Inferring identical by descent sharing of sample ancestors promotes high resolution relative detection

Monica D. Ramstetter*, Thomas D. Dyer†, Donna M. Lehman‡, Joanne E. Curran†, Ravindranath Duggirala†, John Blangero†, Jason G. Mezey*,§, and Amy L. Williams*,1

*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA
†South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA and Edinburg, TX 78539, USA
‡Department of Medicine, University of Texas Health San Antonio, San Antonio, Texas 78229, USA
§Department of Genetic Medicine, Weill Cornell Medicine, New York, NY 10065, USA

Abstract

As genetic datasets increase in size, the fraction of samples with one or more close relatives increases rapidly, resulting in sets of mutually related individuals. We present DRUID—Deep Relatedness Utilizing Identity by Descent—a method that works by inferring the identical by descent (IBD) sharing profile of an ungenotyped ancestor of a set of close relatives. Using this IBD profile, DRUID infers relatedness between these unobserved ancestors and more distant relatives, thereby combining information from multiple samples to remove one or more generations between the deep relationships to be identified. DRUID constructs sets of close relatives by detecting full siblings and also uses a novel approach to identify the aunts/uncles of two or more siblings, recovering 95.4% of real aunts/uncles with zero false positives. We used DRUID to infer relatedness among individuals in both real and simulated data by applying it to close relatives consisting of full siblings or siblings and their aunts/uncles. Compared to PADRE, DRUID correctly infers up to 10% more relatives when using data from two sets of distantly related siblings, and 10–30% more relatives given two sets of siblings and their aunts/uncles. DRUID frequently infers relationships either correctly or within one degree of the truth, with PADRE classifying 43–58% of tenth degree relatives in this way compared to 78–96% using DRUID.

Introduction

Pedigree relationships are fundamental to genetics, with segments of each individual's genome necessarily transmitted across successive generations of ancestors in order to reach the present day. The dynamics of inheritance within pedigrees not only underlie each person's genetic heritage, they affect observed variation in ways that are not fully captured by most population genetic models[1]. The inappropriateness of traditional models has generally had limited impact on analyses to date, largely due to the low rates of closely related samples in most prior studies. Recently however, numerous efforts have been undertaken and are ongoing to generate genetic datasets on a massive scale[2–4]. These substantial studies provide opportunities to better understand human disease genetics[3–5] and to characterize fine-scale population structure and recent

[1] Correspondence: alw289@cornell.edu

demography[6]. Yet to perform these and effectively all genetic analyses, it is necessary to account for the relatedness structure within the sample data[1,7–9].

Given these considerations, relatedness estimation is becoming both more important and more precise because of the widespread relatedness in large samples. A striking illustration of the scope of this is in the ∼500,000 sample UK Biobank data in which almost one-third of the individuals are related by a third degree (e.g., first cousin) or closer relationship to another sample[2]. While this is a higher rate of close relatives than expected by random sampling[2], the proportion of individuals with one or more close relatives in a dataset grows sharply with its size, eventually such that all samples have these relationships.

While association studies account for relatedness using kinship estimates that do not require knowledge of pedigree relationships[10,11], population genetic analyses have the potential to benefit from explicit modeling of segregation patterns using known meiotic distances. As an example, family-based phasing and imputation can achieve near perfect accuracy, with trio phasing used as the gold standard to evaluate accuracy in population phasing methods[12]. Besides these inferences, pedigree samples and their relationships are needed to perform studies of *de novo* mutation[13–15] and recombination[16–19]. Thus, recovering pedigree relationships from genetic data has the potential to empower population genetic analyses in general, and to enhance studies of the two fundamental sources of genetic variation through mining existing datasets for the required family data.

In order to make use of the multi-way relatedness present in large samples, a key question is how best to combine information among the individuals. Two recently developed methods use composite likelihood approaches that enable them to outperform pairwise relatedness inference[20,21]. Results from these methods demonstrate that leveraging multi-way relatedness signals does improve accuracy, but their reliance upon composite likelihoods may be suboptimal. In order to learn the best strategies for relatedness inference, we recently analyzed the accuracy of 12 pairwise relatedness methods and found that: (1) identical by descent (IBD) segment-based methods perform best for classifying a broad range of degrees of relatedness, and (2) the overall accuracy of all methods is highest for close relationships[22].

Building on these findings, we present DRUID, Deep Relatedness Utilizing Identity by Descent, a method that infers relatedness between distant relatives and an ungenotyped ancestor of a set of close relatives. DRUID leverages the fact that, besides the descendants of an individual, all relatives arise through a relationship with at least one parent of each sample. Indeed, a given individual's parents necessarily transmit all the genetic segments they co-inherited with their non-descendant relatives—the so-called IBD segments between these individuals[23]. Yet because a parent transmits only half of his or her genome to each child, and because this transmission is random, the amount of DNA shared with a distant relative has non-trivial variance, with an increasing coefficient of variation for more distant relatives[24]. In consequence, and consistent with results from real data[22], relatedness inference accuracy improves by considering the IBD sharing of a parent with some distant relative.

DRUID applies to full siblings and—through a high specificity approach for locating aunts and uncles—can also incorporate available aunts and uncles of a set of siblings in order to infer the IBD sharing of an ungenotyped grandparent. The method for locating aunts and uncles works on the basis that such individuals, being full siblings of an ungenotyped parent of the set of siblings, necessarily share a sizable amount of DNA that is IBD on both haplotype copies with that parent. This new approach recovers 95.4% of real aunts and uncles (97.5% in simulations) with no false positives.

When sampling of relatives is very dense, there may be two sets of close relatives that are mutually related to one another. In this case, DRUID performs inference of the IBD sharing between two ungenotyped individuals that are the respective ancestors of the two close relative sets. This removes up to four generations of distance between the individuals whose relationships are to be inferred, yielding substantial accuracy gains.

We compared DRUID to the multi-way relatedness method PADRE[20] and to a pairwise inference approach that uses IBD segments called by Refined IBD[22,25]. Both PADRE and DRUID perform much better than pairwise analysis: using simulated data for two sets of full siblings that are distantly related to each other, the multi-way methods classify 12–36% more samples to their exact degree of relatedness. DRUID is also more accurate than PADRE, with up to 10–30% more real and simulated relatives inferred exactly, and with

greater accuracy for nearly all analyses we considered. Turning to relatedness inference that is within one degree of the truth—a metric that is most relevant for more distant relationships—DRUID detects 78–96% of simulated tenth degree relatives while PADRE classifies 43–58% of such relatives in this way.

Methods

DRUID begins by taking input IBD segments inferred from genotype data and performs relatedness inference in two stages. First, it infers the pedigree structure of samples that are connected through first degree relationships—relationships that are very likely to be inferred correctly [22]. In cases where DRUID identifies two or more siblings and only one or neither of their parents, it also locates aunts and/or uncles of these siblings. Second, DRUID performs its primary relatedness inference algorithm, combining IBD information from individuals in the inferred pedigrees when possible to calculate the expected genome-wide IBD sharing proportion between an ungenotyped ancestor and a more distant relative (Figure 1A). Using this quantity, the method then infers the likely degree of relatedness between that ancestor and the distant relative—who may also be an ungenotyped individual. When the relationship to the distant relative arises through one of the close relatives' ancestors (rather than through descent from one of these individuals), the ungenotyped ancestor will be more closely related to the distant sample than the genotyped individuals are.

Because most genotype datasets contain samples that were only collected relatively recently, we make the assumption that the distant relatives do arise through an ancestor and not via descent. DRUID also assumes that there are no errors in the detected IBD segments and that there is no consanguinity among either the close or more distant relatives. That is, we assume that the two parents of any set of siblings are not related to each other and that all IBD segments shared with a sample not in a set of close relatives descend from only one of the close relatives' common ancestors. We have found that, although errors in detected IBD segments do occur, their overall effect is limited when using a relatively accurate IBD detection method. We implemented DRUID to utilize IBD segments detected using Refined IBD [25], but the approach is generally applicable to any method that reports whether samples share one or two IBD segments at a given position. Throughout, we refer to regions in which two individuals share zero, one, or two IBD segments with each other as IBD0, IBD1, and IBD2, respectively. We also use the term *distant relative* to refer to a sample that is not a member of a given set of close relatives and has third degree or more distant relatedness to at least one individual in that set. For first and second degree relatives, DRUID performs inference using pairwise IBD information. DRUID infers up to 13th degree relatedness between any pair it considers, including ungenotyped individuals, labeling others as unrelated. It propagates higher than 13th degree relatedness to descendant individuals in the pedigree structures.

Inferring pedigree structures

To infer the sets of closely related samples and their pedigree structures, DRUID generates a graph in which nodes represent samples and edge labels indicate the relationship type between the linked pair. The input IBD segments are informative about the relationships between the samples, and we use these to estimate $k_{ij}^{(1)}$ and $k_{ij}^{(2)}$, the proportion of their genome that samples i and j share IBD1 and IBD2, respectively. These estimates are simply the sum of the lengths of the IBD1 and IBD2 segments shared between i and j divided by the total genome length, with all lengths in genetic units (e.g., centiMorgans). From this, we derive the estimated kinship coefficient between i and j as $\hat{\phi}_{ij} = \frac{1}{2} \times \hat{k}_{ij}^{(2)} + \frac{1}{4} \times \hat{k}_{ij}^{(1)}$ and deduce their likely relationship type based on this coefficient and $\hat{k}_{ij}^{(2)}$ using the values in Table 1. Initially, the method considers only parent-child, full sibling, and monozygotic (MZ) twin relationships. In the case of MZ twins, DRUID analyzes IBD information from only one of the twins and provides relatedness estimates for the omitted twin(s) that are identical to the analyzed individual.

