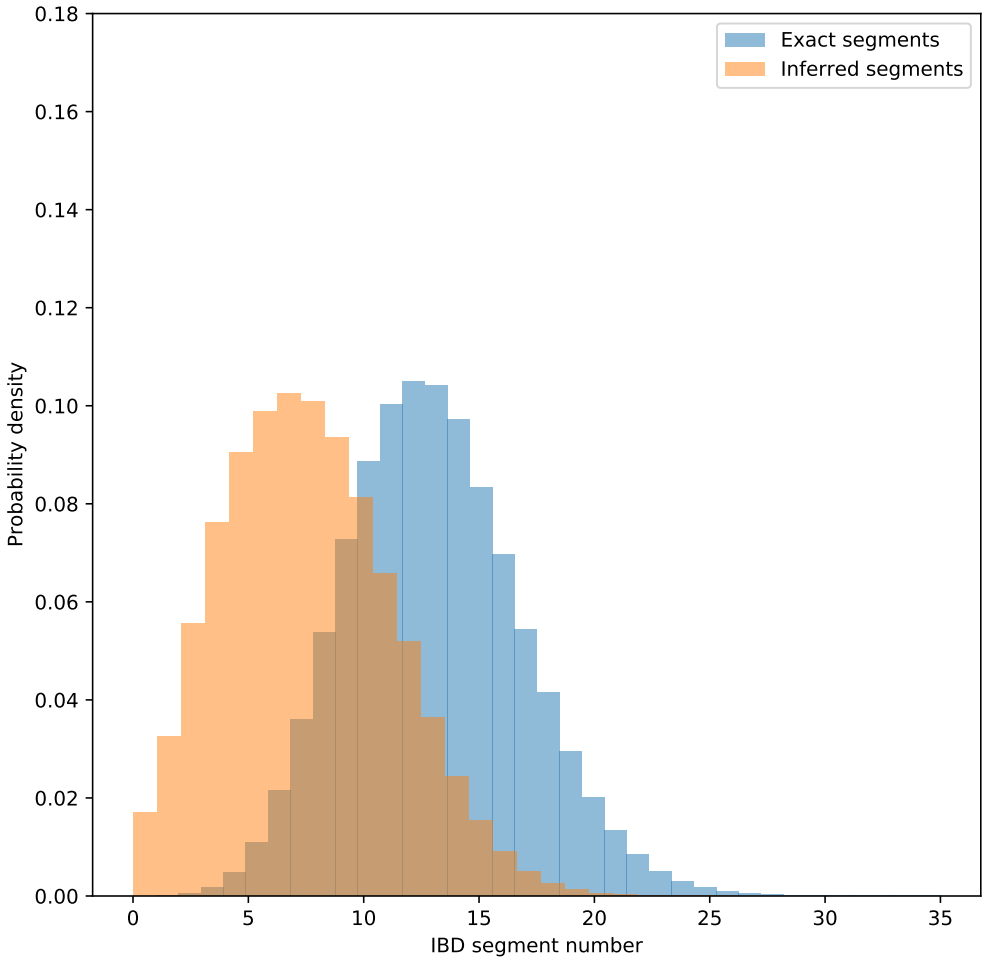

Figure A.4: Distributions of exact and inferred segment numbers in fifth degree pairs.

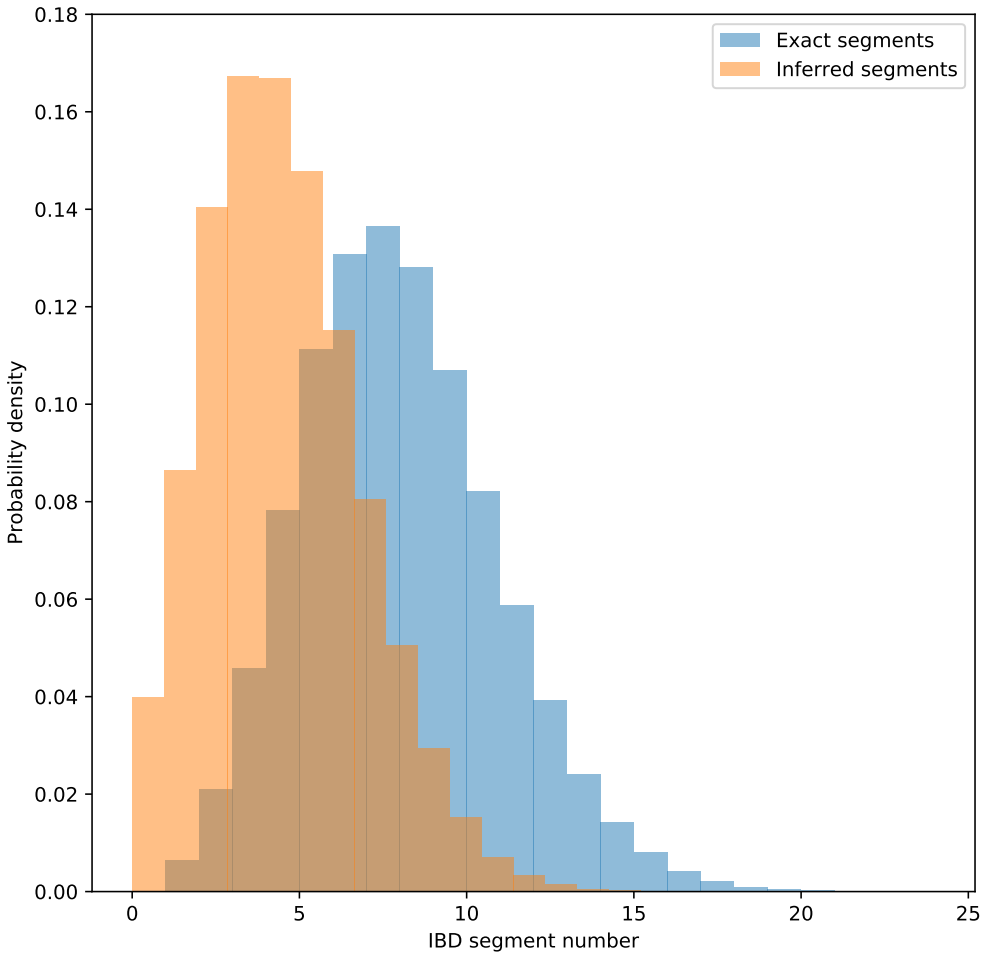

Figure A.5: Distributions of exact and inferred segment numbers in sixth degree pairs.

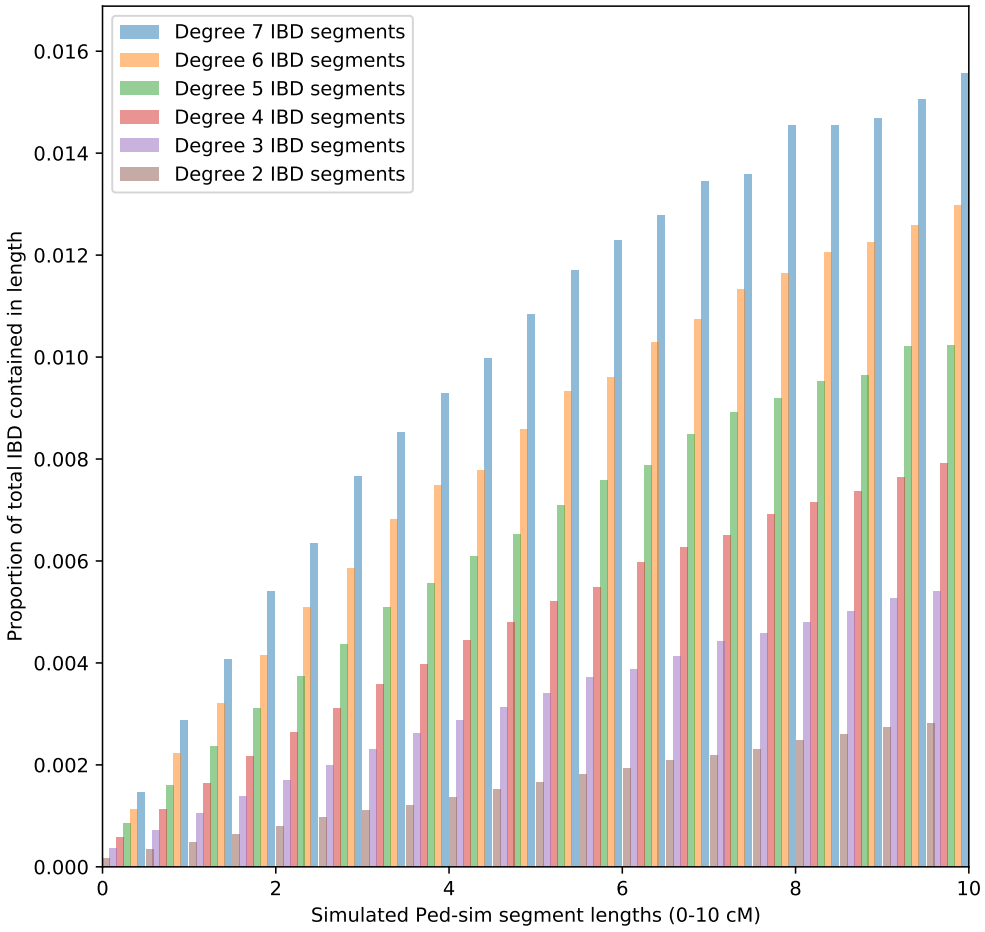

Figure A.6: Average proportions of pairwise total IBD length contained in exact segments of lengths 0-10 cM for relatives of the indicated degrees. Proportions calculated over 33,000 pairs from each degree.

