ARTICLE

Bonsai: An efficient method for inferring large human pedigrees from genotype data

Ethan M. Jewett,[1,*] Kimberly F. McManus,[1] William A. Freyman,[1] the 23andMe Research Team, and Adam Auton[1]

Summary

Pedigree inference from genotype data is a challenging problem, particularly when pedigrees are sparsely sampled and individuals may be distantly related to their closest genotyped relatives. We present a method that infers small pedigrees of close relatives and then assembles them into larger pedigrees. To assemble large pedigrees, we introduce several formulas and tools including a likelihood for the degree separating two small pedigrees, a generalization of the fast DRUID point estimate of the degree separating two pedigrees, a method for detecting individuals who share background identity-by-descent (IBD) that does not reflect recent common ancestry, and a method for identifying the ancestral branches through which distant relatives are connected. Our method also takes several approaches that help to improve the accuracy and efficiency of pedigree inference. In particular, we incorporate age information directly into the likelihood rather than using ages only for consistency checks and we employ a heuristic branch-and-bound-like approach to more efficiently explore the space of possible pedigrees. Together, these approaches make it possible to construct large pedigrees that are challenging or intractable for current inference methods.

Introduction

The ability to infer complex multi-generational pedigrees from genotype data has many applications ranging from genealogical research to the study of diseases. As human genotyping datasets continue to grow in size, there is increasing interest in computational methods that can reconstruct large pedigrees efficiently and accurately.

Although the problem of pedigree inference has been studied extensively, the majority of pedigree inference methods are designed for non-human species. A major challenge for pedigree reconstruction in non-human populations is that pairwise relationships can be difficult to infer with high accuracy, even when the degree of a relationship is small because high-quality genotype data may be unavailable. As a result, methods typically require that all or most individuals in a pedigree are sampled so that pedigrees can be assembled by connecting strings of parent-child, full-sibling, or half-sibling pairs.[1–10] Although it is possible to connect slightly more distant relationships,[11,12] the majority of existing pedigree inference algorithms can be characterized as methods for either jointly or independently inferring pairwise parent-child pairs and full or half sibling sets, which are then consistent with a pedigree structure when assembled together.

In contrast to non-human pedigrees, genotype data for human populations is comparatively abundant and close relationships, such as parent-offspring or sibling pairs, can be inferred with a high degree of accuracy. The major challenge of pedigree inference in human populations is the fact that pedigrees are often sparsely sampled, with few genotyped sibling and parent pairs and few genotyped individuals beyond the most recent two or three generations. In human datasets, including direct-to-consumer genetic databases, genotyped individuals may have only a small number of genotyped relatives within a radius extending to first or second cousins and it is common for an individual's closest relative to be more distant than a second cousin. As a result, it is difficult to construct solid frameworks of close relatives and their genotyped ancestors into which other genotyped individuals can be placed.

There are currently two state-of-the-art methods for inferring complex human pedigrees from genotype data, both of which are maximum likelihood approaches that attempt to find a pedigree that maximizes the sum of log likelihoods of pairwise relationships, given observed patterns of identity-by-descent (IBD) sharing. The two methods differ primarily in the approaches they take to find the maximum likelihood pedigree.

The earlier method, PRIMUS,[13] explores the space of possible pedigrees by starting with a seed individual and then iteratively adding individuals to the pedigree. Each time an individual is added, the method considers all possible positions that are consistent with the estimated pairwise relationships. When adding an individual to the pedigree, each pedigree at the previous step serves as a seed pedigree onto which the individual can be added in multiple ways. By constructing a large set of pedigrees in this way, the algorithm efficiently explores the space of pedigrees that are compatible with the estimated pairwise relationships.

In contrast to PRIMUS, the more recent CLAPPER method[14] begins by connecting all individuals together into an initial guess of a pedigree. Then, at each
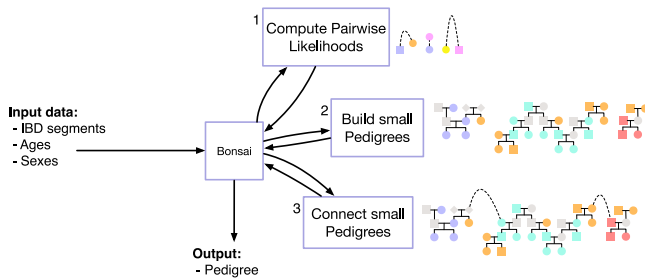
Figure 1. Overview of the full Bonsai method
Details of methods 1, 2, and 3 are presented in Algorithms 1, 2, and 4, respectively, in the Supplemental methods.

subsequent step, the CLAPPER algorithm rearranges the relationships in the pedigree. This update step is done using a Markov chain Monte Carlo (MCMC) approach in which there are many different possible moves that can be made, such as adding or subtracting a degree of relatedness between two individuals, swapping the labels of two nodes, or pruning off an individual and their descendants and attaching them somewhere else.

PRIMUS and CLAPPER make it possible to infer pedigrees in which pairs of genotyped relatives are separated by several ungenotyped relatives. However, neither approach was designed to infer the large and sparse pedigrees that are common in direct-to-consumer genetic datasets where the degree of relationship separating a pair of genotyped individuals may be large, verging on degrees where individuals frequently share no detectable IBD. For such pedigrees, searching a broad pedigree space using the approach of PRIMUS or CLAPPER is computationally infeasible. Although a recent study applied PRIMUS to reconstruct more than 12,000 pedigrees in a large dataset, the greatest degree of relationship between a pair of genotyped individuals in a pedigree was restricted to two.[15]

The PADRE method of Staples et al.[16] partly addresses the problem of building large pedigrees by inferring the founders through which two distantly related pedigrees are connected and their degree of separation. PADRE solves a key problem of large pedigree inference in an elegant way. However, the PADRE method does not subsequently apply these inferences to assemble small pedigrees into large pedigrees.

Here, we introduce a method, Bonsai, for inferring large and sparse pedigrees. To make inference efficient and accurate, we first infer small pedigrees of closely related individuals using an approach that efficiently explores the space of possible pedigrees. This approach is similar to PRIMUS, but differs in key ways that make the search of the pedigree space both more efficient and more thorough. The small pedigrees are then assembled into larger pedigrees using several techniques, including a generalized version of the DRUID method of Ramstetter et al.,[17] which allows the method to link distantly related individuals into large and sparsely sampled pedigrees. We refer to the first stage as "small Bonsai" and to the second stage as "big Bonsai" (Figure 1). We first describe the small and big Bonsai

methods, then use both simulated and real data to investigate the performance of the methods and their components.

Subjects and methods

Overview of the Bonsai method
The Bonsai method is summarized in Figure 1. The input to the method consists of ages and sexes for a set of putatively related individuals, along with IBD segments inferred between each pair of individuals. The method then proceeds through three stages in sequence.

First, the relationship between each pair of genotyped individuals is inferred using age and pairwise IBD data. The likelihoods of many other possible relationships are also computed and stored for each pair. Next, small pedigrees of closely related individuals are inferred from these pairwise likelihoods. Finally, the inferred small pedigrees are assembled into large and sparse pedigrees.

Constructing small pedigrees and combining them together makes it possible to use information in small pedigree structures to improve the accuracy with which more distant relationships are inferred. This approach makes it possible to more precisely infer the ancestral lineages through which small pedigrees are connected, the number of common ancestors shared by each pair of individuals, and segments of so-called background IBD that do not reflect recent ancestry. Each of these additional pieces of information makes it possible to proactively reduce the space of possible pedigrees that must be searched, making inference tractable for large and sparse pedigrees.

Stage 1: Inferring pairwise relationships
The first stage of the Bonsai method is to infer the likelihoods of many possible relationships between each pair of putative relatives. To make the computation of the likelihood efficient without large sacrifices in accuracy, we use a composite likelihood that is the product of the likelihoods of different IBD summary statistics and the likelihoods of the pairwise age differences between the individuals. The genetic component $\mathcal{L}_{\mathcal{R}}^{g}$ of the likelihood, computed from IBD, is multiplied by the age component $\mathcal{L}_{\mathcal{R}}^{a}$ of the likelihood to obtain the final likelihood $\mathcal{L}_{\mathcal{R}}$ of a given relationship type, $\mathcal{R}$:

$$\mathcal{L}_{\mathcal{R}} = \mathcal{L}_{\mathcal{R}}^{g} \mathcal{L}_{\mathcal{R}}^{a}. \qquad \text{(Equation 1)}$$

The likelihood is composite, rather than exact, because we do not model the joint distribution of the IBD count and length summary statistics whose product is $\mathcal{L}_{\mathcal{R}}^{g}$ and because there is an underlying joint distribution of IBD sharing and age difference that is not captured by the product of the two likelihoods $\mathcal{L}_{\mathcal{R}}^{g}$ and $\mathcal{L}_{\mathcal{R}}^{a}$. A quick reference to variables defined in the paper can be found in Table S1.

Pairwise genetic likelihoods
To compute the genetic component of the composite pairwise relationship likelihood, we consider regions of the genome shared identically by descent in a haploid fashion on just one chromatid in each individual, as well as regions shared IBD in a diploid fashion on both chromatids. We use the terms "IBD1 segment" and "IBD2 segment" to refer to regions of haploid and diploid IBD, respectively. The genetic component of the pairwise likelihood is computed using the total length of IBD1 segments, the total length of IBD2 segments, the total number of IBD1 segments, and the total number of IBD2 segments.

It is possible to compute the probability of an observed shared pattern of IBD analytically, at least in approximation. However, in practice we find that error in IBD inference leads to differences between the empirical and analytical IBD distributions for each relationship type, especially for close relationships. Thus, we use likelihoods obtained as moment-fitted Poisson and Gaussian approximations of simulated distributions.

Let $T_1^{i,j}$ and $T_2^{i,j}$ be the total lengths of IBD 1 and 2, respectively, for a pair of individuals (i and j) and let $C_1^{i,j}$ and $C_2^{i,j}$ be the counts of the number of IBD 1 and 2 segments shared between the two individuals. We follow the convention that uppercase variables T_1, T_2, C_1, C_2, etc. denote random variables and their lowercase counterparts, t_1, t_2, c_1, c_2, etc. denote their observed values. The genetic component of the composite likelihood for a given relationship type, $\mathcal{R}$, between a pair of individuals i and j is then computed as

$$\mathcal{L}_{\mathcal{R}}^g(i,j) \approx f_{\mathcal{R}}(t_1)f_{\mathcal{R}}(t_2)\mathbb{P}_{\mathcal{R}}(c_1)\mathbb{P}_{\mathcal{R}}(c_2), \qquad \text{(Equation 2)}$$

where $f_{\mathcal{R}}(t_1) \equiv f_{T_1^{i,j}}(t_1; \mathcal{R})$ is the probability density function of the sum of lengths of all IBD1 segments for a relationship of type $\mathcal{R}$ and $\mathbb{P}_{\mathcal{R}}(c_1) \equiv \mathbb{P}(C_1 = c_1; \mathcal{R})$ is the probability mass function for the total number of segments of IBD1 for a relationship of type $\mathcal{R}$. The quantities $f_{\mathcal{R}}(t_2)$ and $\mathbb{P}_{\mathcal{R}}(c_2)$ are defined analogously for segments of IBD2.

In Equation 2, the quantities $f_{\mathcal{R}}(t_1)$ and $f_{\mathcal{R}}(t_2)$ are modeled as Gaussian distributions and the distributions $\mathbb{P}_{\mathcal{R}}(c_1)$ and $\mathbb{P}_{\mathcal{R}}(c_2)$ are Poisson with means given by the expected numbers of IBD1 and IBD2 segments, respectively, between two individuals of relationship type $\mathcal{R}$. In practice, the Poisson distribution did not provide a good fit for segment counts for close relatives so the segment count data were also modeled as Gaussian. The mean and variance of $T_i^{\mathcal{R}}$ and the mean of $C_i^{\mathcal{R}}$ for a relationship of type $\mathcal{R}$ were obtained empirically using simulations. Details of the simulations used to obtain these moments are provided in Simulations and fitting of empirical pairwise genetic likelihood distributions.

Pairwise age likelihoods

The pairwise age likelihood for a given relationship type, $\mathcal{R}$, was obtained by moment-fitting a Gaussian distribution to the differences between the ages of 23andMe customers who self-reported to be of relationship type $\mathcal{R}$ (Figure S1). We required that the self-reported relationship between each pair of individuals could be verified through a string of inferred parent-child or full-sibling relationships. For example, a self-reported first-cousin relationship between individuals i and j was verified if i and j each had inferred parents in the 23andMe database, and if these parents in turn had the same pair of inferred parents or were inferred to be full siblings.

For two individuals, i and j with ages a_i and a_j, the age component of the likelihood for relationship type $\mathcal{R}$ was modeled as a Gaussian distribution with the empirically observed mean and variance:

$$\mathcal{L}_{\mathcal{R}}^a(i,j) = \frac{e^{-\left[(a_i - a_j) - \mu_a^{\mathcal{R}}\right]^2 / 2\left(\sigma_a^{\mathcal{R}}\right)^2}}{\sigma_a^{\mathcal{R}}\sqrt{2\pi}}. \qquad \text{(Equation 3)}$$

In Equation 3, $\mu_a^{\mathcal{R}}$ and $\sigma_a^{\mathcal{R}}$ are the mean and standard deviation of the empirical age difference for all pairs in our training set with the pairwise relationship, $\mathcal{R}$. Note that the probability $\mathcal{L}_{\mathcal{R}}^a(i,j)$ is not symmetrical in the ages a_i and a_j. This is useful for determining the directionality of the relationship between two people, such as parent-child or nephew-aunt when age information is available.