Starting with an empty graph, DRUID adds nodes and edges corresponding to all inferred full sibling relationships. Following this, the method ensures that for all connected components, the nodes contained in it are all directly connected to one another as full siblings. If any member of a connected component does not have a full sibling relationship with another individual in the component, DRUID checks whether the

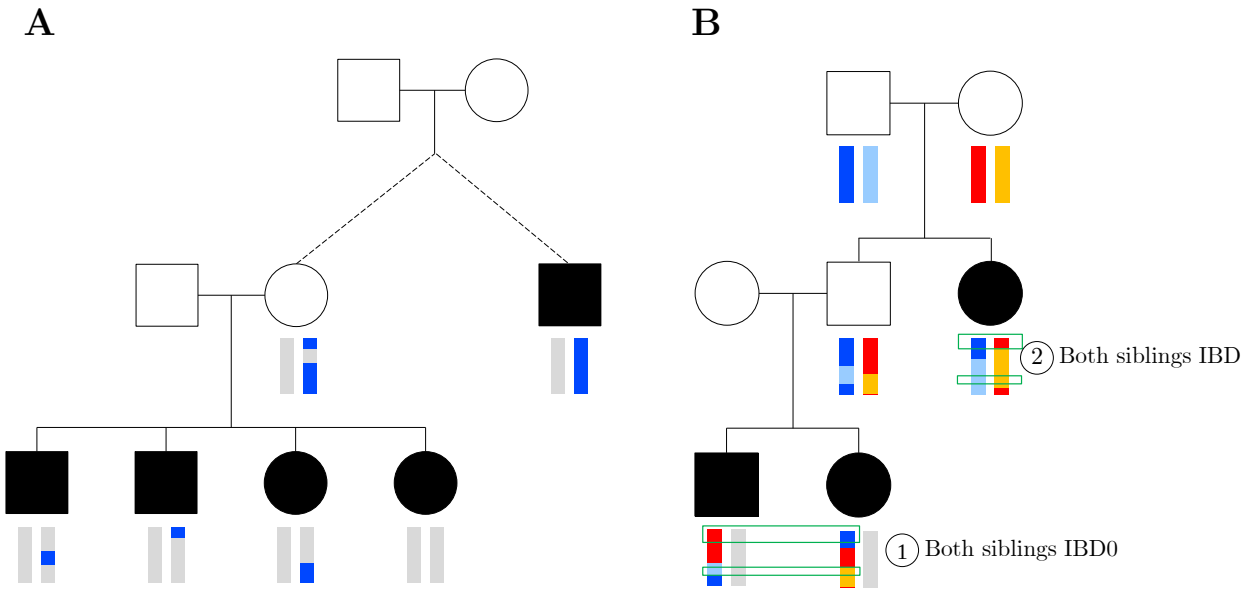

Figure 1: (**A**) Pictorial depiction of relatedness inference approach used in DRUID. Genotyped individuals are shown as filled shapes and haplotypes are colored vertical bars below analyzed samples; the dashed line indicates the number of generations to the most recent common ancestors between the full siblings and the distant relative on the right is unknown. The blue regions in the full siblings represent IBD segments shared with the distant relative on the right. DRUID infers the ungenotyped mother's IBD profile as the union of her children's IBD segments. (**B**) Example haplotype transmissions from grandparents to two full sibling grandchildren (bottom generation) as well as the siblings' father and aunt, with regions of the same color descended from the corresponding grandparent chromosome (we ignore the gray chromosomes for simplicity). Given genotype data for the two siblings and their aunt, the $\text{IBD}^{(011)}$ regions are those where the two siblings are (1) IBD0 with each other, and (2) both are IBD with the aunt, indicated by green boxes. These are positions where the siblings' ungenotyped father is IBD2 with the aunt.

	DRUID		DRUID$_C$	
	$\hat{k}^{(2)}$	$\hat{\phi}$	$\hat{k}^{(2)}$	$\hat{\phi}$
MZ twin	$[\frac{1}{2^{1/2}}, 1]$	$[\frac{1}{2^{3/2}}, 1)$	$[\frac{1}{2^{1/2}}, 1]$	$[\frac{1}{2^{3/2}}, 1)$
Full sibling*	$[\frac{1}{2^{11/4}}, \frac{1}{2^{3/2}})$	$[\frac{1}{2^{7/2}}, \frac{1}{2^{3/2}})$	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Parent-child	$[0, \frac{1}{2^{9/2}})$	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	$[0, \frac{1}{2^{9/2}})$	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Double cousin	$[\frac{1}{2^{9/2}}, \frac{1}{2^{11/4}})$	$[\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$	$[0.1, 1]$	$[\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
Second degree	-	$[\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$	-	$[\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
D^{th} degree	-	$[\frac{1}{2^{(2D+3)/2}}, \frac{1}{2^{(2D+1)/2}})$	-	$[\frac{1}{2^{(2D+3)/2}}, \frac{1}{2^{(2D+1)/2}})$

Table 1: Relationship classification rules used by DRUID and DRUID$_C$. The ranges of $\hat{\phi}$ and their mapping to relationships are based on those used by KING [26]. First degree relatives that do not fall into a given $\hat{k}^{(2)}$ range are not used in DRUID's pedigree reconstruction. MZ twin: monozygotic twin. * We lowered the full sibling $\hat{k}^{(2)}$ value (and consequently its $\hat{\phi}$) relative to that suggested in the KING paper as this yielded better results in simulations.

majority (rounded up in cases of even numbers) of sibling pairs are labeled as full siblings of that individual; in that case, it labels any unconnected pairs as full siblings as well, and otherwise, it breaks the minority full sibling connections. DRUID can also be run so that close relative inferences are made more conservatively, which we refer to as DRUID$_C$. In this mode, for non-fully connected components, DRUID successively removes edges to nodes that are the least connected (randomly selecting a node in the case of a tie) until the resulting components are fully connected. While we expect the standard DRUID method to be applicable in most settings, DRUID$_C$ may be useful in datasets where less common relationship types exist, such as three-quarter siblings.

Next, DRUID incorporates parent-child relationships into the graph, and, when possible, determines which individual is the parent using analysis of relatedness to full siblings. Full sibling relationships are informative about which sample is the parent in the following way. Suppose i and j are inferred to have a parent-child relationship. If i has at least one full sibling in the graph, then those full siblings either all have a parent-child relationship to j, in which case DRUID assigns j as the parent of all the full siblings. Alternatively, DRUID assigns j as the child of i if it has a second degree relationship to all full siblings of i. In all other cases, DRUID labels i and j as general first degree relatives.

As described below, DRUID is able to infer aunts and uncles of a set of siblings, but because mistakes in the inferred pedigree structures can lead to systematic errors in later classification, we do not currently infer half-siblings or other types of relatives. However, DRUID does utilize user-specified relationships, including half-siblings and known directions for parent-child relationships. To incorporate this information, DRUID performs initial inference of a pedigree using first degree relationships and also includes second degree relatives with non-specific second degree relationship edge types. It then compares the inferred pedigree with the user specifications, prints a warning for any discrepancies, and changes the relationships to match the specifications. Following this, the algorithm proceeds to infer aunts and uncles.

Identifying aunts and uncles of a set of siblings

In principle, the length of IBD segments shared between second degree relatives are informative about their underlying relationship type: grandparent-grandchild, half-sibling, avuncular, or double cousin. However, the method RELPAIR, which implements an approach based on this idea, has limited ability to discriminate between these relationships, with previously reported classification true positive rates ranging from 37% to 72% for the different relationship types (excluding double cousins which it does not consider) [27].

In DRUID, we take a different approach based on the IBD sharing patterns among a set of three samples consisting of a pair of full or half-siblings and a second degree relative, determining whether that relative

is an aunt or uncle of the siblings. A distinguishing feature of full siblings is that they share a substantial fraction of their genome IBD2 with each other. As such, the ungenotyped parent of a pair of siblings will share some regions IBD2 with the siblings' aunt or uncle but not with other types of second degree relatives. This marks a unique sharing pattern that allows us to pinpoint these relatives with high specificity. As depicted in Figure 1B, at positions in which two siblings share no IBD segments with each other or are IBD0, they will have inherited distinct haplotype copies from their parents. If the two siblings each also share an IBD segment with another relative in these regions, one of their parents must share two distinct haplotypes IBD with that relative. (This ignores double cousins to whom both parents are related, a case we address below.) Appreciable levels of this IBD pattern among the three samples are a strong indicator that the second degree relative is a full sibling of an ungenotyped parent and therefore an aunt or uncle of the two siblings.

