Published in final edited form as:

Genet Epidemiol. 2016 February ; 40(2): 161–171. doi:10.1002/gepi.21948.

A Pipeline for Classifying Relationships Using Dense SNP/SNV Data and Putative Pedigree Information

Zhen Zeng^{1,*}, Daniel E Weeks^{1,2}, Wei Chen³, Nandita Mukhopadhyay², and Eleanor Feingold^{1,2,*}

¹Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

²Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

³Division of Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, Pennsylvania, United States of America

Abstract

When genome-wide association studies (GWAS) or sequencing studies are performed on family-based datasets, the genotype data can be used to check the structure of putative pedigrees. Even in datasets of putatively unrelated people, close relationships can often be detected using dense single-nucleotide polymorphism/variant (SNP/SNV) data. A number of methods for finding relationships using dense genetic data exist, but they all have certain limitations, including that they typically use average genetic sharing, which is only a subset of the available information. Here, we present a set of approaches for classifying relationships in GWAS datasets or large-scale sequencing datasets. We first propose an empirical method for detecting identity by descent segments in close relative pairs using un-phased dense SNP data and demonstrate how that information can assist in building a relationship classifier. We then develop a strategy to take advantage of putative pedigree information to enhance classification accuracy. Our methods are tested and illustrated with two datasets from two distinct populations. Finally, we propose classification pipelines for checking and identifying relationships in datasets containing a large number of small pedigrees.

Keywords

relationship classification; identity-by-descent; putative pedigree structures; support vector machine

*Correspondence to: Zhen Zeng, Department of Biostatistics, University of Pittsburgh, Parran Hall, 130 DeSoto Street, Pittsburgh, PA 15261. zhz43@pitt.edu; Eleanor Feingold, Department of Human Genetics, University of Pittsburgh, Crabtree Hall A300, 130 DeSoto Street, Pittsburgh, PA 15261. feingold@pitt.edu.

All the authors have no conflict of interest to declare.

Introduction

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

A number of methods are available for testing relationships based on likelihoods and hypothesis testing, such as PREST [McPeck and Sun, 2000] and RELPAIR [Epstein et al., 2000]. These methods usually require sparse and uncorrelated genetic markers. Most of the existing relationship inference tools for dense single-nucleotide polymorphism/variant (SNP/ SNV, we will use SNP for short in the rest of the paper) data as from genome-wide association studies (GWAS) or sequencing studies, such as PLINK [Purcell et al., 2007], need a strong assumption of a homogeneous population. More recent additions, KING [Manichaikul et al., 2010] and REAP [Thornton et al., 2012], are robust in the presence of population structure and admixture. However, although these methods are powerful for detecting first-, second-, and third-degree relationships, none of them can effectively separate second-degree relatives, i.e., grandparent-grandchild, half-siblings, and avuncular pairs, from each other. This is due to the fact that existing algorithms focus on estimating measures of average genetic sharing, such as kinship coefficients and probabilities of identity by descent (IBD) sharing, and the above-mentioned second-degree relatives share the same expected values for these quantities.

Average genetic sharing is only part of the information available in genomic data. In principle, grandparent-grandchild, half-siblings, and avuncular pairs are separable if spatial information on genetic sharing is also considered. IBD states along chromosomes can be described as Markov processes with transition rates λ and 2λ for grandparent-grandchild and half-siblings. For avuncular relationship, the process is non-Markov, but the transition rate is known to be $5/2\lambda$ (λ can be interpreted as the unit of genetic length) [Feingold, 1993]. In other words, the expected sojourn length in different IBD states, a summary of the spatial IBD information, is different for different relationships. The observed times of transition can therefore help classify relationships. Several existing algorithms for detecting segmental sharing of IBD, such as PLINK [Purcell et al., 2007], fastIBD [Browning and Browning, 2011], GERMLINE [Gusev et al., 2009], PARENTE, and PARENTE2 [Rodriguez et al., 2015], can be used to generate such summary statistics of spatial information, but they all have certain limitations: PLINK and PARENTE do not model SNP dependency and require SNP pruning; fastIBD has to phase genotypes and call IBD segments simultaneously; GERMLINE needs correctly phased genotype data as input; and PARENTE2 requires a phased training dataset.

