# Pedigree Reconstruction using Identity by Descent

B. Kirkpatrick[1], S. C. Li[2], R. M. Karp[3], and E. Halperin[4]

[1] Electrical Engineering and Computer Sciences, University of California, Berkeley, and International Computer Science Institute, Berkeley, bbkirk@eecs.berkeley.edu.
[2] International Computer Science Institute, Berkeley, scli@icsi.berkeley.edu.
[3] Electrical Engineering and Computer Sciences, University of California, Berkeley, and International Computer Science Institute, Berkeley, karp@cs.berkeley.edu.
[4] Tel Aviv University, Tel Aviv, Israel, and International Computer Science Institute, Berkeley, heran@icsi.berkeley.edu.

**Abstract.** Can we find the family trees, or pedigrees, that relate the haplotypes of a group of individuals? Collecting the genealogical information for how individuals are related is a very time-consuming and expensive process. Methods for automating the construction of pedigrees could stream-line this process. While constructing single-generation families is relatively easy given whole genome data, reconstructing muti-generational, possibly inbred, pedigrees is much more challenging.

This paper addresses this central question by introducing two multi-generational pedigree reconstruction methods: one for inbreeding relationships and one for outbreeding relationships. In contrast to previous methods that focused on the independent estimation of relationship distances between every pair of typed individuals, here we present methods that aim at the reconstruction of the entire pedigree. We show that both our methods out-perform the state-of-the-art and that the outbreeding method is capable of reconstructing pedigrees at least six generations back in time with high accuracy.

## 1 Introduction

Pedigrees are important in computer science and in genetics. Even thirty years after the development of some of the first pedigree algorithms [15, 9], pedigree graphical models continue to be a challenging graphical model to work with. Known algorithms for likelihood calculations are either exponential in the number of individuals or exponential in the number of loci [16]. There have been numerous and notable attempts to increase the speed of these calculations [20, 1, 10, 5, 11, 17, 7]. Recent work from statistics has focused on fast and efficient calculations of linkage that avoid the likelihood calculations [3, 25]. Recent contributions to genetics from pedigree calculations include fine-scale recombination maps for humans [6], discovery of regions linked to Schizophrenia [18], discovery of regions linked to rare Mendelian diseases [19], and insights into the relationship between cystic fibrosis and fertility [12].

There are both well established and recent practical methods to reconstruct pedigrees from human data. The current state-of-the-art method is an HMM-based approximation of the number of meioses separating a pair of individuals [21]. In this approach the hidden states of the HMM represent the identity-by-descent (IBD) of a pair of individuals. Two individuals are identical-by-descent for a particular allele if they each contain a copy of the same ancestral allele. The probability of the haplotype data is tested for a particular type of relationship. The main draw-back of this approach is that it may estimate a set of pair-wise relationships that are inconsistent with a single pedigree relating all the individuals.

One of the oldest pedigree reconstruction methods is essentially a structured machine learning approach where the aim is to find the pedigree graph that maximizes the probability of observing the data [24]. It should be noted that this approach is limited to extremely small families since the algorithms for computing the likelihood of a fixed pedigree graph are exponential and there are an exponential number of pedigree graphs to consider [22].

Thatte and Steel [23] examined the problem of reconstructing arbitrary pedigree graphs from a synthetic model of the data. Their method used an HMM model for the ancestry of each individual to show that the pedigree can be reconstructed only if the sequences are sufficiently long. Notice that this paper uses an unrealistic model of recombination where every individual passes on a trace of their haplotypes to all of their descendants. For single-generation relationship estimation and for animal microsatellite data, Berger-Wolf's method generates relationship constraints by considering the combinatorial options for IBD [2, 4].

Our contribution to pedigree reconstruction is two algorithms that avoid the exponential likelihood calculations. We do this by estimating the length of genome regions that are shared identical-by-descent. In two related individuals, a region of the genome is identical-by-descent (IBD) if and only if a single ancestral haplotype sequence was the source of the sequence inherited in the two individuals. The length of IBD regions gives a statistic that accurately detects sibling relationships at multiple generations. We have two algorithms: one for constructing inbred pedigrees (CIP) and one for constructing outbred pedigrees (COP). For our outbreeding algorithm the statistic is testable in polynomial time. For our inbreeding algorithm, the statistic is computable in time dependent on the number of meioses in the predicted pedigree. Our outbreeding method works to reconstruct at least six generations back in time. Both methods are more accurate than the state-of-the-art method by Stankovich, et al [21].

## 2 A Lower Bound for Out-breeding

In order to shed light on the problem we first provide a lower bound on the best that one could do in pedigree reconstruction. Speed and others [21] have been able to detect up to 3rd cousins (or relationships of 8 total meioses). We claim that this should be near optimal in the case of an infinite population size. Notice that in the infinite population size, there is no inbreeding. Therefore, the graph relating people has a path-like subgraph connecting every pair of individuals (i.e. the subgraph is either a path with a single founder or almost a path having exactly two founders whose adjacent edges can be contracted to form a simple path). This implies that in order to improve on this we need to use the fact that there is inbreeding, multiple offspring and/or that the population size is small; we will exploit these characteristics in the following sections where we develop our algorithms.