The likelihood of a pedigree

The composite likelihood, $\mathcal{L}_{\mathcal{P}}$, of a pedigree $\mathcal{P}$ is computed as the product of genetic and age likelihoods (Equation 1) for all pairs of individuals in the pedigree,

$$\mathcal{L}_{\mathcal{P}} = \prod_{i,j \in \mathcal{P}} \mathcal{L}_{\mathcal{R}}^g(i,j)\mathcal{L}_{\mathcal{R}}^a(i,j). \qquad \text{(Equation 4)}$$

where $\mathcal{R}$ is the relationship between i and j implied by the pedigree structure. This likelihood is efficiently computed as each new individual is added to the pedigree by inductively extending the existing relationships of the parents and/or children of the newly added person to obtain the relationships of the new person to all existing individuals in the pedigree. We then add the log likelihoods of each of these new pairwise relationships to the log likelihood of the pedigree without the new individual.

The "small" Bonsai method

To construct a pedigree from pairwise likelihoods, the small Bonsai method begins by placing a focal individual by itself in the pedigree. This focal individual is typically the person with the closest average degree of relationship to all other individuals in the putatively related set, but any individual can be chosen. At each subsequent step of the small Bonsai algorithm, the next individual to be placed is chosen to be the unplaced individual with the closest inferred degree of relationship with one of the individuals already placed in the pedigree, where ties are broken by the total amount of IBD shared. Because each pair of individuals has many possible relationships, we determine the order in which individuals are added using the most likely pairwise relationship for each pair.

The next individual to be placed is considered in all ways that are consistent with the most likely inferred pairwise relationships to individuals already placed. In particular, we consider the top r most likely pairwise relationships between the new individual and their closest relative in the set of placed individuals and we place the individual in all ways that are compatible with each of these r most-likely relationships. The result of each placement is a copy of the pedigree with the individual placed in one possible way. At the end of each step of the method, we have a set of putative pedigrees representing different ways of placing an individual.

To avoid a rapid expansion in the number of pedigrees at each step, we employ a heuristic branch-and-bound-like procedure in which we discard each pedigree at the end of each step that is very unlikely, compared with the most likely pedigree. In particular, we discard all pedigrees whose likelihoods are less than a fraction f_ℓ of the likelihood of the most-likely pedigree. In practice, when individuals are closely related, there are only a few pedigrees that have high likelihoods and the rest can be discarded. As a result, the likelihood threshold has a relatively low impact on accuracy while serving to speed up pedigree building.

This heuristic branch-and-like procedure is repeated until no unplaced individual has a pairwise point-estimated degree that is within a degree δ of any placed individual. At this point, the small Bonsai algorithm is terminated. If unplaced individuals remain, a new focal individual is chosen from among the unplaced individuals and the small Bonsai algorithm is applied again. The small Bonsai algorithm is applied repeatedly, choosing a new focal individual each time, until all individuals have been placed into some pedigree.

Figure 2 shows an example sequence for constructing a pedigree using the small Bonsai method. In the first row of the figure, a focal individual (shaded yellow square) is placed into a pedigree on their own. Grey diamonds indicate their parents, whose sexes are unspecified. In the second row, the unplaced individual
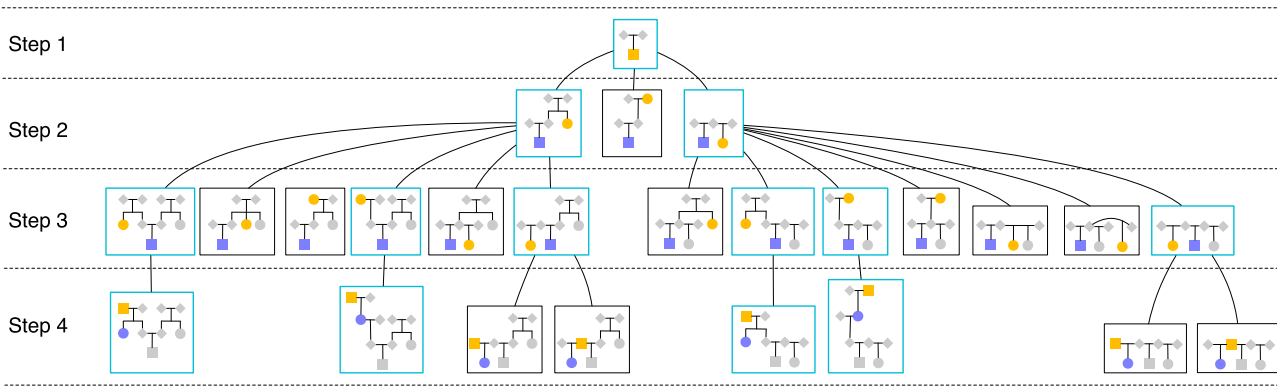
Figure 2. The small Bonsai method
An example of the sequence of steps for building a small pedigree is shown. The sequence proceeds from top to bottom in the figure. The *i*th row of pedigrees in rectangles represents the *i*th step of the small Bonsai algorithm in which the *i*th individual is added to a pedigree. The individual being placed at any given step is shown in yellow. Their closest placed relative is shown in blue. Blue boxes indicate pedigrees that are retained and carried forward to the next step. Black boxes indicate pedigrees with low likelihoods that are discarded.

(yellow circle) with the closest degree of relationship to the placed individual (now shaded in blue), is placed into the pedigree. The new individual is placed in all ways that are consistent with the top r most-likely relationships inferred in the pairwise relationship inference step (Stage 1: Inferring pairwise relationships). Here, we have chosen $r = 3$. These $r = 3$ most-likely relationships happen to be "avuncular," "grandparental," and "half-sibling" in the example shown in step 2 of Figure 2. This is the "branch" step of the heuristic branch-and-bound-like procedure.

Before placing the next individual, we evaluate the likelihood of each pedigree, computed as the product of pairwise likelihoods of the relationships induced by the pedigree. We retain only those pedigrees whose likelihoods are at least a fraction f_ℓ of the likelihood of the most likely pedigree. This is the "bound" step of the heuristic branch-and-bound-like procedure.

When two or more pedigrees formed by adding an individual would be topologically identical, we construct only one of the pedigrees. For example, in the second row of Figure 2, because the sexes of the parents are unknown and there are no placed relatives except the focal individual that can be used for triangulation, adding an avuncular relative through the right parent is topologically identical to adding them through the left parent. Therefore, we only build one of these pedigrees (the one on the far left of the second row).

In the third row of the diagram, the unplaced individual (yellow circle) with the closest degree of relationship to a placed individual is added to all pedigrees that were carried forward from the previous step. The new individual is added to each pedigree in all ways that are consistent with the top r most-likely relationships to their closest placed relative (blue square). Again, these relationships happen to be "avuncular," "grandparental," and "half-sibling" in the example. We then perform the bound step, retaining only those pedigrees whose likelihoods are at least a fraction f_ℓ of the likelihood of the most-likely pedigree.

In the fourth row, we show one final iteration of the procedure. Again, the unplaced individual (yellow square) is added in all ways that are consistent with the top r most-likely pairwise point estimated relationships with their closest relative (blue circle). In this case the most likely point-estimated relationship happens to be "parental." Because parent-child relationships are inferred with near certainty, we have only placed the individual as a parent in the diagram, omitting the next two most-likely relationships which will be considerably less likely.

The "big" Bonsai method
Overview of the big Bonsai method

When building a pedigree containing distantly related individuals, the small Bonsai method is first applied repeatedly to build sets of small non-overlapping pedigrees. The union of individuals in these small pedigrees is equal to the set of individuals in the full pedigree. The big Bonsai method is then applied to combine the small pedigrees together, one pair at a time, with the two pedigrees sharing the most total IBD combined at each step.

The big Bonsai method relies on several methods we introduce that are useful for different aspects of combining pedigrees together. The first method is a generalized version of the DRUID estimator[17] for inferring the degree of relatedness separating the common ancestors of two small pedigrees. The DRUID estimator was derived for specific pedigree structures, such as a set of siblings and their avuncular relatives connected to another such pedigree through the common grandparental ancestors of the two pedigrees. Here, we generalize the DRUID estimator to any pair of outbred pedigrees and, in Appendix A: Re-rooting the DRUID estimator, we further generalize the DRUID estimator to the case in which two pedigrees are connected through two individuals who are not the common ancestors of their respective pedigrees.

The second tool we introduce is an approximation of the likelihood of the degree separating two pedigrees, given the total IBD shared between the two pedigrees. This likelihood, which was inspired by the DRUID estimator, makes it possible to evaluate the relative likelihoods of different degrees separating two pedigrees in addition to obtaining a point estimate of the degree.

The third tool we introduce is a test for detecting segments of background IBD. Background IBD segments are regions of the genome that are shared identically-by-state (IBS) that did not arise by transmission from a single shared common ancestor. Instead, these segments arose because of demographic or evolutionary processes, such as a population bottleneck. They are long regions of IBS with hidden recombination events and they can provide misleading information about the degree of relationship between a pair of individuals. Background IBD segments can lead to mis-inferred pedigrees, particularly when pedigrees are sparsely genotyped.

The fourth tool we introduce is a method for determining the correct ancestral lineages through which two or more pedigrees are
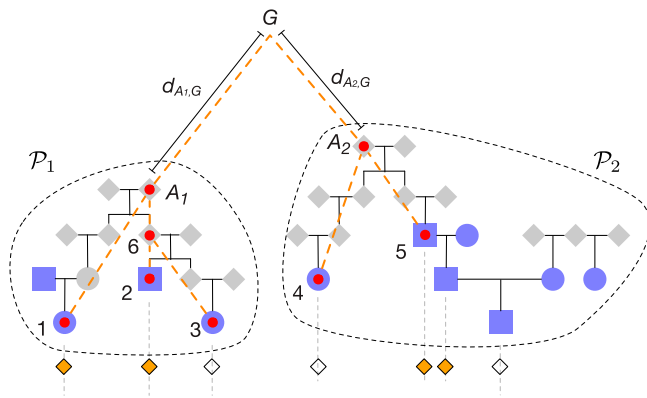
Figure 3. Example of an observed pattern of presence and absence of an ancestral allele
Genotyped individuals are shaded in purple. Filled and empty diamonds below indicate the presence or absence of the allele from G. Red dots on purple genotyped individuals indicate the set of genotyped individuals with no direct genotyped ancestors. Red dots on gray ungenotyped individuals indicate the most recent common ancestors transmitting the segments to the genotyped individuals. Dashed orange lines indicate the paths by which the allele is transmitted from common ancestor G. The number of meioses separating A_1 and A_2 from a common ancestor, G, are $d_{A_1,G}$ and $d_{A_2,G}$.

connected. This approach relies on detecting overlapping IBD segments that are inconsistent with certain lineage combinations.

We also derive a recursive formula for computing the probability of an observed presence-absence pattern of an ancestrally transmitted allele in a set of descendants. This formula is useful for developing the generalized DRUID estimator and the likelihoods for degree estimation and background IBD detection.

Together, the tools we introduce can be used to identify the ancestors through which two small pedigrees are connected, infer the degree separating the two ancestors, and identify and discard individuals whose IBD sharing patterns appear to be background IBD. By using these inference tools to identify highly likely ways of connecting pedigrees, the space of possible pedigrees can be reduced. We now describe each of these approaches in detail.

The probability of a presence-absence pattern of an ancestral allele
Consider two pedigrees $\mathcal{P}_1$ and $\mathcal{P}_2$ of genotyped individuals $\mathcal{N}_1$ and $\mathcal{N}_2$, related through a common ancestor (or pair of ancestors), G (Figure 3). Let A_1 be the common ancestor of $\mathcal{N}_1$ in $\mathcal{P}_1$ and let A_2 be the common ancestor of $\mathcal{N}_2$ in $\mathcal{P}_2$.

Consider an allele transmitted from one chromatid in G to its descendants. We begin by deriving the probability of the observed pattern of presence and absence of the ancestral allele among descendants of A_1 and A_2. Let $d_{A_1,G}$ and $d_{A_2,G}$ be the degrees separating A_1 and A_2 from the set of most recent common ancestors, G, of the pedigree. G corresponds to two individuals if A_1 and A_2 are descended from an ancestral couple and G corresponds to a single common ancestor if A_1 and A_2 are descended from a pair of half siblings. We do not consider cases of endogamy, where G corresponds to more than one ancestor other than a mate pair. To simplify the derivation, we also exclude the case where A_1 and A_2 are full siblings, so that they share at most one ancestral allele from G.

Figure 3 shows a presence-absence pattern of an inherited allele among genotyped individuals in the two small pedigrees $\mathcal{P}_1$ and $\mathcal{P}_2$. The probability of the observed presence and absence pattern can be computed recursively by conditioning on whether the

allele was observed in the ancestor of each individual. This approach is similar to Felsenstein's tree pruning algorithm.[18]

Let O_i be a random variable describing the event that a copy of the allele is transmitted to descendant i and is observed. We set $O_i = 1$ if the allele is observed in individual i and $O_i = 0$ if it is not observed. Let D_i denote the presence-absence pattern at the descendants $\mathcal{N}_i$ of node i.

Defining

$$p_{i,0} \equiv \mathbb{P}(D_i \mid O_i = 0), \ p_{i,1} \equiv \mathbb{P}(D_i \mid O_i = 1), \qquad \text{(Equation 5)}$$

we show in Appendix A: The probability of a pattern of IBD that the probabilities can be computed using the recursion

$$p_{i,0} = \prod_c p_{c,0}$$

$$p_{i,1} = \prod_c \left[2^{-d_{c,i}} p_{c,1} + \left(1 - 2^{-d_{c,i}}\right) p_{c,0} \right], \qquad \text{(Equation 6)}$$

where the products are taken over all child nodes, c, of i. The base conditions at a leaf l with state s are $p_{l,0} = \mathbb{1}_{s,0}$ and $p_{l,1} = \mathbb{1}_{s,1}$. For each allelic copy, g, in G, the probability of an observed IBD sharing pattern $\{O_1, ..., O_k\}$ across k leaf nodes can be computed recursively as $p_{g,1}$ using Equation 6.