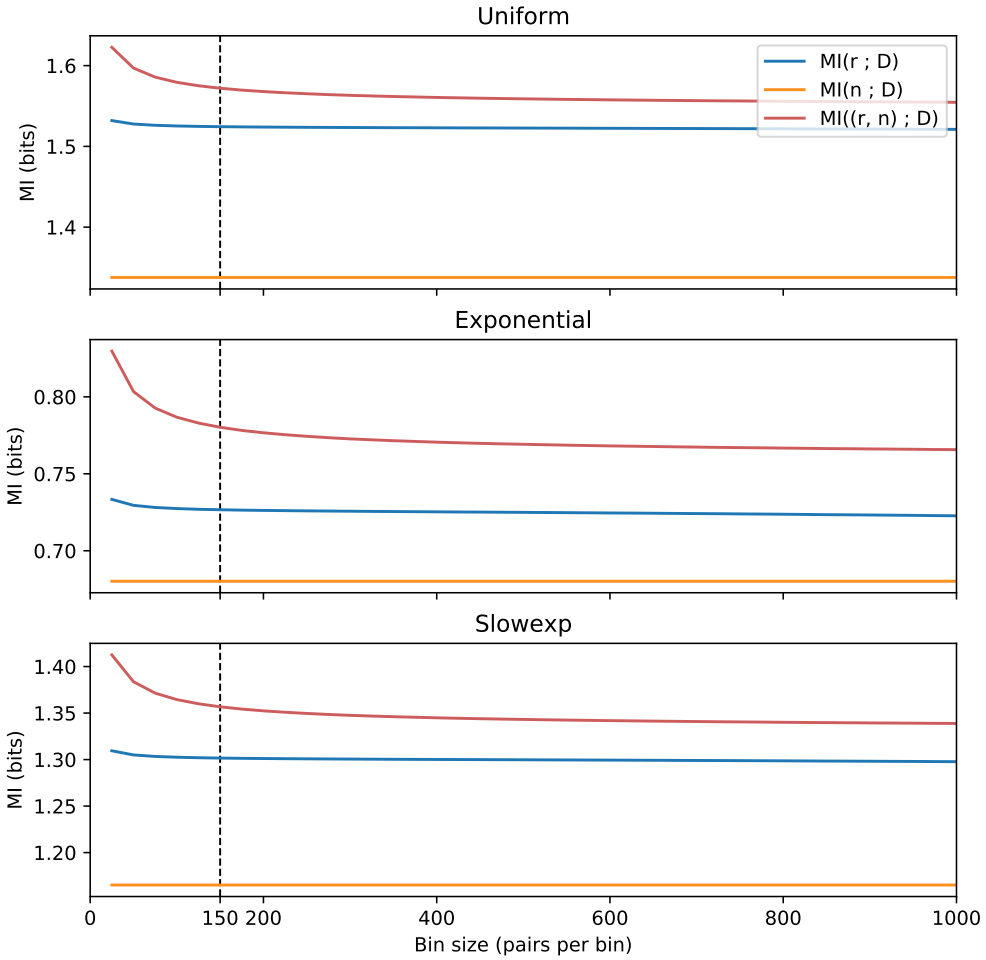

Figure A.7: MI of different feature sets as a function of bin size (pairs per bin), averaged over 80 independent simulations of exact segments from each of the distribution shapes. Slowexp corresponds to the slow-exponential distribution.

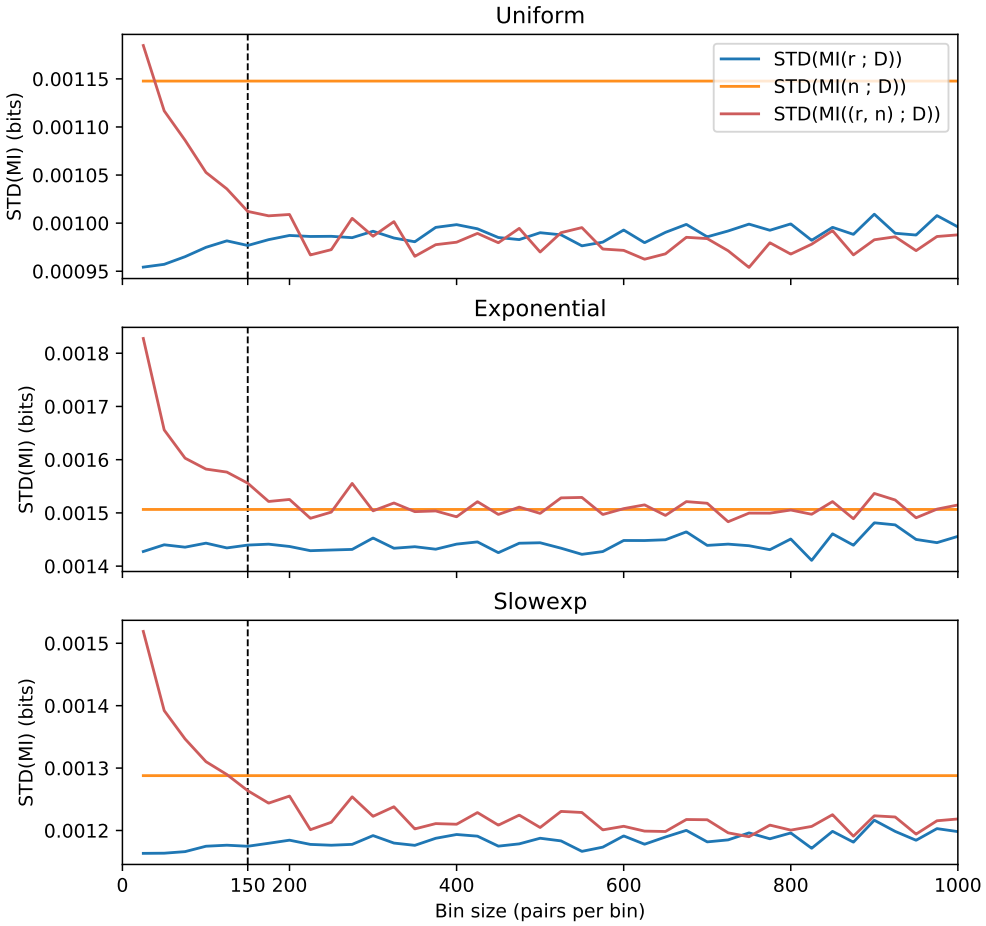

Figure A.8: Standard deviations of MI of different feature sets as a function of bin size (pairs per bin), averaged over 80 independent simulations of exact segments from each of the distribution shapes. Slowexp corresponds to the slow-exponential distribution.

