ANCESTRAL HAPLOTYPE RECONSTRUCTION IN ENDOGAMOUS POPULATIONS USING IDENTITY-BY-DESCENT

Kelly Finke[1,2] **Michael Kourakos**[1] **Gabriela Brown**[1] **Yuval B. Simons**[3] **Alejandro A. Schäffer**[4]

Rachel L. Kember[5] **Maja Bućan**[5] **Sara Mathieson**[6,†]

ABSTRACT

In this work we develop a novel algorithm for reconstructing the genomes of ancestral individuals, given genotype or sequence data from contemporary individuals and an extended pedigree of family relationships. A pedigree with complete genomes for every individual enables the study of allele frequency dynamics and haplotype diversity across generations, including deviations from neutrality such as transmission distortion. When studying heritable diseases, ancestral haplotypes can be used to augment genome-wide association studies or compute polygenic risk scores for the reconstructed individuals.

The building blocks of our reconstruction algorithm are segments of Identity-By-Descent (IBD) shared between two or more genotyped individuals. The method alternates between finding a source for each IBD segment and assembling IBD segments placed within each ancestral individual. After each iteration we perform conflict resolution to remove IBD segments that do not align with well-reconstructed haplotypes and upweight the probability that these segments should be placed in other individuals. We repeat this process until we are no longer successfully reconstructing additional ancestral haplotypes. Unlike previous approaches, our method is able to accommodate complex pedigree structures with hundreds of individuals genotyped at millions of SNPs.

We apply our method to an Old Order Amish pedigree from Lancaster, Pennsylvania, whose founders came to the United States from Europe during the early 18th century. The pedigree includes 1338 individuals from the past 10 generations, 394 with genotype data. The motivation for reconstruction is to understand the genetic basis of diseases segregating in the family through tracking haplotype transmission over time. Using our algorithm `thread`, we are able to reconstruct an average of 230 ancestral individuals per autosome. `thread` was developed for endogamous populations, but can be applied to any extensive pedigree with the recent generations genotyped. We anticipate that this type of practical ancestral reconstruction will become more common and necessary to understand rare and complex heritable diseases in extended families.

***K*eywords** Ancestral inference · Haplotype reconstruction · Pedigree

[1] Department of Computer Science, Swarthmore College, Swarthmore, PA
[2] Department of Biology, Swarthmore College, Swarthmore, PA
[3] Department of Genetics, Stanford University, Stanford, CA
[4] Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health, Bethesda, MD
[5] Department of Genetics, University of Pennsylvania, Philadelphia, PA
[6] Department of Computer Science, Haverford College, Haverford, PA
[†] Corresponding author: Sara Mathieson, `smathieson@haverford.edu`

1 Introduction

Pedigree structures and associated genetic data provide a wealth of information for studying recent evolution. Nuclear families (parents and children) and other small pedigrees have been used to estimate mutation and recombination rates in humans [6, 8, 27, 44] and other species [24, 41, 46]. Pedigrees have informed breeding of domesticated animals [33], enabled the study of short-term evolution in natural populations [9], and can be used to study heritable diseases [4].

Genetic studies of rare, recessive traits pose a challenge to researchers when individuals expressing these traits are too sparse or too scattered to obtain sufficient genetic data. Endogamous populations with detailed pedigree records provide an important exception. Endogamous populations, defined by the practice of marrying within a social, ethnic, or geographic group, are often characterized by small effective population sizes with limited external admixture. These groups are of great interest to geneticists because a single small population can provide enough data to inform rare trait and rare variant studies with worldwide implications [32, 37]. Endogamous populations are also informative for common disease [16, 45].

Extended pedigrees from endogamous populations provide a valuable system for studying heritable disease, but genetic data is typically limited to recent generations. If genetic information from every individual in the pedigree were available, we would be in a better position to understand the transmission of causal variants throughout the history of the population. More specifically, we often know the disease phenotypes of ancestral individuals, but cannot obtain their genetic information. In these cases, reconstructed haplotypes allow us to augment genome-wide association studies (GWAS), where large sample sizes are essential. In addition, reconstructed genomes would enable the computation of polygenic risk scores (PRS) [25, 48] for ancestral individuals.

Reconstructed ancestral haplotypes also allow us to study genome dynamics over short time scales, including inheritance patterns and haplotype transmission. In populations with large nuclear families, transmission distortion [11, 34] and other deviations from neutrality are particularly visible. Understanding which parts of the genome are over- or under-represented in the recent generations could help us identify forms of deleterious variation. From a theoretical perspective, there has been relatively little work on the question of how much ancestral reconstruction is possible given genetic information from contemporary individuals (example from a small livestock pedigree in Hayes et al. [18]).

Previous work on ancestral reconstruction has typically been applied to small pedigrees with no loops (marriage of close relatives). One of the earliest examples comes from the Lander-Green algorithm [28], which uses a hidden Markov model (HMM) with inheritance vectors as the hidden state and genotypes as the observed variables. Methods such as `SimWalk2` [43] and `Merlin` [2] use descent graphs and sparse gene flow trees (respectively) to extend the idea of likelihood-based computation to larger pedigrees. However, these methods do not perform reconstruction explicitly and also do not handle loops, as tree-based intermediate steps are common to both algorithms. With millions of loci and hundreds of individuals, the runtimes of these methods are prohibitive (see [42] for a runtime overview). Other HMM-based approaches such as `HAPPY` [35], `GAIN` [31], and `RABBIT` [51] reconstruct genome ancestry blocks, but do not tie them to specific individuals. `HAPLORE` [50] quantifies possible ancestral haplotype configurations but does not incorporate recombination, and the Bayesian approach in Fishelson et al. [14] is more suitable for haplotyping.

Lindholm et al. [30] reconstructed ancestral haplotypes for the purpose of identifying regions that contain susceptibility genes for schizophrenia. However, their pedigree was much smaller (with no loops), many fewer markers (450) were used, and several of the reconstruction steps were done by inspection or by hand, which does not scale to our scenario. Another study [22] reconstructed the African haplotype of an African-European individual who migrated to Iceland in 1802 and had 788 descendants, 182 of which were genotyped. However, this scenario is much simpler, as the regions of African ancestry within each descendant were easily identified and all belonged to the same individual.

The problem studied here is different from *pedigree reconstruction*, where genetic information is used to reconstruct (previously unknown) family relationships. See [20, 21, 23, 26, 39, 47] for discussions of pedigree reconstruction.