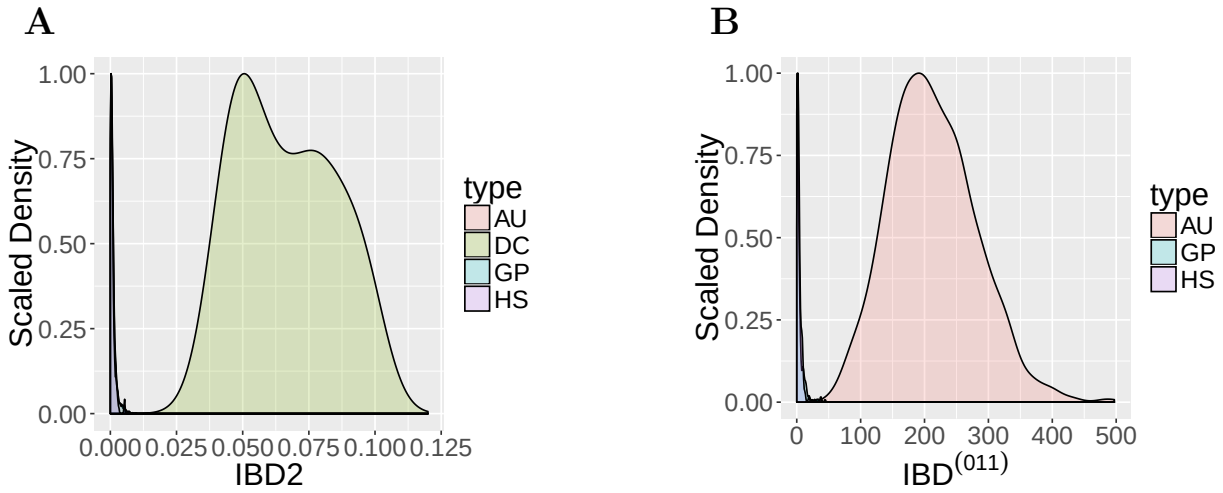

Figure 2: (**A**) Scaled density showing the genome proportion shared IBD2 between real second degree relative pairs from the SAMAFS data. Abbreviations: 'AU', avuncular pairs; 'DC', double cousins; 'GP', grandparent-grandchild pairs; and 'HS' half-siblings. (**B**) Length of genome found to be $\text{IBD}^{(011)}$ between two full siblings and a second degree relative using real data from SAMAFS. Abbreviations as in (A). Double cousins are filtered based on their IBD2 proportion and therefore not shown.

To infer aunts and uncles, DRUID locates all pedigrees containing two or more full or half-siblings that lack data for one or both parents. It then finds all samples that are inferred as second or (to increase power) third degree relatives of at least one of these siblings (Table 1) and considers the $\hat{k}^{(2)}$ level between each sibling and that relative. Double cousins are second degree relatives that are expected to share $\frac{1}{16}$ of their genome IBD2 with each other, and these relatives can also have non-trivial proportions of the IBD pattern otherwise indicative of aunts and uncles. DRUID infers any relative r with whom one of the siblings s shares $\hat{k}^{(2)}_{sr} > \frac{1}{2^{9/2}}$ (or $\hat{k}^{(2)}_{sr} > \frac{1}{10}$ for DRUID_C) as a likely double cousin, and it ignores these in later analyses (Figure 2A, S1A).

For the remaining second/third degree relatives, the method first randomly selects a pair of siblings s_1 and s_2 and calculates the total genetic length of the regions they are inferred to be IBD0 with each other and where both are inferred to be IBD with the other relative r. (We count both IBD1 and IBD2 sharing between each sibling and r, though the total IBD2 levels tend to be small.) This gives a value we refer to as $\text{IBD}^{(011)}_{s_1,s_2,r}$, which is a lower bound on the length of the IBD2 sharing between r and the parent of s_1 and s_2 (assuming all the IBD segments descend from only one parent). Analyses of both real and simulated second degree relatives indicate that when $\text{IBD}^{(011)}_{s_1,s_2,r} > 50$ cM, the second/third degree relative is effectively always an aunt or uncle of the siblings (Figure 2B, S1B). These data also indicate that when $\text{IBD}^{(011)}_{s_1,s_2,r} < 20$ cM, the

second/third degree relative is very unlikely to be an aunt/uncle. The algorithm makes inference using these length values as cutoffs, and when 20 cM $\leq \text{IBD}^{(011)}_{s_1,s_2,r} \leq$ 50 cM for a given sibling pair, it moves to consider a different pair of siblings to get more definitive results. If analyses of all sibling pairs fall in this ambiguous range, the relative is not inferred as an aunt/uncle. Upon inferring an individual as an aunt/uncle of a set of siblings, DRUID adds her/him to the graph as the indicated relationship, and also incorporates avuncular connections with any known full siblings of that aunt/uncle. If the input specifications include avuncular relationships, DRUID compares these to the inferred avuncular relationships at this stage, and it prints a warning if it did not infer any input relationship. In all cases, it trusts the user specifications and performs later inference using these.

With the pedigree relationships between sets of close relatives determined, DRUID undertakes the second stage of its analysis by reconstructing the IBD profiles of the ancestors of these sets. This inference focuses on two configurations of close relatives: siblings (including half-siblings) and siblings together with their aunts/uncles.

Inferring IBD sharing of a parent using data from siblings

A parent transmits to each child a random portion of the IBD segments he/she shares with any relative. Whereas a single child inherits only half of each parents' genome, data for additional children include a more complete subset of the genomes of their parents, including receiving a larger fraction of the IBD segments they each carried. In particular, if $T_{\tilde{p}}$ corresponds to the amount of DNA a parent $\tilde{p}$ transmitted to a set S of full sibling children, then assuming the haplotype transmitted at each locus follows a binomial distribution, $E[T_{\tilde{p}}|S] = 1 - \frac{1}{2^{|S|}}$.

Given the assumption that the parents of S are unrelated, only one parent will have transmitted all the IBD segments that those siblings share with any distant relative d. Thus, although genetic data for the two parents are unobserved, the union of all IBD segments shared by the siblings with d constitutes a partial set of the IBD regions one of the parents shared with d (Figure 1A, S2A). Notably, which parent transmitted these IBD segments is unknown, but this information is not needed to determine the degree of relatedness between that ungenotyped parent $\tilde{p}$ and d.

We relate the proportion of the genome that an ungenotyped parent $\tilde{p}$ and d share IBD1 to the observed IBD segments in S as:

$$k^{(1)}_{\tilde{p}d} \times T_{\tilde{p}} = \mathit{Length}\left(\bigcup_{s \in S} I_{sd}\right) \times \frac{1}{L}. \tag{1}$$

Here, I_{sd} is a set containing the markers that are detected as IBD between a given sibling s and d, where we include both IBD1 and IBD2 segments—the latter arising infrequently and due to error when our assumptions hold. The $\mathit{Length}(I)$ function gives the genetic length of all regions containing sequential markers that are called IBD, and L is the total length of the genome, both in cM. As the union of all IBD regions in the siblings contain only a proportion $T_{\tilde{p}}$ of the parent's IBD regions, we scale the parent's IBD sharing proportion $k^{(1)}_{\tilde{p}d}$ by this quantity. This implicitly models the sharing proportion in the unobserved regions of the parent's genome as being the same as in the transmitted regions.

Because our aim is to provide an estimate of the degree of relatedness between all samples, we compute a point estimate of $k^{(1)}_{\tilde{p}d}$ by calculating its expectation rather than modeling its full distribution. Conveniently, $k^{(1)}_{\tilde{p}d}$ (the fraction of DNA $\tilde{p}$ shares with d) is independent of $T_{\tilde{p}}$, so dividing the right hand side of Equation 1 by $E[T_{\tilde{p}}|S]$ provides this expectation. The estimated kinship coefficient between $\tilde{p}$ and d is then $\hat{\phi}_{\tilde{p}d} = \frac{1}{4} \times E[k^{(1)}_{\tilde{p}d}|S]$, from which we infer a degree of relatedness (Table 1).

An alternative to dividing by the expectation of $T_{\tilde{p}}$ is to estimate of its realized value by analyzing the observed IBD sharing between the siblings. Specifically, at positions where full siblings are all IBD2 with one another, both parents will have transmitted only one haplotype copy, or half of their genome. Conversely, at positions where at least two children are IBD0 with each other, each parent will have transmitted both haplotype copies or all their genetic material. We applied this logic and compared the performance to that

of using the expectation. The results of both approaches are similar but the expectation yields slightly higher accuracy (not shown) and is also more computationally efficient to calculate. This is possibly due to uncertainty in the transmission status of each parent when the full siblings are all IBD1 with each other: at such locations we model the transmission rates of both parents as the expectation of 75%. Another possibility is that false negative or false positive IBD segments adversely affect this estimation.

DRUID analyses relatedness using the union of full siblings and any of their half-siblings that it infers as equally related to d. It considers a set of half-siblings H (that are full siblings of one another) to be equally related to d along with the primary full siblings in S (Figure S2B) if

$$\max_{h \in H} \hat{\phi}_{hd} > \frac{3}{4} \min_{s \in S} \hat{\phi}_{sd}. \tag{2}$$

For the purposes of this analysis, given two or more sets of half-siblings, we take S to be the set containing the individual s^* with the highest overall $\hat{\phi}_{s^*d}$ value. If the above equation does not hold, DRUID analyses H independently of S.

Inferring IBD sharing of a grandparent using siblings and aunts/uncles

When data are available for a set of siblings together with some number of their aunts and uncles, the IBD segments these individuals share with a distant relative descend from a grandparent (Figure S2C). The expected proportion of a grandparent $\tilde{g}$'s genome transmitted to these individuals is $E[T_{\tilde{g}}|A,S] = 1 - \frac{1}{2^{|A|}} + \frac{1}{2^{(|A|+1)}} \times E[T_{\tilde{p}}|S]$, where A is the sets of aunts/uncles of the siblings in S, and $\tilde{p}$ is their ungenotyped parent. Here, $1 - \frac{1}{2^{|A|}}$ is the expected amount of the grandparent's genome transmitted to his/her children in A, and the final term gives the expected amount of DNA transmitted to $(|A|+1)^{\text{st}}$ child multiplied by the expected genome proportion that child transmitted to the grandchildren in S.

We estimate the proportion of the genome that the ungenotyped grandparent $\tilde{g}$ shares IBD with a distant relative d in an analogous way to that of ungenotyped parents:

$$k^{(1)}_{\tilde{g}d} \times T_{\tilde{g}} = \mathit{Length}\left(\bigcup_{r \in R} I_{rd}\right) \times \frac{1}{L}, \tag{3}$$

where $R = A \cup S$. As above, we take the expectation of the left hand side and solve for $E[k^{(1)}_{\tilde{g}d}|A,S]$.

