PLOS | GENETICS

RESEARCH ARTICLE

Composite likelihood method for inferring local pedigrees

Amy Ko[1]*, **Rasmus Nielsen**[1,2,3]

1 Department of Integrative Biology, University of California, Berkeley, Berkeley, California, United States of America, **2** Department of Statistics, University of California, Berkeley, Berkeley, California, United States of America, **3** Museum of Natural History, University of Copenhagen, Copenhagen, Denmark

* amyko@berkeley.edu

OPEN ACCESS

Citation: Ko A, Nielsen R (2017) Composite likelihood method for inferring local pedigrees. PLoS Genet 13(8): e1006963. https://doi.org/10.1371/journal.pgen.1006963

Editor: Ellen Wijsman, University of Washington, UNITED STATES

Received: January 10, 2017

Accepted: August 7, 2017

Published: August 21, 2017

Data Availability Statement: The Greenlandic Inuit data set is available upon request, subject to IRB approval constraints from the Greenland Research Ethics Committee (nun@nanoq.gl), and can be obtained by contacting Dr. Torben Hansen (torben.hansen@sund.ku.dk).

Funding: This study is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No DGE 1106400 (AK). This study is based upon work supported by the National Health Institute under Grant No 2R1400322 (RN). The funders had no role in study design, data collection and analysis,

Abstract

Pedigrees contain information about the genealogical relationships among individuals and are of fundamental importance in many areas of genetic studies. However, pedigrees are often unknown and must be inferred from genetic data. Despite the importance of pedigree inference, existing methods are limited to inferring only close relationships or analyzing a small number of individuals or loci. We present a simulated annealing method for estimating pedigrees in large samples of otherwise seemingly unrelated individuals using genome-wide SNP data. The method supports complex pedigree structures such as polygamous families, multi-generational families, and pedigrees in which many of the member individuals are missing. Computational speed is greatly enhanced by the use of a composite likelihood function which approximates the full likelihood. We validate our method on simulated data and show that it can infer distant relatives more accurately than existing methods. Furthermore, we illustrate the utility of the method on a sample of Greenlandic Inuit.

Author summary

Pedigrees contain information about the genealogical relationships among individuals. This information can be used in many areas of genetic studies such as disease association studies, conservation efforts, and for inferences about the demographic history and social structure of a population. Despite their importance, pedigrees are often unknown and must be estimated from genetic information. However, pedigree inference remains a difficult problem due to the high cost of likelihood computation and the enormous number of possible pedigrees that must be considered. These difficulties limit existing methods in their ability to infer pedigrees when the sample size or the number of markers is large, or when the sample contains only distant relatives. In this report, we present a method that circumvents these computational challenges in order to infer pedigrees of complex structure for a large number of individuals. Using simulations, we find that the method can infer distant relatives much more accurately than existing methods. Furthermore, we show that even pairwise inferences of relatedness can be improved substantially by consideration of the pedigree structure with other related individuals in the sample.

decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Pedigree information is used in many areas of genetic analysis, including discovery of disease-related markers in co-segregation analysis and family-based association studies [1], pedigree-informed haplotype and genotype imputation [2], and in estimating variance components for quantitative traits (e.g. heritability) [3]. At the population level, pedigrees can elucidate the social organization and behavior of a group, such as mating patterns and variance in reproductive success among individuals [4]. Furthermore, pedigree information can be used to infer population parameters such as migration rates between subpopulations at very recent time scales. Most population genetic inference methods are based on coalescence theory, which models the genealogical relationships among samples of genetic data at a time scale of N generations, where N is the effective population size. However, standard coalescence models, such as Kingman's coalescent [5–7] ignore pedigree structure. Simulation studies have shown that the coalescent is a poor approximation of the genealogical process over short time frames ($< log_2 N$ generations, where N is the population size), potentially leading to inaccurate inferences at these time scales [8, 9]. Therefore using the pedigree, which contains more detailed information about the genealogical history of the samples, should provide more power in inferring population parameters for the very recent past.