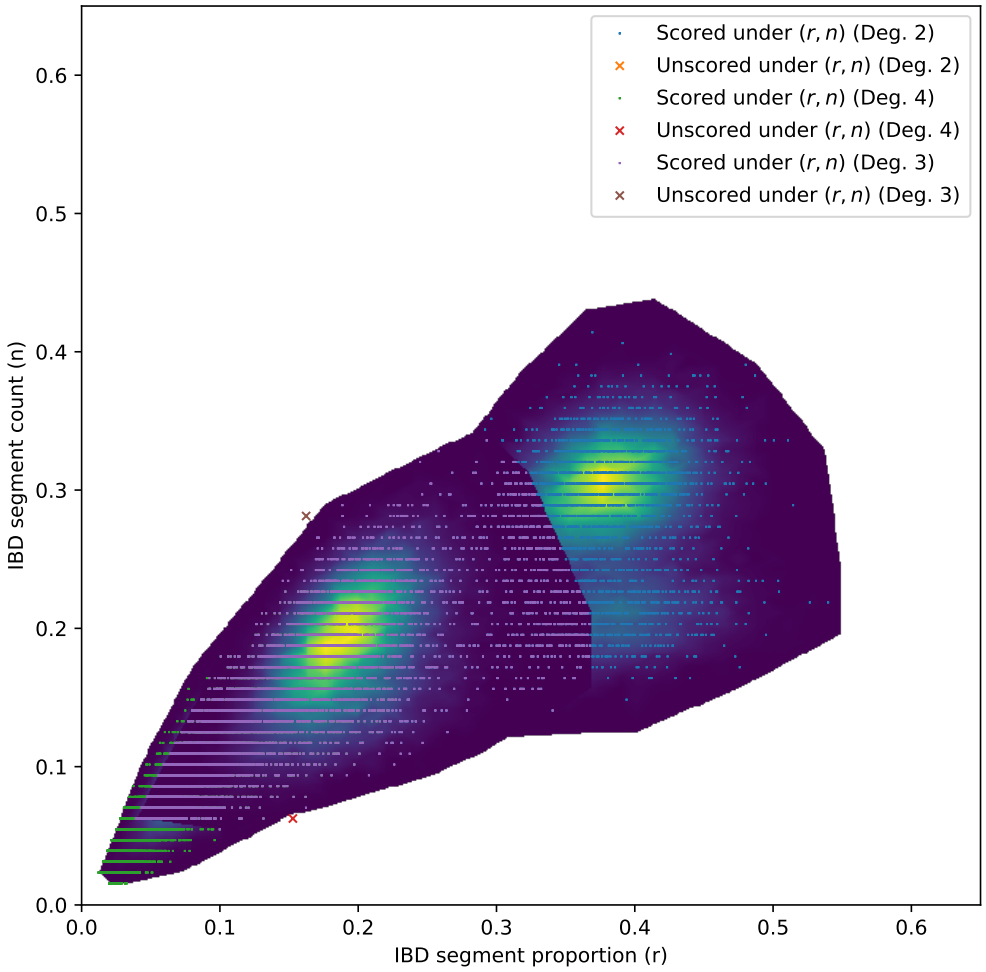

Figure A.9: Heat maps depicting posteriors $\hat{p}(D|f, \vec{T})$ for inferred IBD segments for several values of D. Generated using the `griddata` two-dimensional interpolation on $\hat{p}(f|D)$ calculated from training data. Overlaid are corresponding testing data points colored by their classification. Here, the IBD segment number n has been normalized to unity. Probabilities and points from higher degrees are plotted on top of those from lower degrees.

APPENDIX B

SUPPLEMENTARY FOR CHAPTER 3

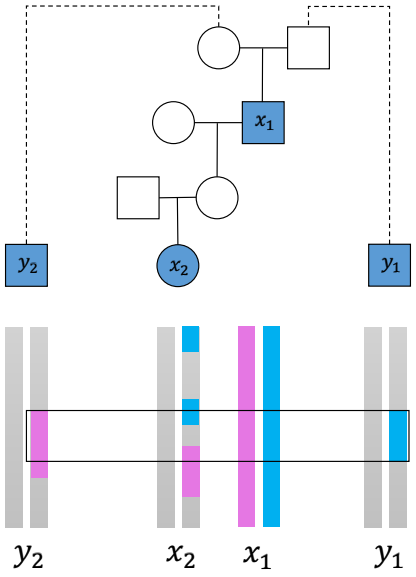

Figure B.1: Example IBD sharing between a GP pair and their mutual relatives on both the maternal and paternal sides of the grandparent. The mutual relatives y_1 and y_2 are related to the GP pair x_1 and x_2 through the grandparent's mother and father, respectively. The blue or purple regions represent either one haplotype of x_1 or IBD segments other individuals share with those haplotypes. The black box outlines the regions CREST deems as being IBD2 between x_1 and the mutual relatives.

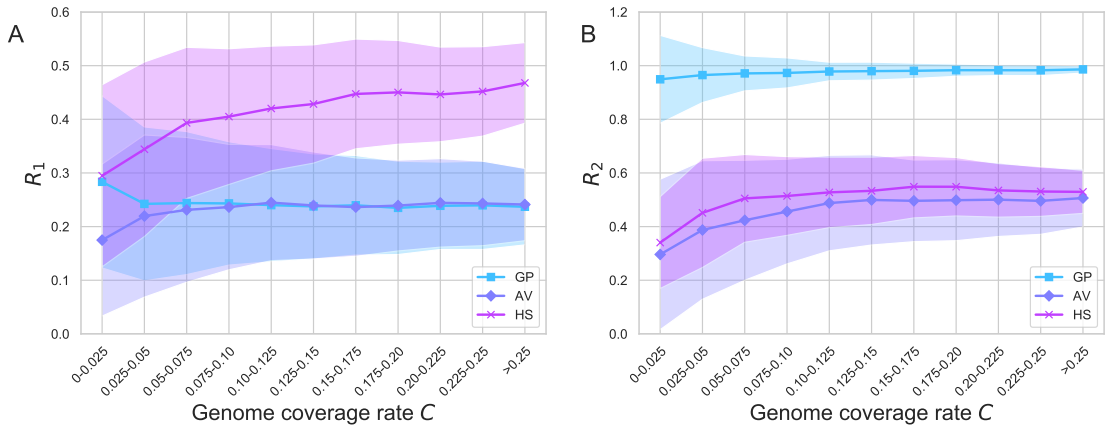

Figure B.2: The variance of ratios R_1 and R_2 decrease as the genome coverage rate increases. (A) R_1 and (B) R_2 values across bins of genome coverage rates. The dots show the mean value in each bin, and the shaded regions span one standard deviation from the mean. Results are from simulated data, with IBD segments detected in genotype data, for all three types.

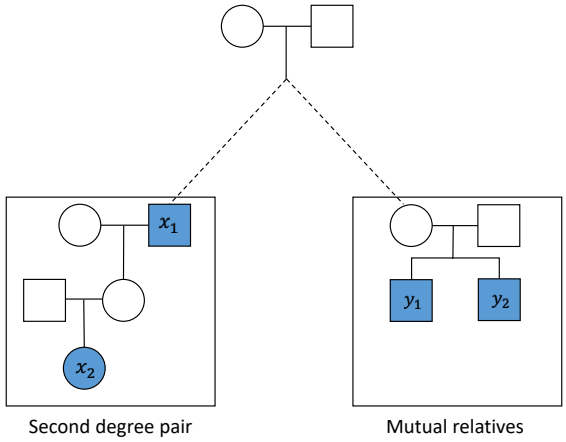

Figure B.3: The structure of the simulated pedigrees used to evaluate CREST's relationship type classification. The left side shows an example target second degree pair, which is either a GP, AV, or HS pair. The right side depicts example mutual relatives, which include one or more individuals that are related to the second degree pair and to each other (Methods). Genotyped samples are shown as filled shapes. The dashed line connects the second degree pair and the mutual relatives to their unknown MRCA.

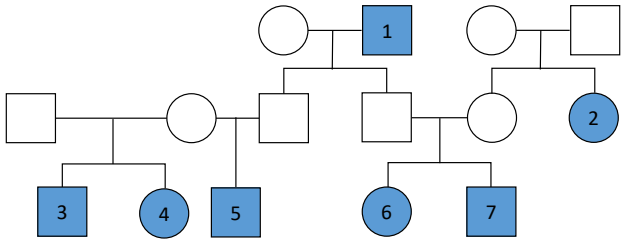

Figure B.4: Example pedigree with individuals contained in multiple second degree pairs. Sample 1 and each of samples 5, 6, and 7 are three GP pairs, while sample 2 and samples 6 and 7 are two AV pairs. The real data results average the sensitivity and specificity among all samples with the same genetically older sample for these types, so the three GP pairs would each contribute a count of $\frac{1}{3}$ to the GP metrics, and the two AV pairs would contribute $\frac{1}{2}$ to the AV metrics. In turn, sample 5 and samples 3 and 4 form two HS pairs with the same common parent, and the real data results similarly include average scores for such pairs, in this case weighting each by $\frac{1}{2}$. Note that sample 5 is a member of both a GP and HS pair, and the results consider each type separately, incorporating the average metrics for all pairs within each type.