The generalized DRUID estimator
The probability of a presence-absence pattern can be used to obtain a fast and accurate point estimator of the degree separating A_1 and A_2 by accounting for all IBD shared among their descendants. Because two genealogically related individuals may share little or no IBD, it is helpful to leverage IBD segments shared among close relatives of the two individuals when inferring their degree of relatedness. Figure S2 illustrates the utility of considering IBD segments among groups of individuals rather than pairwise IBD when the degree of relatedness is not small. In particular, individuals 3 and 4 in Figure S2 share no IBD segments. Thus, one cannot infer their degree of relatedness without additional information. However, if close relatives of 3 and 4 do share IBD with one another, and if pedigrees can be inferred relating these close relatives to 3 and 4, then we can use the IBD in these relatives to estimate the degree of relationship between 3 and 4.

Two approaches have been used to leverage IBD among close relatives to infer the degree of relationship between a pair of common ancestors. They are illustrated in Figure S2. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two sets of genotyped individuals; for example, sets $\mathcal{N}_1 = \{2, 3\}$ and $\mathcal{N}_2 = \{4, 5, 6\}$ in Figure S2. Let A_1 and A_2 be any two most recent common ancestors of $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively, and let $d(A_1, A_2)$ denote the degree between A_1 and A_2. The approach implemented by Staples et al.[16] in their PADRE method is to compute the probability of the observed IBD between each pair of individuals, with one individual in $\mathcal{N}_1$ and the other in $\mathcal{N}_2$ (Figure S2A). For a given degree $d(A_1, A_2)$, the composite likelihood is then computed by taking the product of pairwise likelihoods. In the PADRE method, the pairwise probabilities are computed using the ERSA method of Huff et al.,[19] which gives the probabilities of the lengths and counts of shared segments. Staples et al.[16] found that this approach yielded improved accuracy for inferring $d(A_1, A_2)$ compared with the likelihood for a single pair of individuals.

The second approach, implemented by Ramstetter et al.[17] in their DRUID method, is to first obtain a point estimate of the total amount of IBD shared between A_1 and A_2 and then use this point estimate to infer the degree between A_1 and A_2 (Figure S2B). The DRUID estimator of $d(A_1, A_2)$ is obtained by first merging all IBD segments observed between $\mathcal{N}_1$ and $\mathcal{N}_2$. The total merged IBD is

then converted into a point estimate of the amount of IBD shared between the common ancestor A_1 and the common ancestor A_2. The amount of IBD shared between A_1 and A_2 is estimated by considering the fraction ϕ_1 of the genome of A_1 that is passed on to its genotyped descendants in $\mathcal{N}_1$ and the fraction ϕ_2 of the genome of A_2 that is passed on to its genotyped descendants in $\mathcal{N}_2$. If $IBD(A_1, A_2)$ is the amount of IBD shared between A_1 and A_2, then the expected amount shared between $\mathcal{N}_1$ and $\mathcal{N}_2$ is $IBD(\mathcal{N}_1, \mathcal{N}_2) = \phi_1 \phi_2 IBD(A_1, A_2)$. Solving for $IBD(A_1, A_2)$ yields a point estimator of $IBD(A_1, A_2)$ in terms of the observed quantity $IBD(\mathcal{N}_1, \mathcal{N}_2)$.

The primary advantage of PADRE is that it is accurate and can be used to obtain the likelihoods of different degrees separating pedigrees as well as different choices of ancestors through which pedigrees are connected. The advantage of DRUID is that it is fast and produces estimates that are similar to the maximum likelihood estimate as we demonstrate in Degree estimation.

Ramstetter et al.[17] derived formulas for ϕ_1 and ϕ_2 for specific pedigree configurations, such as sets of siblings or siblings together with avuncular relatives. Here, we generalize the DRUID estimator to general outbred pedigrees.

The fraction ϕ_i of the genome of A_1 that is passed on to some descendant in $\mathcal{N}_i$ can be computed as

$$\phi_i = 1 - p_{i,1}\left(\left\{O_1 = 0, ..., O_k = 0\right\}\right), \qquad \text{(Equation 7)}$$

where $p_{i,1}(\{O_1 = 0, ..., O_k = 0\})$ is the probability that a given allele is observed in no leaf descendant of node i and is computed recursively using Equation 6. Thus, an estimate of the amount of IBD shared between A_1 and A_2 is

$$\widehat{IBD}(A_1, A_2) = \frac{IBD(\mathcal{N}_1, \mathcal{N}_2)}{\phi_1 \phi_2}. \qquad \text{(Equation 8)}$$

Using the expression $\widehat{\phi} = \widehat{IBD}(A_1, A_2)/4L_{genome}$ for the kinship coefficient when all IBD is of type 1, we obtain the generalized DRUID estimator

$$d_D(A_1, A_2) = d : \frac{1}{2^{d+3/2}} \leq \frac{IBD(\mathcal{N}_1, \mathcal{N}_2)}{4\phi_1 \phi_2 L_{genome}} < \frac{1}{2^{d+1/2}}, \qquad \text{(Equation 9)}$$

where the bounds come from Manichaikul et al.[20] and are the ones used for the DRUID estimator presented in Ramstetter et al.[17]

In Appendix A: Re-rooting the DRUID estimator, we demonstrate how the DRUID estimator can be further generalized to the case in which A_1 is descended from one of the individuals in $\mathcal{N}_2$, or from an internal node of the induced subtree that is a descendant of A_2. Thus, we obtain a version of the DRUID estimator that can be applied to general outbred pedigrees.

The likelihood of the degree of relatedness among groups of individuals

Using the DRUID principle, we can develop a likelihood estimator of the pairwise degree of relatedness between the common ancestors A_1 and A_2, given the observed total IBD $T_{1,2}$ between the genotyped descendants of A_1 and A_2.

Consider again the scenario depicted in Figure 3 in which two sets of genotyped individuals, $\mathcal{N}_1$ and $\mathcal{N}_2$, are related through a common ancestor or pair of ancestors, G. The probability that a given allele from G is observed IBD between $\mathcal{N}_1$ and $\mathcal{N}_2$ can be obtained by conditioning on the events that it is observed in A_1 and A_2. Let $\mathcal{I}$ denote the event that the allele is observed IBD. Then

$$\mathbb{P}(\mathcal{I}) = \phi_1 \mathbb{P}(O_{A_1} = 1) \phi_2 \mathbb{P}(O_{A_2} = 1)$$

$$= \phi_1 \phi_2 2^{-\left(d_{A_1,G} + d_{A_2,G}\right)}, \qquad \text{(Equation 10)}$$

where ϕ_i is computed using Equation 7.

If A_1 and A_2 had exactly one common ancestor with one allele to transmit, then Equation 10 would be the fraction of the genome in which we expect to find some segment shared IBD between some member of $\mathcal{N}_1$ and some member of $\mathcal{N}_2$. However, we must account for the fact that each common ancestor of A_1 and A_2 in G carries two allelic copies and that there can be either one or two such common ancestors.

We consider the case in which A_1 and A_2 are not full siblings. In this case, the event that they are IBD for a given ancestral allele in G is mutually exclusive of the event that they are IBD for any other ancestral allele in G. Therefore, if $|G|$ denotes the number of ancestors, then the probability that A_1 and A_2 are IBD for some ancestral allele is $2|G|\mathbb{P}(\mathcal{I})$.

We can use the probability of observing an allele IBD to obtain an approximate likelihood of the total length $T_{1,2}$ of IBD observed between descendants of A_1 and A_2. The mean of this distribution is simply the expected length of the genome in a state of IBD between the two pedigrees, which is

$$E[T_{1,2}] = 2|G|\mathbb{P}(\mathcal{I})L_{genome}, \qquad \text{(Equation 11)}$$

where L_{genome} is the haploid genome length. An approximation of the variance of $T_{1,2}$ is derived in Appendix A: Approximating the variance of T1,2 and is given by

$$\text{Var}(T_{1,2}) \approx 2|G|\mathbb{P}(\mathcal{I})L_{genome} \frac{E\left[L_{1,2}^2\right]}{E[L_{1,2}]}, \qquad \text{(Equation 12)}$$

where $L_{1,2}$ is the length of any given IBD segment between A_1 and A_2 formed by merging all IBD segments between leaf nodes in A_1 and A_2 that overlap one another. The moments $E[L_{1,2}^m]$ are derived in Appendix A: Approximating the variance of $T_{1,2}$ and can be computed using Equations A11 or A12.

If the segments, $L_{1,2}$ were each exponentially distributed, then $T_{1,2}$ would have a gamma distribution. Thus, we can approximate the distribution of $T_{1,2}$ by

$$T_{1,2}|T_{1,2} > 0 \sim \text{Gamma}(k_{1,2}, \theta_{1,2}),$$

where $k_{1,2}$ and $\theta_{1,2}$ are found by matching the mean and variance of the gamma distribution with $E[T_{1,2}]$ and $\text{Var}(T_{1,2})$. Thus, we obtain

$$T_{1,2}\Big|T_{1,2} > 0 \sim \text{Gamma}\left(\frac{E[L_{1,2}]^2}{\text{Var}(L_{1,2})}, \frac{\text{Var}(L_{1,2})}{E[L_{1,2}]}\right), \qquad \text{(Equation 13)}$$

where $E[L_{1,2}]$ and $E[L_{1,2}^2]$ are given by Equation A12.

If every IBD segment has some length, we can assume that $T_{1,2}$ is only identically zero when there are no IBD segments. The distribution of the number of segments can be modeled as a Poisson random variable with mean $E[N_{1,2}]$ equal to the expected number $N_{1,2}$ of merged segments shared between $\mathcal{N}_1$ and $\mathcal{N}_2$. The probability that there are no segments is then $e^{-E[N_{1,2}]}$. Thus, we have the approximation

$$f_{T_{1,2}}(t_{1,2}) \approx \begin{cases} \dfrac{t_{1,2}^{k-1}}{\Gamma(k)\theta^k} e^{-t_{1,2}/\theta}\left(1 - e^{-E[N_{1,2}]}\right) & \text{if } t_{1,2} > 0 \\[2mm] e^{-E[N_{1,2}]} & \text{if } t_{1,2} = 0. \end{cases},$$

$$\text{(Equation 14)}$$

where $k = E[L_{1,2}]^2/\text{Var}(L_{1,2})$, $\theta = \text{Var}(L_{1,2})/E[L_{1,2}]$, and $E[N_{1,2}]$ is given in Equation A8. Figure S3 shows analytical values computed using Equations 11 and 12 compared to empirical values from simulations. Figure S4 shows the approximate analytical distribution computed using Equation 14 compared to the empirical distribution computed from simulations. Although the gamma distribution in Equation 14 provides a good fit to the empirical distribution, a Gaussian distribution can be more robust in practice and ultimately appears to give better accuracy for pedigree inference. In practice, we use the Gaussian distribution for inference.

A maximum likelihood estimator of the degree between A_1 and A_2 can be obtained by determining the degree $d_L(A_1, A_2)$ between A_1 and A_2 for which value of the distribution in Equation 14 is maximized. This gives the maximum likelihood estimator

$$d_L(A_1, A_2) = \arg\max_d f_{T_{1,2}}(t_{1,2}; d). \qquad \text{(Equation 15)}$$

Determining the ancestral branches through which to connect pedigrees

One difficulty in constructing large pedigrees is determining the ancestors through which two sets of gentoyped individuals are related. A simple fundamental question is whether two lineages are both on the maternal side of an individual, both on the paternal side, or on opposite parental sides. Without genotyped parents, the side through which a lineage passes can be difficult to determine, although sex chromosomes and mitochondrial haplotypes can be used to resolve the parent of origin in some cases.

We consider the problem of inferring whether two distant sets of relatives are related through the same parent of a focal individual, or through different parents. The scenario we consider is illustrated in Figure S5. Even if the purple and red pedigrees in Figure S5 shared no IBD, they could still be related to individual 1 through the same parent by passing through different grandparents. However, if the red and purple pedigrees are related to the focal individual 1 through the same parent, the IBD segments the purple pedigree shares with individual 1 cannot spatially overlap with the segments the red pedigree shares with individual 1. This is because two overlapping segments would have undergone recombination in the parent (i.e., individual 10). The result will either be a spliced segment (Figure S5) or the replacement of one segment by the other with possible reduction in segment size.

In the big Bonsai method, when there are multiple possible grandparents through which we can connect a focal node in a focal pedigree $\mathcal{P}$ to two distantly related pedigrees $\mathcal{P}_1$ and $\mathcal{P}_2$, we examine whether the IBD segments between $\mathcal{P}_1$ and the focal node overlap with the IBD segments between $\mathcal{P}_2$ and the focal node. The efficacy of checking segment overlaps is discussed in Segment overlap detection using simulated data.

Likelihoods for identifying background IBD

Another challenge in identifying the proper ancestral lineages through which to connect pedigrees comes from segments that are shared identically by state by chance between two individuals with no recent common ancestor. These segments, which can be confounded with IBD, are referred to as background IBD.

Background IBD can result in the placement of distant relatives onto incorrect ancestral lineages. The result is often an imbalanced pedigree with many distant lineages connected to the same side or ancestral lineage of another pedigree. We present a method for detecting background IBD and correcting the ancestral lineages through which pedigrees are connected. The likelihood of the degree of relatedness among groups of individuals

Summary of the big Bonsai algorithm

We combine the tools previously described (The probability of a presence-absence pattern of an ancestral allele; The generalized DRUID estimator; The likelihood of the degree of relatedness among groups of individuals; Determining the ancestral branches through which to connect pedigrees; Likelihoods for identifying background IBD) to obtain the big Bonsai method presented in Algorithm 4 in Supplemental methods. The input for the big Bonsai method consists of small pedigrees inferred using the small Bonsai method. It assembles these small pedigrees into a large and sparsely sampled pedigree by iteratively combining the two pedigrees that share the greatest total length of IBD until all pedigrees have been agglomerated into a single pedigree or discarded because they cannot be combined in a reasonable way.