Considerations of pedigree structure is becoming increasing relevant as the size of population genetic samples increases, as these samples may have an increasing probability of including cryptic relatives. The likelihood of seeing cryptic relatives in population samples depends on the sample size, effective population size, and breeding structure. For example, Moltke [10] found that due to the small population size in Greenland, even a relatively small sample size of 584 Inuits contained many close relatives, and about half of the samples had to be removed to form an unrelated set. Other examples include the HapMap Phase III data in which Pemberton [11] found 166 pairs of cryptic close relatives (i.e. third degree relatives or closer) among the sample population of about 1400; and the San Antonio Family Studies in which Sun [12] found 4 cryptic relative pairs among 154 putatively unrelated samples. Performing association studies on samples harboring cryptic relatedness may result in spurious associations [13]. In such cases, pedigree information can be used to remove related samples or explicitly model relatedness to increase the power of association studies [14].

Pedigree information is undoubtedly valuable. In many cases, however, pedigrees are not directly observable and must be inferred from genetic data, which is the topic of this paper. However, we note that using estimated pedigrees as a replacement for known pedigrees may not be an optimal procedure in many cases, if the statistical uncertainty in the estimation of the pedigree is ignored. For example, the consequences of using estimated pedigrees in linkage analyses are largely unknown and we warn against the use of such methods without further studies of their properties.

Although numerous pedigree inference methods have been developed to date, most are limited to inferring very close relationships or require a prior knowledge of the sample structure. Many existing methods support only single- or two-generation samples. The single-generation methods are sibship inference algorithms which partition the sampled individuals into sibship clusters [15–18]. The parentage inference methods for two generations find the best parent-offspring combinations from a set of offspring and candidate parents [19–21]. Several methods that can support more than two generations have been developed [22–28]. But they are either limited in the number of markers that can be analyzed [23, 28]; do not support polygamous pedigrees [26, 27]; assume a complete sample (i.e. every member in the pedigree is sampled) [24, 25, 29]; or assume all sampled individuals belong to a single generation [26, 27]. The state-of-the-art method, PRIMUS [30], is the most flexible of the existing methods; it accommodates

missing data and is able to infer multi-generational, polygamous pedigrees. Although PRIMUS is a notable improvement from other methods, its accuracy decreases significantly as the number of missing individuals increases. This is problematic as we expect samples to contain only a small fraction of pedigree members unless the sample represents a large portion of the total population or is specifically designed to include close family members. Extending the work of PRIMUS, PADRE [31] connects PRIMUS-reconstructed family networks to estimate distant relatives. However, PADRE estimates only the degree of relationship between the founders connecting the family networks, which is not equivalent to estimating the pedigree.

The difficulty in pedigree inference comes from three sources. First, the number of possible pedigrees is enormous even for a small sample size [32, 33], making naive enumeration of pedigrees in search for the best one infeasible. Second, computing the likelihood of a pedigree is very expensive. Algorithms for computing the likelihood of a pedigree are either exponential in the number of loci [34], or in the number of individuals [35], which makes the likelihood computation of large pedigrees at many loci prohibitively slow. Finally, inference of pedigree relationships from genetic relationships, measured by the proportion of the genome shared by identical-by-descent (IBD), has high uncertainty. As the pedigree relationship between two individuals becomes more distant, the coefficient of variation and the magnitude of skew in genome sharing become larger [36]. For example, the distribution of genome sharing between second cousins overlaps significantly with that of third cousins, making these two pedigree relationships difficult to distinguish based on pairwise genome sharing alone.

In this report, we present CLAPPER (Composite Likelihood Approach to Pedigree Reconstruction), a method that estimates the unknown pedigree from the genotype data of a sample of individuals. Note that our parameter of interest is the pedigree, which is not equivalent to the set of all pairwise relationships. In fact, pairwise relationships do not necessarily define a unique pedigree. Our new inference method addresses the drawbacks of the existing methods. More specifically, our method can utilize many markers genome-wide, support multi-generational pedigrees (up to 5 generations) and polygamous reproduction, and allows many missing individuals in the sample. We assume that all individuals are outbred and that the pedigrees do not create cycles, except in the case of full-sibs. To increase computation efficiency, we use a composite likelihood to approximate the full likelihood based on pairwise likelihoods, and use simulated annealing as a heuristic optimization algorithm for maximizing the composite likelihood. We validate our method on simulated data and show that it outperforms existing methods for inferring distant relatives. Furthermore, we demonstrate our method's application to real data on a sample of Greenlandic Inuit.