In this study we apply our method to an Old Order Amish population from Lancaster, Pennsylvania who can trace their ancestry to founders who came from Europe to Philadelphia in the early 18th century (see Figure 3 of [29] for an analysis of the contributions of the 554 founders). The Amish are an ethno-religious group in the Anabaptist tradition, with a history of detailed record keeping and marriage within the Amish community [13]. In this work, we study an unpublished pedigree of 1338 individuals, augmented [3] from a pedigree of 784 individuals originally described in the Amish Study of Major Affective Disorder [10, 15]. Roughly one third of the individuals in the original pedigree display some form of mood disorder, and about 19% have been diagnosed with bipolar disorder specifically [25]. Bipolar disorder in a broad sense is roughly 80% heritable in this pedigree [25], and recent work has focused on understanding the genetic basis of this disease [15]. The availability of genetic data from 394 contemporary individuals from this pedigree gives us an opportunity to use reconstruction as another lens on inheritance patterns of mood disorders.

Here we present a novel algorithm, `thread`, for reconstructing ancestral haplotypes given an arbitrary pedigree structure and genotyped or sequenced individuals from the recent generations. `thread` can be applied in a variety of scenarios including pedigrees with loops, inter-generational marriage, and remarriage. More ancestral chromosomes will be reconstructed as the percentage of individuals with genetic data increases, but our method can be applied even when this fraction is modest. This work represents a key step towards understanding the limits of quantifying the genomes of ancestral individuals in the absence of ancient DNA. `thread` is available as an open-source software package: `https://github.com/mathiesonlab/thread`.

2 Methods

Problem statement: The first input to `thread` is a pedigree structure $\mathcal{P}$. For each individual $p \in \mathcal{P}$, we have information about the mother $p^{(m)}$ and father $p^{(f)}$, which are also members of $\mathcal{P}$. In the case of founders or married-in individuals, we let $p^{(m)}$ and $p^{(f)}$ be 0. The pedigree may contain loops, meaning that the parents of a child share a recent common ancestor. The second input is a dataset of phased haplotypes (e.g. in Variant Call Format, VCF) from a subset of individuals in the pedigree, typically from the most recent generations. Phasing assigns the alleles of each individual to parental haplotypes. Our aim is to reconstruct the haplotypes of as many ancestral individuals in the pedigree as possible. An illustration of the problem is shown in Figure 1.

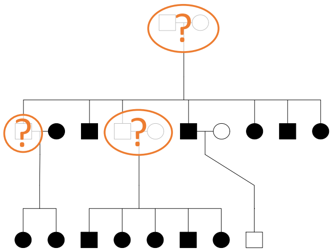

Figure 1: *Problem statement illustration. Squares represent males and circles represent females. Horizontal lines create couples and show sibling relationships. Parents and offspring are connected by vertical lines. Filled in symbols represent individuals who have been genotyped. Our aim is to reconstruct all ungenotyped individuals (orange question marks) who have genotyped descendants.*

High level description: `thread` is built upon the idea of Identity-By-Descent (IBD). IBD segments are long stretches of DNA shared by a *cohort* of two or more individuals due to descent from a common ancestor (*source*). Each segment is analyzed independently (as opposed to working sequentially along the chromosome as an HMM would). We attempt to find the source of each IBD segment, as well as individuals who are on descendance paths from this ancestor to the cohort. After this step we proceed through each individual, clustering and assembling their associated IBD segments into haplotypes. During this grouping step we identify IBD segments that have been poorly placed – in the next iteration we will update their common ancestors. We alternate the process of analyzing IBD segments and individuals until we are no longer building new haplotypes. A schematic of `thread` is shown in Figure 2, and pseudocode is given in Algorithm S1 (Supplementary Material).

Input pedigree: The Amish pedigree under study was developed from several sources, including the book *Descendants of Christian Fisher* [5], the Anabaptist Genealogy Database (AGDB) [3] and associated software `PedHunter` [29], and the Amish Study of Major Affective Disorder [10]. The AGDB is covered by an IRB-approved protocol at the NIH. All work contained within this study was approved by the IRB of the Perelman School of Medicine at the University of Pennsylvania. The complete pedigree structure is shown in Figure S1 (created with the `kinship2` R package [40]).

Step 1: We first read in the pedigree structure. We do not require that individuals be separated into generations, and we allow inter-generational marriage and loops. Let t be the total number of individuals in the pedigree (here $t = 1338$), and n be the number of genotyped individuals (here $n = 394$). We further define m to be the number of ungenotyped

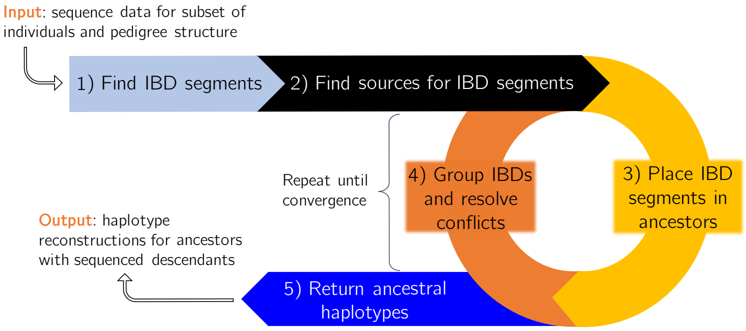
Input: sequence data for subset of individuals and pedigree structure
1) Find IBD segments
2) Find sources for IBD segments
Repeat until convergence
4) Group IBDs and resolve conflicts
3) Place IBD segments in ancestors
Output: haplotype reconstructions for ancestors with sequenced descendants
5) Return ancestral haplotypes

Figure 2: *Algorithm overview. In the first two steps we identify IBD segments and find a list of potential sources for each one. In the iterative phase, we alternate between choosing sources for each IBD and grouping the IBDs that are placed within each individual. IBD segments that conflict with strong haplotypes are rejected and must be assigned a different source. When we are no longer building more haplotypes, we return the reconstructed chromosomes.*

individuals with genotyped descendants (here $m = 686$). In our case, this leaves 258 individuals with no genotyped descendants; we do not expect to be able to reconstruct these individuals.

Genotypes for each genotyped individual were obtained from Illumina Omni 2.5M SNP arrays, and then phased into haplotypes using `SHAPEIT2` [12]. We identify IBD segments between pairs of genotyped individuals using `GERMLINE` [17], although `IBD-Groupon` for detecting IBD in groups could be used instead [19]. For each IBD segment I, we combine pairs until we obtain a cohort of individuals who share this segment, $C \in \{2, n\}$. Here, the size of C ranged from two to 180 individuals. The *descendance path* of an IBD segment includes all descendants of the source who also passed down the IBD to reach the cohort descendants. Table 1 shows the number of unique IBD segments found on each chromosome.

Step 2: In the next phase of `thread`, sources for each IBD segment are identified independently. By the end of this step we will have enumerated all possible individuals who could have been the source of each IBD segment I, given its associated cohort C. This process is done only once and is not part of the iterative phase. When searching for *all* common ancestors of a cohort, each previous generation doubles the number of ancestors to search. `thread` maximizes efficiency in this exponential problem by merging overlapping paths using a modified breadth-first search algorithm (explained in detail below and in pseudocode in Algorithm S2).