Another important piece of relationship information is the putative pedigree that is typically generated based on subject interviews. Depending on when, where, and how the information is collected and the complexity of the pedigrees, these pedigree structures may contain errors. The errors can be caused by incorrect knowledge of pedigree structure, for example,

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

In this paper, we propose a set of approaches for classifying relationship types in GWAS datasets or large-scale sequencing datasets. We first present a new empirical algorithm for finding regions of IBD in closely related individuals using un-phased dense SNP data. A summary of IBD spatial information, observed recombination number (M), is generated. We then demonstrate how that information can be used in principle to distinguish relationships. We also build a classifier and develop novel approaches taking advantage of information from putative pedigree structures. All the methods are tested and illustrated with two different datasets. Finally, we propose classification pipelines for checking and identifying relationships aimed at datasets containing a large number of small pedigrees. Computational tools for implementing our methods are provided.

Methods

Datasets

Our methods were applied to two datasets from two distinct populations. One dataset consists of a US sample (non-Latino whites) from the Center for Oral Health Research in Appalachia (COHRA) Project (dbGaP accession number phs000095.v3.p1). The other consists of a Guatemalan sample from the Gene-Environment Association Studies (GENEVA) Guatemala Dental Caries Project (dbGaP accession number phs000440.v1.p1). Both datasets were genotyped using Illumina Human610-Quad v1_B BeadChip (Illumina, Inc., San Diego, CA, USA), and were cleaned to have genotyping rate per individual larger than 0.9, genotyping rate per SNP larger than 0.9, minor allele frequency larger than 0.01, and Hardy-Weinberg equilibrium test P -value larger than 10^{-4} . Approximately 540,000 autosomal SNPs were included in each dataset. Principal component analysis has previously identified the Guatemalan sample as more stratified than the US sample. The Guatemalan sample has higher level of ancestry admixture and distant population background relatedness, so it can be deemed as a nonhomogeneous population, while the US sample is more homogeneous. Both datasets contain abundant close relationships. Pedigree errors have been previously detected and cleaned manually by experts. All the data and associated study information can be obtained on dbGaP. More details about pedigree cleaning are available

upon request. We selected a number of pairs of individuals with confident relationships for the following categories as the gold standard to build two separate classification pipelines, one for each population: monozygotic twins (MZ), full-siblings (FS), parent-offspring (PO), grandparent-grandchild (GG), half-siblings (HS), avuncular pair (AV), first-cousins (FC), and unrelated pair (UN). Duplicate GG pairs (a grandchild and each of the grandparents are duplicated pairs) and any problematic relationships identified during the current analyses were removed from the training data. The training data sizes by relationship categories are summarized in Table 1.

Algorithms for Inferring IBD Segments

Our algorithm is based on rules relating IBD and identity by state (IBS): assuming no genotyping error, in IBD = 0 regions, the IBS state could be 0, 1, or 2; in IBD = 1 regions, the IBS state could be either 1 or 2; in IBD = 2 regions, the IBS state can only be 2. Under these assumptions, large IBD segments can be identified simply by eye (Fig. 1). The algorithm essentially automates this visual inspection process. The intuition behind the algorithm is to call the IBD state in small chromosomal segments and then fill any low-information gaps and filter out small IBD segments. We investigated eight variations on this algorithm, considering different ways to define chromosomal segments, whether to use a sliding window, and whether to use a reference panel consisting of UN pairs for calling segmental IBD states. After careful comparison (see Results), the final algorithm defines chromosomal segments with a fixed number of SNPs, and neither uses a sliding window nor a reference panel (algorithm 5 in Table 2). Steps for inferring IBD segments for a unilineal pair of individuals are shown in Figure 2 and described as follows.