Materials and methods

Composite likelihood

CLAPPER is based on the idea of forming a composite likelihood function based on marginal likelihood functions calculated for pairs of individuals. While even pairwise likelihoods are slow to calculate for full genomic data, they can be tabulated and stored in computer memory. It is thereby possible to estimate pedigrees, based on a composite likelihood function, by only calculating the likelihood function between pairs of individuals once. This makes our method potentially applicable to large data sets containing thousands of individuals. As we will later discuss, using some heuristics, the method may even be applicable to large GWAS data sets.

We define a pedigree as undirected graphs where a node represents an individual and an edge represents a parent-offspring relationship (S1 Text). Each individual has a sex and is associated with 0, 1 or 2 edges connecting the individual to its parents, which must be of different sexes if the individuals has two identified parents. An individual in the pedigree may or may

not be represented in the sample, but if individual i is represented in the sample it is associated with genotype vector, X_i. For each pedigree, the set of k sampled individuals is denoted by H, and the composite likelihood for such a pedigree is defined as

$$CL(H) = \begin{cases} P(X_i), & \text{if } k = 1 \\ \frac{\prod_{(i,j)\in H} P(X_i, X_j|R_{i,j})}{\prod_{i\in H} P(X_i)^{k-2}}, & \text{otherwise} \end{cases} \quad (1)$$

where $R_{i,j}$ is the relationship between i and j induced by the pedigree. For a pedigree consisting of one individual, the likelihood is simply the probability of the individual's observed genotypes. For $k > 1$ the composite likelihood is obtained as the product of marginal pairwise likelihoods. However, to obtain a more natural scaling of the composite likelihood we note that the probability of the data for each individual has been calculated $k - 1$ times and we therefore divide the composite likelihood function with the marginal likelihood of each individual $k - 2$ times. This has several desirable properties such as convergence of the composite likelihood to the true likelihood as the relatedness among individuals goes to zero. Another way to think of this composite likelihood function is in terms of products of conditional likelihoods. We can factor the full likelihood as

$$P(X_1, \cdots, X_k|H) = P(X_1)P(X_2|X_1, H) \cdots P(X_k|X_1, \cdots, X_{k-1}, H).$$

Since computing the conditional likelihoods $P(X_i|X_1, \cdots, X_{i-1}, H)$ is difficult, we approximate them with

$$P(X_i) \prod_{j=1}^{i-1} \frac{P(X_i|X_j, H)}{P(X_i)}.$$

That is, we multiply the marginal probability of our current observation $P(X_i)$ by the likelihood ratio $\frac{P(X_i|X_j,H)}{P(X_i)}$ for each previous observation X_j. If the previous observation informs our current observation, then $\frac{P(X_i|X_j,H)}{P(X_i)} \neq 1$, so the likelihood of the current observation increases or decreases accordingly. Using this approximation, we arrive at Eq (1). Note that $P(X_i|H) = P(X_i)$ since $P(X_i)$ is simply the likelihood of observing the genotypes X_i, which is independent of the pedigree, H.

The pairwise likelihood $P(X_i, X_j|R_{i,j})$ can be computed efficiently using the Hidden Markov Model (HMM) approximation by [37], which is used in this study. However, we note that any other definition of the pairwise likelihood function could have been used. For a set of possible outbred relationships in a 5-generation pedigree (See S1 Table), the pairwise likelihood for each pair (i, j) is precomputed and stored in memory. The total pre-computation time for $\binom{n}{2}$ pairs of individuals, s types of relationships, and L loci, therefore, is $O(n^2sL)$. Since the composite likelihood of a pedigree is a simple function of the pairwise and marginal likelihoods, it can be computed fast by accessing the precomputed values stored in memory. The full composite likelihood for a set of local pedigrees is then computed by taking the product of the composite likelihood for each local pedigree.

It is worthwhile to note alternative ways to construct a composite likelihood. Another, perhaps more intuitive, formulation that also ensures that the composite likelihood converges to the true likelihood as the relatedness among individuals goes to zero, is

$$\prod_{i\neq j} P(X_i, X_j)^{\frac{1}{n-1}}, \quad (2)$$

which scales the product of pairwise likelihoods by $\frac{1}{n-1}$ to account for the multiple counting of each sample. However, as we will discuss in the Results section, this formulation leads to a worse approximation of the full likelihood function.