In an infinite population, consider two individuals $i$, and $j$, where their most recent common ancestor is $g$ generations ago. For instance, if $g = 1$ they are siblings. Note that they have two common ancestors in this case. For general $g$, each individual has $2^g$ ancestors, where exactly two of them are shared across $i$ and $j$; this is where we use the fact that the population is infinite, since the probability of having more than two shared ancestors is zero.

Each of the ancestors of $i$ and $j$ has two haploids. Each of the haploids arrived from a different pedigree. Consider only the haploids that arrived from the shared pedigree (the case $g = 1$ is different since there there is IBD sharing on both haploids of $i$ and $j$). These haploids of $i$ and $j$ are generated by a random walk over the ancestors of $i$ and $j$ in the $g$th generation. The total number of *haploid* ancestors in that generation is $2^g$ for each of $i$ and $j$. Out of those, four are shared across $i$ and $j$ (two shared ancestors, each has two haploids). Let $k$ be the number of meioses separating individuals $i$ and $j$, where $k = 2(g-1)$. For this reason, the expected number of bases shared between $i$ and $j$ is $\frac{4L}{2^k} = \frac{L}{2^{k-2}}$, where $L$ is the length of the genome.

On the other hand, we can calculate the average length of a shared region between the two haploids. The number of recombinations across all generations is Poisson distributed with parameter $krL$, where $r$ is the recombination rate, $L$ is the length of the genome. Let $C = rL$ be the expected number of recombinations after one generation ($C \approx 30$). Now, the length, $X$, of a shared region that originated from one of the four shared haploids is $X_1 + X_2$ where $X_i \sim exp(kr)$. So the expected length, $E[X]$, is $\frac{2}{kr}$. Since the probability to move from one shared haploid to another is negligible, we get that this is the expected length of a shared region.

Now, if $t_k$ is the expected number of regions shared between two individuals separated by $k$ meioses, we know that $t_k \frac{2}{kr} = \frac{L}{2^{k-2}}$, and therefore, $t_k = \frac{krL}{2^{k-1}}$. Therefore, $t_{10} < 1$, and it is impossible to detect a pair-wise relationship with high probability between 4th cousins.

## 3 Algorithm for Constructing Inbred Pedigrees (CIP)

Given a set of extant individuals with haplotype information available, we want construct a pedigree, using the assumption that individuals are possibly inbred. This algorithm will be referred to as CIP. We construct the pedigree recursively, one generation at a time. For example, the first iteration consists of deciding which of the extant individuals are siblings. The next iteration would determine which of the parents are siblings (yielding cousin relationships on the extant individuals).

Clearly, we can obtain any pedigree by making the right decisions at each generation. Here we define a "regular" pedigree as one where mating options are restricted to individuals in the same generation. From this point on, we assume that the pedigree is regular and monogamous.

Let us consider the first generation of relationship decisions (i.e. full siblings). We build a compatibility graph where the nodes are individuals and edges represent possible sibling relationships. We then use a heuristic to partition the graph into disjoint sets of vertices with the property that each set in the partition has many edges connecting its vertices while there are few edges connecting vertices from separate sets in the partition. This is somewhat similar to the approach taken by Berger-Wolf, et al [2], where they tried to partition the graph in order to maximize the size of the sibling groups.

For successive generations of decisions, we also employ the compatibility graph. But now the edges are given by comparing two test results. For individuals $k$ and $l$ in generation $g$, and their respective descendants $i$ and $j$, we consider one case (see Figure 1). The triangles represent the inferred sub-pedigree containing all the descendants of the individual at the point of the triangle, and individuals at the base of the triangle are extant individuals. Note that the triangles may overlap, indicating shared ancestry at an earlier generation (i.e. inbreeding).

Let $X_{i,j}$ be the length of a shared region based on the pedigree structure of the model. In order to estimate this quantity, we can sample random walks in the space of inheritance possibilities. Specifically, consider the inheritance of alleles at a single position in the genome. When there are $n$

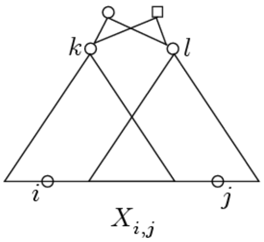

**Fig. 1. Test Case.** Specific individuals in the pedigree are indicated with either circles or squares. The triangle represents all the descendants of a particular individual. This represents the case where individuals $i$ and $j$ are cousins via the oldest generation.

non-founder individuals, define an inheritance vector as a vector containing $2n$ bits, where each pair of bits, $2i$ and $2i+1$, represents the grand-parental origin of individual $i$'s two alleles. Specifically, bit $2i$ represents the maternal allele and is zero if the grand-paternal allele was inherited and is one otherwise. Similarly, bit $2i+1$ represents the paternal allele of individual $i$. The set of possible inheritance vectors comprise the $2^{2n}$ vertices of a $2n$-dimensional hypercube, where $n$ is the number of non-founders in the pedigree. A random walk on the hypercube represents the recombination process by choosing the inheritance vectors of neighboring regions of the genome.

Given an inheritance vector, we can model the length, in number of positions, of the genomic region that is inherited according to that inheritance vector. The end of that genomic region is marked by a recombination in some individual, and constitutes a change in the inheritance vector. The random walk on the hypercube models the random recombinations, while the length of genomic regions are modeled using an exponential distribution. This model is the standard Poisson model for recombinations. Details can be found in Section 3.1. There is an analytical approach to obtaining the distribution of $X_{i,j}$ using eigen-vectors and eigen-values. However, the analytical calculation is at worst exponential. We choose to simulate the random walk and estimate the distribution.