As the siblings may have aunts/uncles both through their mother and their father, we group the aunts/uncles that are inferred to be siblings of one another to form up to two sets of aunts/uncles associated with a sibling set S, denoting these as $A^{(1)}$ and $A^{(2)}$. To infer relatedness to a distant relative d, DRUID includes a given set of aunts/uncles $A^{(i)}$ of a sibling set S when

$$\max_{a \in A^{(i)}} \hat{\phi}_{ad} > \min_{s \in S} \hat{\phi}_{sd}. \tag{4}$$

When this inequality holds for both sets $A^{(1)}$ and $A^{(2)}$, we use the set with the higher average $\hat{\phi}_{ad}, a \in A^{(i)}$. When no set fits this criteria, we continue the analysis using only the sibling set.

Inferring IBD sharing between two sets of close relatives

In sufficiently large datasets or those with family-based recruitment, DRUID will infer several sets of closely related samples. When this occurs, distant relatedness may exist between two close relative sets and not merely to a single distant relative d. Inferring the amount of IBD shared between two ungenotyped ancestors from the two pedigrees enables inference at greater resolution than the potential alternative of using a single member of one of the pedigrees. Given two pedigrees 1 and 2 with corresponding sets A_1, A_2 of aunts/uncles and S_1, S_2 of siblings, we estimate the IBD sharing between two ungenotyped grandparents $\tilde{g}_1$ and $\tilde{g}_2$ as

$$k^{(1)}_{\tilde{g}_1\tilde{g}_2} \times T_{\tilde{g}_1} \times T_{\tilde{g}_2} = \mathit{Length}\left(\bigcup_{\substack{r_1 \in R_1 \\ r_2 \in R_2}} I_{r_1 r_2}\right) \times \frac{1}{L}, \tag{5}$$

where $R_i = A_i \cup S_i$ for $i \in \{1, 2\}$. Analogous equations apply for estimating relatedness between ungenotyped parents $\tilde{p}_1$ and $\tilde{p}_2$ when both pedigrees only have sibling sets, and for cases where only one pedigree has an aunt/uncle set. When IBD segments exist between at least one member of two sibling sets S_1 and S_2, DRUID performs this analysis.

To determine whether to include available aunt/uncle sets, we consider the two pedigrees separately. For example, for pedigree 1, suppose siblings S_1 have two aunt/uncle sets $A_1^{(1)}, A_1^{(2)}$ (through their two parents). Then, similar to Equation 4, we include a given set $A_1^{(i)}$ in the inference if:

$$\max_{\substack{a \in A_1^{(i)} \\ s_2 \in S_2}} \hat{\phi}_{as_2} > \min_{\substack{s_1 \in S_1 \\ s_2 \in S_2}} \hat{\phi}_{s_1 s_2}. \tag{6}$$

As before, if this inequality holds for both aunt/uncle sets, we include the set with higher average $\hat{\phi}_{as_2}$ for $a \in A_1^{(i)}, s_2 \in S_2$. The inequality for pedigree 2 is analogous.

Given our assumptions, the only relatives that will share regions IBD2 with each other are full siblings. Therefore, although distant relatives themselves will not share IBD2 regions, there are circumstances where their ancestors will have IBD2 sharing. One example of this is two sibling sets that are first cousins of each other: their parents are full siblings that consequently share non-trivial amounts of IBD2. It is important to account for any IBD2 when estimating the kinship coefficient between these ungenotyped ancestors. Extending the IBD$^{(011)}$ concept from inferring aunts/uncles to apply to first cousins, we infer regions that are IBD2 between the ungenotyped parents of two full sibling sets in the following way. First, for every pair of siblings $s_{11}, s_{12} \in S_1$, we locate regions in which they are IBD0 with each other. We then identify places these siblings are each IBD to two different siblings $s_{21}, s_{22} \in S_2$. So long as this IBD inference is correct, these regions must be IBD2 in the full sibling parents of S_1 and S_2. Similar to Equation 5, we compute $E[k^{(2)}_{\tilde{p}_1\tilde{p}_2}|S_1, S_2]$ as the total length of these inferred IBD2 regions divided by the genome length L, and also divided by $E[T_{\tilde{p}_1}|S_1] \times E[T_{\tilde{p}_2}|S_2]$. We use this as the $k^{(2)}$ term when computing $\hat{\phi}_{\tilde{p}_1\tilde{p}_2}$, adjusting for the fact that these IBD2 regions were also counted in the $k^{(1)}$ term from Equation 5. We perform this inference only when the initial $k^{(1)}$-based estimate of $\hat{\phi}_{\tilde{p}_1\tilde{p}_2}$ corresponds to a third degree or closer relationship. This approach is general and applies to any pair of sibling sets—including any sets of aunts/uncles—whenever they each have two or more members.

Determining degrees of relatedness for all sample pairs

To perform relatedness inference between all samples, DRUID must determine when to use standard pairwise relatedness measures and when to apply its multi-way inference. Prior to inferring the pedigree structures for close relatives, DRUID infers a pairwise-only degree of relatedness (Table 1). When this value is less than or equal to two for a given pair, DRUID reports that degree; additionally, if neither sample is in the graph (i.e., neither is in a close relative set), DRUID reports the pairwise relatedness estimate. Otherwise, the method determines whether a parent or grandparent of a set of close relatives is in the graph, and if so, whether that ancestor has the same or higher IBD sharing level with the other sample. If so, it successively moves up to older generations until arriving at two samples whose relatedness DRUID is to estimate.

Let i, j be the two samples with relatedness to be inferred where at least one is a member of a close relative set. If neither i nor j have any full siblings, half-siblings, or aunts/uncles, DRUID reports the pairwise degree of relatedness between these samples and deduces from this the relatedness degrees between them and all the descendants in the pedigrees to which they each belong. Otherwise, when one or both of the samples have full siblings, it uses these for inference, and also checks the relevant inequalities to determine whether to use any half-siblings and aunts/uncles in the inference.

Simulated relatives

We developed Ped-sim, a simulator for generating sample data from a specified pedigree structure that can have any number of generations and that is only (currently) constrained such that parents must marry

either founders or samples from the same generation. Thus, Ped-sim is able to generate common relationship types including full and half-siblings, as well as less common types such as double cousins and various forms of inbred individuals. Ped-sim randomly samples founders from phased input samples and generates non-founders by simulating recombination events between the two haplotypes of all parents in the pedigree. The program uses either a sex-averaged genetic map or male- and female-specific genetic maps, with the sexes of parents randomized in the presence of sex-specific maps. The current version uses a Poisson model of recombination, which does not account for crossover interference[18]. However, models of crossover interference exist[28,29] and simulation from these models is both feasible and planned for an upcoming version.

In the present study, we simulated pedigree samples using data from European-descent individuals collected from throughout Europe, the United States, Australia, and New Zealand[30]. We began by using all samples and SNPs listed (in files distributed by the European Genome-phenome Archive) as included by the original study following quality control procedures[30]. We further removed SNPs that failed a Hardy-Weinberg equilibrium filter ($P < 10^{-4}$), had more than 5% missingness, or were outside the regions covered by the HapMap genetic map[31], yielding a total of 10,299 individuals and 463,366 SNPs. Following phasing of all samples jointly with Beagle[32] 4.1 using the HapMap genetic map, we ran PLINK[33] v1.90b2k to estimate pairwise relatedness and input these estimates to FastIndep[34] to find a set of unrelated individuals. For purposes of filtering samples to use in this simulation, we considered all sample pairs with kinship $\hat{\phi} < \frac{1}{2^{11/2}}$, (i.e., fifth degree or more distant) as "unrelated." Taking the 8,955 resulting phased samples as input, and using recently generated sex-specific genetic maps[35], we simulated pedigrees with Ped-sim v0.87.1b. Because the sex-specific genetic maps cover a slightly smaller region than that spanned by the input SNPs, the simulated individuals included 462,828 SNPs.

The pedigree structures we simulated consist of distant relatives that are all some form of cousins. That is, their common ancestors are two grandparents from some number of generations ago depending on the degree of relatedness, as depicted in Figure S3. The four general pedigree types we simulated are: (1) five full siblings and one D^{th} degree relative, $D \in \{3, ..., 10\}$; (2) two sets of five full siblings such that the sets are D^{th} degree relatives of one another, $D \in \{3, ..., 10\}$; (3) five full siblings, two of their aunts/uncles, and one D^{th} degree relative, $D \in \{4, ..., 10\}$; and (4) two sets that each contain five full siblings and two of their aunts/uncles such that the siblings in the two sets are D^{th} degree relatives of one another, $D \in \{5, ..., 10\}$. As DRUID uses pairwise inference for first and second degree relatives, we chose D such that none of the distant relatives had relationships closer than third degree.

For all four pedigree types and the corresponding degrees of relatedness D, we simulated 120 replicate pedigrees, each randomized with respect to the assigned founder individuals and the sex of the parents. Specifically, we simulated the data in 10 batches, each batch containing 12 replicates of the various pedigree types and degrees. For any given run, Ped-sim assigns an input individual as a founder only once, but independent runs will in general assign some of the same input samples as founders. Thus, while the batches contain some overlapping founders, any specific pedigree structure will be extremely likely to have different permutations of founders across batches.