Simulated annealing

Because the number of possible pedigrees grows very rapidly with sample size, an exhaustive search for the most likely pedigree is infeasible for even a moderate number of individuals. Therefore, we use simulated annealing [38] to maximize the composite likelihood function. In this algorithm, a perturbation of the pedigree is generated by locally modifying the edges and nodes of the current pedigree (S1 Text). We explore the pedigrees with high likelihoods by always accepting proposals with higher likelihoods and occasionally accepting those with lower likelihoods to avoid getting stuck in local maxima. We implemented 22 different perturbations (moves) detailed in S1 Text. These moves can be broadly categorized into three classes. The first class of moves involves choosing two individuals and modifying their pairwise relationship. These moves include transitions between: parent-offspring and full siblings; parent-offspring and half siblings; uncle-nephew and nephew-uncle; grandparent-grandchild and half siblings; and full siblings and self. Related to these are moves that add or subtract an edge between two nodes. For example, adding an edge causes parent-offspring relationships to become grandparent-grandchild relationship, whereas subtracting an edge has the opposite effect. The motivation for this class of moves is that these pairs of relationships have similar IBD coefficients, hence similar likelihoods. So these perturbations allow transitions between pedigrees with similar likelihoods.

The second class of moves allows bigger perturbations in the current pedigree. These moves include splitting a pedigree into two, joining two pedigrees into one, or the combination of splitting and joining. Splitting a pedigree can be done in two ways: we can either detach a chosen individual's sub-pedigree (i.e. its descendant and itself) from its ancestors, or split off a randomly selected subset of its children to form a new pedigree. Joining two pedigrees involves creating a common ancestor between two individuals that belong to different local pedigrees.

The last class of moves is designed to transition between similar pedigrees when sex or age information is missing. For example, one move allows an individual and its descendant to swap places if age information is not present to resolve the directionality of the relationship. Another move changes the sex of an individual if sex information is not available, which in turn switches the sex of its potential spouses.

All of these transitions modify a small part of the current pedigree to generate a new configuration. Since the composite likelihood is a function of the pairwise and marginal likelihoods, the likelihood of the new configuration can be computed fast by adjusting the old likelihood by the changes made to the modified part of the pedigree.

The outline of the simulated algorithm is described below:

Initialization: Let each individual be a singleton pedigree (i.e. everyone is unrelated). Compute and store the composite likelihood of the current configuration.

Recursion:

1. Choose one of the 22 moves at random and generate a new configuration accordingly.
2. If the new configuration is an invalid pedigree, reject and go back to step 1. If it is a valid pedigree, compute the composite likelihood $CL(H_{new})$ for the new configuration. Accept with probability $min[(CL(H_{new})/CL(H_{old}))^t, 1]$, where t is the annealing temperature.
3. Repeat steps 1-2 C times.

4. Decrease the temperature to t/f, where $f > 1$ and go to step 1.

Termination: Terminate after I iterations or when the change in composite likelihood is less than e.

The tuning parameters C, f, I, and e were optimized to achieve a balance between convergence and computational efficiency using a number of trial runs on different simulated data sets. S2 Table shows an example of the composite likelihood score at different stopping times determined by the maximum number of iterations. We run multiple instances of the algorithm with different random seeds. The algorithm then reports the pedigree with the highest likelihood encountered among all runs.