Let $\hat{s}_{ij}$ be the observed average length of shared segments between haplotyped individuals $i$ and $j$. This can be computed directly from the given haplotype data and need only be computed once for all generations. Now, for a pair of individuals $k$ and $l$ in the oldest reconstructed generation, recall that $X_{i,j}$ is the random variable for the length of a shared region for individuals $i, j$ under the model. Now, consider the test value

$$v_{k,l} = \frac{1}{|D(k)||D(l)|} \sum_{i \in D(k)} \sum_{j \in D(l)} \frac{(\hat{s}_{ij} - \mathbb{E}[X_{ij}])^2}{var(X_{ij})} \tag{1}$$

where $D(k)$ is the set of extant individuals descended from ancestor $k$. We compute $v_{k,l}$, making edges when $v_{k,l} < c$ for all $k, l$ in the oldest generation, $g$, for some threshold $c$. Now, we proceed to greedily assign parents to the individuals in generation $g$ by iteratively finding clusters of nodes in the graph, making that cluster a set in our partition, and removing those nodes from the graph. This process results in a partition of the graph into sibling groups. We choose the the threshold, $c$, empirically by simulating many pedigrees and choosing the threshold which provides the best reconstruction accuracy.

The random-walk simulation which allows for inbreeding causes this algorithm to have an exponential running-time. The number of states in the IBD process is exponential in the number of meioses in the graph relating individuals $i$ and $j$. So, the random-walk simulation is exponential in the size of the inferred pedigree.

### 3.1 Sampling Details

Given a pedigree and individuals of interest $i$ and $j$, we will compute the distribution on the length of shared regions. Here we mean sharing to be a contiguous region of the genome for which $i$ and $j$ have at least one IBD allele at each site.

*Poisson Process* We can model the creation of a single zygote (i.e. haplotype) as a Poisson process along the genome where the waiting time to the next recombination event is exponentially distributed with intensity $\lambda = -ln(1-\theta)$ where $\theta$ is the probability of recombination per meiosis (i.e. per generation, per chromosome) between a pair of neighboring loci. For example, if we think of the genome as being composed of 3000 blocks with each block being 1MB in length and the recombination rate $\theta = 0.01$ between each pair of neighboring blocks. Then we would expect 30 recombinations per meiosis, and the corresponding intensity for the Poisson process is $\lambda = 0.01$.

Now, we have $2n$ meioses in the pedigree, with each meiosis creating a zygote, where $n$ is the number of non-founder individuals. Notice that at a single position in the genome, each child has two haplotypes, and each haplotype chooses one of the two parental alleles to copy. These choices are resented in an inheritance vector, a binary vector with $2n$ entries. The $2^{2n}$ possible inheritance vectors are the vertices of a $2n$-dimensional hypercube. We can model the recombination process as a random walk on the hypercube with a step occurring each time there is a recombination event. The waiting time to the next step is drawn from $exp(2n\lambda)$, the meiosis is drawn uniformly from the $2n$ possible meioses, and a step taken in the dimension that represents the chosen meiosis. The equilibrium distribution of this random walk is uniform over all the $2^{2n}$ vertices of the hypercube.

*IBD Process* Recall that we are interested in the distribution of the length of a region that is Identical-by-Descent (IBD). IBD is defined as the event that a pair of alleles are inherited from the same founder allele. For individuals $i$ and $j$, let $D$ be the set of hypercube vertices that result in $i$ and $j$ sharing at least one allele IBD. Given $x_0$ a hypercube vertex drawn uniformly at random from $D$, we can compute the hitting time to the first non-IBD vertex by considering the random walk restricted to $D \cup \{d\}$ where $d$ is an aggregate state of all the non-IBD vertices. The hitting time to $d$ is the quantity of interest. In addition, we also need to consider the length of the shared region before reaching $x_0$, which is the time reversed version of the same process.

The transition matrix for this IBD process is easily obtained as $Pr[x_{i+1} = u | x_i = v] = \frac{1}{2n}$ when vertices $u$ and $v$ differ by exactly one coordinate, and $Pr[x_{i+1} = u | x_i = v] = 0$ otherwise. Transitions to state $d$ are computed as $Pr[x_{i+1} = c | x_i = u] = 1 - \sum_{v \in D} Pr[x_{i+1} = v | x_i = u]$.

Now we can either analytically compute the hitting time distribution or estimate the distribution by simulating paths of this random walk. Since the number of IBD states may be exponential, it may be computationally infeasible to find eigenvectors and eigenvalues of the transition matrix [8].

We choose to simulate this random walk and estimate the distribution. This simulation is at worst exponential in the number of individuals.

## 4 Algorithm for Constructing Outbred Pedigrees (COP)

Given a set of extant individuals with haplotype information available, we want construct a pedigree, using the assumption that individuals are outbred. This algorithm will be referred to as COP. Again, we construct the pedigree recursively, one generation at a time.