To analyze the performance of aunt/uncle inference, we further simulated various forms of second degree relatives and included these in the 10 simulation batches along with the all other pedigree types. Specifically, each batch included: 12 sets of two full siblings and one aunt/uncle; 12 sets of two full siblings and one half-sibling; 6 sets of two full siblings and their two grandparents (with analysis done with respect each grandparent separately); and 12 sets of two full siblings and one double cousin.

To aid in the phasing step that precedes IBD detection, we incorporated an additional 1,674 unrelated samples in each batch, yielding a total of 5,022 individuals per batch. We used a Ped-sim option to output samples that were not used as founders in each batch, so the additional samples are drawn from the same European-descent sample data as the founders. Overall, one-third of each batch consists of these unrelated individuals (more precisely, fifth degree or more distant samples).

We inferred IBD segments separately in each batch by running Refined IBD (part of Beagle 4.1 and using the HapMap genetic map) three times and reporting IBD regions as the union of identified segments across these runs (as recommended by the authors[25]). We used these segments in all analyses of both the pairwise relatedness method we refer to as Refined IBD, and for running DRUID. Our analyses range over different

sizes of full sibling sets S, with $|S| \in \{2, 3, 4, 5\}$, and we randomly removed one sibling to obtain the set of siblings used for analysis of $|S| - 1$ siblings. Thus the included samples with smaller numbers of siblings are proper subsets of those with more. In a similar way, when analyzing the pedigrees that contain sets of aunts/uncles A, we test with $|A| \in \{0, 1, 2\}$, and ensure that we include the same aunt/uncle for the range of sibling numbers $|S|$ whenever $|A| = 1$. (When $|A| \in \{0, 2\}$ the included aunts/uncles are necessarily identical for all values of $|S|$.)

PADRE[20] requires input from PRIMUS[36], a method for reconstructing pedigree structures using estimates of genome-wide IBD sharing of all pairs of samples. To infer these pedigrees, we ran PRIMUS v1.9.0 on each simulation batch (including the extra 1,674 unrelated samples) in two stages. First, we used the `--no_IMUS` and `--no_PR` options so that PRIMUS only ran PLINK v1.90b2k to calculate genome-wide IBD estimates and did not perform pedigree reconstruction. We then subdivided the resulting `.genome` file (from PLINK) into files containing only pairs of samples contained in each distinct pedigree structure (including the siblings, aunts/uncles, and all deep relatives). Afterwards we ran PRIMUS on each of these files, this time such that it inferred pedigrees, but with the subdivided files preventing PRIMUS from searching for relatives across distinct pedigree structures.

PADRE also requires initial likelihoods for pairwise degrees of relatedness from ERSA[37], which itself requires IBD segments inferred by GERMLINE[38]. As part of the Refined IBD analysis noted above, we simultaneously inferred phase (i.e., using Beagle 4.1) in each batch, and then used GERMLINE version 1.5.1 (with options `-err_het 2 -err_hom 1 -min_m 1` as recommended by the ERSA authors[39]) to detect the IBD segments we provided to ERSA 2.0. Although PADRE's documentation suggests using non-default parameters for ERSA, we find slightly poorer performance using those parameters as opposed to the defaults (not shown), and so we used the defaults. We then provided the ERSA results and the output from PRIMUS as input to PADRE. By default, PADRE infers only up to ninth degree relatives, and we modified this to enable it to infer up to degree 13.

Real data

Besides simulated data, we evaluated DRUID and the other methods on real SNP array genotypes from the San Antonio Mexican American Family Studies (SAMAFS)[40–42]. Specifically, we analyzed 2,485 individuals typed at 521,184 SNPs from one of several dozen pedigrees. We previously described our quality control procedures for these data[22], but in brief, we set sites with Mendelian errors to missing and mapped the SNP probes to GRCh37. Following SNP filters based on dbSNP and other auxiliary databases, we filtered SNPs with more than 2% missingness and samples with more than 10% missingness. The samples were typed on several Illumina SNP arrays, and our filters resulted in the inclusion of SNPs that are common to all arrays. We also excluded from consideration 2,618 pairs of samples (out of more than 3 million) that have evidence of being descended from parents that are cryptically related.

To analyze these data, we phased and detected IBD in all 2,485 individuals using Refined IBD three times (phase generated simultaneously using Beagle 4.1) and performed analysis using the union of the segments identified in all runs. As in the simulated data, we ran PRIMUS in two stages, first to obtain genome-wide IBD sharing estimates, and second on subdivided (PLINK generated) `.genome` files. These subdivided files restrict PRIMUS to analyzing two sets of close relatives to be analyzed (described next). We also ran ERSA on the phased data for all 2,485 samples using recommended settings. Finally, we ran PADRE using the output of PRIMUS and ERSA and with the modification to enable it to infer up to relatedness degree 13.

We analyzed the performance of DRUID and PADRE on the SAMAFS data by analyzing two reported close relative sets that are also reported to be distantly related to one another. Any misreported relationships have the potential to confound this analysis, though they should equally affect both DRUID and PADRE and in general be fairly infrequent. We analyzed two distantly related sets of full siblings, and when one or two sets of aunts/uncles associated with these sibling sets are available, we also analyzed the same samples after adding one (choosing randomly if there are two) and then both sets of aunts/uncles.

To evaluate the performance of DRUID's aunt/uncle inference procedure, we considered all reported sets

of ≥ 2 full siblings and their reported second degree relatives. Once again, misreported relationships may confound this analysis, although our results suggest most relationships are correct. We randomly selected two siblings from each full sibling set for analysis, repeatedly sampling non-overlapping pairs of siblings when more than two siblings are available. Then, for every selected pair of siblings, we analyzed their $k^{(2)}$ and $\text{IBD}^{(011)}$ levels to all their reported second degree relatives. Note that whenever multiple pairs of siblings from the same sibling set exist, this considers the same second degree relatives several times.

As a more conservative analysis, we also ran DRUID and Refined IBD on a restricted set of samples consisting of confidently inferred sets of full siblings and their aunts/uncles together with only one distant relative. For this, we filtered reported full siblings to those whose inferred pairwise degree of relatedness is one, obtaining sets in which all pairs of siblings have a first degree relationship. To ensure the reported aunts/uncles of these full sibling sets are correct, we check whether the pairwise relationship of a given aunt/uncle with each full sibling is second degree. If this is case, we accepted this individual as an aunt/uncle. Because inference of second degree relatives has a slightly lower power than first degree inference [22], for each aunt/uncle verified in this manner, we also include any validated full siblings (based on the first degree relationship criteria) of this person as aunts/uncles. We also performed an analysis using half-siblings, and for this purpose, we generated confident sets of half-siblings in the same manner as for aunts/uncles. To increase sample size, we analyzed the inference rates using each set of close relatives and (one by one) all their available distant relatives of a specified degree. Therefore, results include multiple analyses of the close relative sets but treats them as independent.

Results

To evaluate DRUID, we examined the rate at which it correctly classifies the degree of relatedness—both exactly and to within one degree of the truth—between simulated and real samples in comparison to PADRE and Refined IBD. The latter approach uses the IBD regions detected by Refined IBD to infer a pairwise degree of relatedness. This is the same approach DRUID employs to infer first and second degree relatives and is among the top performing pairwise relatedness methods [22].

PADRE uses a composite likelihood method to infer relatedness between two networks of samples, with each network containing individuals whose relationship to one another must be represented in an input pedigree that PRIMUS inferred. In the presence of ambiguous relationships, PRIMUS outputs multiple possible structures together with their likelihoods and PADRE considers each of these likelihoods and their corresponding structures in its analysis. PADRE also takes as input pairwise relatedness likelihoods inferred by ERSA 2.0 and uses these to identify the degrees of relatedness that maximize the overall composite likelihood. The full composite likelihood for a given pair of pedigree structures and relatedness degree between distant relatives is the product of the pedigree likelihoods (from PRIMUS) and all pairwise relationship likelihoods (from ERSA) that exist across the two networks. Because PADRE is intended to apply to two non-singleton networks of samples, we ran it on the simulated pedigree structures that contain two non-singleton close relative sets (i.e., types 2 and 4). Results from running DRUID and Refined IBD on the other types of simulations are in Figures S4, S5, S8, and S9.

Below we briefly describe DRUID's performance for inferring aunts/uncles and then proceed to compare it to Refined IBD and PADRE for distant relatedness inference in simulated and real data. For all analyses of distant relatives, we report classification rates only for pairs in the sibling sets (i.e., excluding the aunts/uncles). This enables a direct comparison of results for the same degree of relatedness with and without the inclusion of aunts/uncles. For simulated pedigree types 1 and 3, which have only one distant relative, the target number of pairs of distantly related samples to classify is $|S|$. For pedigree types 2 and 4, there are $|S|^2$ pairs to classify. We compute the classification rate for a given collection of distant relatives by averaging over the indicated number of pairs; we then average these rates over all distant relative collections (i.e., over each instance of two distantly related sets of close relatives). In simulations, this average is over the 120 replicate structures, and in real data, we report the numbers of structures.