Background relatedness

Since the composite likelihood function is based on pairwise likelihood values, any inference based on it is limited by the quality of the pairwise likelihoods. One important factor that confounds the likelihood computation is linkage disequilibrium (LD), which often causes relationships to be overestimated [39]. Unrelated pairs of individuals often have higher likelihoods for being distantly related (S1 Fig), which leads to false detection of relatives. The method of [37] attempts to correct for LD by conditioning on nearby markers. However, in our experience residual effects of LD will still tend to bias inferences when markers are in high LD. One way to further reduce the effects of LD is pruning, or thinning, of markers. However, there is no consensus on how best to choose a set of markers that contains minimal LD and yet harbors enough information to detect distant relatives. To get a better sense of the effects of LD pruning on relationship inference, we simulated various pairwise relationships (i.e. second cousins, third cousins, unrelated) at linked loci. We pruned the markers based on LD in 100 unrelated founders and measured the pairwise prediction accuracy for the test pair. We repeated this procedure under different levels of LD pruning to choose an appropriate level of pruning threshold (See Results). In addition to LD pruning, we further controlled for false detection of relatives by adding a regularization term to the composite likelihood. The regularizer was designed to weight against individuals from forming family clusters, motivated by the fact that in large data sets there are so many potential pedigree relationships for each individual, that most individuals will be inferred to have some pedigree relationship to at least one individual in the sample, even when they are unrelated. This is essentially a multiple testing problem in which an increasing number of individuals in the sample implies a reduced probability of inferring an individual to be unrelated to all individuals in the sample. There are natural ways of addressing this problem in a Bayesian framework that we might also be able to appeal to in the current framework. In particular, we will assign a probability distribution on the number of local pedigrees inferred. More specifically, we used the regularized composite likelihood

$$CL^*(X) = CL(X)Pr(Q = q)^{\beta}, \tag{3}$$

where q is the number of local pedigrees and $\beta > 0$. We chose a Poisson distribution with mean n, the sample size, as the distribution of Q. This regularization is conservative in the sense that it favors every individual to remain a singleton unless there is strong evidence otherwise. Our choice to use the Poisson distribution was made, in part, for computational convenience but, as we will discuss in the Results section, resulted in good statistical properties of the method.