Following on the pedigree reconstruction algorithm given in [13], we discuss an algorithm that assumes out-breeding and monogamy. In that paper, the author discussed the equivalence between a labeled pedigree graph and the posets of the pedigree graph. We can define a partial ordering of the individuals in a pedigree as follows. For every individual, make a list containing itself and all its descendants; this is a *descendant set* for that person. These descendant sets and the subset operator give a partial ordering on the vertices. If we have the descendant sets for all the individuals, the pedigree graph can be reconstructed since there is a fixed in-degree of two for every node in the graph. Notice that this property holds for any graph having a fixed in-degree.

The algorithm we propose now is one that builds the descendant sets at each generation with the assumption that there is a single path between each ancestor and its descendants. The challenge is to find descendant sets that are consistent with one pedigree graph.

At each generation, $g$, build a graph $H^g$ of descendant relationships where the nodes of the graph are extant descendants and the edges indicate a relationship between a pair of individuals at generation $g$. Once we obtain such a graph, we employ a partitioning method similar to the one used for the previous algorithm. To obtain the edges in the graph, we do a test for relationship-pairs of the form shown in Figure 2. If a pair of extant individuals are related at generation $g$ via a single ancestor at that generation, then the length of the regions they share IBD will be distributed according to the sum of two exponential variables, specifically, $exp(2(g-1)\lambda)$. This is the waiting time, where time corresponds to genome length, for a random walk to leave the state of IBD sharing.

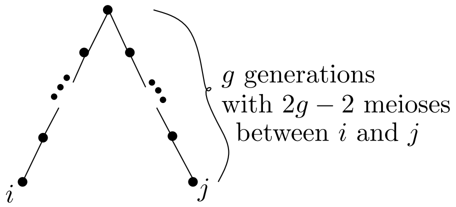

**Fig. 2. Pair of Individual Related at Generation $g$.** To test whether individuals $i$ and $j$ are related at generation $g$, we again look at the distribution on the length of genetic regions shared IBD.

Notice that if we are able to get 'exact' answers for the relationship-pair test, the outbred, monogamous pedigree will have connected components in each graph $H^g$, and the connected components from generations $H^{g-1}$ will be nested inside those from graph $H^g$. Until we reach generation $c$, for which the entire graph $H^c$ is a connected component representing the most recent common ancestor of all the extant individuals. This follows directly from the nesting property of the descendant sets.

Due to the nesting, rather than consider graph $H^g$ with nodes being extant individuals, we construct graph $G^g$ where the nodes are the individuals at generation $g$, as was done in the CIP algorithm. algorithm. For $G^g$, we take all the edges that pass a test based on Fig. 2. Specifically, the test is

$$w_{k,l} = \frac{1}{|D(k)||D(l)|} \sum_{i \in D(k)} \sum_{j \in D(l)} \frac{(\hat{s}_{ij} - \mathbb{E}[X])^2}{var(X)} \tag{2}$$

where $X = X_1 + X_2$ where $X_i \sim exp(2(g-1)\lambda)$. Here $w_{k,l} < c$ for some threshold $c$ means that there is an edge between $k$ and $l$ in the graph. Again, we set the threshold empirically with a fixed value for all the generations.

The next step is to partition the vertices of the graph. For each set in the partition, assume that it represents two monogamous parents. To guarantee that we produce descendant sets that are compatible from generation to generation, similar to the CIP algorithm, we cluster the individuals at generation $g$.

A heuristic algorithm is used to partition the vertices, $V(G^g)$, of graph $G^g$, into a partition $P = P_1, P_2, ..., P_k$, where $P_i \cap P_j = \emptyset$ for all $i, j$, and $V(G^g) = \cup_{i=1}^{k} P_i$. For a given partition set, let $E_i$ be the edges of the subgraph induced by vertices $P_i$. We wish to find a partition such that each set in the partition is a clique or quasi-clique of vertices. The objective function is to find a partition that maximizes $\sum_{i=1}^{k} a|E_i| - \binom{|P_i|}{2}$ where $a = 0.1$ is a parameter of the algorithm.

First, we partition the graph randomly such that within each set of the partition every vertex has a path length at most two to other vertices within the set. Then three heuristics are implemented to improve the score of the partition: (1) vertices are moved between sets; (2) sets are merged or split if necessary; and (3) if several sets in the partition have many edges between them, (for example, sets $P_i$ and $P_j$ have the edges induced by $P_i \cup P_j$ after excluding $V_i$ and $V_j$), then a dynamic programming algorithm is used to redistribute the vertices into an equivalent number of new clusters that optimize the objective function.

These three heuristics are applied iteratively until there are no further improvements in the score. We assessed the accuracy of this method for identifying distorted cliques. Graphs of disjoint cliques were simulated and edges were randomly removed and added with some probability. This heuristic partition method was able recover the original cliques in 90% of the simulations.

The running-time of this graph-partitioning heuristic largely determines the running-time of the pedigree reconstruction algorithm. The partitioning algorithm runs in polynomial time in the size of the graph, if the size of each set in the partition is constant. The step of creating the graph is polynomial in the size of the previous generation graph. Clearly it is possible, if no relationships are found, for the size of the graph at each generation to double. So, in the worst case, this algorithm is exponential. However, in practice this method performs quite well for constructing eight-generation pedigrees on large inputs.