We assume that a pair of pedigrees, $\mathcal{P}_1$ and $\mathcal{P}_2$, can only be combined in ways that connect individuals who share IBD. When combining two pedigrees, the big Bonsai method identifies the sets $\mathcal{N}_1$ and $\mathcal{N}_2$ of genotyped nodes in each pedigree that share at least one IBD segment with an individual in the other pedigree. It is possible that some nodes in the set $\mathcal{N}_1$ are not truly related to the set $\mathcal{N}_2$ and vice versa due to background IBD. In this case, the set $\mathcal{N}_i$ may not have a single common ancestor. If the set $\mathcal{N}_i$ does not have a single common ancestor or a single pair of common ancestors who are partners, we attempt to find the subset of $\mathcal{N}_i$ that has a common ancestor and shares the most IBD with the other set. To accomplish this, we find the set $\tilde{\mathcal{A}}_i$ of most recent ancestral nodes whose descendants comprise $\mathcal{N}_i$. The pair of ancestors $A_1 \in \tilde{\mathcal{A}}_1$ and $A_2 \in \tilde{\mathcal{A}}_2$ whose descendants share the greatest total length of IBD is then determined and we redefine $\mathcal{N}_1$ and $\mathcal{N}_2$ to be the genotyped descendants of A_1 and A_2, respectively.

Our objective is to identify pairs of individuals through which $\mathcal{P}_1$ and $\mathcal{P}_2$ can be connected in such a way that all individuals in $\mathcal{N}_1$ are related to all individuals in $\mathcal{N}_2$. This is accomplished if and only if the sets $\mathcal{N}_1$ and $\mathcal{N}_2$ share at least one common ancestor. Sets $\mathcal{N}_1$ and $\mathcal{N}_2$ will be connected through a common ancestor if their respective common ancestors, A_1 and A_2, share a common ancestor or if A_1 is descended from any individual in $\mathcal{N}_2$ or from any ancestor on the induced subtree Λ_2 of pedigree $\mathcal{P}_2$ relating $\mathcal{N}_2$ to one another. Similarly, sets $\mathcal{N}_1$ and $\mathcal{N}_2$ will have a common ancestor if A_2 is descended from any individual in $\mathcal{N}_1$ or from any ancestor on the induced tree Λ_1 of pedigree $\mathcal{P}_1$ relating $\mathcal{N}_1$.

We present a generalized DRUID estimator in Appendix A: Rerooting the DRUID estimator for connecting pedigrees through individuals A who are not common ancestors of $\mathcal{N}_1$ or $\mathcal{N}_2$. However, connecting pedigrees $\mathcal{P}_1$ and $\mathcal{P}_2$ through all possible pairs can be computationally inefficient. Instead, we accept a certain loss in accuracy and allow pedigrees to be connected only through common ancestors. We find that this approach works well in practice, generating pedigrees that are nearly as accurate as those constructed by connecting $\mathcal{P}_1$ and $\mathcal{P}_2$ in all possible ways.

Let A_1 be a most recent common ancestor of $\mathcal{N}_1$ and let A_2 be a most recent common ancestor of $\mathcal{N}_2$. For each pair of possible ancestors (A_1, A_2), we compute the generalized DRUID estimate $d_D(A_1, A_2)$ of the degree using Equation 9. We then perform the test for background IBD described in Likelihoods for identifying background IBD, which potentially results in a new pair of common ancestors A_1' and A_2' whose descendants do not share detectable background IBD. If the pair (A_1', A_2') differs from the original pair (A_1, A_2), we replace A_1 and A_2 with A_1' and A_2' and recompute the generalized DRUID estimate $d_D(A_1, A_2)$. At the end of

these steps, we have a set of possible ancestral pairs through which $\mathcal{P}_1$ and $\mathcal{P}_2$ can be connected, along with point estimates, $d_D(A_1, A_2)$, of the total degree separating each pair.

It remains to evaluate the likelihood of each pair and degree. Following the notation of Ko and Nielsen,[14] denote the relationship between a pair of individuals A_1 and A_2 with common ancestor (or ancestral pair) G by (d_1, d_2, n), where d_1 is the number of meiotic events separating A_1 from G, d_2 is the number of meiotic events separating A_2 from G, and $n = |G|$ is the number of common ancestors. For a given estimate $d_D(A_1, A_2)$ of the degree between A_1 and A_2 and a number of common ancestors n, we consider all relationship types (d_1, d_2, n) corresponding to degree $d_D(A_1, A_2)$; in other words, we consider all relationship types such that $d_1 + d_2 = d_D(A_1, A_2) + n - 1$.

For a given pair of ancestors A_1 and A_2, and for each relationship (d_1, d_2, n), we connect A_1 and A_2 through all such relationships and we evaluate the composite likelihoods of the resulting pedigrees computed using Equation 4. All pedigrees whose likelihoods are at least a fraction f_ℓ of that of the most-likely pedigree are stored and the rest are discarded. We also apply the test in Determining the ancestral branches through which to connect pedigrees for incompatible ancestral lineages to each retained pedigree and we retain only those pairs that pass the test.

Here, we have considered the procedure for combining two pedigrees $\mathcal{P}_1$ and $\mathcal{P}_2$. However, the output of the small Bonsai method is a set of high-likelihood pedigrees S and the input to the big Bonsai method is a list $\overrightarrow{S} = [S_1, \ldots, S_K]$ of K such sets, if K small pedigrees have been inferred. Let $\mathcal{N}_S$ denote the genotyped node set corresponding to the pedigree set S; in other words, $\mathcal{N}_S$ is the genotyped node set of every pedigree $\mathcal{P} \in S$. If $\mathcal{N}$ is the set of genotyped nodes in the full pedigree, then $\cup_{i=1}^{K} \mathcal{N}_{S_i} = \mathcal{N}$.

At each step of the big Bonsai method, we compare each pair of genotyped sets $\mathcal{N}_{S_i}$ and $\mathcal{N}_{S_j}$ ($1 \leq i, j \leq K$) to determine the pair with the greatest shared total amount of IBD. Here, the total amount of IBD is the total length of IBD obtained by merging the segments shared between all pairs of individuals $(u, v) \in \mathcal{N}_{S_i} \otimes \mathcal{N}_{S_j}$. We then identify the subsets $\mathcal{N}_i \subseteq \mathcal{N}_{S_i}$ and $\mathcal{N}_j \subseteq \mathcal{N}_{S_j}$ that share IBD and we combine each pair of pedigrees $(\mathcal{P}_i, \mathcal{P}_j) \in S_i \otimes S_j$ through all pairs of possible most recent common ancestors of $\mathcal{N}_i$ and $\mathcal{N}_j$. The full algorithm is presented in Algorithm 4 in supplemental methods.

It is possible to mis-infer relationships early in the process of pedigree building that lead to conflicts several steps later in the process. The downstream effects of a misplaced individual can be difficult to predict and prevent without a bird's-eye view of the pedigree, but misplaced pairs of relatives can often be detected after the pedigree is built. In practice, we include a final step in the pedigree building process to detect internal inconsistencies by comparing the final pairwise relationships implied by the pedigree structure to the initial pairwise likelihood predictions. When the inferred relationships have low pairwise likelihoods, we rebuild the pedigree, iteratively expanding the number of pedigrees that are retained at each step to increase the chances that the correct pedigree is explored. We also correct pairwise point estimates that are likely to be incorrect when viewed in the context of a fully built pedigree before attempting to re-infer the pedigree.

Putting together the point estimator, the small Bonsai method, and the big Bonsai method, we obtain the full Bonsai method shown in Figure 1. Outlines of the three primary stages of Bonsai are shown in Algorithms 1, 2, and 4 in supplemental methods. The Bonsai method performs these stages in series.

Subjects and simulations

Our empirical analyses are based on simulated data, as well as a dataset comprised of the pedigrees of 23andMe research participants. All simulations and analyses that used real genotype data were performed using individuals consented for research according to the 23andMe research protocol, which is approved by Ethical & Independent Review Services, a review board accredited by the Association for the Accreditation of Human Research Protection Programs. The study is in accordance with U.S. Federal Policy for the Protection of Human Subjects.

Overview of simulations

Simulations were carried out using two different general approaches. In one approach, no genotype or customer data were used and IBD segments were known with certainty, their positions and lengths being recorded during the simulation process. In the second simulation approach, the full-genome genotypes of research-consented 23andMe customers were used for the pedigree founders and genotypes were simulated for individuals in all subsequent generations through cross-over events. Identical-by-descent segments were then inferred between each pair of individuals using an in-house method for inferring IBD on unphased data,[21] which is similar to that of Seidman et al.[22]

In all simulations, the number of cross-over events in each meiosis was drawn such that the expected number of events was one per 100 cM and the locations of cross-overs were sampled uniformly along chromosomes.

Validated real pedigrees

To evaluate Bonsai on real pedigrees, we constructed 718 pedigrees for individuals in the 23andMe database that were known with a high degree of certainty because a very large fraction of individuals were genotyped. In particular, we identified sets of individuals in which each individual was connected to every other individual through a chain comprised of first-degree relationships (parental or full-sibling). We considered a pair of individuals to be parent and child if they shared at least 3,400 cM of IBD1 and at most 100 cM of IBD2, and if their ages were at least 17 years apart. We considered a pair of individuals to be full siblings if they shared at least 2,400 cM of IBD1, at least 400 cM of IBD2, and at most 3,000 cM of IBD2 and if their parents identified by the aforementioned criteria were exactly the same. We further required that any pair of inferred parents were of opposite sexes. Pedigrees identified in this way allowed us to know the true pedigree structure with a high degree of certainty because parent-offspring and full-sibling pairs can be identified with nearly perfect accuracy.

We identified the set of the largest such pedigrees in each of ten populations. The population of a pedigree was taken to be the computationally inferred population of the majority of pedigree members, where population membership was predicted for a given individual using the approach described in Campbell et al.[23] We considered only pedigrees that contained at least 10 genotyped individuals, resulting in 101 European, 104 North European, 31 South European, 56 African American, 88 Ashkenazi, 57 East Asian, 16 South Asian, 25 Middle Eastern, 101 Latino, and 139 "other" pedigrees.

For analyses comparing Bonsai with the state-of-the-art method PRIMUS, we subsampled to a smaller set of pedigrees to allow the analysis to complete in a reasonable amount of time. For these analyses, we downsampled more heavily in over-represented populations to attain greater uniformity in the numbers of pedigrees from different populations. For these analyses we considered the largest 40 pedigrees from each of 8 computationally inferred populations, except when there were fewer than 40 pedigrees in a population,

in which case, we considered all pedigrees. We retained 40 European, 40 African American, 40 Ashkenazi, 40 East Asian, 16 South Asian, 25 Middle Eastern, 40 Latino, and 40 other pedigrees.

Self-reported pedigrees

The Family Tree feature provided by 23andMe allows users to edit and validate relationships in their pedigrees. We considered a set of such pedigrees where users had either verified or changed relationships, indicating that they knew the correct relationships for at least a subset of individuals in the pedigree. We considered only individuals in these pedigrees who were consented for research and inferred the pedigree using only the subset of research-consented individuals. The inferred relationships in the pedigree could then be compared with the user-verified relationships.

Simulations and fitting of empirical pairwise genetic likelihood distributions

The distribution of the total length of IBD1 and IBD2, the distribution of lengths of IBD1 and IBD2 segments, and the distribution of the total counts of IBD1 and IBD2 segments for a specified relationship type $\mathcal{R}$ were obtained by simulating full genomes for 100 pairs of individuals of the relationship type. For each simulation replicate, a pedigree was specified containing the relationship of interest and cross-over events were simulated within the pedigree.

Gaussian distributions were then fitted to the observed data by moment matching. In practice, we fitted Gaussian distributions to both the total lengths and segment counts, rather than fitting Poisson distributions to the segment count data because the Gaussian distribution provided a better fit for segment counts for close relatives.

The IBD inference algorithm we used operates on unphased data. For such data, it is natural to consider a third class of IBD, which is "IBD1 or IBD2," in other words contiguous regions of the genome in which any IBD is detected. We denote this form of IBD by "IBD3." In practice, the likelihoods used for the analyses in this paper were fitted to the distributions of total lengths and segment counts of IBD3 and IBD2 rather than those of IBD1 and IBD2. The use of IBD3 instead of IBD1 can reduce the variance in segment counts because true IBD1 segments can be broken up by stretches of IBD2. The use of IBD3 segments instead of IBD1 primarily affects the inference of full sibling relationships, which are the most common non-consanguineous relationships with IBD2 and which are easily distinguished from other relationships.

Let T_3 denote the total genome-wide length of IBD3 and let T_2 denote the total genome-wide length of IBD2. Let C_3 denote the total number of segments of IBD3 and let C_2 denote the total number of segments of IBD2. Over the 100 simulation replicates, we computed the mean μ_Q and standard deviation σ_Q of the quantities $Q = T_3, T_2, C_3$, and C_2. The means and standard deviations of these quantities for the simulated relationships are provided with the Bonsai software and are used for inference. Users may prefer to use different IBD quantities and distributions. Instructions for replacing the Bonsai distributions with user-generated ones are provided in the documentation for Bonsai.

Large simulated pedigrees

The 718 validated customer pedigrees described in Validated real pedigrees are often small enough that the small Bonsai method is capable of building them without relying heavily on the big Bonsai method. To evaluate the big Bonsai method, we required considerably larger pedigrees whose structures were known with certainty. Although many pedigrees for 23andMe research-consented customers are large, the relationships within them are typically not known with certainty. Therefore, we simulated large pedigrees to evaluate the big Bonsai method.

Exact IBD was simulated for pedigrees with a depth of five generations by choosing a focal individual and building the "cone" of ancestors comprised of 2 parents, 4 grandparents, 8 great-grandparents, and 16 great-great-grandparents. For each individual in the ancestral cone, a second partner was added with probability 0.2. Two children were created for every pair of partners in the pedigree. Two children were repeatedly created for every pair with no children until the generation with the focal individual was reached. An example of a pedigree generated by this approach is shown in Figure S6.