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

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

Step 3: Call IBD states in each chromosomal segment. Uncertain small gaps are then filled according to their flanking IBD status (Fig. 2). For regions where SNPs are particularly sparse, such as centromeres, the IBD states are labeled as unknown.

Our algorithm should work for genome-wide data where markers are dense and relatively evenly distributed across genome, such as whole-genome SNP array data and whole-genome deep sequencing data. There are two tuning parameters in our method: the rate of IBS errors caused by genotyping errors P and the number of SNPs in each chromosomal segment. These can be chosen to accommodate the data features. The number of SNPs can be chosen to break the autosomes into 2,500 to 3,500 segments (roughly equal to the total genetic length of autosomes in centi-Morgan). A reasonable value for P can range from 0.0001 to 0.001 for SNP array data and 0.001 to 0.01 for whole genome sequencing data, depending on the actual data quality. Alternatively, it can be easily estimated from PO pairs. Because

the IBD state is always 1 for PO pairs, the ratio of the number of SNPs with IBS = 0 and the number of SNPs with IBS = 1 is a direct estimate of this parameter.

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

Quantifying the Accuracy of the Algorithms by Simulation and Comparison

To choose the best algorithm from different combinations of strategies and evaluate the accuracy of the final algorithm, the eight proposed algorithms (Table 2) were evaluated in several ways. Our first comparison was based on simulated data. We used the third generation Rutgers Combined Linkage-Physical Map of The Human Genome [Matise et al., 2007] to estimate the genetic position for each SNP. Assuming recombination is a Poisson process on chromosomes, we generated random variables from an exponential distribution with mean 100 as the distances between recombination events (i.e., length of IBD segments). Artificially synthesized IBD data were simulated by joining the IBD = 0 and IBD = 1 segments sampled from 30 UN pairs and 30 PO pairs randomly selected from the US dataset. Synthesized IBD data on chromosome 1 for 1,000 artificial pairs were simulated. Each of the eight proposed algorithms was used to infer IBD for the simulated data and the results were compared with the truth to estimate the false-negative and false-positive rates.

Beside simulation, we evaluated the accuracy of algorithms by quantifying the concordance between duplicated GG pairs that share exactly the same IBD patterns but in opposite phase. We also compared each variant of algorithms with a recently developed recombination detection method. This unpublished method assumes the relationships are known and utilizes the relationship information to call recombination. Although it is not a “gold standard,” the use of more information presumably results in better estimates for IBD segments.

Calculating the Observed Recombination Number

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

Estimating IBD Scores

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

SVM Classification and Cross-Validation

A support vector machine (SVM) was used to build classifiers for unilineal relatives (AV/HS, FC, GG, PO, and UN). Unadjusted classifiers without putative pedigree information were based on k_0 and N only. The k_1 score was not included as a feature because for unilineal relatives, k_1 and k_0 are collinear. To adjust for systematic difference in N among populations due to different population background relatedness, we subtract from N the mean of UN pairs in each population, and set to 0 if it becomes negative after adjustment.

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

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

where $x_i^T = [k_0, N, I_1, \dots, I_n]$, and

$$P = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & S & & \\ & & & \ddots & \\ & & & & S \end{bmatrix}_{(n+2) \times (n+2)}$$

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

Comparing With PREST

PREST is to our knowledge the only existing tool that can distinguish among second-degree relationships with un-phased genome-wide SNP data, so we ran PREST (PREST-plus Version 4.09) on the US and Guatemalan datasets for a comparison. PREST implements a maximum likelihood ratio test (MLRT) for 11 different null relationship types, calculating a P -value for each of the null relationships. To run MLRT, we first used PLINK to prune SNP data to obtain an independent subset. Then, about 2,000 SNPs were randomly selected to further thin the SNP subset as recommended for PREST. We obtained the genetic map of the

selected SNPs from the third generation Rutgers Combined Linkage-Physical Map of The Human Genome [Matisse et al., 2007]. We performed MLRT for all of the training pairs in the two populations (US and Guatemala) and the plausible relationship type with the largest P -value was reported as the predicted relationship for each pair of individuals [Sun and Dimitromanolakis, 2014].