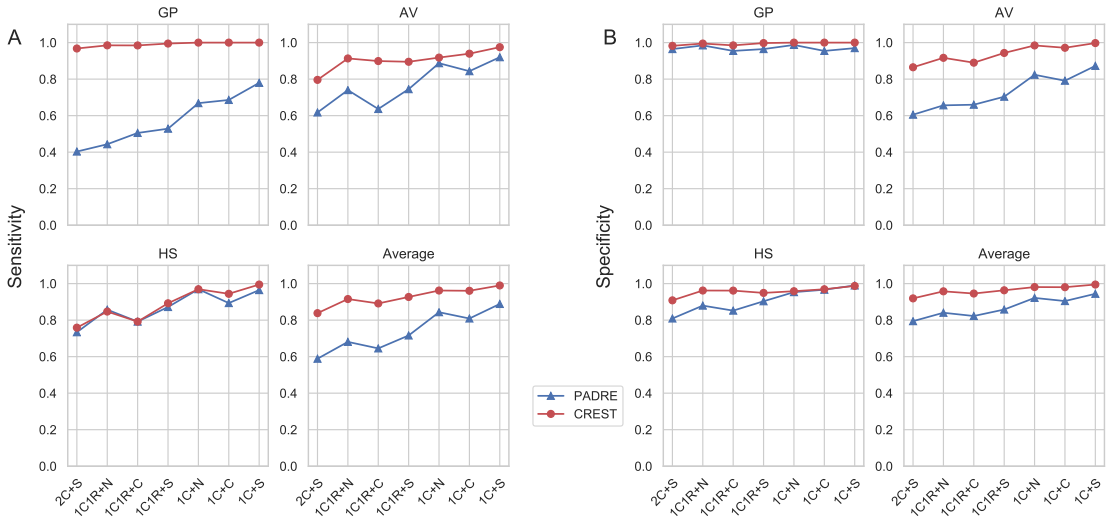

Figure B.5: Performance of CREST and PADRE for second degree relationship type classification of pairs both tools classify. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types, along with the average of these rates across the three relationships. The x-axis indicates the mutual relatives types included in the analysis, with each target relationship type and mutual relative combination including data only for those pairs (out of 200 per data point) that both PADRE and CREST classify.

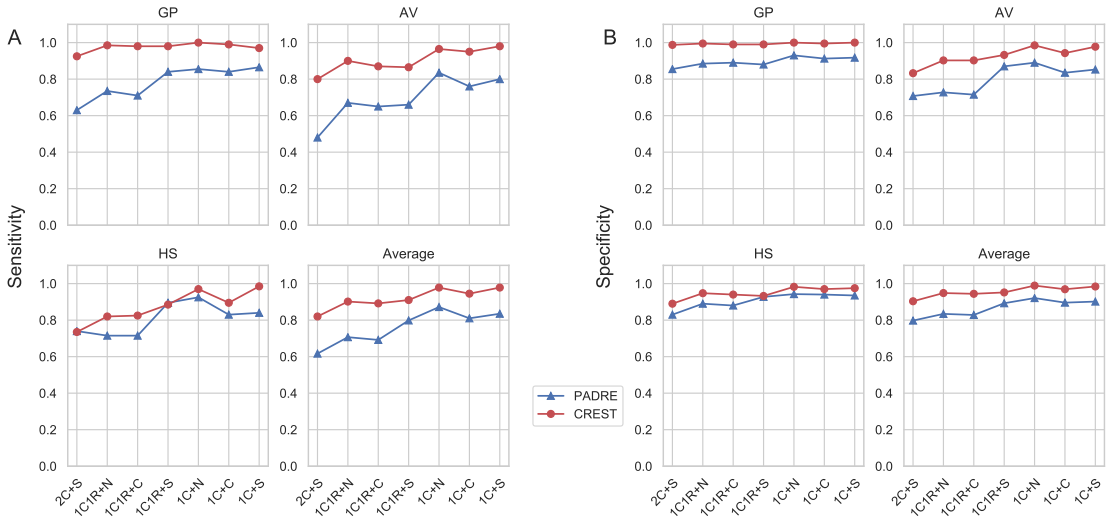

Figure B.6: Performance of CREST and PADRE for second degree relationship type classification where PADRE used perfect haplotypes. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types, along with the average of these rates across the three relationships. The x-axis indicates the mutual relatives types included in the analysis, with each target relationship type and mutual relative combination including data from simulated phased haplotypes of 200 pairs.

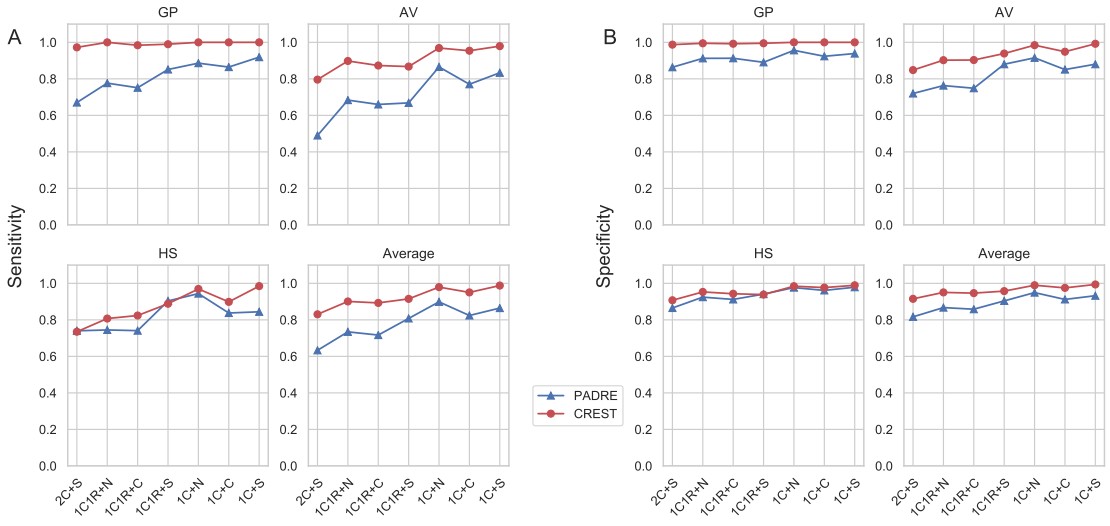

Figure B.7: Performance of CREST and PADRE for second degree relationship type classification of pairs both tools classify and where PADRE used perfect haplotypes. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types, along with the average of these rates across the three relationships. The x-axis indicates the mutual relatives types included in the analysis, with each target relationship type and mutual relative combination including data only for those pairs (out of 200 per data point) that both PADRE and CREST classify.