First all the individuals in the cohort are added to a queue. For example, in Figure 3,

$$C = \{1, 2, 5, 7, 8\}$$

so we would start out with $Q = (1, 2, 5, 7, 8)$. We then pop the first individual off the queue, p_0. If p_0 is an ancestor of all individuals in the cohort, we add p_0 to a set of possible sources. Either way, we add p_0's parents to the back of the queue and keep processing individuals (even if p_0 is an ancestor, its parents may be ancestors via paths that do not include p_0). In this example we would consider individual 1 first. Since it has not been processed, we add its parents:

$$Q = (2, 5, 7, 8, a, b).$$

Each time we add an individual p to the queue, we keep track of how many paths exist from p to the members of the cohort, using a multiset M_p. For the members of the cohort, $M_p = \{p\}$ (just one path to themselves). When we add a parent to the queue, we concatenate the multisets of the individual's children. For individual a in this example, its multiset would become $M_a = \{1, 2\}$, indicating one path to individual 1 and one path to individual 2. Going further up the pedigree, individual ℓ has two children, h and e with $M_h = \{1, 2, 5, 7, 8\}$ and $M_e = \{5\}$. Concatenating these two multisets, we obtain the multiset $M_\ell = \{1, 2, 5, 5, 7, 8\}$, indicating that there are two possible paths from ℓ to cohort member 5. As soon as an individual's multiset contains all members of the cohort, the individual can be a source.

There are two post-processing phases to the source-finding algorithm. (1) We trim redundant sources: a redundant source is an ancestor of another source without adding any unique descendance paths. In other words, we do not want to include individuals if *all* their paths to the cohort go through another source. If the cardinality of an individual's multiset is not greater than the maximum cardinality of the multisets of its children, it is redundant (for example, k is a redundant ancestor since $|M_k| = |M_h|$). (2) We merge couples into a single source, as typically we will not be able to resolve the source of an IBD segment beyond the couple level. Spouses with different multiset cardinality (usually caused by remarriage) are an exception. Individual ℓ is an example; we do not consider couple $k\ell$ a source because $|M_k| < |M_\ell|$ due to ℓ's remarriage to m. If the cardinalities had been the same (and not redundant), we would have considered $k\ell$ a source.

In the Figure 3 example, we identify three potential sources: $S = \{gh, \ell, pq\}$. Note that we cannot stop processing the queue when we get to source gh, as there exist sources further up the pedigree that contain unique paths.

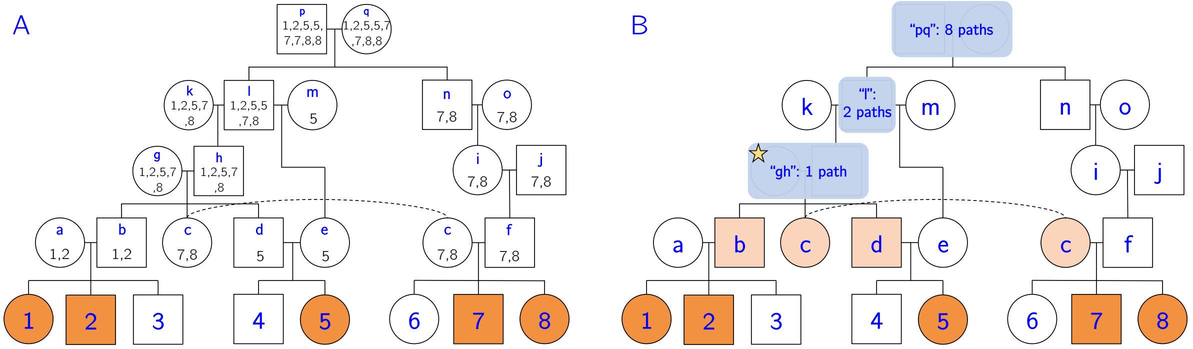

Figure 3: *Source-finding illustration. A) Let individuals 1–8 be the genotyped individuals of this pedigree. Let* $C = \{1, 2, 5, 7, 8\}$ *(orange individuals) be the cohort sharing IBD segment* I*. Note that this pedigree contains a loop, since* c *and* f *share recent ancestors. The multiset* M_p *for each ancestral individual* p *is shown below the node name.* M_p *is formed by concatenating the multisets of* p*'s children, and it represents the number of paths from ancestor* p *to each member of the cohort. B) After trimming redundant ancestors and merging couples, we obtain a set of putative sources for the IBD segment. In this case, we have three potential sources:* $S = \{gh, \ell, pq\}$*. We begin the iterative phase by selecting the source with the fewest descendance paths, which in this case is* gh *(starred). We place the IBD segment in individuals that are on all paths from* gh *to the cohort. In this case we would add the IBD segment to individuals* b*,* c*, and* d *(light orange).*

The use of multisets allows us to quickly determine the number of descendance paths from each source to the cohort. For each source s and each individual c in the cohort, let $m_s(c)$ be the multiplicity of c in M_s. For example, in M_ℓ, the multiplicity of individual 5 is two, meaning that there are two paths from ℓ to individual 5. The total number of descendance paths (d) from source s to cohort C (sharing IBD I) is the product of all the multiplicities:

$$d(s) = \prod_{c \in C} m_s(c)$$

In this example, we obtain $d(gh) = 1$, $d(\ell) = 2$, and $d(pq) = 8$. A few of these descendance paths are shown in blue in Figure 4 for clarity.

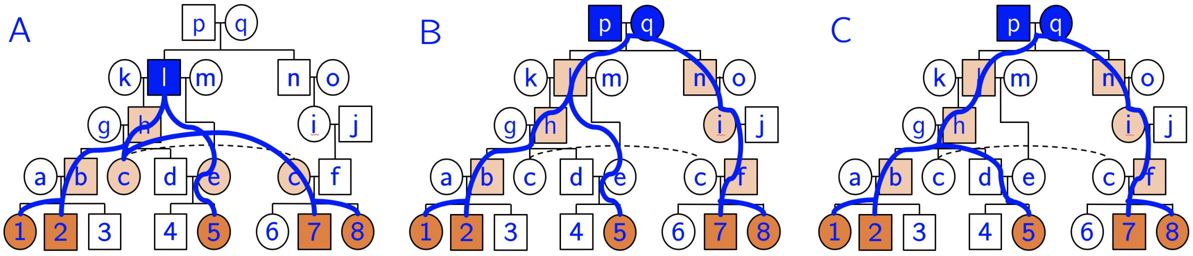

Figure 4: *Example descendance paths. Given a cohort of five individuals sharing an IBD segment (orange), we often obtain multiple sources (blue nodes) and multiple descendance paths (blue lines) from each source. In this example we have 11 total paths from three sources. After we choose a source, we assign the IBD segment to ancestors along* all *descendance paths (light orange). A) One path from source* ℓ*. B-C) Two different descendance paths from the same source* pq*. We do not assign the IBD to* d *and* e *since they are not on all paths from this source.*

Before moving into the iterative part of the algorithm, we take note of individuals that are on *all paths from all sources*. For example, individual b happens to be on all 11 paths from the sources, so we know that individual b should have the IBD segment.