Putative pedigree information was not considered during the comparison, so the accuracy was based on no putative relationship available. Note that when such information is available, correct putative pedigree information can help increase our classification accuracy, but has no effect on PREST MLRT inference.

Because PREST considers three more relationship types than our methods: half-avuncular (HAV), half-first-cousin (HFC), and half-sib+first-cousin (HSFC), its accuracy was presented both including and excluding these predictions. If a prediction was a tie between two relationship types, it was counted as half for each.

Results

Comparison of Different Strategies for the IBD Detection Algorithm

Table 2 shows metrics of accuracy and computational time for the eight combinations of algorithm strategies. Algorithm 5, which uses fixed number of SNPs to define chromosomal segments and does not use a sliding window or a reference panel, is among the best for all metrics and is faster than others. Thus, algorithm 5 was chosen as our final algorithm. We investigated algorithm 5's performance by examining all the IBD segments omitted or mistakenly identified in the simulation. One hundred fifty IBD segments were false negatives and 21 were false positives. We found all the false-positive segments were in the same region and were from a single pair sampled repeatedly in the simulation. The distribution of genetic length of false-negative segments indicates the omitted segments are usually quite short (Fig. 3). This is natural because our algorithm filters out small uncertain regions. The filtering caused more false negatives than false positives and therefore introduced a small bias, which can be seen from the mean differences between inferred N and the truth in the simulation (Table 2). However, the bias is reasonably small. Also, it should be noted that the false positives were either due to genotyping errors within $IBD = 1$ regions or due to population background relatedness in $IBD = 0$ regions, which should be prevented aggressively, while false negatives are small IBD segments usually due to close double-recombination. In reality, the double-recombination interference results in fewer small IBD segments compared to the simulation with Poisson process, so our algorithm should have even less bias for real data.

Classifying Relationships Using N and k_0

N and k_0 were used as two features to train the classifiers. Figure 4 shows the scatter plots of N and k_0 for the US and Guatemala training data and visualizes the two classifiers. AV and HS cannot be distinguished in most cases, so these two relationships were pooled together and treated as one category (see Discussion). Cross-validation results are shown in Table 3. The prediction accuracy was greater than 90% for all the relationship categories, except for

GG (84.5% in the US sample). Cross-population prediction results, i.e., using the classifier built in one population to predict the training data in the other population, were also satisfactory, with the accuracy for all relationship categories better than 80% (Table 4). The adjustment for N with the mean of UN pairs is crucial. Essentially, the two populations have very different background relatedness. The Guatemalan sample has inflated N compared to the US sample. In other words, excessive shared IBD segments were observed between Guatemalan unrelated individuals, probably due to background distant relatedness in the population. Therefore, in practice, obtaining the mean N from a set of UN pairs to adjust for N is a useful extra step to enhance the robustness of our classifiers.

Incorporating Putative Relationships

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

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

Figure 5 shows the improvement of classification accuracy and the decrease of recovery rates as s value increases in the two samples. To balance the gain and loss, a value of 0.025 for s is recommended because the improvement is substantial (prediction accuracy >96% for all relationship categories in both samples), while the recovery rates of all types of misspecification are above 80% except for GG being misspecified as AV/HS in the US samples. However, this type of misspecification is presumably rare in most cases.