Large simulated pedigrees to evaluate the effect of background IBD detection on pedigree accuracy

To test the effects of background IBD on pedigree inference accuracy, we required pedigrees with realistic levels of background IBD. The real pedigrees described in Validated real pedigrees were often too small to require considerable amounts of assembly using big Bonsai, where the test for background IBD is performed. The pedigrees in Large simulated pedigrees did not contain background IBD because all IBD was exact. Therefore, we repeated the simulations in Large simulated pedigrees, but this time using the full-genome genotypes of research-consented 23andMe customers as the pedigree founders. Genotypes were simulated for individuals in all subsequent generations through cross-over events and IBD was detected as described in Overview of simulations. We simulated 100 pedigrees for each of ten populations. For a given population, the pedigree founders were research-consented 23andMe customers who were computationally predicted to be from that population.

Simulated pedigrees for testing degree inference

The approach for simulating pedigrees for degree inference was similar to that in Large simulated pedigrees; however, the pedigree structure was different. For these pedigrees, we were interested in inferring the degree between a pair of common ancestors A_1 and A_2, given IBD observed between their descendants $\mathcal{N}_1$ and $\mathcal{N}_2$.

For this analysis, we created two identical small pedigrees: $\mathcal{P}_1$ and $\mathcal{P}_2$. Each small pedigree had the same structure comprised of the common ancestor, A_1 or A_2, their partner, their two children, and four grandchildren, where the grandchildren were comprised of two children for each child of A_1 or A_2. The ancestors A_1 and A_2 were then connected by degree $d(A_1, A_2)$ through a pair of common ancestors, where the degree d varied from 1 to 13. Exact IBD for 200 pedigrees was simulated for each degree.

Results

We considered both simulated and real data to investigate the performance of the small and big Bonsai methods and their components.

Degree estimation

To evaluate the accuracy of degree inference using the likelihood estimator (Equation 15) and the generalized DRUID estimator (Equation 9), we applied these estimators to infer the degree between common ancestors A_1 and A_2 of two small pedigrees $\mathcal{P}_1$ and $\mathcal{P}_2$ (Simulated pedigrees for testing degree inference).

Figure 4 shows the accuracy of the likelihood estimator d_L and the generalized DRUID estimator d_D for inferring the degree $d(A_1, A_2)$, conditional on the event that any IBD at all was observed between the leaf nodes in $\mathcal{P}_1$ and $\mathcal{P}_2$. From Figure 4, it can be seen that both the maximum
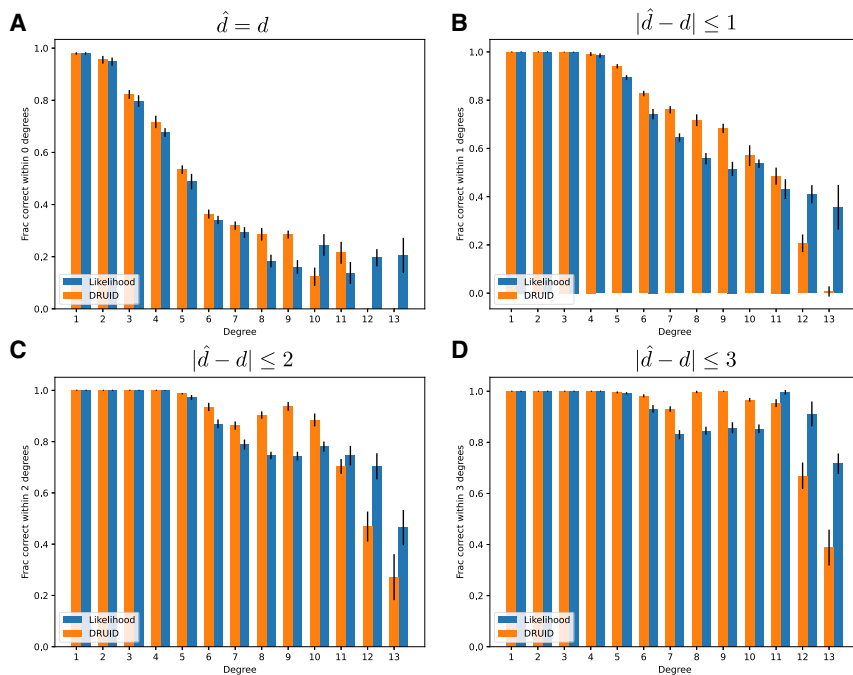
Figure 4. The accuracy of the likelihood method (Equation 15) and the generalized DRUID method (Equation 9) for inferring the degree between a pair of common ancestors
The accuracy of the estimate is shown for four different tolerances: (A) exactly equal to the true degree, (B) within one degree of the true degree, (C) within two degrees of the true degree, and (D) within three degrees of the true degree. Error bars are symmetrical with total lengths equal to twice the standard deviation.

likelihood estimator d_L and the generalized DRUID estimator d_D have similar accuracies for inferring the degree $d(A_1, A_2)$, although the DRUID estimator was more accurate for moderate degrees whereas the likelihood estimator was able to infer higher degrees. This difference is likely due to the choice of distribution used in the approximation of the likelihood, which appears mis-calibrated for moderate degrees, but permits inference for high degrees. In practice, because the DRUID and likelihood estimators are similar, we use the generalized DRUID estimator for inferring the degree of separation between two small pedigrees for reasons of computational efficiency.

To compute the bar heights and standard deviations in Figure 4, we performed ten replicates in which we subsampled four nodes without replacement from $\mathcal{P}_1$ and four nodes without replacement from $\mathcal{P}_2$ within each of the 200 pedigrees for a given degree. We then computed the variance across these ten replicates.

Segment overlap detection

We evaluated the degree to which overlapping IBD segments can be informative about the ancestors through which two pedigrees are connected using the large simulated pedigrees described in Large simulated pedigrees. For each pedigree, we considered the four grandparents of a focal individual and the leaves descended from all lineages extending up from each of the four grandparents. In the example large pedigree shown in Figure S6, the focal individual is one of the yellow leaf nodes and the grandparental clades corresponding to the four leaf sets are colored in green, cyan, red, and magenta.

For a pair of leaf sets related to the focal individual through an ancestral couple, we expect to see no overlap in the IBD segments shared with the focal individual. For a pair of leaf sets related to the focal individual through two grandparents who are not a couple, we expect to observe overlapping segments occasionally.

Figure 5 shows the rate at which segments from one leaf set overlapped segments from another leaf set by more than a fraction f of the total IBD observed between the two leaf sets, combined, for $f \in 0.01, 0.05, 0.1, 0.15, 0.2$. Each bar in Figure 5 was computed using 100 pedigrees, each with four pairs of leaf sets related to individual 1 through a pair of grandparents who were not a couple. Only identical-by-descent segments greater than 5 cM in length were considered.

Let i denote the focal individual. For leaf sets $\mathcal{N}_1$ and $\mathcal{N}_2$ with total amounts of IBD to the focal individual denoted by $T_{i,\mathcal{N}_1}$ and $T_{i,\mathcal{N}_2}$, let $T_{i,\mathcal{N}_1 \cup \mathcal{N}_2}$ denote the total length of merged segments between focal individual i and either set. We recorded an overlap in segments if the following relationship was satisfied: $T_{i,\mathcal{N}_1} + T_{i,\mathcal{N}_2} - T_{i,\mathcal{N}_1 \cup \mathcal{N}_2} > f T_{i,\mathcal{N}_1 \cup \mathcal{N}_2}$. Figure 5 indicates that even with few sampled leaves from each leaf set, it is possible to detect overlapping identical-by-descent segments a large fraction of the time when the leaves are related through grandparents who are not a couple.

Timing and accuracy of small Bonsai, compared with PRIMUS

To evaluate the accuracy and running time of Bonsai in comparison with PRIMUS, we applied PRIMUS and Bonsai to a set of 281 pedigrees comprised of research-consented 23andMe customers (Validated real pedigrees) for which the true pedigree was known with a high degree of certainty because a large fraction of individuals were genotyped.

Pedigrees in which all individuals have been genotyped are simple to infer by connecting together first-degree relatives. The difficulty is in constructing pedigrees in which only a small fraction of individuals have been genotyped. Therefore, to evaluate the accuracy of Bonsai and PRIMUS, we subsampled the validated pedigrees and performed inference using the subset of individuals, ignoring the remaining individuals. The resulting pedigree could then be compared to the subgraph of the true pedigree
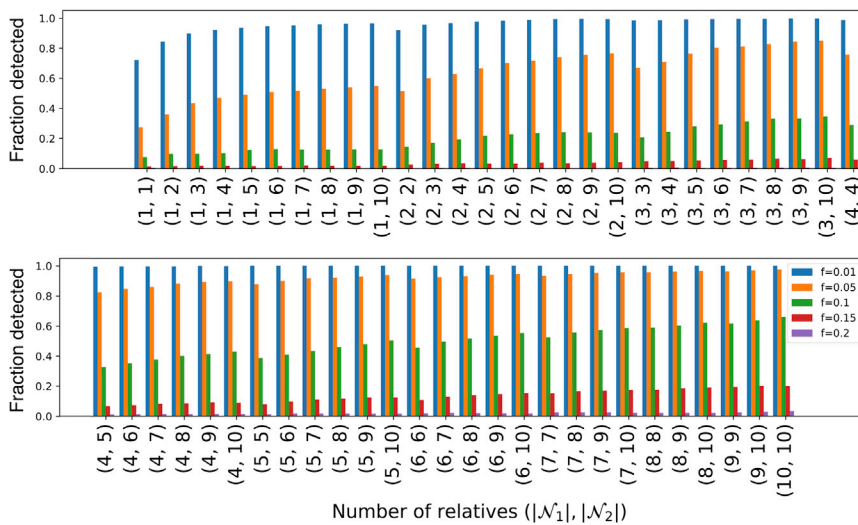
Figure 5. The probability of observing an IBD segment overlap

The plot shows the probability of observing an overlap of at least fraction f ($f = 0.01, 0.05, 0.1, 0.2$) among segments shared identical-by-descent between the focal individual and sets of leaves related to the focal individual through ancestors who are not a couple. Exact IBD segments were simulated for large pedigrees like that shown in Figure S6. IBD was recorded between the focal individual and the leaf nodes of each of the four clades related to the focal individual through each of the four grandparents (colored green, cyan, red, and magenta in Figure S6). An observed identical-by-descent segment overlap was evidence that the lineages were related to the focal individual through a pair of ancestors who were not a couple.

corresponding to the subsampled individuals to determine the accuracy of the inference.

For each pedigree, we considered the set of individuals corresponding to all leaves and their parents. We then subsampled this set in each pedigree to 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of its members with a minimum of at least two individuals sampled per pedigree. Figure 6A shows the fraction of the time the small Bonsai and PRIMUS pedigrees matched the true pedigree. If multiple PRIMUS pedigrees achieve the maximum likelihood, multiple pedigrees are returned. The bars in Figure 6A are labeled "Any PRIMUS" because they show whether any of the highest-scoring PRIMUS pedigrees matched the true pedigree exactly. In comparison, Bonsai returns a single pedigree by default.

We also compared the running time of the Bonsai method to the running time of PRIMUS for the same set of pedigrees described in Validated real pedigrees. Figure 6B shows the running time for small Bonsai compared to the running time for PRIMUS for different percentages of sampled lineages from each of the pedigrees. Because PRIMUS often did not complete for a given pedigree or required a very long running time, we terminated the Bonsai or PRIMUS compute for a pedigree if it took longer than 30 s (Figure S7). Because no Bonsai compute took longer than 30 s, the running times for PRIMUS in Figure 6B are biased downward, whereas the times are shown for all Bonsai pedigrees.

Timing and accuracy of the big Bonsai method

We investigated the ability of the big Bonsai method to accurately infer pedigrees using the very large simulated pedigrees described in Large simulated pedigrees as well as the full set of 718 validated pedigrees described in Validated real pedigrees. The simulated pedigrees allowed us to explore the ability of Bonsai to reconstruct very large pedigrees with distant relationships, and to investigate the effect of pedigree size on running time. The validated pedigrees allowed us to investigate the performance of Bonsai on real data across several populations.

Figure 7 shows timing and accuracy results for reconstructing large five-generation pedigrees simulated using the approach described in Large simulated pedigrees. To evaluate the ability of the big Bonsai method to reconstruct pedigrees with sparsely sampled individuals, we further subsampled 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the non-founder individuals in the most recent two generations. Sampling 10% of these individuals corresponds to sampling approximately 5% of all individuals in the full pedigree and sampling 100% of these individuals corresponds to sampling approximately 50% of all individuals in the pedigree overall. Our sampling scheme presents a challenge to pedigree reconstruction because the samples did not contain ancestral individuals who could provide information about the degrees of distant relationships.

From Figure 7A, it can be seen that the running time is on the order of several seconds per pedigree, even though pedigree sizes were large. Bonsai built pedigrees with more than 100 sampled individuals in tens of seconds.

The big Bonsai method is designed to drop small pedigrees from consideration, rather than combining them with the other pedigrees when an inconsistency is detected. This can occur, for example, if the small pedigree is inferred with a very unlikely relationship despite rerunning with parameter values that search a broader pedigree space and attempting to correct relationships that are judged to be inaccurate. Figure 7B indicates that the fraction of times individuals or small pedigrees were dropped was small, as the number of placed individuals was typically very close to the number of sampled individuals.