Step 3: At this stage we begin the iterative part of the algorithm. Every iteration begins with lists of *reconstructed* individuals and *unreconstructed* individuals. During the first iteration, the *reconstructed* list only includes genotyped individuals. The goal of Step 3 is to select a source for each IBD segment out of the potential sources enumerated in Step 2. We use the greedy approach of choosing the source with the fewest paths, provided that it does not conflict with one of the *reconstructed* individuals. The intuition behind choosing the source with the fewest paths is that this source will (often) be more recent than others, with fewer meioses separating the source from the cohort. For example, in Figure 3, we would choose source gh since it has only one descendance path. Once we select a source, we can begin to look at the individuals that lie on paths from this source. In the case of only one path, all the individuals on the path will be given the IBD segment (b, c, and d in this example), thus augmenting the associated cohort. In the more common situation when we have multiple paths from the source, we give the IBD segment only to individuals that appear on *all* the paths. However, if we try to give this IBD segment to a *reconstructed* individual and it conflicts with both their haplotypes, we reject the source and immediately choose the source with the next fewest paths. These tentative assignments result in potentially conflicting IBDs being assigned to the same individual, which we resolve in Step 4.

Step 4: During Step 3, we analyzed each IBD segment independently, identifying ancestral individuals who likely also share the IBD segment. In Step 4, we analyze the *individuals* independently and assemble the IBDs that have been placed within the individual. Say we are analyzing ancestral individual p with putative set of IBD segments $\mathcal{I}_p$. The goal of assembly is to separate the IBD segments into two haplotypes such that their sequences are consistent within each group. At a high level, this process can be compared to *de novo* genome assembly, where many small reads are stitched together to create contigs and chromosomes. However, we may have misplaced IBD segments, which we will need to identify and remove.

Our grouping algorithm (covered in pseudocode in Algorithm S3) begins by identifying regions of homozygosity within the IBD segments. This is accomplished by condensing all segments in $\mathcal{I}_p$ down into a single sequence with a list of alleles at each site. Any region greater than 300kb with only one allele per site and at least 100 SNPs is declared homozygous. It is important to identify these regions early in the grouping algorithm, otherwise we may assume only one group shares this stretch. Each homozygous region is duplicated so that each chromosome will have a copy, and IBD segments contained within homozygous regions are not used in the next stages.

We process the remaining IBDs (those not incorporated into a group) one by one, from longest to shortest (in kbp). If the IBD does not overlap with any of the current groups, we create a new group initialized by the IBD segment. If the IBD does overlap with one or more groups, we add it to the group with the largest overlap (above a threshold).

At this point in the grouping algorithm, we have a set of homozygous groups, a set of heterozygous groups, and a set of remaining IBDs. If an IBD overlaps two groups, we use it to merge these groups into one. Finally, we merge groups that “line up” with each other – i.e. they do not overlap, but their IBD segments span adjacent SNPs and were likely separated by an ancestral recombination event. At the end of this process, three situations may emerge:

- We have two clear groups (which we denote as *strong*) forming two haplotypes. This is the ideal scenario and it means we have a successful reconstruction of the individual. To determine if a group is *strong*, it must meet a combination of thresholds: a minimum number of IBD segments and a minimum coverage (#SNPs reconstructed/#SNPs genotyped on the chromosome). We use a sliding scale: if the group contains 1-2 IBDs, it must cover 90% of the SNPs. If a group contains 3-9 IBDs, it must cover 70% of the SNPs. And if a group contains 10 or more IBDs, it must cover 50% of the SNPs. These parameters can be customized by the user.
- We have two strong groups, but we also have several weaker ones. This scenario is resolvable, as we can retain the two strong groups as the reconstruction, and reject the other groups. Specifically, the two best groups must meet our *strong* threshold and the rank three group must either have half as many IBD segments or be half as long. The IBD segments from the rejected groups give us a lot of information – since this individual was on *all* paths from the selected source, if the IBD segment does not fit with the reconstructed haplotypes, then we

know the source was incorrect. Throughout Step 4 we collect all IBD segments that have been incorrectly sourced to update in the next iteration.

- In all other situations, we typically cannot resolve the individual's haplotypes. We may have only one group (which could be one of the individual's haplotypes), but we do not declare the individual reconstructed. We could have many groups without two strong ones, or we many not have given the individual any IBDs to group.

At the end of Step 4, we move individuals from the first two scenarios in to the *reconstructed* list. IBDs that did not cause any conflicts are marked as processed and we retain the rest to re-source in the next iteration. An illustration of the grouping algorithm is shown in Figure 5.

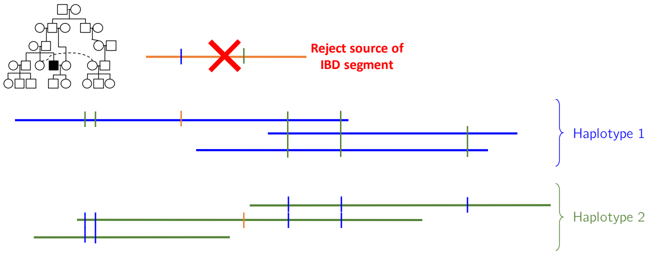

Figure 5: *Grouping algorithm illustration. Each horizontal line represents one IBD segment that we placed within a specific individual (highlighted in the pedigree inset). Each vertical line indicates a difference (heterozygous site) between groups. In this case, the orange IBD segment conflicts with both the blue and green groups, so we would reject its source and attempt to find a new one in the next iteration.*

Iteration and Step 5: At the end of Step 4 we have a set of IBD segments that were incorrectly sourced. We then repeat Step 3: we update the source for each such IBD by selecting the source with the next fewest paths. This allows us to assign the IBD to a new set of individuals. In the next Step 4 we treat reconstructed individuals and unreconstructed individuals differently. If an individual is already marked as reconstructed, we use each additional IBD to strengthen its groups or reject the new source of the IBD. If an individual has not been reconstructed, we run the grouping algorithm again. We keep iterating Steps 3 and 4 until we are no longer reconstructing new individuals.