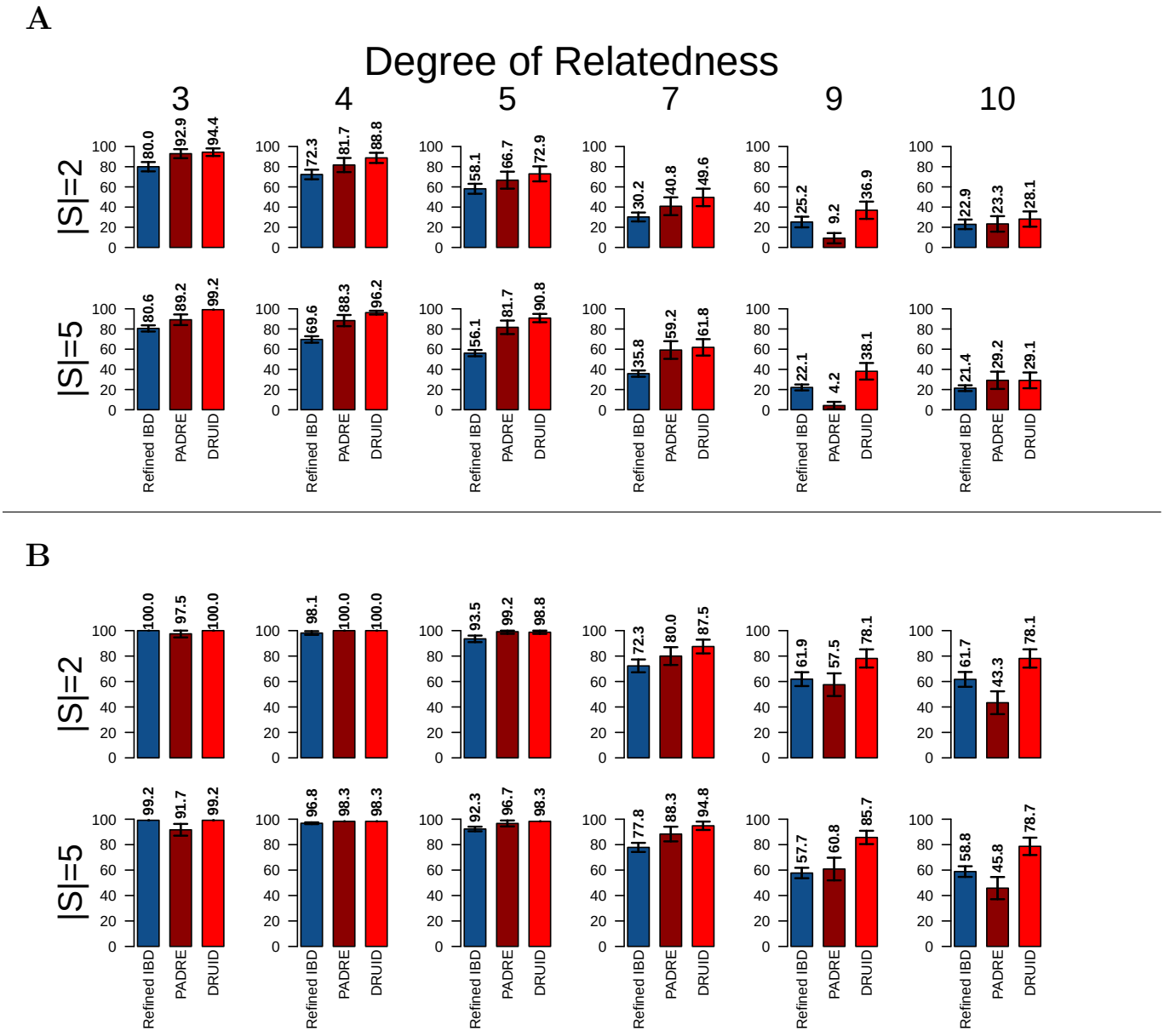

Figure 3: (**A**) For simulated type 2 pedigrees, average percent of distantly related sample pairs from the two sibling sets that are inferred as their true degree of relatedness using Refined IBD, PADRE, and DRUID. Rows of bar plots have the same number of siblings included in both sibling sets, indicated as $|S|$ (left). Columns show results for different degrees of relatedness, with the true degree listed above. (**B**) As in (A), but shows the average percent of distant relatives inferred to be related as the true degree D or as $D \pm 1$. Error bars are bootstrapped (over complete relative sets) 95% confidence intervals.

Inferring aunts and uncles with DRUID

We used the real SAMAFS data as well as simulated individuals to examine DRUID's performance for aunt/uncle classification. As noted earlier, in nearly all cases, a pair of siblings s_1 and s_1 and one aunt/uncle a have more than 50 cM of the $\text{IBD}^{(011)}_{s_1,s_2,a}$ pattern DRUID uses for classification (Figures 2B and S1B). Indeed, considering SAMAFS data for two reported full siblings and one reported second degree relative, DRUID infers 95.4% of all aunt/uncles as such with zero false positives (total 1449 aunts/uncles out of 1893 second degree relatives). Results using simulated data are similar: with two full siblings and various second degree relatives (Methods), DRUID recovers 97.5% of aunts/uncles, also with zero false positives (total 120 aunts/uncles out of 480 second degree relatives). The filter that removes double cousins from consideration (Figure 2A and S1A) is highly effective as it captures all double cousins in both real and simulated data with zero false positives (total 38 real and 120 simulated double cousins).

Distant relative inference in simulated data

Figure 3 depicts results for type 2 pedigrees (i.e., those with two distantly related sets of siblings), subdivided into exact and within-one-degree inference accuracy, for $|S| = 2$ and $|S| = 5$, and for most degrees of relatedness D (full results in Figures S6 and S7). When analyzing type 2 pedigrees, both DRUID and PADRE outperform the pairwise method Refined IBD for $D = 3$ to 9, with between 12–36% more relatives classified exactly for these degrees (Figure 3A). DRUID and PADRE show greater improvement relative to Refined IBD when $|S| = 5$ than $|S| = 2$, consistent with the methods having more information with which to make inference for larger $|S|$. The differences among these methods are more modest for $D = 10$, presumably because there is limited information with which to perform classification at this degree. Indeed, DRUID only classifies 28% of relatives exactly for $D = 10$. Despite this, DRUID's within-one-degree classification accuracy is quite high for $D = 10$ at 78–80% (Figure 3B).

For the multi-way relatedness methods, DRUID shows greater accuracy than PADRE for most values of D, classifying up to 10% more pairs exactly when $D = 3$ to 7 (Figure 3A). Considering within-one-degree classification, DRUID has 100% accuracy up to $D = 5$, with PADRE showing some (though modest) inaccuracies up to this degree (Figure 3B). However, for more distant degrees, PADRE's within-one-degree classification rates drop relative to DRUID. For example, when $|S| = 5$, PADRE infers only 63% and 47% of sample pairs within-one-degree for $D = 9$ and $D = 10$, respectively. By contrast, DRUID infers, respectively, 90% and 80% of these relatives to within one degree of the truth. For exact inference of higher relatedness degrees, PADRE's accuracy drops when $D = 9$, but is more competitive with DRUID when $D = 10$. As noted above, by default PADRE only infers up to ninth degree relatives, and its within-one-degree accuracies are similar when using this default compared to these results (not shown). However, with those settings, it infers 0% of tenth degree relatives exactly. Thus, while there is an apparent bias against ninth degree inference in the settings we used, the alternative always misclassifies tenth degree relatives.

The differences between the methods are even more striking when inferring relatedness within type 4 pedigrees, which includes siblings and up to two aunts/uncles, as shown in Figure 4 (full results in Figure S10 and S11). For $D = 5$ and $D = 7$, when analyzing data for all $|A| = 2$ aunts/uncles and $|S| = 5$ siblings (from both close relative sets), DRUID infers 15% and 32% more pairs exactly correct, respectively (Figure 4A). Considering higher relatedness degrees, the downward bias in PADRE for $D = 9$ remains in this case, resulting in a sizable discrepancy between the two methods. When $D = 10$, and again leveraging all data ($|A| = 2, |S| = 5$), DRUID infers 11% more relatives exactly than PADRE.

One of the key reasons why DRUID is likely to outperform PADRE is its ability to correctly infer the aunts/uncles to their correct relationship type and not as half-siblings or grandparents. The effect of the ambiguity in second degree relatives is especially evident when $|A| = 1$: in most instances, PADRE classifies fewer relatives correctly for $|A| = 1$ than when $|A| = 0$ (i.e., with less data). Given only one second degree relative, PADRE seems to commonly consider that individual as a grandparent, which then sometimes leads to the siblings being classified one degree more distant than they would otherwise be inferred. This effect is diminished for $|A| = 2$ because PRIMUS readily detects these aunts/uncles as full siblings of one another and benefits from the fact that a grandparents' sibling should be one degree more distantly related to the