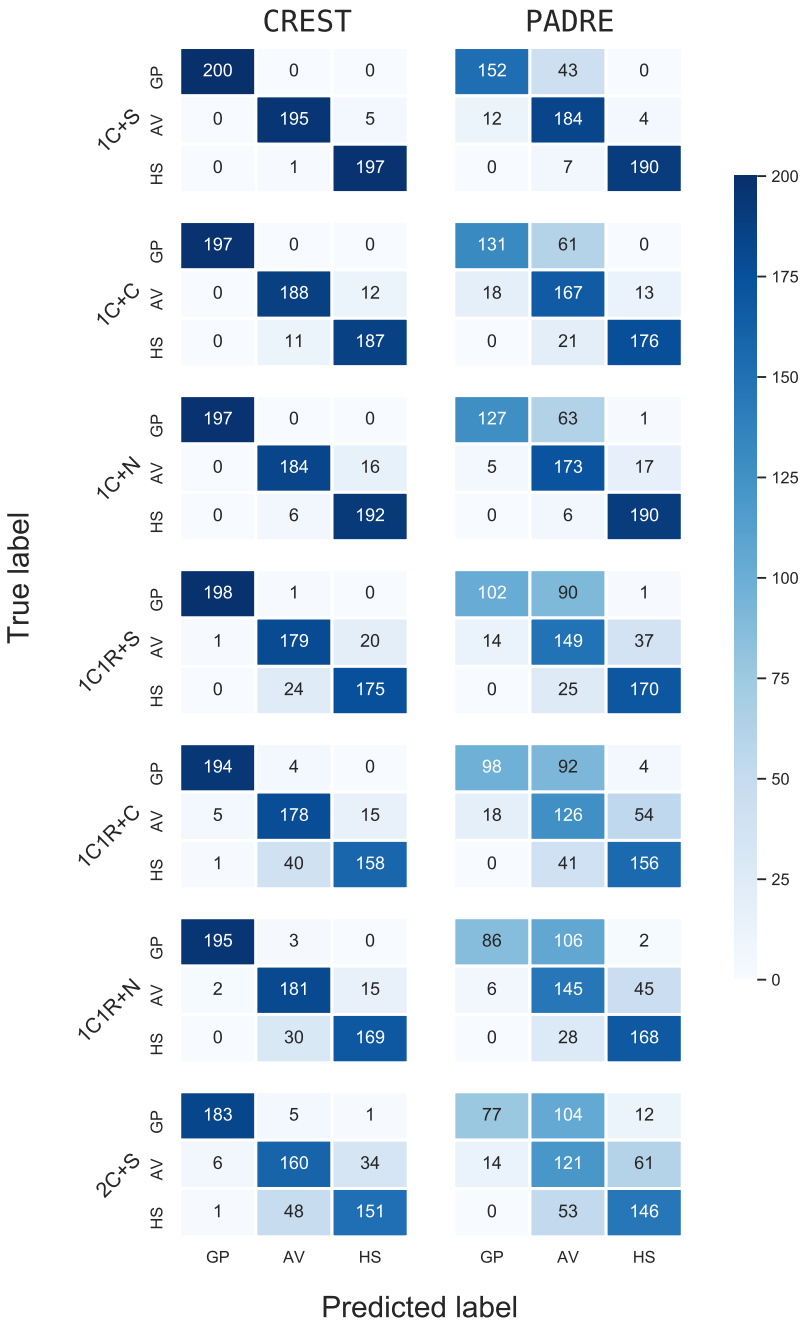

Figure B.8: The confusion matrices from the CREST and PADRE classification results. Analyses of CREST and PADRE include 200 pairs of GP, AV, and HS over different pedigree structures. Labels on the left indicate the mutual relatives in the pedigree structures. The row of each matrix gives the true relationship type and the column is the predicted relationship type. Since a few pairs failed classification by CREST or PADRE (Results), the sums of each row are not always 200.

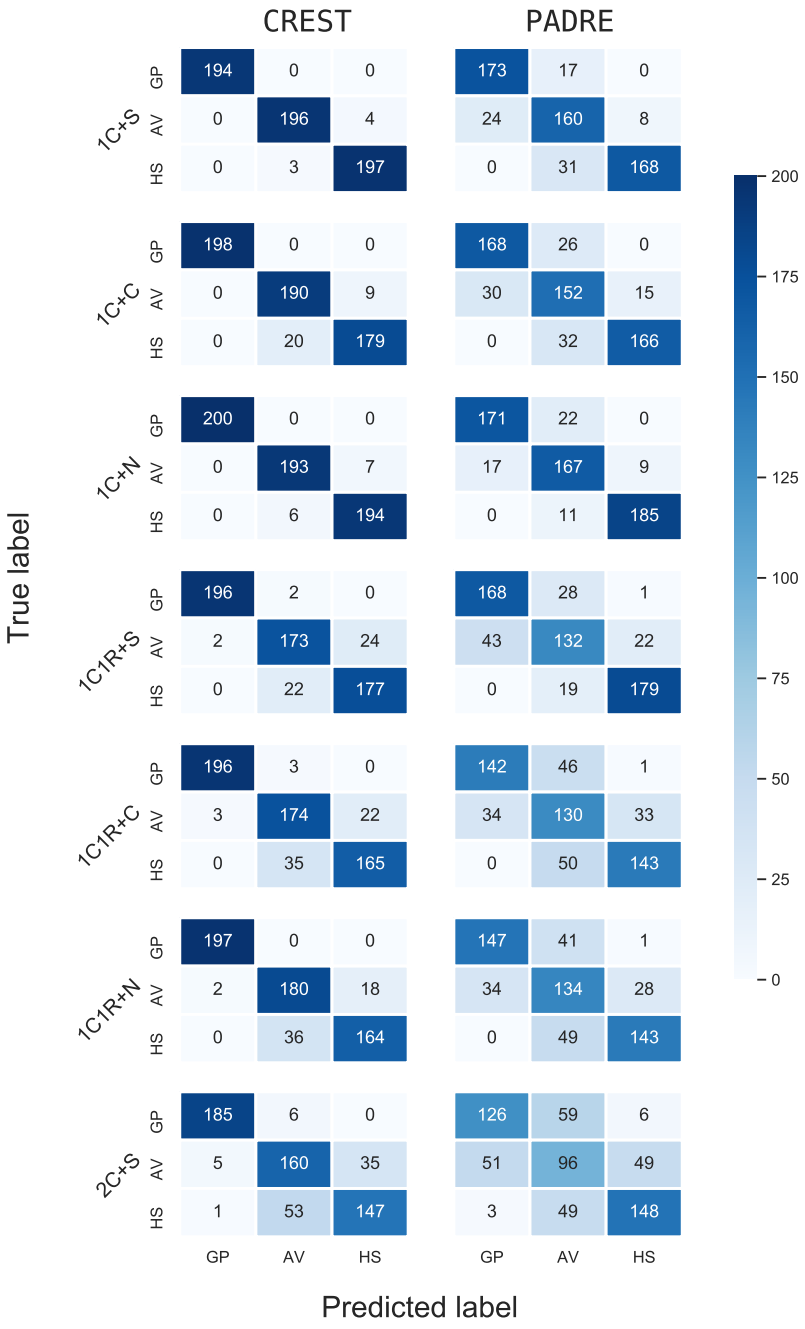

Figure B.9: The confusion matrices from the CREST and PADRE classification results where PADRE used perfect haplotypes. Analyses of CREST and PADRE include 200 pairs of GP, AV, and HS over different pedigree structures. Labels on the left indicate the mutual relatives in the pedigree structures. The row of each matrix gives the true relationship type and the column is the predicted relationship type. Since a few pairs failed classification by CREST or PADRE (Results), the sums of each row are not always 200.

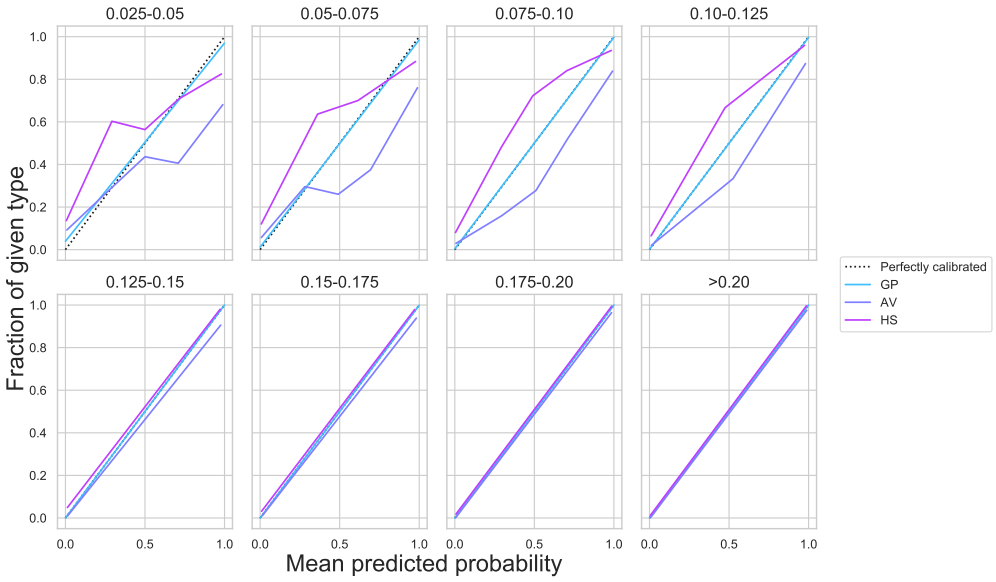

Figure B.10: The calibration curves for classifying second degree relatives over different coverage rates. In each plot, the analysis includes 1,000 pairs of each type. The x-axis shows the per-bin mean predicted probability and the y-axis indicates the proportion of pairs that are of the given type in the corresponding bin. We used five bins where possible, but reduced the number of bins if needed to ensure that each bin includes at least 50 pairs.In all cases, bins are uniformly spaced.

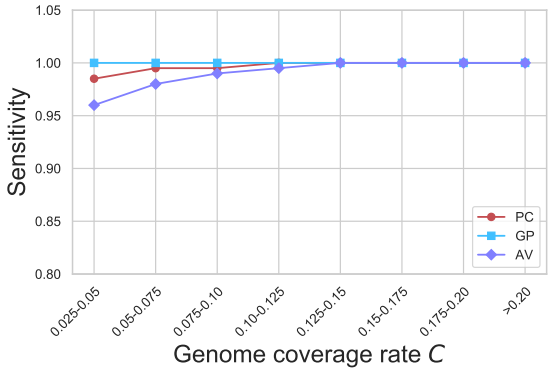

Figure B.11: CREST accurately infers the directionality of GP, AV, and PC pairs. Plot shows the sensitivity across bins of genome coverage rates for 200 pairs in each bin.