The final step is to return the haplotype sequences for the reconstructed individuals. These may contain some gaps, but due to our coverage and length thresholds, if an individual is declared reconstructed, we will return at least half of each haplotype (for the chromosome under consideration).

Simulations: To validate our method, we simulate genetic data from an endogamous population. To generate the levels of IBD sharing seen in the Amish population, we first simulate marriage and offspring between individuals who share a common ancestor three generations in the past. For the founder genomes of these small pedigrees we use haplotypes drawn from European individuals (CEU from the 1000 Genomes Project [1]). This process simulates endogamy pre-immigration to the United States. Then we use these composite individuals as founders and feed them through our exact pedigree structure (of 1338 individuals), simulating meiosis and recombination from a human genetic map (chr 21). We record the genomes of all individuals in this simulated system, but only use the same 394 genotyped individuals when we run `thread`. After reconstruction, we compare the genomes we built to the true underlying genomes (accounting for arbitrary haplotype order).

3 Results

In our validation, we compared the true genomes from our CEU simulations to those reconstructed by `thread`. In the parts we reconstruct, we often see sequence similarity that is either close to 100% or around 70%. On average we see about 84% sequence similarity with the true haplotypes – symmetries between maternal and paternal lineages in the pedigree structure may account for part of the discrepancy. In the simulations we reconstructed 107 individuals, which is lower than for the real Amish data.

Moving to the real data, we began by testing the grouping algorithm on genotyped individuals. Figure 6 shows two chromosomes of a *genotyped* individual that were reconstructed using `thread`. Each horizontal line represents one IBD segment shared with a cohort of other genotyped individuals. IBD segments of the same color represent haplotypes, and have a consistent sequence along the chromosome. For example, if we condensed the orange IBD segments in Figure 6B, a single sequence would emerge. The small vertical lines represent heterozygous sites between the two haplotypes.

In general we found that our grouping algorithm worked very well for genotyped individuals, who typically share many IBD segments with other members of the pedigree. Very occasionally we obtained three groups (example in Figure 6A).

Next we investigated the number of sources per IBD segment and the number of descendance paths per source. These distributions are shown in Figure S4 for chromosome 21.

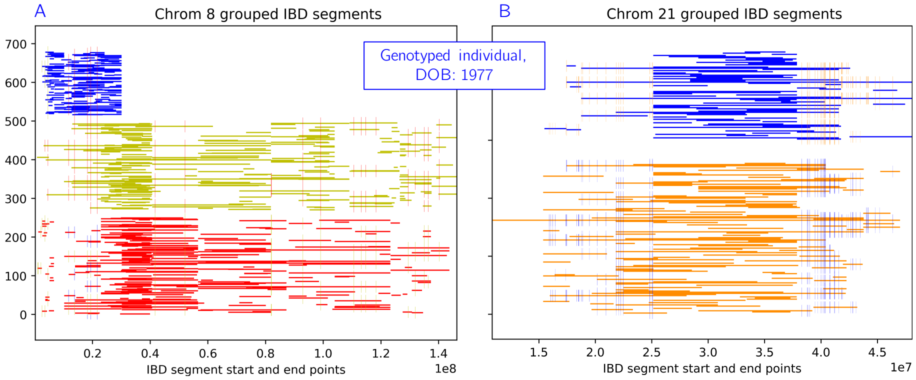

Figure 6: *Example of the grouping algorithm on a genotyped individual. Each horizontal line represents one IBD segment shared with a cohort of other genotyped individuals. IBD segments of the same color represent haplotypes, and have a consistent sequence along the chromosome. Small vertical lines represent heterozygous sites between the two haplotypes. A) Chrom 8: very occasionally we merge groups incorrectly and obtain three groups. B) Chrom 21: we almost always see two clear haplotypes (here we also see a large stretch of homozygosity).*

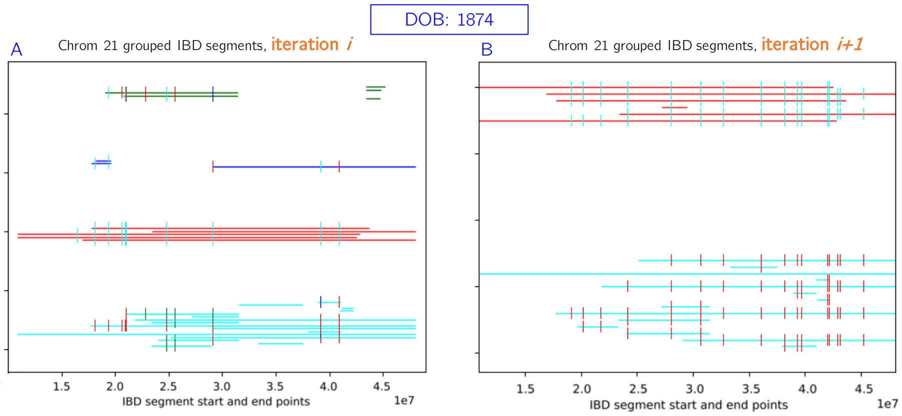

Figure 7: *Conflict resolution example. The blue and green groups are removed, since they are less* strong *than the cyan and red groups. In the next iteration, we retain only strong groups and consider the individual reconstructed. Newly sourced IBDs after this point may not conflict with these reconstructed haplotypes.*

After running `thread` on each autosome using the entire pedigree and all genotyped individuals, we assessed the results in terms of how many individuals were successfully reconstructed (based on the criteria in Section 2). This means that at least half the chromosome can be constructed, with sufficient support in terms of coverage and number of IBD segments. Typically `thread` converged in 6-10 iterations and we were able to reconstruct between 166 and 260 individuals per chromosome (24%-38% of the 686 individuals with genotyped descendants). See Table 1 for the details of each chromosome.

The conflict resolution step was essential for removing misplaced IBD segments and routing them to other sources. An example is shown in Figure 7. In this case, the green and blue groups were removed from this individual, as they were much less *strong* than the cyan and red groups. In the next iteration, we re-source the associated IBDs and consider the individual reconstructed. Examples of successful ancestral reconstructions are shown in Figure 9, for a variety of different chromosomes and generations back in time. As expected, in the more distant generations, we place fewer IBD segments and generally have less coverage over the chromosome.

Although we reconstruct many individuals well in the recent generations, there are many haplotypes we are unable to resolve. A few examples are shown in Figure S3. Sometimes we build one haplotype successfully, but not the other (Figure S3A). Often we have some successful reconstruction, but the groups do not meet our threshold for "two strong" since the third group has too many IBD segments (Figure S3B). Four haplotypes could represent ambiguity between the individual's spouse or close relative (Figure S3C). Sometimes we are placing too many IBDs in this individual, which could arise if they have many descendants (Figure S3D).