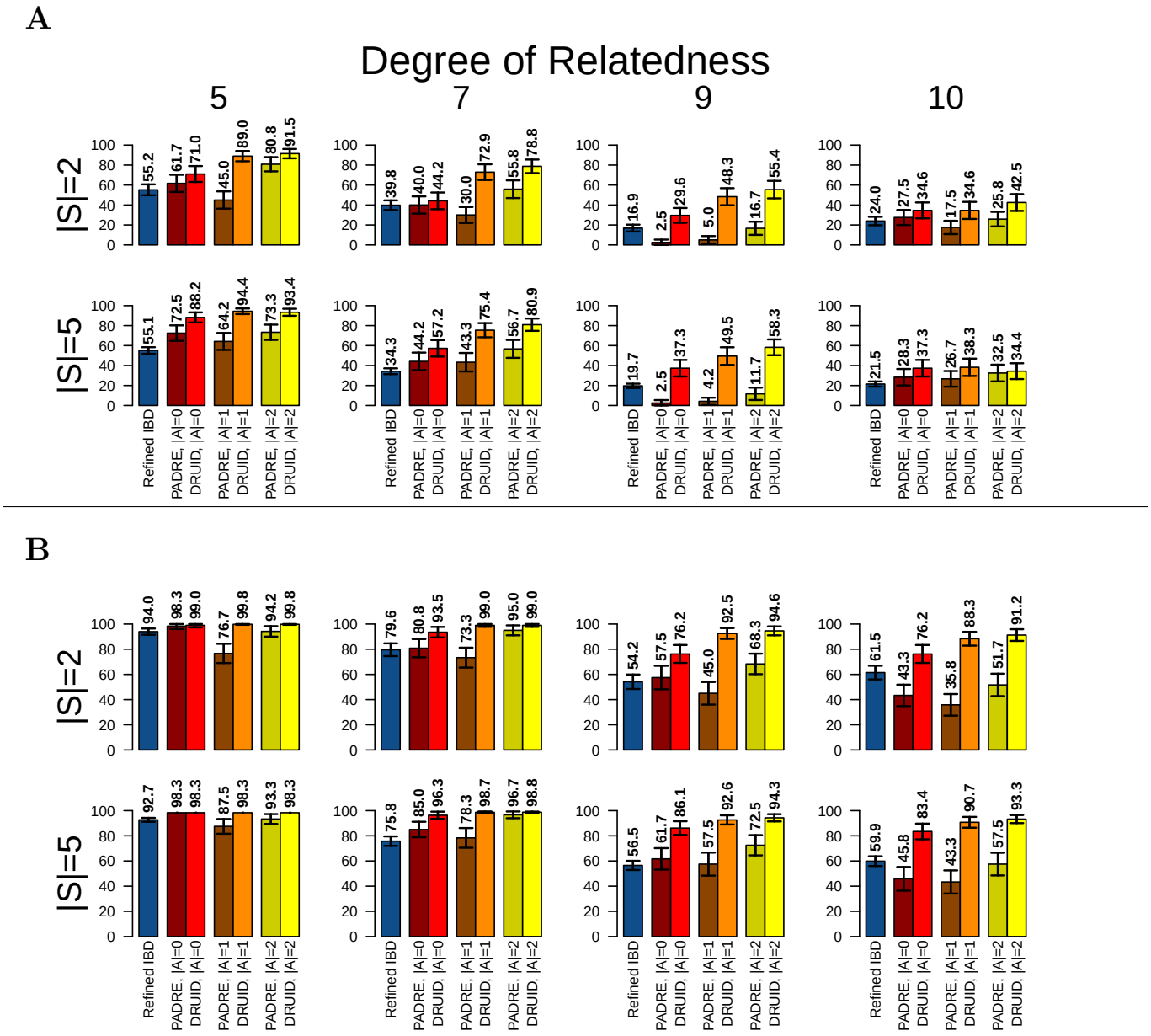

Figure 4: (**A**) For simulated type 4 pedigrees, average percent of distantly related sample pairs from the two sibling sets (bottom generation in Figure S3) that are inferred as their true degree of relatedness using Refined IBD, PADRE, and DRUID. Rows of bar plots have the same number of siblings included in both sibling sets, indicated as $|S|$ (left). Analysis with different numbers of aunts/uncles $|A|$ included shown as individual bar labels. Columns show results for different degrees of relatedness between the two sibling sets, with the true degree listed above. (**B**) As in (A), but shows the average percent of distantly related siblings inferred to be related as the true degree D or as $D \pm 1$. Error bars are bootstrapped (over complete relative sets) 95% confidence intervals.

siblings, whereas aunts/uncles (and half-siblings) should be equally related. Overall, DRUID's aunt/uncle inference is effective and appears to enable sizable gains in inferring relatives to their exact relatedness degree and within one degree of this.

The within-one-degree inference results for type 4 pedigrees show similar patterns to those from the exact inference (Figure 4B). Considering $D = 5$ and $D = 7$, and when $|A| \geq 1$, DRUID infers between 98–100% of all pairs correctly or different by one degree. PADRE is competitive with DRUID for these two degrees when $|A| = 2$, but underperforms for $|A| = 1$. When $D = 9$ and $D = 10$, and while $|A| = 2$ and $|S| = 5$, DRUID classifies 96–98% of pairs to within one degree of the truth; at the same time, PADRE infers 58–73% of these relatives correct or off by one degree. These results indicate that DRUID is highly accurate at inferring deep relatedness to within one relatedness degree for pedigrees containing siblings and their aunts/uncles; it also performs well for this type of inference using two sets of siblings.

Distant relative inference in SAMAFS data

Our initial distant relative analysis in the SAMAFS data focused on the performance of DRUID in comparison to Refined IBD. When using sets of confidently inferred close relatives (Methods) consisting of either one set of full siblings and a distant relative (Figure S12), or siblings, some number of aunts/uncles, and a distant relative (Figure S13), the classification results largely recapitulate those from simulations. In particular, including larger numbers of close relatives leads to greater accuracy gains by DRUID, with inference using aunts/uncles providing dramatic improvement relative to Refined IBD. One distinct feature of the results in these real data compared to simulations is a general increase in accuracy rates for both methods. This may indicate that the simulation results underestimate the accuracy of all the methods, which would be expected if the simulated individuals have higher variance in IBD sharing than occurs in real data. Alternatively, because the SAMAFS data contain a relatively small number of pedigrees, each with many close relatives, the quality of phasing (and therefore of detected IBD segments) may be higher than in the simulated data.

Following this analysis, we used the SAMAFS data to analyze whether including confidently inferred half-siblings provides the same classification rates as including only full siblings. As shown in Figure S14, the accuracy rates are effectively identical when using two or three full siblings as when using one half-sibling together with one or two full siblings. This is as expected since IBD detection between distant relatives should not be heavily influenced by the exact types of close relationships that exist among other samples.

Finally, we compared PADRE to DRUID using relative sets from SAMAFS, relying only upon reported relationships whether these fit the confident inference criteria or not. As with the simulated data, we performed the inferences with two sets of full siblings that are distantly related to each other. We did not restrict the size of the sibling sets, and these range from $|S| = 2$ to 11. Where available, we also included one or two sets of aunts/uncles that are associated with one or both of the sibling sets, denoting the number of included aunt/uncle sets as $\mathcal{A}$. We required all included aunt/uncle sets to have size $|A_i| = 2$, randomly selecting two samples when more than two are available. To increase the sample sizes, we performed inference with the maximum $\mathcal{A}$ and also randomly removed aunt/uncle sets so that we included all pairs of sibling sets that have the needed data for each $\mathcal{A}$ (thus analyses with smaller $\mathcal{A}$ include a superset of the distantly related sibling sets). As in all other analyses, the reported classification rates are only for distant relatives from the two sibling sets.

Figure 5 depicts the results of this analysis. In all cases, the rate of classifying relatives to their exact relatedness degree is higher in DRUID than in PADRE, and in many cases DRUID classifies 10–15% more sample pairs correctly. For third degree relatives, DRUID infers all relatives exactly correct, whereas PADRE infers roughly 14% of pairs differently. This is in part due to PRIMUS misclassifying these third degree relative pairs: results for this analysis improve when PRIMUS only infers up to second degree relatives (not shown). As observed in the simulated data, DRUID outperforms PADRE by a larger margin when using aunts/uncles than using only siblings, though both methods generally improve when given aunt/uncle data. For sixth degree inference, DRUID infers 7.6–11.7% more pairs exactly correct compared to PADRE, with 99.2% of sixth degree relatives inferred exactly when $\mathcal{A} = 2$. Both methods perform well when inferring relatives to within one degree of their true relatedness, with DRUID always inferring 100% of relatives in