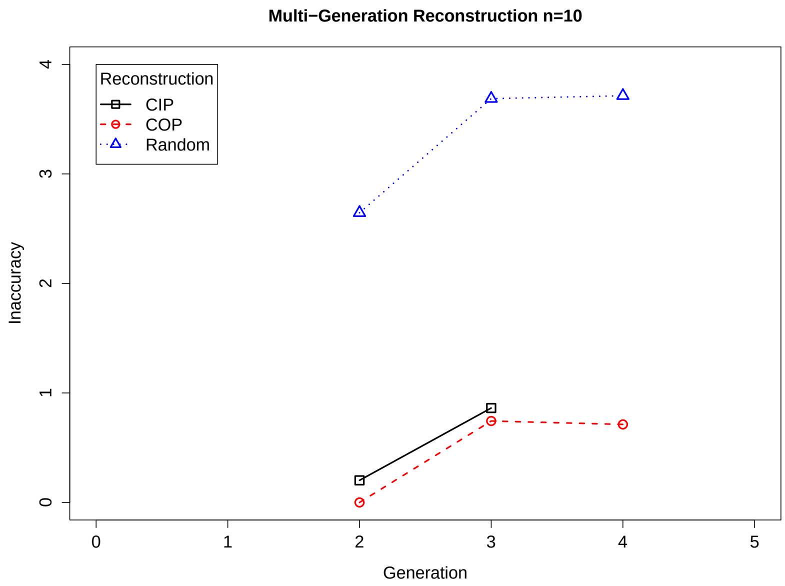

**Fig. 3. Reconstruction under High Inbreeding.** Here the pedigrees were simulated with a fixed population size of $n = 10$ individuals per generation. Over multiple generations, this results in a high level of inbreeding. (Reconstruction accuracy of 50 simulated pedigrees were averaged.)

## 5 Results

Pedigrees were simulated using a variant of the Wright-Fisher model with monogamy. The model has parameters for a fixed population size, $n$, a Poisson number of offspring $\lambda$, and a number of generations $g$. In each generation $g$, the set of $n_g$ individuals is partitioned into $n_g/2$ pairs, and for each pair we randomly decide on a number of offspring using the Poisson distribution with expectation 3.

The human genome was simulated as 3,000 regions, each of length 1MB, with recombination rate 0.01 between each region and where each founder haplotype had a unique allele for each region. The two assumptions here are 1) if two haplotypes share the same alleles in a mega-base region, then that region is identical by descent, and 2) haplotypes can be obtained for input to the pedigree reconstruction methods. Notice that Stankovich, et al. also require haplotype input to their method [21]. The requirement that haplotypes are given is not highly restrictive since our algorithms search for haplotype regions that are shared between individuals, and since we consider regions of length 1Mb (typically $> 1000$ SNPs), it is quite easy to determine whether two individuals have a shared haplotype across 1Mb.

In each experiment we end up having the true pedigree generated by the simulation, as well as an estimated pedigree. We evaluate the accuracy of the estimated pedigree by comparing the kinship matrices of the two pedigrees. Kinship is a model-based quantity defined as the frequency of IBD allele-sharing between a pair of individuals in a pedigree (averaged over the alleles of each individual). Since both pedigrees have the same set of haplotyped individuals, the comparison is actual and $L_1$ distance between the kinship estimate of those individuals. Let $K^P$ and $K^Q$ be the kinship

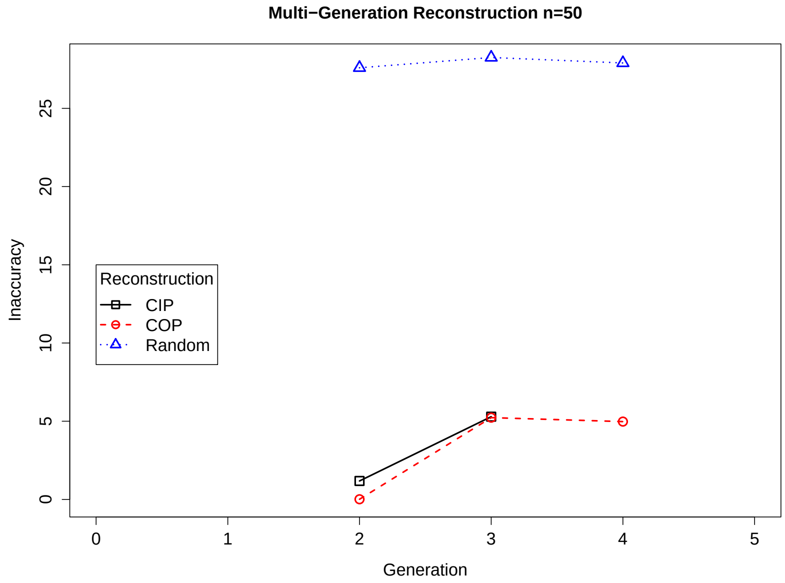

**Fig. 4. Reconstruction under Less Inbreeding.** Pedigrees here were simulated with a population size of $n = 50$. (Reconstruction accuracy of 50 simulated pedigrees were averaged.)

matrices of the actual pedigree $P$ and the estimated pedigree $Q$, respectively. Then the evaluation method is:

$$\sum_{i<j} |K^P_{i,j} - K^Q_{i,j}|$$

for haplotyped individuals $i$ and $j$.