Table 1 and Figure 8 show our results in a wholistic view. Table 1 shows how many individuals we are successfully reconstructing for each chromosome. Figure 8 shows these same results on the family level, broadly indicating which individuals we are reconstructing well. Figure S2 shows these results on the individual level.

Table 1: *Whole-genome ancestral reconstruction results. The second column shows the number of unique IBD segments per chromosome. Each IBD segment is shared with a cohort of 2-180 individuals. The third column shows how many iterations the algorithm needed to converge. The fourth column shows the number of ancestral (ungenotyped) individuals we were able to successfully reconstruct. We require a successfully reconstructed chromosome to have two haplotypes that cover at least half the chromosome, with sufficient IBD support for each haplotype. Finally, the last column shows the runtime in hours.*

chrom	# unique IBDs	# iter	# reconstructed	time (hrs)
1	28359	7	253	28.61
2	26962	7	260	28.00
3	22488	8	254	17.10
4	20980	6	254	13.83
5	19448	8	233	13.62
6	20883	7	242	16.57
7	19370	6	245	9.93
8	16950	8	242	9.34
9	17547	6	244	6.61
10	16822	8	231	9.72
11	15416	6	234	5.03
12	16712	6	229	7.29
13	13296	7	215	3.26
14	11867	7	234	2.51
15	12179	7	233	2.58
16	13010	10	245	3.30
17	11768	6	166	2.69
18	11359	8	228	2.35
19	10702	6	213	1.52
20	9910	7	219	1.73
21	5020	7	204	0.74
22	5773	9	171	0.95

4 Discussion

The methodology behind `thread` represents a new direction for ancestral reconstruction that scales in both the number of individuals and the number of loci. Previous ancestral haplotype reconstruction algorithms have either been too slow to apply, too rigid to accommodate a complex pedigree, perform steps by hand, or consider a more diverse ancestral population. Although a likelihood approach to reconstruction is theoretically possible, our work represents a practical alternative as pedigree size and complexity continues to grow. We note that our method is most suitable when genotyped individuals exhibit high levels of IBD sharing. As effective population size and/or admixture levels increase, this type of method will become less useful.

There are many possible algorithmic improvements to our method. In particular, choosing the source with the fewest paths may bias us toward poor reconstructions in some situations. A more robust probabilistic approach might take other aspects into account, including: (1) the number of generations separating the cohort and the ancestor, (2) the length of the IBD segment, and (3) the location of the IBD segment on the chromosome. Due to recombination events at each generation, all of these factors affect the likelihood that an IBD is passed down, intact, from a certain ancestor to the descendants. In terms of implementation, `thread` could be parallelized across IBD segments and individuals.

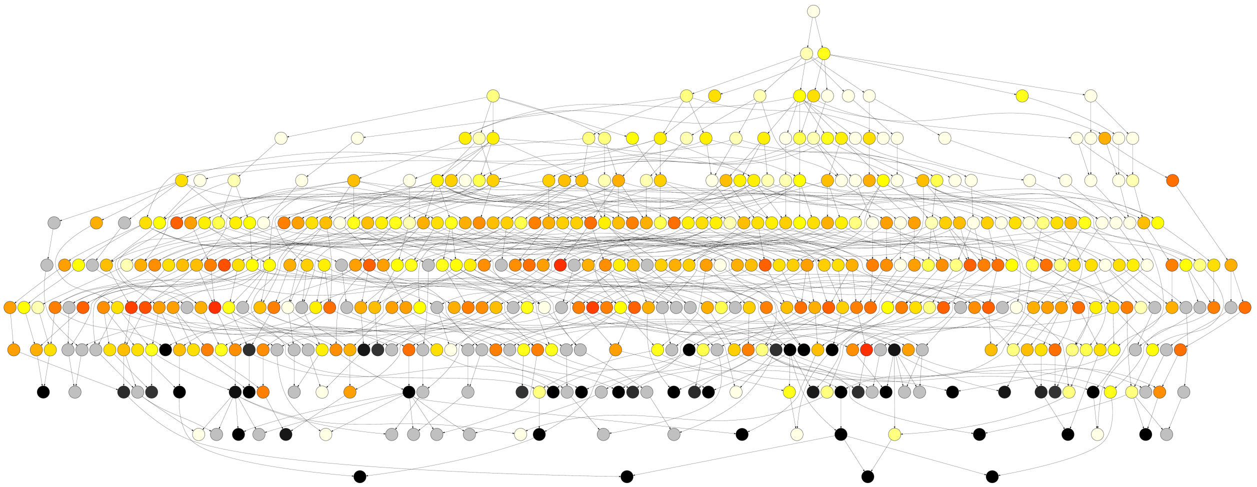

Figure 8: *Nuclear family graph. Each node represents a nuclear family (parents and children). When a child of one family becomes the parent of another, we draw an edge. Black nodes have at least 80% of the family genotyped. Gray nodes have at least 80% of the family without genotyped descendants. Yellow (fewer) – Red (more) node colors represent the average number of chromosomes reconstructed for the individuals in the family.*