Figures 7C–7E show the accuracy for inferring large pedigrees when different fractions of individuals were sampled. Here, we compare each true pairwise relationship to the relationship reconstructed in the most likely Bonsai tree. Close relationships were typically reconstructed accurately, whereas distant relationships were more challenging, yet still generally accurate especially when the fraction of sampled individuals was high.
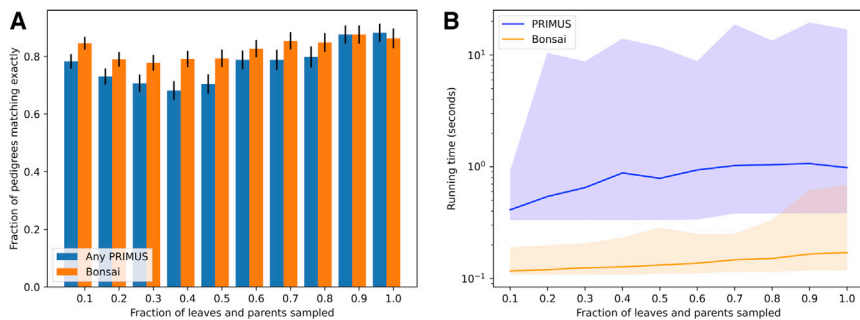
Figure 6. Evaluation of pedigree accuracy and comparison of running times between small Bonsai and PRIMUS

Accuracy and running time were evaluated using 281 pedigrees of 23andMe research participants that were known with a high degree of certainty because most individuals in each pedigree were genotyped. When inferring a pedigree, we subsampled 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% of the set comprised of leaves and parents of leaves uniformly at random without replacement. The subsampled individuals were then used to reconstruct the pedigrees using PRIMUS and Bonsai using the same pairwise relationship likelihoods.

(A) Comparison of overall pedigree accuracy for all placed individuals. The bars labeled "Any PRIMUS" show the rate at which any of the highest likelihood pedigrees returned by PRIMUS correctly matched the true pedigree. Symmetrical error bars have lengths given by the twice the standard deviation of a normalized binomial random variable with n given by the total number of pedigrees and p given by the fraction of pedigrees with the correct topology.

(B) Running time for PRIMUS and Bonsai. Bands show the range between the minimum and maximum running times.

Note that, because the ages of individuals in the pedigree conformed to average age differences between generations, it was sometimes possible to distinguish distant half relationships from distant full relationships. For example, a pair of individuals of the same age related by four degrees of separation is more likely to be a pair of half first cousins, rather than a full first cousin once removed. Half relationships are likely to be more challenging to infer in practice, given that age differences may differ from expectation.

To investigate the accuracy of the big Bonsai method on real data, we inferred 718 customer pedigrees that were known with a high degree of confidence because a large number of individuals had been genotyped. Again, to re-create realistic sampling conditions, we subsampled these pedigrees to 50% of their genotyped leaves and the parents of the leaves. Figure S8 shows the number of pedigrees for which the true pedigree was recovered exactly. The rate was relatively high, given that the inferred pedigree did not match the true pedigree unless all relationships were correctly inferred.

Reconstruction of self-reported pedigrees using big Bonsai

We also compared relationships inferred by Bonsai with self-reported relationships using 265 pedigrees for which the relationships between two or more individuals had been self-reported by the focal individual for whom the pedigree was built (Self-reported pedigrees).

Figure 8 shows the correspondence of each inferred relationship type with the self-reported relationship type. The plots show the fraction of times the self-reported and inferred relationships agreed exactly in that their relationship tuples (up, down, number of ancestors) were the same. The plots also show the fraction of times the relationships agreed in degree, the fraction of times the relationships agreed within one degree, and the fraction of times the relationships agreed within two degrees.

The inferred and self-reported relationships typically agreed for close relationships up to first cousins. However,

the inferred relationship often differed from the self-reported relationship for distant relationship types, and occasionally for relatives as close as siblings or parents. For parent-child and full sibling pairs, it is possible to check whether the self-reported relationship is correct because the identical-by-descent sharing patterns for these relationships are very distinct from other relationship types. It is of interest to note that in all but one case in which the inferred and self-reported relationships differed for a parent-child or full sibling pair, the self-reported relationship was, in fact, incorrect due to impossible levels of shared IBD. In these cases, it was frequently the case that a self-reported parent-child pair shared no IBD, or that a self-reported full sibling pair shared no IBD2 and instead had an IBD sharing pattern that was more consistent with a half sibling or a cousin. In only one case was the self-reported relationship type consistent with the IBD sharing pattern, and in this case one individual had a self-reported age much greater than 100 years, leading to a strong contribution from the age component of the likelihood and an incorrectly inferred relationship type.

For distant relationships, we observed greater disparities between the self-reported and inferred values. However, the inferred degree was often within one or two degrees of the self-reported relationship, even for relationships as distant as seventh degree or higher in some cases. Moreover, relationships for which the self-reported and inferred degrees differed by more than two degrees typically had few self-reported pairs (Figure 8). This relatively high accuracy for distant relationship degree is consistent with our analysis of the accuracy of the generalized DRUID estimator.

Discussion

We have presented a method for inferring large pedigrees quickly and accurately, even when the fraction of genotyped individuals in a pedigree is low and the distance between an individual and their closest relative can be moderate or large. Our method has three
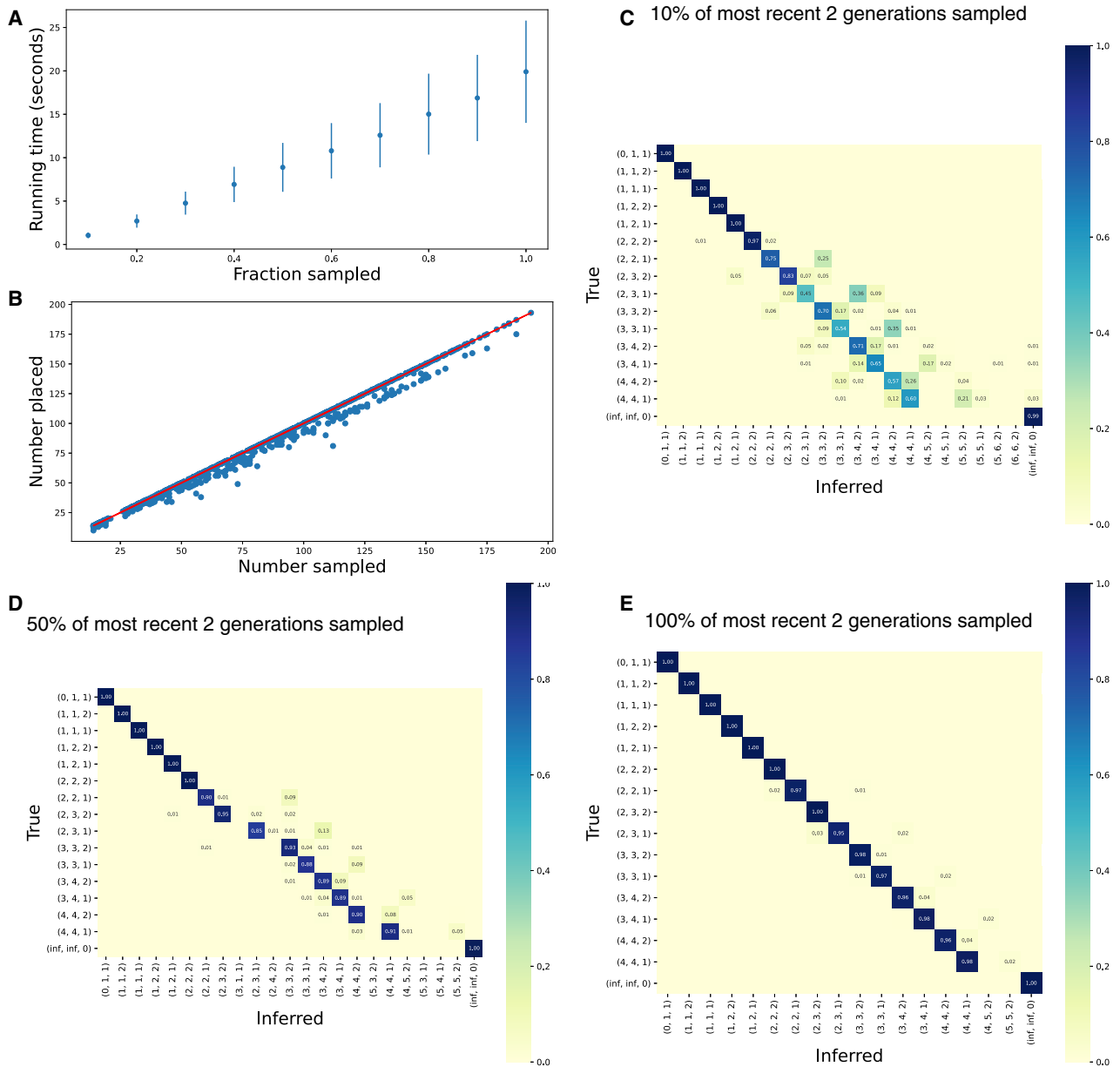
Figure 7. Timing and accuracy of the big Bonsai method

Large pedigrees were simulated with a depth of five generations and two offspring per pair as described in Large simulated pedigrees. To capture the sparsity of pedigrees observed in direct-to-consumer pedigree data, we sampled only a fraction of individuals in the most recent two generations of each pedigree and used these to infer the pedigree.

(A) Running time for big Bonsai as a function of the fraction of sampled individuals in the most recent two generations.

(B) The number of sampled individuals from each pedigree and the number placed.

(C–E) The fraction of pairs with a given relationship type that were inferred to have each other relationship type. The inferred pairwise relationships were those reconstructed in the most likely Bonsai tree. Tuples $(d_{i,G}, d_{j,G}, |G|)$ indicate a specific relationship type between individuals i and j using the notation of Ko and Nielsen:[14] (up, down, number of common ancestors). The tuple (inf, inf, none) indicates unrelated individuals.

component algorithms that are applied in sequence: (1) a method to infer the likelihoods of pairwise relationships between each pair of individuals using both age and IBD data, (2) a method for inferring pedigrees of small-to-moderate size, and (3) a method for combining small pedigrees together into large and sparsely sampled pedigrees.

The small Bonsai algorithm efficiently explores the space of possible pedigrees using a constructive approach. This approach is similar to that of PRIMUS,[13] but it employs several features that make it more efficient and more accurate than PRIMUS, including incorporating ages directly into the likelihoods, expanding the set of pedigrees that are explored, and introducing a branch-and-bound-like
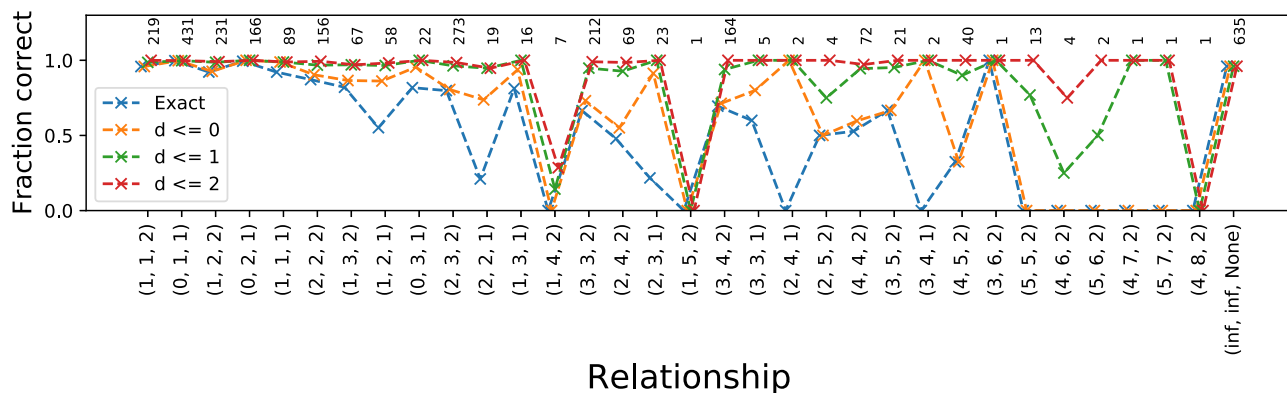
Figure 8. Comparison with self-reported pedigrees

Comparison of predicted relationships with self-reported relationships. Blue markers show the fraction of relationship pairs for which the inferred and self-reported relationships agreed exactly. The orange, green, and red markers show the fraction of pairs for which the degrees of the inferred and self-reported relationships differed by at most 0, 1, or 2 degrees, respectively. The number of pairs for each relationship is shown above the curves. Dashed lines are included to improve visibility.

method for exploring the space of pedigrees more efficiently.

The methodological approaches implemented in the small Bonsai method provide a pedigree inference algorithm with improved accuracy and performance. However, the primary novelty of the Bonsai method is in the big Bonsai algorithm, which combines small pedigrees together into large and sparsely sampled pedigrees. This algorithm makes it possible to construct pedigrees that are much bigger than the sizes that can be constructed by current approaches.

The construction of large and sparse pedigrees requires a fundamentally different approach from combining individuals one at a time as is done in PRIMUS or small Bonsai, or by searching a broad pedigree space by rearranging pedigrees as is done in CLAPPER. Because the space of possible pedigrees is large, it is useful to proactively narrow the set of possible pedigrees to include only the pedigrees with the highest likelihoods.

Combining small pedigrees together into large and sparse pedigrees, as is done in the PADRE and DRUID methods, makes it possible to leverage information in the previously inferred small pedigrees to identify the most likely ways in which the small pedigrees can be connected together. Leveraging information across small pedigrees allows us to more accurately infer the degree of relatedness between two small pedigrees, to identify background IBD, and to identify likely lineages through which the pedigrees are combined together.

We have introduced three tools for combining pedigrees together. First, we have generalized the DRUID method of Ramstetter et al.[18] to apply to general outbred pedigrees, rather than specific pedigree structures. We have also extended the method to allow pedigrees to be connected through pairs of individuals who are not common ancestors. We have shown that the generalized DRUID estimate is similar to the approximate maximum likelihood estimate. Thus, rather than exploring multiple ways of con-

necting two pedigrees and selecting the most likely pedigree, we can simply connect the two pedigrees through the DRUID point estimate and achieve a similar result, speeding up the inference process.