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

Considering Sex Information of Meiosis for GG

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

Comparison to PREST

Prediction results for PREST are shown in Table 5. The overall accuracy of our within-population cross-validation and cross-population prediction was superior to PREST (compare to the diagonal elements in Table 3 and the accuracy rows in Table 4), regardless of whether we considered the three additional relationship types (HAV, HFC, and HSFC). It is important to note that cross-population prediction accuracy is more appropriate for comparison (Table 4), because the results are more generalizable. Specifically, except for the prediction of FC in the US population, where PREST-plus accuracy is 88.9% after excluding those predicted as HAV, HFC, and HSFC, and the Guatemala predict US cross-population accuracy is 83.5%, our methods resulted in better accuracy for all other comparable comparisons. Our methods were also three orders of magnitude faster than PREST: PREST took 12–15 min to carry out MLRT for a pair of individuals, while our methods only took about one second per pair to calculate N .

Pipelines for Classifying Relationships With and Without Prior Pedigree Information

Combining our methods with existing approaches, we propose general pipelines for relationship classification (Fig. 6). In general, the pipelines can be summarized into four stages.

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

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

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

Stage 4: Carry out the classification with the SVM classifier (using either one of the two provided or a user-defined population-matched classifier). One could adjust the classifiers using prior pedigree information. If putative

relationships are used, those predictions disagreeing with corresponding putative relationships should be reclassified by unadjusted classifiers. Based on the final classification results, corrections can be made to the pedigree file.

Discussion

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

HMM-based methods are available for detecting IBD segments and/or differentiating relationships [Browning and Browning, 2011; Kyriazopoulou-Panagiotopoulou et al., 2011; Purcell et al., 2007]. Our method differs from them in many ways and has a slightly different goal. HMM-based methods are accurate and have potentials to separate distant relationship categories, but they require many more assumptions and information as input. Their accuracy depends on the validity of the assumptions as well as the availability and accuracy of pertinent information. Some of these methods require phased genotype data. Phasing is not a trivial problem when studying a population without a reference panel or when the study sample size is small. Also, some need allele and/or haplotype frequencies, which are unavailable without a reference population. In addition, some require SNP pruning. In reality, the needs of more steps, assumptions, or information mean being more error-prone. Finally, the computational intensity of HMM-based methods is usually prohibitive for very large datasets. In contrast, our method requires few assumptions and less external information, and our goal is to handle datasets consisting of many small pedigrees fast and accurately instead of a few large and complex pedigrees.

Although many relationships have different expected length of shared IBD segments and therefore different expected observed recombination number N , the large variances of the IBD segment length make differentiation among them challenging [Thompson, 2013]. This can be seen from the estimated standard deviations in Table 1. This variance may compromise the usage of N in relationship classification, particularly for distant relatives. Also, due to randomness, very short IBD segments that are not detectable by our algorithm may present, which could potentially affect our estimates of N . However, we showed that our algorithm and classifier have quite good accuracy for the relative types that we consider.

We demonstrated our methods with two real datasets, one from the US population and the other from the Guatemalan population. Systematic differences between the two populations were observed. Basically, inflated observed recombination number N 's were observed for all relationship categories in the Guatemalan sample (Table 1). The inflation manifests the presence of sporadic shared chromosome segments in the population, presumably due to distant population background relatedness. In practice, if possible, we encourage investigators to collect user-defined training data and build population-specific classifiers. However, as long as a number of UN pairs can be obtained and their mean is used to adjust

N , the cross-population prediction accuracy is quite satisfactory, as seen in our US and Guatemalan samples. It is noteworthy that except for the background relatedness, population stratification and ancestry admixture in the Guatemalan sample did not affect the results of our methods.

We were unable to show classification improvements by considering sex of meiosis. Limited by training data sizes, we only attempted separate-sex classification for GG pairs. Even so, the training sample sizes of maternal meiosis GG (GGm) and paternal meiosis GG (GGp) were quite small. In theory, sexes of meiosis could be considered for several other relationships. Table 6 lists all the relationships that can be divided into subtypes by sexes of their pertinent relatives and be modeled in the same way as GG. Thus, more advanced classifiers could be built accordingly if there were enough data. So, despite our failure to show improvement, sex information still has potential to enhance the classification performance, and is worth further investigation in the future. It should be noted that sex information is also putative in practice, because it is obtained from putative pedigree information. Effects of sex misspecification should also be investigated.