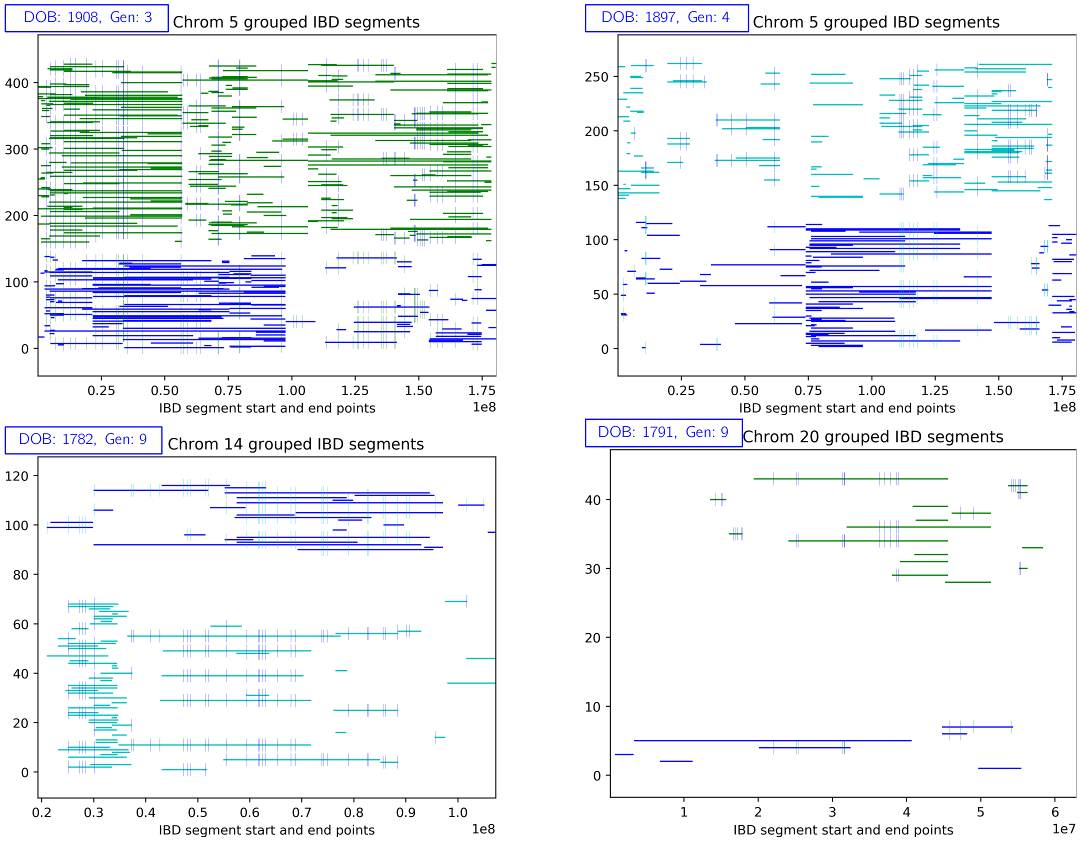

Figure 9: *Successful ancestral reconstructions of ungenotyped individuals, from a variety of chromosomes and generations (back in time). As we go back in time, we generally have fewer IBD segments to group.*

The grouping algorithm could make use of the genetic map to merge groups at recombination hotspots. More realistic simulations could model crossover interference and sex-specific recombination maps, as in Caballero et al. [7].

Individual-level reconstruction opens the door for many types of downstream analysis. Using reconstructed genomes to augment GWAS could increase sample sizes by hundreds of individuals when the phenotype is known. More generally, quantifying allele frequency changes, transmission distortion, and un-reconstructable ("lost") regions of the genome allows us to model genome dynamics on a recent time scale. `thread` could be applied to other genetically characterized endogamous populations with high levels of recessive traits, such as Mennonites and Hutterites [36]. Our method would also be suitable for model organisms and domestic animals, where extensive pedigree records are common.

Our results could also be used to find individuals of clinical significance in cases where a gene-inhibiting drug may provide a therapeutic option for a disease. More specifically, loss of function (LoF) mutations in some genes have shown to protect against disease [38, 49]. As gene inhibition as not been extensively studied in humans, identifying individuals who are already heterozygous null or homozygous null could be extremely valuable.

Acknowledgments

The authors would like to thank Jeff Knerr for providing invaluable computational support, Amy Williams for guidance with simulations, Iain Mathieson for feedback on the method and manuscript, and the participants in the Amish Study of Major Affective Disorder (ASMAD). This research is supported in part by the National Institutes of Health, National Cancer Institute.

References

[1] 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*, **526,**(7571) 68–74.

[2] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. 2002. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30,** 97–101.

[3] Agarwala, R., Schäffer, A. A., and Tomlin, J. F. 2001. Towards a complete North American Anabaptist genealogy II: Analysis of inbreeding. *Human Biology*, **73,**(4) 533–545.

[4] Beaty, T., Kwiterovich Jr, P., Khoury, M., White, S., Bachorik, P., Smith, H., Teng, B., and Sniderman, A. 1986. Genetic analysis of plasma sitosterol, apoprotein B, and lipoproteins in a large Amish pedigree with sitosterolemia. *American Journal of Human Genetics*, **38,**(4) 492.

[5] Beiler, K. *Descendants of Christian Fisher*. Pequea, 4th edition, 2009.

[6] Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *The American Journal of Human Genetics*, **63,**(3) 861–869.

[7] Caballero, M., Seidman, D. N., Sannerud, J., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Carmi, S., and Williams, A. L. 2019. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *bioRxiv*, page 527655.

[8] Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O'Roak, B. J., Sudmant, P. H., Shendure, J., et al. 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, **44,**(11) 1277.

[9] Chen, N., Juric, I., Cosgrove, E. J., Bowman, R., Fitzpatrick, J. W., Schoech, S. J., Clarke, A. G., and Coop, G. 2019. Allele frequency dynamics in a pedigreed natural population. *PNAS*, **116,**(6) 2158–2164.

[10] Coriell Institute for Medical Research. Amish major affective disorders, 2019. URL `https://www.coriell.org/`.

[11] de Villena, F. P.-M. and Sapienza, C. 2001. Nonrandom segregation during meiosis: the unfairness of females. *Mammalian Genome*, **12,**(5) 331–339.

[12] Delaneau, O., Zagury, J.-F., and Marchini, J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, **10,** 5–6.

[13] Elizabethtown College. Young center for Anabaptist and Pietist studies, 2019. URL `https://www.etown.edu/centers/young-center/`.

[14] Fishelson, M., Dovgolevsky, N., and Geiger, D. 2005. Maximum likelihood haplotyping for general pedigrees. *Human Heredity*, **59,**(1) 41–60.

[15] Georgi, B., Craig, D., Kember, R. L., Liu, W., Lindquist, I., Nasser, S., Brown, C., Egeland, J. A., Paul, S. M., and Bućan, M. 2014. Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genetics*, **10,**(3) e1004229.

[16] Grant, S. F., Thorleifsson, G., Frigge, M. L., Thorsteinsson, J., Gunnlaugsdóttir, B., Geirsson, Á. J., Gudmundsson, M., Vikingsson, A., Erlendsson, K., Valsson, J., et al. 2001. The inheritance of rheumatoid arthritis in Iceland. *Arthritis & Rheumatism*, **44,**(10) 2247–2254.

[17] Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, **19,** 318–326.

[18] Hayes, B. J., Lewin, H. A., and Goddard, M. E. 2013. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*, **29,**(4) 206–214.

[19] He, D. 2013. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise ibd relationships. *Bioinformatics*, **29,**(13) i162–i170.

[20] He, D., Wang, Z., Han, B., Parida, L., and Eskin, E. 2013. IPED: inheritance path-based pedigree reconstruction algorithm using genotype data. *Journal of Computational Biology*, **20,**(10) 780–791.

[21] He, D., Wang, Z., Parida, L., and Eskin, E. 2017. IPED2: Inheritance path based pedigree reconstruction algorithm for complicated pedigrees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **14,**(5) 1094–1103.

[22] Jagadeesan, A., Gunnarsdóttir, E. D., Ebenesersdóttir, S. S., Guðmundsdóttir, V. B., Thordardottir, E. L., Einarsdóttir, M. S., Jónsson, H., Dugoujon, J.-M., Fortes-Lima, C., Migot-Nabias, F., Massougbodji, A., Bellis, G., Pereira, L., Másson, G., Kong, A., Stefánsson, K., and Helgason, A. 2018. Reconstructing an African haploid genome from the 18th century. *Nature Genetics*, **50,** 199–205.