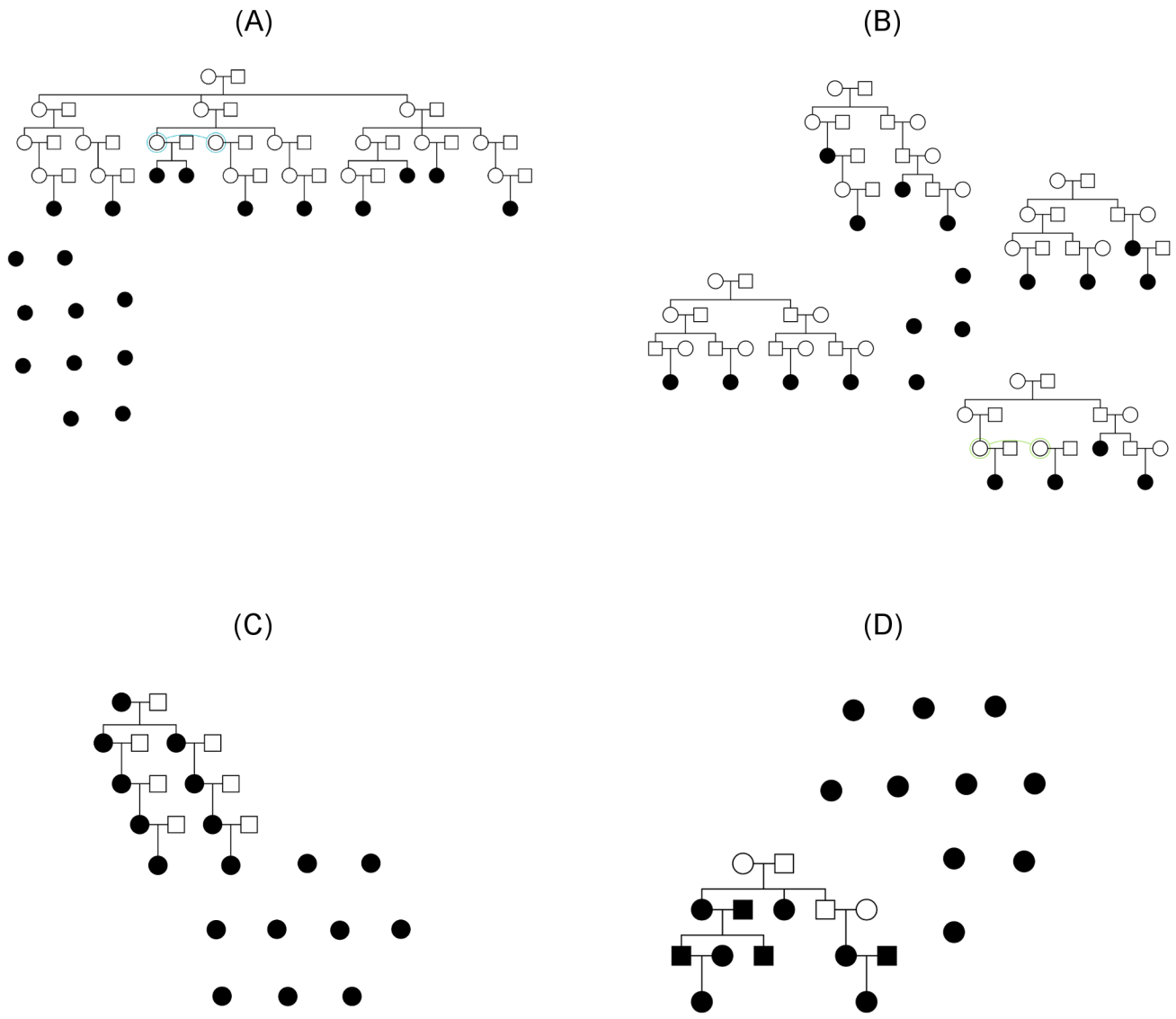

Fig 1. Simulated pedigrees. Shaded nodes indicate sampled individuals for which we have genotype data and unshaded nodes indicate unsampled individuals. (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D.

https://doi.org/10.1371/journal.pgen.1006963.g001

Simulated dataset

We tested the performance of our method on simulated pedigrees. We generated human autosomal haplotypes using msprime [40] with effective population size of 10,000, average recombination rate of 1.3e-8, and mutation rate of 1.25e-8. Using these founder haplotypes, we simulated four pedigree structures shown in Fig 1.

Simulation A consisted of 10 singletons and a 45-person family that spanned 5 generations. Of the 45 family members, 10 were sampled and 35 were missing. The kinship coefficients of the sampled relative pairs ranged from 1/4 (e.g. full siblings) to 1/256 (e.g. third cousins). Simulation B was designed to study the performance of our method on smaller family clusters. It consisted of 4 family clusters and 4 singletons. Each family cluster contained 15 to 18 members, of which only 4 of them were sampled. The sampled individuals spanned multiple generations and formed pairwise relationships with kinship coefficients

ranging from 1/4 to 1/256. Simulation C was designed the test the method on pedigree structures in which every sampled individual, excluding singletons, has at least one close relative in the data. It consisted of 9 singletons and a 16-person pedigree that spanned 5 generations. The 16-person pedigree contained 7 missing individuals and 9 sampled individuals, where each sampled individual formed a parent-offspring relationship with at least one other sample. Finally, simulation D was designed to test the method on a pedigree that is relatively easy yet more difficult to infer than simulation C. Whereas every sample was connected by parent-offspring relationships in simulation C, some samples in simulation D were connected only by an avuncular relationship.

Each simulation scenario was replicated 100 times. For each sampled individual, we simulated genotyping error by switching each allele to the alternate allele with probability 0.01. To reduce the level of LD among markers, we used PLINK [41] to prune the original set of markers at $r^2 = .05$, resulting in about 10,000 markers. The sex of each sample was assumed known, whereas the age was assumed unknown.

Empirical dataset

We applied our method to reconstruct the previously unreported pedigrees of 100 individuals in Tasiilaq villages in Greenland which had been genotyped [10] using the Illumina CardioMetaboChip, consisting of 196,224 SNPs. Since the European admixture into the Greenlandic population can confound relationship inference, we selected individuals from Tasiilaq villages, which showed one of the lowest levels of European admixture in the sample. In particular, the 100 individuals we selected were estimated to have European admixture proportion of 5 percent or less. To reduce the effects of LD, with pruned the markers using PLINK at $r^2 = 0.05$. Due to the unusually high level of LD in the Greenlandic population, we were left with 2173 SNPs after LD-pruning.

Competing methods for comparison

We compared the performance of our method on simulated data to PRIMUS (v1.9.0), arguably the state-of-the-art pedigree reconstruction method. Although many pedigree inference methods exist, we chose to use PRIMUS as a benchmark since it is the most flexible of the existing methods in the types of pedigrees it can infer. More specifically, PRIMUS supports the inference of multi-generational, polygamous pedigrees and allows for missing individuals. PRIMUS reconstructs pedigrees that are consistent with pairwise IBD estimates and reports high-scoring configurations.