Two issues regarding the classification accuracy should be noted. One is the composition of the data to be tested. In our classifiers, some categories contain subtypes, such as GG (comprises GGm and GGp) and AV/HS (comprise AV and HS), and the results are combined, ignoring subtypes. When the prediction accuracy differs among subtypes, composition of test data would influence the prediction accuracy of a category. For example, GGp is inherently classified better than GGm (because the N of GGm is closer to AV/HS). The more paternal GG in the test data, the better the classification results of GG will be. The other issue is the number of instances of each category in the training datasets. Because parameter selection is based on the overall prediction accuracy, small categories will automatically sacrifice for larger ones, i.e., the classifier will be trained to be more accurate for classes with higher frequency in the dataset. Because we do not have balanced training data sizes for all relationship categories, our classifiers may have a preference for the categories with larger training data size. This issue can be solved easily when more training data are available. We suggest using confirmed relationships as additional training data to build better classifiers when possible.

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

Our methods were implemented in R. In terms of computational efficiency, the most time-consuming step is calculating N . It took 441 sec system time to compute N for the US sample (546 pairs) and 378 sec for the Guatemalan sample (488 pairs) with two quad-core

2.93 GHz CPUs and 24 GB of memory. Basically, the computing time increases linearly with the number of pairs to be tested. Also, the time required to read in the genotype data is not trivial when the dataset is very large. Data size is proportional to both the number of individuals and the number of SNPs. For computational efficiency, we recommend eliminating irrelevant individuals from the genotype file in the data cleaning step and transforming the data to a better format before processing it with R. Example code is given for transforming the data with PLINK and shell commands and can be found in the documentation of our R package. In addition, we recommend removing most of the confidently unrelated pairs with a conservative k_0 score to reduce the number of pairs to be tested as suggested in the pipelines (Fig. 6). Our algorithm can be easily parallelized by both chromosomes and individual pairs to deal with extremely large datasets.

Our IBD transition detection algorithm is developed for both whole-genome SNP array data and whole-genome deep sequencing data. We have also tried on whole-exome sequence data, which lie in between whole-genome and targeted sequencing (data not shown). The results show that directly applying our algorithm to whole-exome sequence data can be problematic.

Our relationship classification pipelines focus on generating accurate pairwise relationships. It is also important to reconstruct the pedigrees with individual relationship pairs. Of note, some relationships may conflict with each other during pedigree reconstruction, which implies classification errors. It might be of interest to consider modeling relationship classification and pedigree construction together so that such errors can be avoided while further improving the relationship classification accuracy. A recent pedigree constructing tool has made use of such a notion, but it treated all second-degree relationships as one category [Staples et al., 2014]. By dividing second-degree relationships with our methods, more accurate pedigrees might be reconstructed. As a possible direction of our future work, more features may be included and training data of different third-degree relationships may be collected for building more advanced classifiers that can distinguish third-degree relationships, which could help reconstruct more complex pedigrees. We implemented the putative relationship extraction tool, IBD transition detection algorithm, and relationship classifiers in an R package (available through <http://relcla.sourceforge.net/>).

Acknowledgments

We sincerely thank Dr. Mary L. Marazita for providing the COHRA (US) and GENEVA (Guatemala) datasets. We also want to thank the reviewers for their valuable comments and suggestions. The development of these datasets was funded by grants R01-DE014899, R01-DE016148, and U01-DE018903. The work of Z.Z. and E.E on this paper was funded by R01-HD038979.