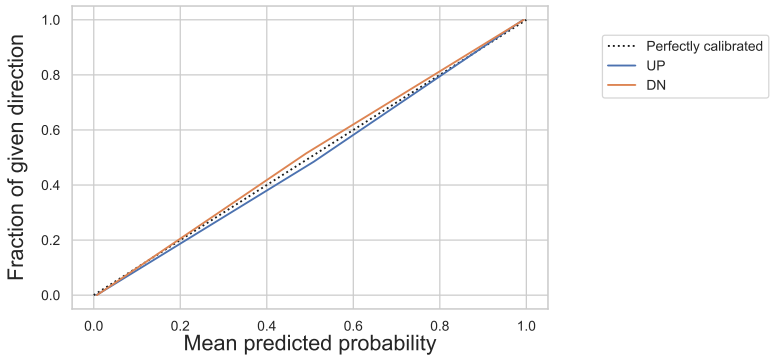

Figure B.12: The calibration curves for inferring the relationship directionality of GP, AV, and PC pairs. The x-axis shows the per-bin mean predicted probability and the y-axis indicates the proportion of pairs that are of the given direction in the corresponding bin.The analysis includes 300 pairs of each direction. Plot includes three uniformly spaced bins.

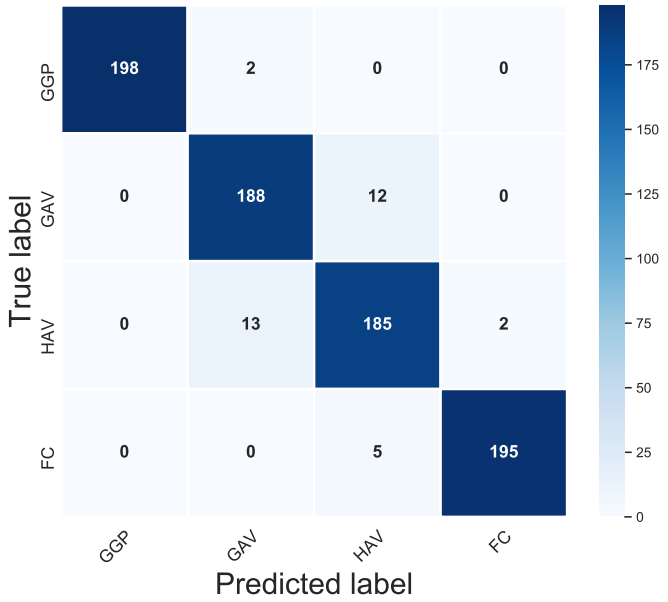

Figure B.13: The confusion matrix for classifying third degree relatives. The rows correspond to the true relationship type and the column is the predicted type. The analysis includes 200 simulated pairs of each type (didn't use inferred degrees).

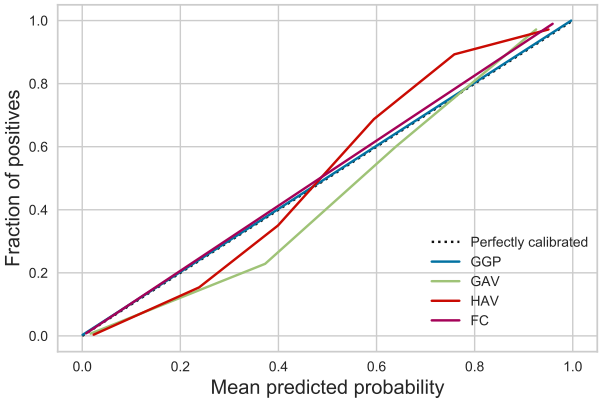

Figure B.14: The calibration curves for classifying third degree relatives. The x-axis shows the per-bin mean predicted probability and the y-axis indicates the proportion of pairs that are of the given type in the corresponding bin. The analysis includes 200 pairs of each type. Plot includes five uniformly spaced bins.

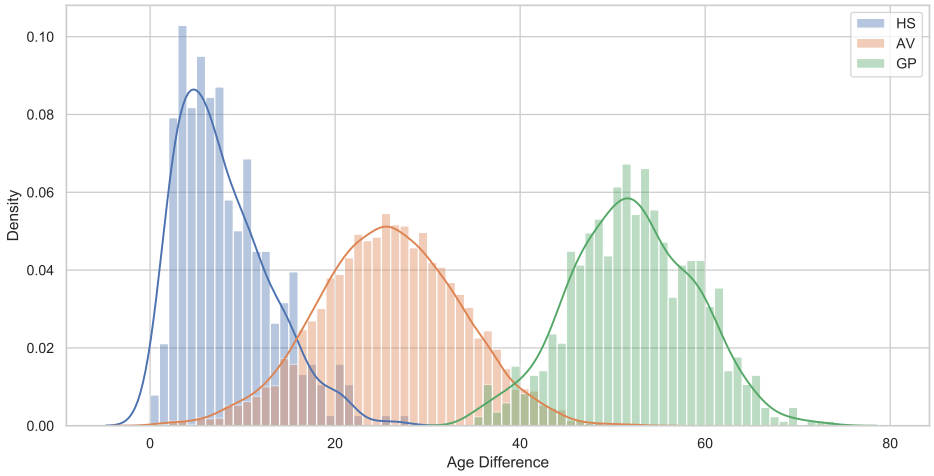

Figure B.15: The distribution of age differences of second degree relatives in GS dataset. Histograms of the absolute value of age differences of all GP, AV, and HS pairs in the GS dataset.

APPENDIX C

SUPPLEMENTARY FOR CHAPTER 4

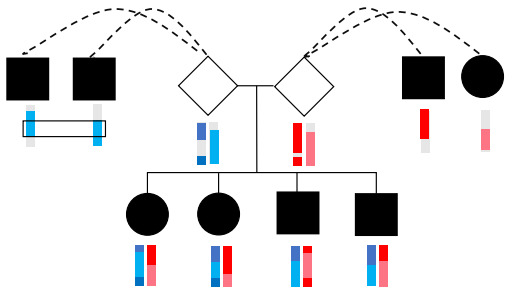

Figure C.1: Example IBD sharing between siblings and their relatives on both the maternal and paternal sides. The regions in different colors represent either haplotype of siblings or IBD segments other individuals share with those haplotypes. The black box outlines the regions that are also IBD segments between two relatives. Relatives on the same side may or may not share IBD segments with each other.

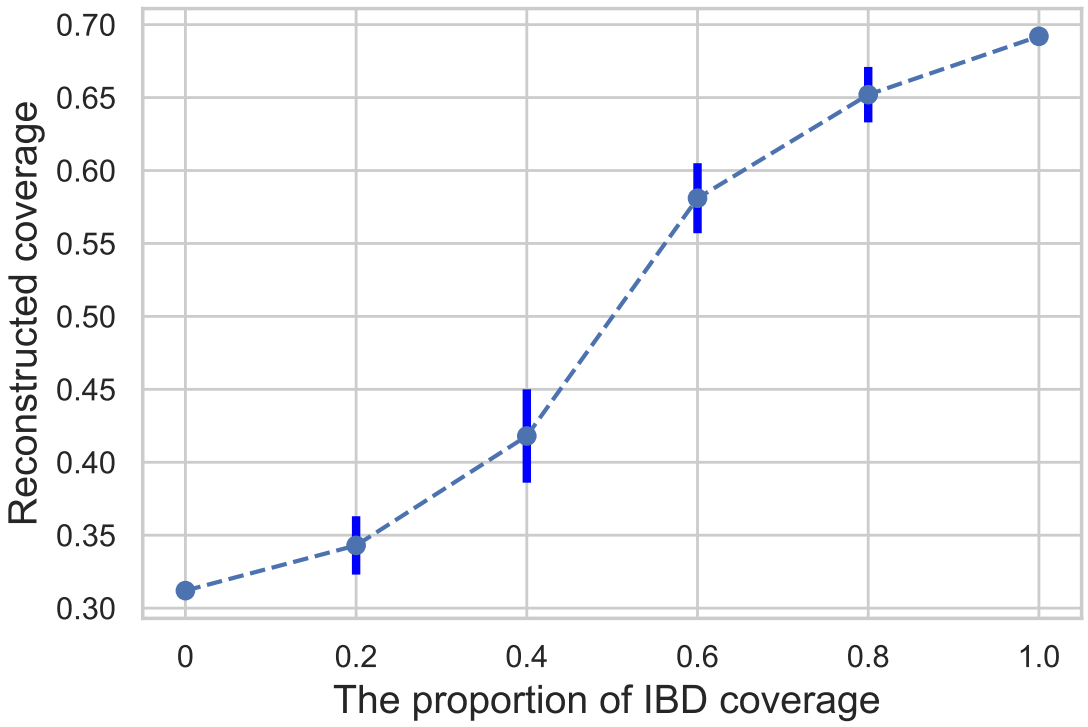

Figure C.2: The reconstructed coverage varies as the proportion of IBD regions change. The average reconstructed genotypes coverage and standard deviation over 50 samplings for given the proportion of IBD regions.