We compare the COP and CIP methods on inbred pedigree simulations with high and moderate inbreeding, respectively $n = 10$ and $n = 50$, in Figures 3 and 4. These figures show the kinship-based inaccuracy on the y axis and the number of generations in the reconstructed pedigree on the x axis. As the depth of the estimated pedigree increases the error in the kinship of the estimated pedigree increases. However the accuracy is still much better than the accuracy of a randomly constructed pedigree, which is the highest, i.e. worst, line in each figure. Both methods perform better on the smaller population size.

Both the COP and CIP methods can reconstruct pedigree with four generations. The COP method for outbred pedigrees can reconstruct pedigrees going back to the most-recent common ancestor of the extant individuals. Provided with enough individuals, the method can construct pedigrees many generation deep. For example, given 400 individuals the method can construct 6 generations. As Figure 5 shows, the performance relative to a random reconstruction method is very good, and so is the variance of the COP reconstruction method.

We compare our two methods with the state-of-the-art method, called GBIRP, by Stankovich, et al [21]. Since GBIRP is limited to small pedigrees, we compare the methods on three-generation simulated pedigrees with population size $n = 10$. GBIRP predicts meiosis distance, $g_{ij}$, between pairs of individuals, $i, j$, without inferring pedigree relationships. In order to compare GBIRP with the

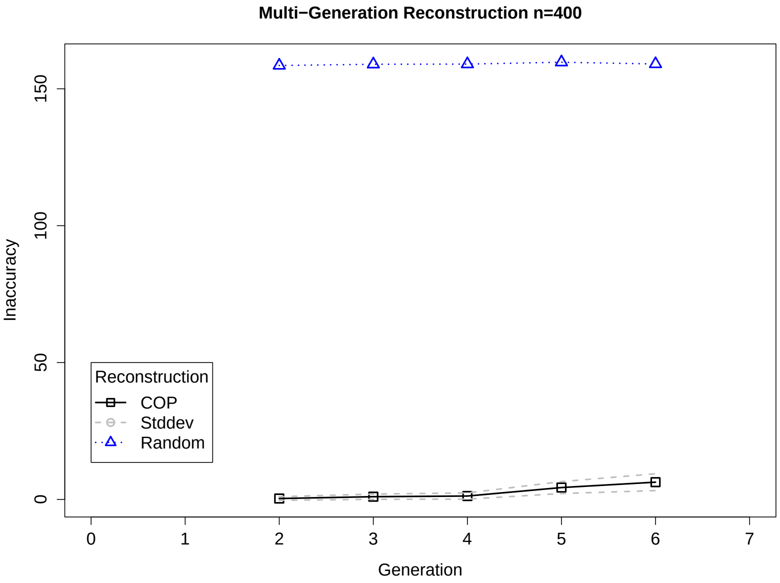

**Fig. 5. Reconstruction for Deep Pedigrees.** Pedigrees here were simulated with a population size of $n = 400$. (Reconstruction accuracy of 400 simulated pedigrees were averaged.)

actual pedigree, we extract the minimum number of meiosis, $a_{ij}$, separating every pair of individuals $i$ and $j$ in the simulated pedigree. From our predicted pedigrees, we again extract a minimum meiosis distance $p_{i,j}$. Now can compute $L_1$ distances between the actual and predicted meiosis distances. These quantities are $\sum_{i<j:g_{i,j}\neq\infty} |a_{i,j} - g_{i,j}|$, and $\sum_{i<j:p_{i,j}\neq\infty} |a_{i,j} - p_{i,j}|$. These two quantities are plotted above the line $y = 0$ in Figures 6 and 7. We plot the number of unpredicted relationships below the line $y = 0$. Specifically this is $\sum_{i<j:g_{i,j}=\infty} 1$, and $\sum_{i<j:p_{i,j}=\infty} 1$. Notice also that under this measure of accuracy, CIP performs better than COP, whereas the kinship accuracy would lead to the impression that COP performs better than CIP.

Figure 6 was done with the simulation method described above. However, in Figure 7, to obtain pedigrees with even more outbreeding, a large population size was simulated and a sub-pedigree with the desired number of extant individuals was extracted from the large simulation. Notice that with more inbred pedigrees, under this measure of accuracy, the CIP algorithm performs superior to both the COP and the GBIRP methods. The accuracy of all of the methods improve when given outbred simulation data, with both CIP and COP performing very well. However, COP performs the best with outbred input data, as expected by the modeling assumptions of the method.

Taking haplotype data from HapMap, we ran the COP algorithm on unrelated individuals. Given the parents of the CEU and YRI trios, the algorithm discovers no relationships for eight generations. The CIP algorithm, on a subset of the CEU and YRI individuals (due to running time constraints), similarly finds no relationships for three generations. Taking the individuals from the Wellcome Trust data that have at least 85% IBS with some other individual, we ran COP to construct an eight-generation pedigree. There were no relationships inferred for the first seven generations, and there were several relationships inferred at the 8th generation (i.e. seventh cousin relationships).