References

- Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. *Am J Hum Genet.* 2011; 88(2):173–182. [PubMed: 21310274]
- Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol.* 2011; 2(3):1–27.
- Epstein MP, Duren WL, Boehnke M. Improved inference of relationship for pairs of individuals. *Am J Hum Genet.* 2000; 67(5):1219–1231. [PubMed: 11032786]

- Feingold E. Markov-processes for modeling and analyzing a new genetic-mapping method. *J Appl Prob.* 1993; 30(4):766–779.
- Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* 2009; 19(2):318–326. [PubMed: 18971310]
- Hill WG, White IMS. Identification of pedigree relationship from genome sharing. *G3.* 2013; 3(9): 1553–1571. [PubMed: 23893739]
- Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. A high-resolution recombination map of the human genome. *Nat Genet.* 2002; 31(3):241–247. [PubMed: 12053178]
- Kyriazopoulou-Panagiotopoulou S, Kashef Haghighi D, Aerni SJ, Sundquist A, Bercovici S, Batzoglou S. Reconstruction of genealogical relationships with applications to Phase III of HapMap. *Bioinformatics.* 2011; 27(13):i333–i341. [PubMed: 21685089]
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010; 26(22):2867–2873. [PubMed: 20926424]
- Matisse TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, Hyland FC, Kennedy GC, Kong X, Murray SS, et al. A second-generation combined linkage physical map of the human genome. *Genome Res.* 2007; 17(12):1783–1786. [PubMed: 17989245]
- McPeck MS, Sun L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet.* 2000; 66(3):1076–1094. [PubMed: 10712219]
- Morrison J. Characterization and correction of error in genome-wide IBD estimation for samples with population structure. *Genet Epidemiol.* 2013; 37(6):635–641. [PubMed: 23740691]
- Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006; 2(12):e190. [PubMed: 17194218]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81(3):559–575. [PubMed: 17701901]
- Ray A, Weeks DE. Relationship uncertainty linkage statistics (RULS): affected relative pair statistics that model relationship uncertainty. *Genet Epidemiol.* 2008; 32(4):313–324. [PubMed: 18205206]
- Rodriguez JM, Bercovici S, Huang L, Frostig R, Batzoglou S. Parente2: a fast and accurate method for detecting identity by descent. *Genome Res.* 2015; 25(2):280–289. [PubMed: 25273070]
- Staples J, Qiao D, Cho MH, Silverman EK, Nickerson DA, Below JE. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet.* 2014; 95(5): 553–564. [PubMed: 25439724]
- Sun L, Dimitromanolakis A. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. *BMC Proc.* 2014; 8(Suppl 1):1–6.
- Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics.* 2013; 194(2):301–326. [PubMed: 23733848]
- Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. *Am J Hum Genet.* 2012; 91(1):122–138. [PubMed: 22748210]

Figure 1.

An example illustrating the IBD segments identified by our algorithm on chromosome 1 for a pair of individuals. Each dot represents a SNP. Red bars indicate IBD = 0 segments. Blue bars indicate IBD = 1 segments. Gray bars indicate uncertain regions where SNPs are too sparse. In this example, the observed recombination number (N) is 5.

Figure 2.

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

Figure 3.
Genetic length of 150 false-negative IBD segments (red) and 3,784 true segments (blue).

Figure 4. SVM classifiers with features N and k_0 for the US and Guatemalan populations. Colored areas illustrate different relationship categories. Training data are plotted with circles. Cross symbols indicate the support vectors. AV and HS are grouped as one category.

Figure 5.

Classification accuracy and recovery percentage as functions of s value in the US and Guatemalan populations. The red, blue, and green solid curves represent the classification accuracy based on 1,000 fivefold cross-validation for relationship categories AV/HS, FC, and GG, respectively, when including relationship indicators as features and the putative relationships are correct. Dashed curves represent the percentage of pairs being recovered when the putative relationships are misspecified as shown. Classification accuracy for relationship categories not shown and recovery percentage of types of misspecification not shown are 100% across different s values.