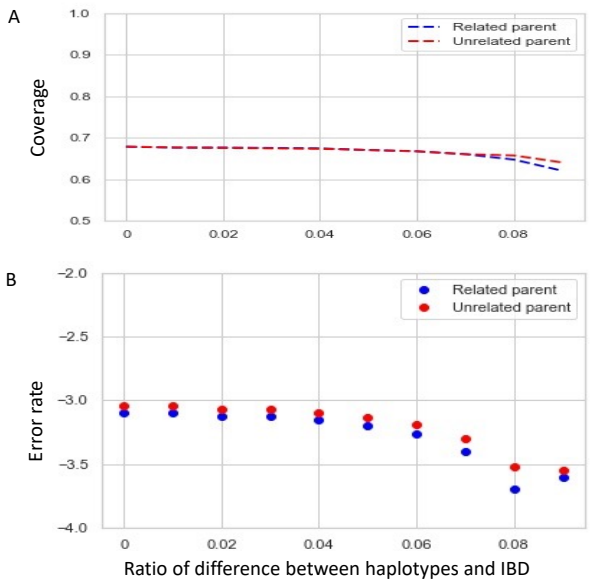

Figure C.3: The coverage and error rates of reconstructed genotypes under different ratios. (A) The coverage and (B) error rates of reconstructed genotypes for two parents with four children and their relatives under different ratios of difference between haplotypes and IBD regions. The blue line or dot represents the results of parent that is related to the relatives; The red line or dot represents the results of the other parent that is unrelated to the relatives.

BIBLIOGRAPHY

[1] William Astle and David J. Balding. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, 24(4):451 – 471, 2009.

[2] Mohamed Bennasar, Yulia Hicks, and Rossitza Setchi. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42(22):8520–8532, 2015.

[3] Claude Bhérer, Christopher L Campbell, and Adam Auton. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature Communications*, 8, 2017.

[4] Brian L. Browning and Sharon R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009.

[5] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[6] Sharon R Browning and Brian L Browning. Identity-by-descent-based heritability analysis in the northern finland birth cohort. *Human genetics*, 132(2):129–138, 2013.

[7] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham,

Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[8] Madison Caballero, Daniel N Seidman, Ying Qiao, Jens Sannerud, Thomas D Dyer, Donna M Lehman, Joanne E Curran, Ravindranath Duggirala, John Blangero, Shai Carmi, et al. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genetics*, 15(12):e1007979, 2019.

[9] Christopher L. Campbell, Nicholas A. Furlotte, Nick Eriksson, David Hinds, and Adam Auton. Escape from crossover interference increases with maternal age. *Nature Communications*, 6:6260, Feb 2015.

[10] Christopher C. Chang, Carson C. Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, 2015.

[11] Charles Y.K. Cheung, Elizabeth A. Thompson, and Ellen M. Wijsman. Gigi: An approach to effective imputation of dense genotypes on large pedigrees. *The American Journal of Human Genetics*, 92(4):504–516, 2013.

[12] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011.

[13] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[14] Apostolos Dimitromanolakis, Andrew D Paterson, and Lei Sun. Fast and accurate shared segment detection and relatedness estimation in unphased genetic data via truffle. *The American Journal of Human Genetics*, 105(1):78–88, 2019.

[15] Ravindranath Duggirala, John Blangero, Laura Almasy, Thomas D Dyer, Kenneth L Williams, Robin J Leach, Peter O'Connell, and Michael P Stern. Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *American Journal of Human Genetics*, 64(4):1127–1140, 1999.

[16] Michael P. Epstein, William L. Duren, and Michael Boehnke. Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, 67(5):1219–1231, Nov 2000.

[17] Luke M Evans, Rasool Tahmasbi, Matt Jones, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Douglas W Bjelland, Teresa R de Candia, Jian Yang, Michael E Goddard, et al. Narrow-sense heritability estimation of complex traits using identity-by-descent information. *Heredity*, 121(6):616–630, 2018.

[18] Kelly Finke, Michael Kourakos, Gabriela Brown, Huyen Trang Dang, Shi Jie Samuel Tan, Yuval B Simons, Shweta Ramdas, Alejandro A Schäffer, Rachel L Kember, Maja Bućan, et al. Ancestral haplotype reconstruction in endogamous populations using identity-by-descent. *PLoS computational biology*, 17(2):e1008638, 2021.

[19] William A Freyman, Kimberly F McManus, Suyash S Shringarpure, Ethan M Jewett, Katarzyna Bryc, The 23, Me Research Team, and Adam Auton. Fast and Robust Identity-by-Descent Inference with the Templated

Positional Burrows–Wheeler Transform. *Molecular Biology and Evolution*, 38(5):2131–2151, 12 2020.

[20] Alexander Gusev, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, 2009.

[21] Bjarni V. Halldorsson, Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, Asmundur Oddsson, Gisli H. Halldorsson, Florian Zink, Sigurjon A. Gudjonsson, Michael L. Frigge, Gudmar Thorleifsson, Asgeir Sigurdsson, Simon N. Stacey, Patrick Sulem, Gisli Masson, Agnar Helgason, Daniel F. Gudbjartsson, Unnur Thorsteinsdottir, and Kari Stefansson. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363(6425), 2019.

[22] D. He, Z. Wang, L. Parida, and E. Eskin. Iped2: Inheritance path based pedigree reconstruction algorithm for complicated pedigrees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(5):1094–1103, Sep. 2017.

[23] Brenna M. Henn, Lawrence Hon, J. Michael Macpherson, Nick Eriksson, Serge Saxonov, Itsik Pe'er, and Joanna L. Mountain. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS ONE*, 7(4):e34267, 04 2012.

[24] WG Hill and BS Weir. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(01):47–64, 2011.

[25] William G. Hill and Ian M. S. White. Identification of pedigree relationship from genome sharing. *G3: Genes, Genomes, Genetics*, 3(9):1553–1571, 2013.

[26] Nazrul Hoque, Dhruba K Bhattacharyya, and Jugal K Kalita. Mifs-nd: A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14):6371–6385, 2014.

[27] EA Housworth and FW Stahl. Crossover interference in humans. *The American Journal of Human Genetics*, 73(1):188–197, 2003.

[28] Chad D. Huff, David J. Witherspoon, Tatum S. Simonson, Jinchuan Xing, W. Scott Watkins, Yuhua Zhang, Therese M. Tuohy, Deborah W. Neklason, Randall W. Burt, Stephen L. Guthery, Scott R. Woodward, and Lynn B. Jorde. Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome Research*, 21(5):768–774, 2011.

[29] Kelly J Hunt, Donna M Lehman, Rector Arya, Sharon Fowler, Robin J Leach, Harald HH Göring, Laura Almasy, John Blangero, Tom D Dyer, Ravindranath Duggirala, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans. *Diabetes*, 54(9):2655–2662, 2005.

[30] Liang-Dar Hwang, Justin D Tubbs, Justin Luong, Mischa Lundberg, Gunn-Helen Moen, Geng Wang, Nicole M Warrington, Pak C Sham, Gabriel Cuellar-Partida, and David M Evans. Estimating indirect parental genetic effects on offspring phenotypes using virtual parental genotypes derived from sibling and half sibling pairs. *PLoS genetics*, 16(10):e1009154, 2020.

[31] Anuradha Jagadeesan, Ellen D Gunnarsdóttir, S Sunna Ebenesersdóttir, Valdis B Guðmundsdóttir, Elisabet Linda Thordardottir, Margrét S Einarsdóttir, Hákon Jónsson, Jean-Michel Dugoujon, Cesar Fortes-Lima, Florence

Migot-Nabias, et al. Reconstructing an african haploid genome from the 18th century. *Nature Genetics*, 50(2):199–205, 2018.

[32] Ethan MacNeil Jewett, Kimberly F McManus, William A Freyman, and Adam Auton. Bonsai: An efficient method for inferring large human pedigrees from genotype data. *bioRxiv*, 2021.