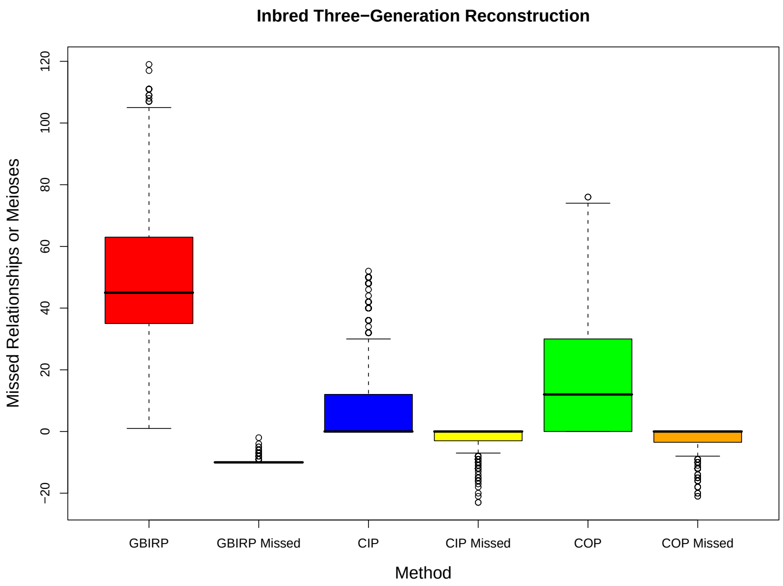

**Fig. 6. Comparison with GBIRP on Inbred Simulations.** The three-generation pedigrees here were simulated with a population size of $n = 10$, since GBIRP could not process larger pedigrees. The accuracy of 1000 simulated pedigrees were computed and plotted.

## 6 Discussion

The reconstruction of pedigrees from haplotype data is a undoubtedly a natural question of interest to the scientific community. Reconstructing very small families, or first generation relationships is a relatively easy task, but the reconstructing a full inbred pedigree involving a few generations is inherently difficult since the traces left in our genomes by an ancestor drops exponentially with the distance to the ancestor. Here, we proposed a reconstruction method for pedigrees given haplotype data from the most recent generation. We use a generation-by-generation pedigree reconstruction approach that takes haplotype data as input and finds the pedigree(s) that relate the individuals. Notably, our methods are the first to reconstruct multi-generational pedigrees, rather than a set of pair-wise relationships which may not be consistent with each other.

We present two methods of inferring the pedigrees that relate the input haplotypes. Both our methods proceed from the bottom of the pedigree towards the top. The main difference between our methods is that in CIP we assume an inbreeding model, and in COP we assume an outbreeding model. We show that our methods perform considerably better than the state of the art.

One of the basic questions that we ask is how many generations back would it be possible to reconstruct a pedigree. By simulations, we show that one can reconstruct at least fifth cousins with some accuracy. Furthermore, we obtain a lower bound showing that given two individuals with the most-recent-common ancestor being five generations back there is a constant probability for the two not to share any genomic region inherited from the common ancestor. This bound obviously does not apply to inbred pedigrees or to multi-way relationships (i.e. rather than pair-wise relationships, consider relationships on a set of individuals). One of the open problems naturally arising from this is

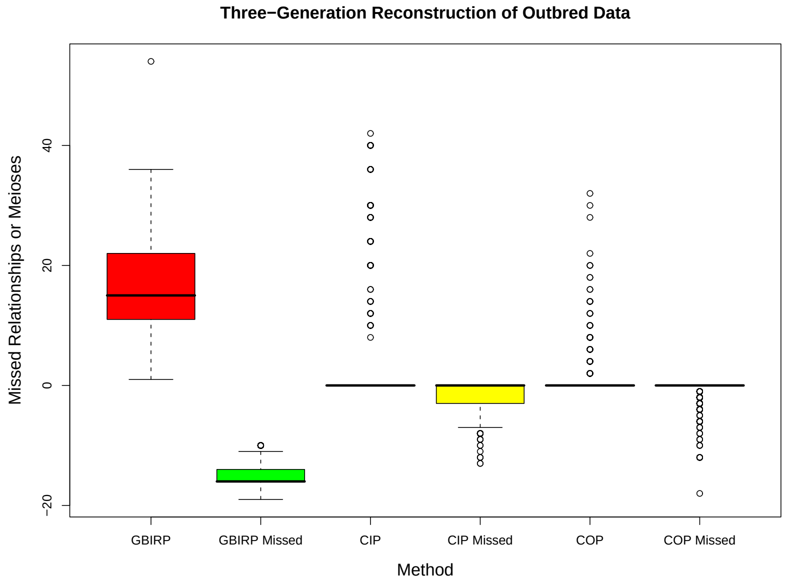

**Fig. 7. Comparison with GBIRP on Outbred Simulations.** The three-generation pedigrees here were simulated with a population size of $n = 10$, since GBIRP could not process larger pedigrees. The accuracy of 1000 simulated pedigrees were computed and plotted.

whether our lower bound can be extended to the case of inbreeding and to multi-way relationships. More generally, a major challenge would be to understand what are the limitations of pedigree reconstruction and under which conditions.

We note that our methods and analysis are limited to a restricted scenario in which there is monogamy and the generations are synchronous. If monogamy is broken then our approach will not work since the partition of the sibling relationship graph at each level will not be a simple partition. It is plausible that a different graph formulation may still provide an accurate solution to more complex pedigrees, however the exact formulation that will resolve such pedigrees is currently unknown and is left as an open challenge.