Relationship classification pipeline without prior pedigree information

Relationship classification pipeline with prior pedigree information

Figure 6. Classification pipelines with or without prior pedigree information. UN*, within-family founder pairs; AIM, ancestrally informative marker.

Table 1

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

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

^aNot adjusted by the mean of UN pairs.

PO, parent-offspring; GG, grandparent-grandchild; HS, half-siblings; AV, avuncular pair; FC, first-cousins; UN, unrelated pair; GGp and HSp, paternal meiosis GG and HS; GGm and HSm, maternal meiosis GG and HS.

Table 2

Comparison of different proposed algorithm strategies

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

Table 3

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

		True relationship				
	Predicted	AV/HS	FC	GG	PO	UN
US	AV/HS	93.8 (92.5, 95.0)	5.8 (4.4, 7.7)	15.5 (10.9, 21.7)	0	0
	FC	0	94.2 (92.3, 95.6)	0	0	0
	GG	6.2 (5.0, 7.5)	0	84.5 (78.3, 89.1)	0	0
	PO	0	0	0	100 (100, 100)	0
	UN	0	0	0	0	100 (100, 100)
Guatemala	AV/HS	95.5 (94.4, 96.3)	0.4 (0, 2.6)	4.8 (2.8, 5.6)	0	0
	FC	1.2 (0.6, 1.9)	99.6 (97.4, 100)	0	0	0
	GG	3.3 (3.1, 3.8)	0	95.2 (94.4, 97.2)	0	0
	PO	0	0	0	100 (100, 100)	0
	UN	0	0	0	0	100 (100, 100)

Table 4

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

		True relationship				
Predicted	AV/HS	FC	GG	PO	UN	
US predicts Guatemala	AV/HS	157	1	14	0	0
	FC	2	38	0	0	1
	GG	1	0	58	0	0
	PO	0	0	0	100	0
	UN	0	0	0	0	99
Accuracy		98.1%	97.4%	80.6%	100%	99%
Guatemala predicts US	AV/HS	182	14	1	0	0
	FC	0	76	0	0	0
	GG	18	0	45	0	0
	PO	0	0	0	100	0
	UN	0	1	0	0	100
Accuracy		91%	83.5%	97.8%	100%	100%

Table 5

Prediction results of PREST for the US and Guatemalan datasets

PREST-plus prediction		True relationship				
		AV/HS	FC	GG	PO	UN
Guatemala	AV/HS	145	2	28	0	0
	FC	2	20.5	0	0	0
	GG	8	1	43	0	0
	PO	0	0	0	100	0
	UN	0	0	0	0	82
	HAV	1	9.5	1	0	0
	HFC	0	6	0	0	18
US	HSFC	4	0	0	0	0
	Accuracy	90.6%	52.6%	59.7%	100.0%	82.0%
	Accuracy ^a	93.5%	87.2%	60.6%	100.0%	100.0%
	AV/HS	169.5	5	14	0	0
	FC	1	48	0	0	0
	GG	26.5	1	31	0	0
	PO	0	0	0	100	0
	UN	0	0	0	0	89
	HAV	2	27	1	0	0
	HFC	0	9	0	0	11
	HSFC	1	1	0	0	0
	Accuracy	84.8%	52.7%	67.4%	100.0%	89.0%
	Accuracy ^a	86.0%	88.9%	68.9%	100.0%	100.0%

AV, avuncular pair; HS, half-sibling FC, first-cousin; GG, grandparent-grandchild; PO, parent-offspring; UN, unrelated; HAV, half-avuncular; HFC, half-first-cousin; HSFC, half-sib + first-cousin. If a prediction is a tie between two relationship categories, it is counted as half for each.

^a After excluding those predicted as HAV, HFC, and HSFC.

Table 6
Relationships for which the sex of pertinent relatives can be used to create subcategories

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