[33] Mark A Jobling and Peter Gill. Encoded evidence: Dna in forensic analysis. *Nature Reviews Genetics*, 5(10):739–751, 2004.

[34] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–354, Apr 2010.

[35] Manfred Kayser and Peter De Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192, 2011.

[36] Daniel Kling, Christopher Phillips, Debbie Kennett, and Andreas Tillmar. Investigative genetic genealogy: Current methods, knowledge and practice. *Forensic Science International: Genetics*, page 102474, 2021.

[37] Amy Ko and Rasmus Nielsen. Composite likelihood method for inferring local pedigrees. *PLOS Genetics*, 13(8):e1006963, 08 2017.

[38] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.

[39] Jaesung Lee and Dae-Won Kim. Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications*, 42(4):2013–2025, 2015.

[40] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R. Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the haplotype reference consortium panel. *Nat Genet*, 48(11):1443–1448, Nov 2016.

[41] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[42] Braxton D Mitchell, Candace M Kammerer, John Blangero, Michael C Mahaney, David L Rainwater, Bennett Dyke, James E Hixson, Richard D Henkel, R Mark Sharp, Anthony G Comuzzie, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. *Circulation*, 94(9):2159–2170, 1996.

[43] Reka Nagy, Thibaud S. Boutin, Jonathan Marten, Jennifer E. Huffman, Shona M. Kerr, Archie Campbell, Louise Evenden, Jude Gibson, Carmen Amador, David M. Howard, Pau Navarro, Andrew Morris, Ian J. Deary, Lynne J. Hocking, Sandosh Padmanabhan, Blair H. Smith, Peter Joshi, James F. Wilson, Nicholas D. Hastie, Alan F. Wright, Andrew M. McIntosh, David J. Porteous, Chris S. Haley, Veronique Vitart, and Caroline Hayward. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine*, 9(1):23, Mar 2017.

[44] Ardalan Naseri, Junjie Shi, Xihong Lin, Shaojie Zhang, and Degui Zhi. Raffi: Accurate and fast familial relationship inference in large scale biobank studies using rapid. *PLoS Genetics*, 17(1):e1009315, 2021.

[45] Jurg Ott, Jing Wang, and Suzanne M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5):275–284, May 2015.

[46] Wenbin Qian and Wenhao Shu. Mutual information criterion for feature selection from incomplete data. *Neurocomputing*, 168:210–220, 2015.

[47] Ying Qiao, Jens G Sannerud, Sayantani Basu-Roy, Caroline Hayward, and Amy L Williams. Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps. *The American Journal of Human Genetics*, 108(1):68–83, 2021.

[48] Raheleh Rahbari, Arthur Wuster, Sarah J. Lindsay, Robert J. Hardwick, Ludmil B. Alexandrov, Saeed Al Turki, Anna Dominiczak, Andrew Morris, David Porteous, Blair Smith, Michael R. Stratton, UK 10K Consortium, and Matthew E. Hurles. Timing, rates and spectra of human germline mutation. *Nature Genetics*, 48:126–133, Dec 2015. Article.

[49] Monica D. Ramstetter, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, Jason G. Mezey, and Amy L. Williams. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 207(1):75–82, 2017.

[50] Monica D. Ramstetter, Sushila A. Shenoy, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, Jason G. Mezey, and Amy L. Williams. Inferring identical-by-descent sharing

of sample ancestors promotes high-resolution relative detection. *The American Journal of Human Genetics*, 103(1):30–44, 2018.

[51] Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.

[52] Thomas A. Sasani, Brent S. Pedersen, Ziyue Gao, Lisa Baird, Molly Przeworski, Lynn B. Jorde, and Aaron R. Quinlan. Large, three-generation ceph families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *bioRxiv*, 2019.

[53] Daniel N. Seidman, Sushila A. Shenoy, Minsoo Kim, Ramya Babu, Ian G. Woods, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, and Amy L. Williams. Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification. *The American Journal of Human Genetics*, 106(4):453–466, Apr 2020.

[54] Blair H Smith, Archie Campbell, Pamela Linksted, Bridie Fitzpatrick, Cathy Jackson, Shona M Kerr, Ian J Deary, Donald J MacIntyre, Harry Campbell, Mark McGilchrist, Lynne J Hocking, Lucy Wisely, Ian Ford, Robert S Lindsay, Robin Morton, Colin N A Palmer, Anna F Dominiczak, David J Porteous, and Andrew D Morris. Cohort profile: Generation Scotland: Scottish family health study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology*, 42(3):689–700, 2013.

[55] Matthew Stallard and Jerome de Groot. "Things are coming out that are questionable, we never knew about": DNA and the new family history. *Journal of Family History*, 45(3):274–294, 2020.

[56] Jeffrey Staples, Evan K. Maxwell, Nehal Gosalia, Claudia Gonzaga-Jauregui, Christopher Snyder, Alicia Hawes, John Penn, Ricardo Ulloa, Xiaodong Bai, Alexander E. Lopez, Cristopher V. Van Hout, Colm O'Dushlaine, Tanya M. Teslovich, Shane E. McCarthy, Suganthi Balasubramanian, H. Lester Kirchner, Joseph B. Leader, Michael F. Murray, David H. Ledbetter, Alan R. Shuldiner, George D. Yancoupolos, Frederick E. Dewey, David J. Carey, John D. Overton, Aris Baras, Lukas Habegger, and Jeffrey G. Reid. Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. *The American Journal of Human Genetics*, 102(5):874–889, May 2018.

[57] Jeffrey Staples, Deborah A Nickerson, and Jennifer E Below. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genetic epidemiology*, 37(2):136–141, 2013.

[58] Jeffrey Staples, Dandi Qiao, Michael H. Cho, Edwin K. Silverman, Deborah A. Nickerson, and Jennifer E. Below. Primus: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *The American Journal of Human Genetics*, 95(5):553–564, Nov 2014.

[59] Jeffrey Staples, David J Witherspoon, Lynn B Jorde, Deborah A Nickerson, Jennifer E Below, Chad D Huff, University of Washington Center for Mendelian Genomics, et al. PADRE: Pedigree-aware distant-relationship estimation. *The American Journal of Human Genetics*, 99(1):154–162, 2016.

[60] Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

[61] Lei Sun and Apostolos Dimitromanolakis. PREST-plus identifies pedigree

errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. *BMC Proceedings*, 8(Suppl 1):S23, 2014.

[62] Alun Thomas, Nicola J Camp, James M Farnham, Kristina Allen-Brady, and Lisa A Cannon-Albright. Shared genomic segment analysis. mapping disease predisposition genes in extended pedigrees using snp genotype assays. *Annals of human genetics*, 72(2):279–287, 2008.

[63] Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

[64] Benjamin F Voight and Jonathan K Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLOS Genetics*, 1(3):e32, 2005.

[65] John Wakeley, Léandra King, Bobbi S Low, and Sohini Ramachandran. Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent. *Genetics*, 190(4):1433–1445, 2012.

[66] Bruce S Weir, Amy D Anderson, and Amanda B Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[67] Bruce S Weir, Amy D Anderson, and Amanda B Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[68] Amy L Williams, Giulio Genovese, Thomas Dyer, Nicolas Altemose, Katherine Truax, Goo Jun, Nick Patterson, Simon R Myers, Joanne E Curran, Ravi Duggirala, et al. Non-crossover gene conversions show strong gc bias and unexpected clustering in humans. *Elife*, 4:e04637, 2015.

[69] Amy L. Williams, David Housman, Martin Rinard, and David Gifford. Rapid haplotype inference for nuclear families. *Genome Biology*, 11(10):R108, 2010.

[70] Cole M. Williams, Brooke Scelza, Christopher R. Gignoux, and Brenna M. Henn. A rapid, accurate approach to inferring pedigrees in endogamous populations. *bioRxiv*, 2020.

[71] Alexander I. Young, Michael L. Frigge, Daniel F. Gudbjartsson, Gudmar Thorleifsson, Gyda Bjornsdottir, Patrick Sulem, Gisli Masson, Unnur Thorsteinsdottir, Kari Stefansson, and Augustine Kong. Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics*, 50(9):1304–1310, 2018.

[72] Alexander I. Young, Seyed Moeen Nehzati, Chanwook Lee, Stefania Benonisdottir, David Cesarini, Daniel J. Benjamin, Patrick Turley, and Augustine Kong. Mendelian imputation of parental genotypes for genome-wide estimation of direct and indirect genetic effects. *bioRxiv*, 2020.

[73] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L. Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLOS Genetics*, 9(5):e1003520, 05 2013.

[74] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 44(7):821–824, Jul 2012.