We have also introduced is an approximate likelihood for the degree separating the common ancestors of two pedigrees given the total length of IBD shared by the pedigrees. This likelihood is essentially a reformulation of the generalized DRUID estimator in a likelihood context. This likelihood is used as the foundation for our method for testing whether the IBD shared between two sets of individuals is the result of a true relationship, or whether the IBD is background IBD.

Finally, we have also introduced a method for determining when the connection of pedigrees through certain ancestral branches is consistent with patterns of IBD overlap. This method improves the accuracy of assigning two pedigrees to the correct parental sides of a focal individual in a focal pedigree. Using only information contained in pairwise IBD sharing, inconsistent pedigrees would not be detected, as pedigrees formed by connecting two pedigrees through incompatible grandparental lineages would appear to have the same likelihood as the true pedigree. This approach achieves high sensitivity even when few relatives on each parental side have been sampled.

In addition to detecting segment overlaps, it is likely that ancestral lineage placement could be improved by using IBD detected on sex chromosomes. At present, the Bonsai method uses only autosomal IBD to avoid considering the sexes of ancestral individuals along the paths connecting each pair of individuals when computing the likelihoods of their relationships. Increased sensitivity can also be obtained by using SNP-level information in the test of IBD overlap, such as opposite homozygotes, instead of identical-by-descent segments, as overlaps often occur between segments that are too short to be identified by existing IBD methods.

Compared to previous methods for inferring complex human pedigrees, the Bonsai method yields improvements in both accuracy and computational efficiency and makes it possible to build pedigrees that are considerably larger than those that were possible before. The speed of pedigree building depends on the complexity of the pedigree, the proportion of individuals who are genotyped, and the distribution of these individuals throughout the generations of the pedigree. As a result, it can be difficult to characterize the running time of Bonsai relative to other methods. However, in a comparison of running times on 281 real pedigrees, Bonsai was always faster than the current fastest method PRIMUS and often built pedigrees in a matter of seconds that did not complete when built with PRIMUS.

The faster running time of Bonsai is due in part to efficiencies including the heuristic branch-and-bound-like approach, and in part to the fact that ages are incorporated directly into the likelihoods. The age component of the likelihood often tips the balance in favor of one relationship over another, allowing pedigrees with that relationship to have higher likelihoods. As a result, a greater number of pedigrees can be discarded at each step than if age information were ignored or used only for pairwise checks (e.g., parent older than child).

Age distributions vary somewhat among populations and mis-specification of the age distributions could be a source of bias in the Bonsai estimates. The 23andMe database makes it possible to estimate age difference distributions for many different relationship types. However, these distributions may differ from the age distributions in other populations to which Bonsai may be applied. Although the age difference distributions did not appear to differentially affect accuracy in different populations in the samples we tested, it's possible that differences in accuracy could become apparent in other datasets or when aggregating across many pedigrees, with resulting biases in downstream analyses. Thus, it may be useful to keep in mind that both age and IBD distributions were trained on one particular dataset, albeit a large one. For any particular analysis, both the IBD and age distributions used by Bonsai are customizable by the user as described in the software documentation.

The speed and accuracy of the Bonsai method depend in part on the values of the parameters r, f_ℓ, and δ, with higher values of r and lower values of f_ℓ typically resulting in more accurately inferred pedigrees because they permit a more thorough exploration of the pedigree space and larger values of δ resulting in less accurate pedigrees because they permit the connection of more distant relationship types, which are inferred with a lower degree of accuracy. However, the effects of these parameter values on Bonsai accuracy and running time are not simple to predict. In particular, they all pertain to the small Bonsai algorithm, which is only a subset of the full algorithm that contains additional heuristics for attaching pedigrees, post hoc checks on inferred relationships, and logic for rebuilding pedigrees using different parameter values if the method detects unlikely placements of individuals relative to their pairwise point predictions. Thus, we suggest that users run Bonsai using the default values of these parameters unless they have a good reason to change them.

There is the potential to improve close relationship estimates by using phasing information. Williams et al.[24] have demonstrated that half-sibling, avuncular, and grandparental relationships, which have been difficult to differentiate historically due to the fact that the total amount of expected IBD is the same for each of these relationship types, can be differentiated by making use of long-range phasing information. Phased IBD estimates, obtained from programs such as the PhasedIBD method of Freyman et al.,[25] could provide a considerable boost in accuracy for close relationships.

Close relationship accuracy can also be improved using statistics that capture differences in the spatial distributions of IBD among sets of more than two relatives, such as those presented in Qiao et al.[26] and Ramstetter et al.[17] Improved close relationships would lead to improved distant relationships due to the fact that the small pedigree structures leveraged by the distant degree estimates would be more accurate.

Although the theoretically maximal accuracy with which a pedigree can be inferred differs across human populations due to differences in demographic histories, it is likely that improvements in accuracy can be attained for all populations through improved methodology, such as the improvement of pairwise relationship inference by methods such as deep-learning trained in specific populations, the inclusion of additional consanguineous relationship types, the addition of spatial IBD information, and the inclusion of additional genetic information from sex chromosomes and mitochondrial DNA. By nature, pedigree inference is a complicated problem requiring methods that can handle a wide variety of pedigree structures and input data. However, our results show that inference of large and sparse human pedigrees can be done rapidly, and that accuracy will continue to increase as pedigrees become increasingly densely sampled.

Appendix A

The probability of a pattern of IBD

Consider the induced subtree in a pedigree relating a set of genotyped individuals. This tree is shown with dashed red lines in Figure 3 with nodes of the tree indicated with red dots. Let D_i denote the presence-absence pattern at the leaves descended from i and define $p_{i,s} \equiv \mathbb{P}(D_i|O_i = s)$, where s is the state O_i at node i.

We have, using the approach of Felsenstein,[18]

$$p_{i,0} = \prod_c \Big[\mathbb{P}(O_c = 0 | O_i = 0)p_{c,0} +$$

$$\mathbb{P}(O_c = 1 | O_i = 0)p_{c,1} \Big] = \prod_c p_{c,0} \quad \text{(Equation A1)}$$

where the product is taken over all child nodes c of individual i and the second term is zero because $\mathbb{P}(O_c = 1|O_i = 0) = 0$. Similarly, we have

$$p_{i,1} = \prod_c \left[\mathbb{P}(O_c = 1 \big| O_i = 1) p_{c,1} + \mathbb{P}(O_c = 0 \big| O_i = 1) p_{c,0} \right]$$

$$= \prod_c \left[2^{-d_{c,i}} p_{c,1} + (1 - 2^{-d_{c,i}}) p_{c,0} \right]. \qquad \text{(Equation A2)}$$

In the final lines of Equations A1 and A2, we have used the fact that the probability that an allelic copy is transmitted in one meiosis is $1/2$.

Equations A1 and A2 establish a recursion for computing the probability of an observed presence and absence pattern for a given ancestral allelic copy at a single base of the genome.

$$p_{i,0} \equiv \mathbb{P}(D_i \mid O_i = 0), \; p_{i,1} \equiv \mathbb{P}(D_i \mid O_i = 1),$$

$$p_{i,0} = \prod_c p_{c,0}$$

$$p_{i,1} = \prod_c \left[2^{-d_{c,i}} p_{c,1} + \left(1 - 2^{-d_{c,i}} \right) p_{c,0} \right],$$

For a leaf node l with state s, the base conditions are $p_{l,0} = \delta_{0,s}$ and $p_{l,1} = \delta_{1,s}$ where $\delta_{u,v}$ is the kronecker delta taking value 1 if $u = v$ and the value 0, otherwise.

Approximating the variance of $T_{1,2}$

Here, we derive an approximation of the variance of the total length, $T_{1,2}$, of IBD shared across the genotyped descendants of two ancestral individuals, A_1 and A_2. When we encounter a patch of IBD at a locus, the length of the patch can be approximated as the maximum length of $|\mathcal{N}_1| \times |\mathcal{N}_2|$ different identical-by-descent segments, where $\mathcal{N}_i$ is the set of genotyped nodes below ancestor A_1 at locus m in which the identical-by-descent segment is observed. This approximation comes from conceptualizing IBD sharing among the $|\mathcal{N}_1|$ identical-by-descent segment carrying descendants of A_1 and the $|\mathcal{N}_2|$ identical-by-descent segment carrying descendants of A_2 as $|\mathcal{N}_1| \times |\mathcal{N}_2|$ independent segments with a single point at which all segments overlap. The length of the merged segment to one side of this focal point then has a distribution given by the maximum of $|\mathcal{N}_1| \times |\mathcal{N}_2|$ exponential random variables whose means depend on the degree of separation between the corresponding pairs of leaf individuals. To simplify matters, we assume that the length of the full merged overlapping segment (not just to the left or right) is exponentially distributed.

This approximation is an oversimplification of the identical-by-descent sharing pattern because the segments are not truly independent and need not overlap at a single point. Moreover, under this approximation, the length of the merged segment would be the maximum over sums of identically distributed random variables, representing the sum of the length of a segment to the right of the center point and the length of the segment to the left. However, we are not overly concerned with these drawbacks of the conceptualization because our main goal is to obtain an accurate yet simple approximation of the variance of

the distribution. We also assume that no member of $\mathcal{N}_i$ is the direct ancestor of another member of the set, which holds in practice if we drop all individuals from $\mathcal{N}_i$ who are descended from others.

The length, $\ell_{i,j}$, of an identical-by-descent segment between leaf nodes i and j can be modeled as an exponentially distributed random variable with mean length $\mu_{ij} = L_{genome}/d_{i,j}R$, where $d_{i,j}$ is the number of meioses between them and R is the expected number of recombination events, genome wide, in one meiosis.[19] When the length of the genome is expressed in centimorgans (cM), the expected number of recombination events in the genome is $L_{genome}/100$. Thus, the expected length in cM of an identical-by-descent segment between individuals i and j separated by $d_{i,j}$ meioses is $\mu_{ij} = 100/d_{i,j}$.

Let $L_{1,2}$ denote a random variable describing the length of the segment formed by merging all segments at a given locus m between descendants of A_1 and A_2. If the lengths of all segments at this locus were independent, their merged length in our conceptualization would have a distribution given by the maximum over independent exponentially distributed random variables with means given approximately by $\{\mu_{i,j}\}_{i \in \mathcal{N}_1, j \in \mathcal{N}_2}$.

If the leaf nodes with observed IBD at the marker are $\mathcal{N}_1$ and $\mathcal{N}_2$, then we have $L_{1,2} = \max(\{\ell_{i,j}\}_{i \in \mathcal{N}_1, j \in \mathcal{N}_2})$. Under this condition, the cumulative density function (CDF) $F_L(\ell; \mathcal{N}_1, \mathcal{N}_2)$ of L is

$$F_{L_{1,2}}(\ell; \mathcal{N}_1, \mathcal{N}_2)$$

$$= \mathbb{P}(L_{1,2} < \ell; \mathcal{N}_1, \mathcal{N}_2)$$

$$= \mathbb{P}(\ell_{i,j} < \ell, \text{ for } i \in \mathcal{N}_1, j \in \mathcal{N}_2)$$

$$= \prod_{\substack{i \in \mathcal{N}_1 \\ j \in \mathcal{N}_2}} \mathbb{P}(\ell_{i,j} < \ell)$$

$$= \prod_{\substack{i \in \mathcal{N}_1 \\ j \in \mathcal{N}_2}} \left(1 - e^{-\lambda_{ij}\ell} \right)$$

$$= 1 - \sum_{i \in \mathcal{N}_1, j \in \mathcal{N}_2} e^{-\lambda_{ij}\ell} + \sum_{i,u \in \mathcal{N}_1, j,v \in \mathcal{N}_2} e^{-(\lambda_{ij} + \lambda_{u,v})\ell} \left(1 - \delta_{(i,j),(u,v)} \right)$$

$$\text{(Equation A3)}$$

$$- \sum_{i,u,w \in \mathcal{N}_1, j,v,z \in \mathcal{N}_2} e^{-(\lambda_{ij} + \lambda_{u,v} + \lambda_{z,w})\ell} \left(1 - \delta_{(i,j),(u,v)} \right) \left(1 - \delta_{(i,j),(z,w)} \right)$$

$$\times \left(1 - \delta_{(u,v),(z,w)} \right) + \cdots, \qquad \text{(Equation A4)}$$

where $\lambda_{i,j} = 1/\mu_{i,j} = d_{i,j}/100$ and $\delta_{(a,b),(c,d)}$ is the Kronecker delta between tuples (a,b) and (c,d), which is equal to 1 when $(a,b) = (c,d)$ and 0, otherwise.

The sets $\mathcal{N}_1$ and $\mathcal{N}_2$ are, themselves, random variables. Summing over all sets $\mathcal{N}_1$ and $\mathcal{N}_2$, we have

$$F_{L_{1,2}}(\ell) = \sum_{\mathcal{N}_1, \mathcal{N}_2} F_{L_{1,2}}(\ell; \mathcal{N}_1, \mathcal{N}_2) \mathbb{P}(\mathcal{N}_1) \mathbb{P}(\mathcal{N}_2), \quad \text{(Equation A5)}$$

where the probabilities $\mathbb{P}(\mathcal{N}_1)$ and $\mathbb{P}(\mathcal{N}_2)$ are probabilities of observing IBD in the sets of leaf nodes below A_1 and A_2 and can be approximated using the recursion in Equation 6.

Over the length of the genome, the number $N_{1,2}$ of identical-by-descent segments between the descendants of A_1 and A_2 is approximately Poisson distributed with mean $2|G|\mathbb{P}(\mathcal{I})L_{genome}/E[L_{1,2}]$. This rate comes from the fact that the average total amount of the genome in a patch of IBD is $2|G|\mathbb{P}(\mathcal{I})L_{genome}$ while the average length of any given segment is $E[L_{1,2}]$. When the lengths of IBD are short and far apart, which they are when the degree between A_1 and A_2 is large, this is a reasonable approximation. This is precisely the regime in which the distribution in Equation 14 is most useful.