[23] Jones, O. R. and Wang, J. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular ecology resources*, **10,**(3) 551–555.

[24] Keightley, P. D., Ness, R. W., Halligan, D. L., and Haddrill, P. R. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics*, **196,**(1) 313–320.

[25] Kember, R. L., Hou, L., Ji, X., Andersen, L. H., Ghorai, A., Estrella, L. N., Almasy, L., McMahon, F. J., Brown, C., and Bućan, M. 2018. Genetic pleiotropy between mood disorders, metabolic, and endocrine traits in a multigenerational pedigree. *Translational Psychiatry*, **8,**(218) 1–12.

[26] Kirkpatrick, B., Li, S. C., Karp, R. M., and Halperin, E. 2011. Pedigree reconstruction using identity by descent. *Journal of Computational Biology*, **18,**(11) 1481–1493.

[27] Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nature Genetics*, **31,**(3) 241.

[28] Lander, E. S. and Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *PNAS*, **84,** 2363–2367.

[29] Lee, W.-J., Pollin, T. I., O'Connell, J. R., Agarwala, R., and Schäffer, A. A. 2010. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. *BMC Medical Genetics*, **11,**(68).

[30] Lindholm, E., Åberg, K., Ekholm, B., Pettersson, U., Adolfsson, R., and Jazin, E. E. 2004. Reconstruction of ancestral haplotypes in a 12-generation schizophrenia pedigree. *Psychiatric Genetics*, **14,** 1–8.

[31] Liu, E. Y., Zhang, Q., McMillan, L., de Villena, F. P.-M., and Wang, W. 2010. Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics*, **26,**(12) i199–i207.

[32] Locke, A. E., Steinberg, K. M., Chiang, C. W. K., Service, S. K., Havulinna, A. S., Stell, L., Pirinen, M., Abel, H. J., Chiang, C. C., Fulton, R. S., Jackson, A. U., Kang, C. J., Kanchi, K. L., Koboldt, D. C., Larson, D. E., Nelson, J., Nicholas, T. J., Pietilä, A., Ramensky, V., Ray, D., Scott, L. J., Stringham, H. M., Vangipurapu, J., Welch, R., Yajnik, P., Yin, X., Eriksson, J. G., Ala-Korpela, M., Järvelin, M.-R., Männikkö, M., Laivuori, H., Project, F., Dutcher, S. K., Stitziel, N. O., Wilson, R. K., Hall, I. M., Sabatti, C., Palotie, A., Salomaa, V., Laakso, M., Ripatti, S., Boehnke, M., and Freimer, N. B. 2019. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*, **572,** 323–328.

[33] Mc Parland, S., Kearney, J., Rath, M., and Berry, D. 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. *Journal of Animal Science*, **85,**(2) 322–331.

[34] Meyer, W. K., Arbeithuber, B., Ober, C., Ebner, T., Tiemann-Boege, I., Hudson, R. R., and Przeworski, M. 2012. Evaluating the evidence for transmission distortion in human pedigrees. *Genetics*, **191,**(1) 215–232.

[35] Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C., and Flint, J. 2000. A method for fine mapping quantitative trait loci in outbred animal stocks. *PNAS*, **97,**(23) 12649–12654.

[36] Payne, M., Rupar, C. A., Siu, G. M., and Siu, V. M. 2011. Amish, Mennonite, and Hutterite genetic disorder database. *Paediatrics and Child Health*, **16,**(3) e23.

[37] Peltonen, L., Palotie, A., and Lange, K. 2000. Use of population isolates for mapping complex traits. *Nature Reviews Genetics*, **1,** 182–190.

[38] Pollin, T. I., Damcott, C. M., Shen, H., Ott, S. H., Shelton, J., Horenstein, R. B., Post, W., McLenithan, J. C., Bielak, L. F., Peyser, P. A., et al. 2008. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science*, **322,**(5908) 1702–1705.

[39] Sheikh, S. I., Berger-Wolf, T. Y., Khokhar, A. A., Caballero, I. C., Ashley, M. V., Chaovalitwongse, W., Chou, C.-A., and DasGupta, B. 2010. Combinatorial reconstruction of half-sibling groups from microsatellite data. *Journal of Bioinformatics and Computational Biology*, **8,**(02) 337–356.

[40] Sinnwell, J., Therneau, T., Schaid, D., Atkinson, E., and Mester, C. kinship2: Pedigree functions, 2019. URL `https://CRAN.R-project.org/package=kinship2`.

[41] Smeds, L., Mugal, C. F., Qvarnström, A., and Ellegren, H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS genetics*, **12,**(5) e1006044.

[42] Sobel, E. SimWalk2: Overview, 2004. URL `https://watson.hgen.pitt.edu/docs/simwalk2.html`.

[43] Sobel, E. and Lange, K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58,**(6) 1323–1337.

[44] Sun, J. X., Helgason, A., Masson, G., Ebenesersdóttir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. 2012. A direct characterization of human mutation based on microsatellites. *Nature Genetics*, **44,**(10) 1161.

[45] Sveinbjörnsdóttir, S., Hicks, A. A., Jonsson, T., Pétursson, H., Guðmundsson, G., Frigge, M. L., Kong, A., Gulcher, J. R., and Stefansson, K. 2000. Familial aggregation of Parkinson's disease in Iceland. *New England Journal of Medicine*, **343,**(24) 1765–1770.

[46] Tatsumoto, S., Go, Y., Fukuta, K., Noguchi, H., Hayakawa, T., Tomonaga, M., Hirai, H., Matsuzawa, T., Agata, K., and Fujiyama, A. 2017. Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. *Scientific Reports*, **7,**(1) 13561.

[47] Thatte, B. D. and Steel, M. 2008. Reconstructing pedigrees: a stochastic perspective. *Journal of Theoretical Biology*, **251,**(3) 440–449.

[48] Torkamani, A., Wineinger, N. E., and Topol, E. J. 2018. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, **19,** 581–590.

[49] Whiffin, N., Armean, I. M., Kleinman, A., Marshall, J. L., Minikel, E. V., Karczewski, K. J., Cummings, B. B., Francioli, L., Laricchia, K., Wang, Q., et al. 2019. Human loss-of-function variants suggest that partial LRRK2 inhibition is a safe therapeutic strategy for Parkinsons disease. *bioRxiv*, page 561472.

[50] Zhang, K., Sun, F., and Zhao, H. 2004. HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics*, **21,**(1) 90–103.

[51] Zheng, C., Boer, M. P., and van Eeuwijk, F. A. 2015. Reconstruction of genome ancestry blocks in multiparental populations. *Genetics*, **200,** 1073–1087.