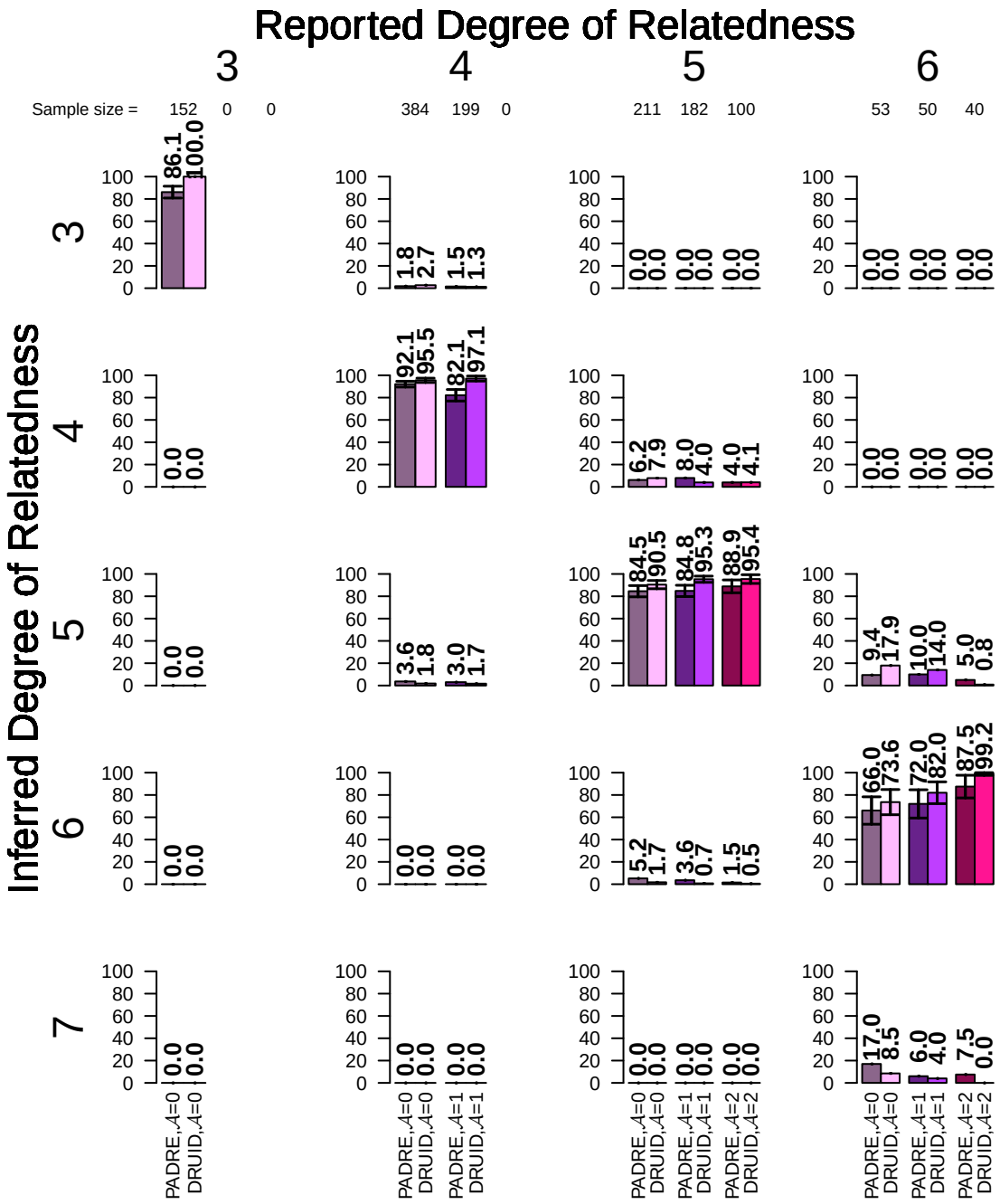

Figure 5: Rates of DRUID and PADRE inferring a range of degrees of relatedness between two full sibling sets in SAMAFS that are distantly related to each other. Analyses consider inclusion of two full sibling sets only ($\mathcal{A} = 0$), two sibling sets and one associated aunt/uncle set ($\mathcal{A} = 1$), and two sibling sets each with an aunt/uncle set ($\mathcal{A} = 2$), as indicated by bar labels. Sample size counts the number of instances of two distantly related relative sets (not the number of siblings contained in these sets). Error bars are bootstrapped (over complete relative sets) 95% confidence intervals.

this way, and PADRE sometimes having discrepancies as it identifies between 87–100% of distant relatives by this metric. Overall, this analysis in SAMAFS demonstrates the effectiveness of performing inference based on the reconstructed IBD sharing profile of an ungenotyped ancestor, as DRUID's accuracy results exceed both multi-way pairwise inference methods.

Discussion

Classifying relatedness between samples is a classical problem in genetics[8], and is relevant to both trait association mapping[7,10,11] as well as population genetics[1]. Recent advances in this domain have come through the analysis of IBD segments[39] instead of only independent markers, and most recently through analyses that leverage signals from multiple individuals[20,21] instead of only pairwise inference.

We developed and evaluated DRUID, an algorithm that uses IBD sharing between multiple close relatives in order to infer the IBD sharing profile of one of their an ungenotyped ancestors. A critically important factor in DRUID's performance is the correct inference of the pedigree relationships among the close relatives. This is so because mistakes in these relationships lead to reconstruction of IBD sharing either for an unexpected individual or one that does not exist—e.g., composed of a combination of IBD segments from a grandparent and a parent. Because the relatedness inferred for the genotyped individuals derive from these unobserved samples, mistakes have the potential to bias the results. To avoid such issues, we focus our inference on full siblings and we developed a new approach to accurately detect aunts and uncles of a pair of siblings. The aunt/uncle inference quality is very high with zero false positives and 95.4% and 97.5% recall in real and simulated data, respectively.

Our analyses show that DRUID and PADRE both perform substantially better than Refined IBD, with up to 36% more relatives classified correctly in simulations when data for two distantly related sets of full siblings are available (Figure 3A). Considering DRUID and PADRE only, for relatedness between two sets of full siblings, DRUID infers up to 10% more pairs correctly in simulations (Figure 3A), and 3.4–14% more relatives correctly in SAMAFS (Figure 5). In turn, for data comprising two sets of full siblings with up to two aunts/uncles, DRUID infers up to 10–30% more simulated pairs to the correct degree of relatedness compared to PADRE (Figure 4A), and up to 15% more relatives in SAMAFS (Figure 5). This dramatic improvement stems from DRUID's approach in general as well as from its aunt/uncle inference.

Ideally, methods would be able to infer the degree of relatedness between samples to their exact values—and DRUID makes possible such specific inference for fifth degree degree relatives given data for two sets of full siblings (73–91% exactly correct in simulations, Figure 4A; 85% exactly correct in SAMAFS, Figure 5). Still, inference to within one degree of the true relationship is also very informative, and DRUID infers between 78–96% of simulated tenth degree relatives in this way (Figures 3B and 4B). This high rate of classification for such distant relatives signals a dramatic shift in the effectiveness of such inference. Because the variance in IBD sharing increases for more distant relatedness, exact inference is most likely to only be achievable for lower relatedness degrees. Yet exact inference isn't required in all settings, and identifying relatedness to within a range of degrees provides information that is useful for further characterization. One noteworthy application area is that of genetic genealogy, a service offered to consumers by several genetic testing companies, and which has become especially popular in recent years as testing prices have fallen.

Going forward, a possible extension of this work is to effectively iterate DRUID's approach: identify and classify the relationships of two sets of distant relatives, and then use that classification to reconstruct the IBD sharing profile between their mutual ancestor(s) and other even more distant relatives. Success in this context will require either modeling uncertainty in which ancestor is being reconstructed, or alternative methodologies that enable more precise knowledge of the pedigree structure that exists between the initial two sets of close relatives. Another tantalizing possibility in this domain is that of directly inferring genetic data (such as haplotypes) for specific ancestors. While DRUID infers information about ancestors, it is ambiguous about which ancestor is being analyzed. Nevertheless, its approach is effective and opens the door to more detailed pedigree-based analyses in datasets that contain many relatives—now including nearly

all new datasets.

Web resources

DRUID:http://github.com/williamslab/druid/
Ped-sim: http://github.com/williamslab/ped-sim/

Acknowledgements

This work was supported by a National Science Foundation Graduate Research Fellowship grant number DGE-1144153 to M.D.R.; Qatar National Research Fund grant NPRP 7-1425-3-370 to J.G.M.; an Alfred P. Sloan Research Fellowship, and a seed grant from Nancy and Peter Meinig to A.L.W. The SAMAFS are supported by NIH grants R01 HL0113323, P01 HL045222, R01 DK047482, and R01 DK053889. 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

References

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

[2] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. Genome-wide genetic data on 500,000 UK Biobank participants. *bioRxiv*, page 166298, 2017.

[3] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H ODonnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

[4] Frederick E Dewey, Michael F Murray, John D Overton, Lukas Habegger, Joseph B Leader, Samantha N Fetterolf, Colm O'Dushlaine, Cristopher V Van Hout, Jeffrey Staples, Claudia Gonzaga-Jauregui, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, 354(6319):aaf6814, 2016.

[5] Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. An atlas of genetic associations in UK Biobank. *bioRxiv*, page 176834, 2017.

[6] Eunjung Han, Peter Carbonetto, Ross E Curtis, Yong Wang, Julie M Granka, Jake Byrnes, Keith Noto, Amir R Kermany, Natalie M Myres, Mathew J Barber, et al. Clustering of 770,000 genomes reveals post-colonial population structure of north america. *Nature Communications*, 8:14238, 2017.

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

[8] 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.

[9] Joshua G Schraiber and Joshua M Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, 2015.

[10] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

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

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

[13] Catarina D. Campbell, Jessica X. Chong, Maika Malig, Arthur Ko, Beth L. Dumont, Lide Han, Laura Vives, Brian J. O'Roak, Peter H. Sudmant, Jay Shendure, Mark Abney, Carole Ober, and Evan E. Eichler. Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, 44:1277–1281, Sep 2012.

[14] Vagheesh M. Narasimhan, Raheleh Rahbari, Aylwyn Scally, Arthur Wuster, Dan Mason, Yali Xue, John Wright, Richard C. Trembath, Eamonn R. Maher, David A. van Heel, Adam Auton, Matthew E. Hurles, Chris Tyler-Smith, and Richard Durbin. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nature Communications*, 8(1):303, 2017.

[15] 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.

[16] F Baudat, J Buard, C Grey, A Fledel-Alon, C Ober, M Przeworski, G Coop, and B De Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–840, 2010.

[17] Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, 2010.

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

[19] 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.

[20] 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.

[21] Amy Ko and Rasmus Nielsen. Composite likelihood method for inferring local pedigrees. *PLOS Genetics*, 13(8):1–21, 08 2017.

[22] 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.

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

[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] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[26] 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.

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

[28] Karl W Broman and James L Weber. Characterization of human crossover interference. *The American Journal of Human Genetics*, 66(6):1911–1926, 2000.

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

[30] 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.

[31] Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[32] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[33] 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):1, 2015.

[34] Kuruvilla Joseph Abraham and Clara Diaz. Identifying large sets of unrelated individuals and unrelated markers. *Source code for biology and medicine*, 9(1):1, 2014.

[35] 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.

[36] Jeffrey Staples, Dandi Qiao, Michael H Cho, Edwin K Silverman, Deborah A Nickerson, Jennifer E Below, University of Washington Center for Mendelian Genomics, et al. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American Journal of Human Genetics*, 95(5):553–564, 2014.

[37] Hong Li, Gustavo Glusman, Hao Hu, et al. Relationship estimation from whole-genome sequence data. *PLOS Genetics*, 10(1), 2014.

[38] 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.

[39] 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, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, 21(5):768–774, 2011.

[40] 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.

[41] 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.

[42] 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.