The total length $T_{1,2}$ of merged IBD among the descendants of A_1 and A_2 is

$$T_{1,2} = \sum_{n=1}^{N1,2} N_{1,2} L_{1,2}^{(n)}, \qquad \text{(Equation A6)}$$

where $L_{1,2}^{(n)}$ is the length of the nth merged segment. We can derive the variance of $T_{1,2}$ using the law of total variance as

$$\text{Var}(T_{1,2}) = E[\text{Var}(T_{1,2}|N_{1,2})] + \text{Var}(E[T_{1,2}|N_{1,2}])$$

$$= E[N_{1,2}\text{Var}(L_{1,2})] + \text{Var}(N_{1,2}E[L_{1,2}])$$

$$= E[N_{1,2}]\text{Var}(L_{1,2}) + \text{Var}(N_{1,2})E[L_{1,2}]^2. \qquad \text{(Equation A7)}$$

Note that because $N_{1,2} \sim \text{Poisson}(2|G|\mathbb{P}(\mathcal{I})L_{genome}/E[L_{1,2}])$, we have

$$E[N_{1,2}] = \text{Var}(N_{1,2}) = 2|G|\mathbb{P}(\mathcal{I})L_{genome}/E[L_{1,2}]. \qquad \text{(Equation A8)}$$

So Equation A7 simplifies to

$$\text{Var}(T_{1,2}) = \frac{2|G|\mathbb{P}(\mathcal{I})L_{genome}}{E[L_{1,2}]}\left[\text{Var}(L_{1,2}) + E[L_{1,2}]^2\right]$$

$$= 2|G|\mathbb{P}(\mathcal{I})L_{genome}\frac{E[L_{1,2}^2]}{E[L_{1,2}]}, \qquad \text{(Equation A9)}$$

where we have used the fact that $\text{Var}(X) = E[X^2] - E[X]^2$.

It remains to find $E[L_{1,2}]$ and $E[L_{1,2}^2]$. Using the CDF of $L_{1,2}$ in Equation A5 and the fact that $E[X^m] = m\int_{\mathbb{R}} x^{m-1}[1 - F_X(x)]dx$, we have

$$E_{\mathcal{N}_1,\mathcal{N}_2}\left[L_{1,2}^m\right] = m \int_{\ell=0}^{\infty} x^{m-1}\left[1 - F_{L_{1,2}}(\ell;\mathcal{N}_1,\mathcal{N}_2)\right]d\ell$$

$$= \sum_{i\in\mathcal{N}_1,j\in\mathcal{N}_2} \int_{\ell=0}^{\infty} m\ell^{m-1}e^{-\lambda_{i,j}\ell}d\ell$$

$$- \sum_{i,u\in\mathcal{N}_1,j,v\in\mathcal{N}_2} \int_{\ell=0}^{\infty} m\ell^{m-1}e^{-(\lambda_{i,j}+\lambda_{u,v})\ell}d\ell$$

$$+ \sum_{i,u,w\in\mathcal{N}_1,j,v,z\in\mathcal{N}_2} \int_{\ell=0}^{\infty} m\ell^{m-1}e^{-(\lambda_{i,j}+\lambda_{u,v}+\lambda_{z,w})\ell}d\ell + \cdots$$

$$= \sum_{i\in\mathcal{N}_1,j\in\mathcal{N}_2} \frac{m}{\lambda_{i,j}^m} - \sum_{i,u\in\mathcal{N}_1,j,v\in\mathcal{N}_2} \frac{m}{(\lambda_{i,j}+\lambda_{u,v})^m}$$

$$+ \sum_{i,u,w\in\mathcal{N}_1,j,v,z\in\mathcal{N}_2} \frac{m}{(\lambda_{i,j}+\lambda_{u,v}+\lambda_{z,w})^m} + \cdots \qquad \text{(Equation A10)}$$

where the integrals in Equation A10 can be evaluated by noting that they are essentially expressions for the moments of exponential random variables with parameters λ_i, $(\lambda_i + \lambda_j)$, $(\lambda_i + \lambda_j + \lambda_k)$, etc.

Thus, we can use Equation A10 to compute

$$E\left[L_{1,2}^m\right] = \sum_{\mathcal{N}_1,\mathcal{N}_1} E_{\mathcal{N}_1,\mathcal{N}_2}\left[L_{1,2}^m\right]\mathbb{P}(\mathcal{N}_1,\mathcal{N}_2), \qquad \text{(Equation A11)}$$

where $\mathbb{P}(\mathcal{N}_1,\mathcal{N}_2)$ is the probability of observing identical-by-descent segments at the leaves $\mathcal{N}_1$ and $\mathcal{N}_2$, and is approximated using the recursion in Equation 6. We then plug Equation A11 in to obtain the variance of $T_{1,2}$ in Equation A9.

In practice, it is too computationally demanding to compute the sums in Equation A11 because the terms $E_{\mathcal{N}_1,\mathcal{N}_2}[L_{1,2}]$ and $E_{\mathcal{N}_1,\mathcal{N}_2}[L_{1,2}^2]$ are not fast to compute in large quantities. However, the probabilities $\mathbb{P}(\mathcal{N}_1,\mathcal{N}_2)$ can be computed quickly enough, allowing us to find the most likely sets of leaf nodes, $\widehat{\mathcal{N}}_1$ and $\widehat{\mathcal{N}}_2$, with observed IBD. Thus, in practice we use an approximation in which we assume that the most likely IBD pattern has been observed and we compute

$$E\left[L_{1,2}^m\right] \approx E_{\widehat{\mathcal{N}}_1,\widehat{\mathcal{N}}_2}\left[L_{1,2}^m\right]. \qquad \text{(Equation A12)}$$

The assumption used in this approximation is that most patterns of observed IBD at the leaves are unlikely compared with the most likely patterns and that most likely patterns of IBD will yield similar moments $E[L_{1,2}^m]$.

Re-rooting the DRUID estimator

In some scenarios, A_2 can be the direct descendant of A_1, or vice versa. This scenario, along with the scenario treated in The generalized DRUID estimator in which $\mathcal{N}_1$ and $\mathcal{N}_2$ are connected through their common ancestors, covers all possible ways in which $\mathcal{N}_1$ and $\mathcal{N}_2$ can be connected such that they are mutually related.

We now describe an approach for computing the generalized DRUID estimate when A_2 is descended from an individual A who is the common ancestor of only a subset of $\mathcal{N}_1$. We consider A to be any node ancestral to some node in $\mathcal{N}_1$, including any member of $\mathcal{N}_1$ itself.

Let $\Lambda_1(A_1)$ denote the induced subtree in pedigree $\mathcal{P}_1$ that relates A_1 and their descendants $\mathcal{N}_1$. To obtain the generalized DRUID estimate when A_2 is descended from A, we re-root the tree $\Lambda_1(A_1)$ at A to obtain a re-rooted

tree $\tilde{\Lambda}_1(A)$ (Figure S9). We then compute the generalized DRUID estimate from The generalized DRUID estimator using the re-rooted tree $\tilde{\Lambda}_1(A)$. The estimate between A and A_2 obtained using Equation 9 applied to $\tilde{\Lambda}_1(A)$ and $\Lambda_2(A_2)$ is then the number of meioses separating A and A_2, except for the considerations described below.

A_2 is descended from both A and a partner A', who may also be an ancestor of one or more of A's genotyped descendants $\mathcal{N}_A \subseteq \mathcal{N}_1$. When A' is also an ancestor of one or more descendants of A, A_2 is more closely related by one degree to $\mathcal{N}_{A'}$ than to the other members of $\mathcal{N}_1$. Moreover, the DRUID estimator must be expanded to consider IBD shared between A_2 and a *pair* of ancestors (A, A'), rather than a single ancestor A.

In this case, denote the probability that a single allele is shared between A and some member of $\mathcal{N}_1$ by ϕ_A and denote the probability that a single allele is shared between A' and some member of N_1 by $\phi_{A'}$. Suppose A_2 shares an allele with either of A or A'. The probability that this allele is shared with an individual in $\mathcal{N}_1$ is then $\phi_A/2 + \phi_{A'}/2$, using the fact that the probability the allele is shared with A or A' is $\mathbb{P}(A) = \mathbb{P}(A') = 1/2$. Thus we obtain

$$\widehat{IBD}((A, A'), A_2) = \frac{IBD(\mathcal{N}_1, \mathcal{N}_2)}{2^{-1}(\phi_A + \phi_{A'})\phi_2}, \qquad \text{(Equation A13)}$$

where ϕ_A is obtained by re-rooting the tree $\Lambda_1(A_1)$ to $\tilde{\Lambda}_1(A)$ and evaluating Equation 7, and where $\phi_{A'}$ is obtained directly from Equation 7 without re-rooting.

If d is the number of meioses separating A_2 from A and A', then the expected amount of IBD shared between A_2 and the tuple (A, A') is 2^{-d+1}. This amount is equivalent to the amount that would be shared between A_2 and A if they were connected by one fewer degree. Treating the tuple (A, A') as a single individual that is connected to A_2 by degree $d - 1$, we obtain a modified DRUID estimator

$$d_D((A, A'), A_2) = d : \frac{1}{2^{d-1+3/2}} \le \frac{IBD(\mathcal{N}_1, \mathcal{N}_2)}{2(\phi_A + \phi_{A'})\phi_2 L_{genome}}$$
$$< \frac{1}{2^{d-1+1/2}}, \qquad \text{(Equation A14)}$$

where the inferred degree d is the number of meioses between A_2 and A.

Data and code availability

The simulated identical-by-descent data generated during this study are available at the bonsaitree github repository (https://github.com/23andMe/bonsaitree). There are restrictions to the availability of genotype and identity-by-descent data due to 23andMe consent and privacy guidelines. These data will not be made available.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.09.013.

Web resources

Bonsai algorithm, https://github.com/23andMe/bonsaitree

References

1. Almudevar, A. (2003). A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theor. Popul. Biol. *63*, 63–75.
2. Almudevar, A., and Anderson, E.C. (2012). A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. Mol. Ecol. Resour. *12*, 164–178.
3. Cowell, R.G. (2009). Efficient maximum likelihood pedigree reconstruction. Theor. Popul. Biol. *76*, 285–291.
4. Cowell, R.G. (2013). A simple greedy algorithm for reconstructing pedigrees. Theor. Popul. Biol. *83*, 55–63.
5. Cussens, J., Bartlett, M., Jones, E.M., and Sheehan, N.A. (2013). Maximum likelihood pedigree reconstruction using integer linear programming. Genet. Epidemiol. *37*, 69–83.

6. Jones, O.R., and Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol. Ecol. Resour. 10, 551–555.

7. Kirkpatrick, B., Li, S.C., Karp, R.M., and Halperin, E. (2011). Pedigree reconstruction using identity by descent. J. Comput. Biol. 18, 1481–1493.

8. Riester, M., Stadler, P.F., and Klemm, K. (2009). FRANz: reconstruction of wild multi-generation pedigrees. Bioinformatics 25, 2134–2139.

9. Sheehan, N.A., Bartlett, M., and Cussens, J. (2014). Improved maximum likelihood reconstruction of complex multi-generational pedigrees. Theor. Popul. Biol. 97, 11–19.

10. Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. Genetics 166, 1963–1979.

11. Anderson, E.C., and Ng, T.C. (2016). Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation. Theor. Popul. Biol. 107, 39–51.

12. Huisman, J. (2017). Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. Mol. Ecol. Resour. 17, 1009–1024.

13. Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., Nickerson, D.A., Below, J.E.; and University of Washington Center for Mendelian Genomics (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. 95, 553–564.

14. Ko, A., and Nielsen, R. (2017). Composite likelihood method for inferring local pedigrees. PLoS Genet. 13, e1006963.

15. Staples, J., Maxwell, E.K., Gosalia, N., Gonzaga-Jauregui, C., Snyder, C., Hawes, A., Penn, J., Ulloa, R., Bai, X., Lopez, A.E., et al. (2018). Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. Am. J. Hum. Genet. 102, 874–889.

16. Staples, J., Witherspoon, D.J., Jorde, L.B., Nickerson, D.A., University of Washington Center for Mendelian Genomics, Below, J.E., and Huff, C.D. (2016). PADRE: Pedigree-aware distant-relationship estimation. Am. J. Hum. Genet. 99, 154-162.

17. Ramstetter, M.D., Shenoy, S.A., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Mezey, J.G., and Williams, A.L. (2018). Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. Am. J. Hum. Genet. 103, 30–44.

18. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

19. Huff, C.D., Witherspoon, D.J., Simonson, T.S., Xing, J., Watkins, W.S., Zhang, Y., Tuohy, T.M., Neklason, D.W., Burt, R.W., Guthery, S.L., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 21, 768–774.

20. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873.

21. Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I., and Mountain, J.L. (2012). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS ONE 7, e34267.

22. Seidman, D.N., Shenoy, S.A., Kim, M., Babu, R., Woods, I.G., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., and Williams, A.L. (2020). Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification. Am. J. Hum. Genet. 106, 453–466.

23. Campbell, C.L., Furlotte, N.A., Eriksson, N., Hinds, D., and Auton, A. (2015). Escape from crossover interference increases with maternal age. Nat. Commun. 6, 6260. https://doi.org/10.1038/ncomms7260.

24. Williams, C.M., Scelza, B., Gignoux, C.R., and Henn, B.M. (2020). A rapid, accurate approach to inferring pedigrees in endogamous populations. bioRxiv. https://doi.org/10.1101/2020.02.25.965376.

25. Freyman, W.A., McManus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., 23andMe Research Team, and Auton, A. (2020). Fast and robust identity-by-descent inference with the templated positional burrows-wheeler transform. Mol. Biol. Evol. 38, 2131-2151.

26. Qiao, Y., Sannerud, J.G., Basu-Roy, S., Hayward, C., and Williams, A.L. (2021). Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps. Am. J. Hum. Genet. 108, 68–83.