There are significant open challenges with pedigree reconstruction. For example, it would be nice to obtain confidence values on the inferred pedigree edges. However this seems very difficult, even if we can draw pedigrees from the posterior distribution of pedigree structures given the data. Since edges in a pedigree are not labeled, obtaining confidence values for a pedigree P would translate to: drawing pedigree samples, Q, from the distribution, identifying the edges in P and Q that provide the same relationships, and scoring the edges of P according to the probability of pedigree Q. As discussed in the companion submission to RECOMB [14], the second step, identifying the edges in P and Q that provide the same relationships, is a hard problem.

## References

1. GR Abecasis, SS Cherny, WO Cookson, and LR Cardon. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30:97–101, 2002.

2. Tanya Y. Berger-Wolf, Saad I. Sheikh, Bhaskar DasGupta, Mary V. Ashley, Isabel C. Caballero, Wanpracha Chaovalitwongse, and S. Lahari Putrevu. Reconstructing sibling relationships in wild populations. *Bioinformatics*, 23(13):i49–56, 2007.
3. C. Bourgain, S. Hoffjan, R. Nicolae, D. Newman, L. Steiner, K. Walker, R. Reynolds, C. Ober, and M. S. McPeek. Novel case-control test in a founder population identifies p-selectin as an atopy-susceptibility locus. *American Journal of Human Genetics*, 73(3):612–626, 2003.
4. Daniel Brown and Tanya Berger-Wolf. Discovering kinship through small subsets. *WABI 2010: Proceedings for the 10th Workshop on Algorithms in Bioinformatics*, 2010.
5. Sharon R. Browning, J. David Briley, Linda P. Briley, Gyan Chandra, Jonathan H. Charnecki, Margaret G. Ehm, Kelley A. Johansson, Brendan J. Jones, Andrew J. Karter, David P. Yarnall, and Michael J. Wagner. Case-control single-marker and haplotypic association analysis of pedigree data. *Genetic Epidemiology*, 28(2):110–122, 2005.
6. Graham Coop, Xiaoquan Wen, Carole Ober, Jonathan K. Pritchard, and Molly Przeworski. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science*, 319(5868):1395–1398, 2008.
7. Duong Doan and Patricia Evans. Fixed-parameter algorithm for haplotype inferences on general pedigrees with small number of sites. *WABI 2010: Proceedings for the 10th Workshop on Algorithms in Bioinformatics*, 2010.
8. Kevin P. Donnelly. The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, 23(1):34 – 63, 1983.
9. R.C. Elston and J. Stewart. A general model for the analysis of pedigree data. *Human Heredity*, 21:523–542, 1971.
10. M. Fishelson, N. Dovgolevsky, and D. Geiger. Maximum likelihood haplotyping for general pedigrees. *Human Heredity*, 59:41–60, 2005.
11. D. Geiger, C. Meek, and Y. Wexler. Speeding up HMM algorithms for genetic linkage analysis via chain reductions of the state space. *Bioinformatics*, 25(12):i196, 2009.
12. Gallego Romero I and Ober C. CFTR mutations and reproductive outcomes in a population isolate. *Human Genet*, 122:583–588, 2008.
13. B. Kirkpatrick. Pedigree reconstruction using identity by descent. *Class project, Prof. Yun Song, 2008. Technical Report No. UCB/EECS-2010-43*, 2010.
14. Bonnie Kirkpatrick, Yakir Reshef, Hilary Finucane, Haitao Jiang, Binhai Zhu, and Richard M. Karp. Algorithms for comparing pedigree graphs. *CoRR*, abs/1009.0909, 2010.
15. E.S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Science*, 84(5):2363–2367, 1987.
16. S. L. Lauritzen and N. A. Sheehan. Graphical models for genetic analysis. *Statistical Science*, 18(4):489–514, 2003.
17. X Li, X-L Yin, and J Li. Efficient identification of identical-by-descent status in pedigrees with many untyped individuals. *Bioinformatics*, 26(12):i191–i198, 2010.
18. Ng MY, Levinson DF, and et al. Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry*, 14:774–85, 2009.
19. Sarah B. Ng, Kati J. Buckingham, Choli Lee, Abigail W. Bigham, Holly K. Tabor, Karin M. Dent, Chad D. Huff, Paul T. Shannon, Ethylin Wang W. Jabs, Deborah A. Nickerson, Jay Shendure, and Michael J. Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, January 2010.
20. E. Sobel and K. Lange. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, 58(6):1323–1337, 1996.
21. J. Stankovich, M. Bahlo, J.P. Rubio, C.R. Wilkinson, R. Thomson, A. Banks, M. Ring, S.J. Foote, and T.P. Speed. Identifying nineteenth century genealogical links from genotypes. *Human Genetics*, 117(2–3):188–199, 2005.
22. Bhalchandra D. Thatte. Combinatorics of pedigrees, 2006.
23. Bhalchandra D. Thatte and Mike Steel. Reconstructing pedigrees: A stochastic perspective. *Journal of Theoretical Biology*, 251(3):440 – 449, 2008.
24. E. A. Thompson. *Pedigree Analysis in Human Genetics*. Johns Hopkins University Press, Baltimore, 1985.
25. T. Thornton and M.S. McPeek. Case-control association testing with related individuals: A more powerful quasi-likelihood score test. *American Journal of Human Genetics*, 81:321–337, 2007.