To estimate the pairwise IBD coefficients for the simulated data, we used two different methods: PLINK and RELATE [37]. To use PLINK, we first estimated the population allele frequencies from 100 founder individuals. We then used PLINK to estimate the IBD coefficients for the individuals in our simulated pedigrees, where the population allele frequency estimates were provided as input. This mimics the inference procedure recommended in the PRIMUS documentation. A similar procedure was used to run RELATE to estimate the pairwise IBD proportions (S2 Text). The IBD estimates were then used by PRIMUS to reconstruct likely pedigrees. We denote the combined method of PLINK and PRIMUS as PP, and Relate and PLINK as RP. Since PRIMUS was designed to reconstruct pedigrees where samples are connected by third-degree relationships or closer, we applied PP and RP only to simulations C and D.

We used PADRE [31] for simulations A and B, where PRIMUS was inappropriate to use due to the presence of samples connected only by distant relationships. PADRE takes as input relationship likelihoods by ERSA [42] and output by PRIMUS, and reports the degree

of relationship for each pair of samples. To generate the results by PRIMUS, we used PP and RP as described before. ERSA uses estimates of IBD segments to compute the pairwise relationship likelihoods. Since RELATE was used to compute the pairwise likelihood of IBD proportions for CLAPPER, we used RELATE also to estimate the pairwise IBD segments to generate the input for ERSA. We denote the combined method of PP and PADRE as PPP, and RP and PADRE as RPP. The command lines used for running the softwares are provided in S2 Text.

Recall that CLAPPER maximizes a statistic that incorporates both the likelihood score and the number of family networks Eq (3). In PP and RP, however, all reported pedigrees have the same number of family networks, which makes maximizing both the likelihood score and the number of family networks equivalent to maximizing the likelihood score alone. The same is true for PPP and RPP, which report a single best estimate of family networks.

We also compared our method to the pairwise inference method. In this method, we used RELATE to compute the pairwise likelihood under each possible relationship (S1 Table) for all pairs of individuals. Then we assigned each pair the relationship with the highest pairwise likelihood. We controlled the false positive rate by multiplying the likelihood of being unrelated by a scalar $c > 0$, in order to provide comparable results between methods. The pairwise inference method produces only the best relationship for each pair, which may not result in a valid pedigree when all pairwise relationships are pieced together. Still, it serves as a useful benchmark to evaluate the accuracy of pairwise predictions by our method.

Measuring the error rate

We measured the performance of our method in two ways: the frequency of estimating the true pedigree; and the distance between the estimated pedigree and the true pedigree in terms of pairwise relationships. We note that since CLAPPER does not consider inbred pedigrees whereas PP and RP do, we pre-processed the output of PP and RP before measuring the error rate to make a fair comparison. More specifically, we removed all inbred pedigrees from the output of PP and RP and measured the error rate using just the remaining pedigrees.

Frequency of estimating the true pedigree configuration. We say that the estimated pedigree is correct if there is a one-to-one mapping between the nodes of the estimated pedigree and the nodes of the true pedigree such that each edge in the estimated pedigree has a corresponding edge in the true pedigree. Note that for PP and RP, which potentially report multiple highest-scoring pedigrees, we say that the estimated pedigree is correct if the true pedigree is in the set of highest-scoring pedigrees.

Pairwise error rate. To measure the error rate of the pairwise method, which estimates pairwise relationships directly, we compared the true relationships to the estimated relationships. Therefore, we define the error rate for each pair as

$$e = \begin{cases} 0, & \text{if } \hat{w}_1 = w_1 \text{ and } \hat{w}_2 = w_2 \\ 1, & \text{otherwise} \end{cases}$$

where w_i is the probability that two individuals share i pairs of alleles IBD at a random locus under the true relationship; and $\hat{w}_i$ is the corresponding probability for the estimated relationship. In other words, the estimated relationship is correct if its three Jacquard coefficients [43] are exactly the same as those of the true relationship.

Furthermore, to measure the distance between the estimated relationship and the true relationship for each pair, we computed the kinship coefficient distance

$$d = \frac{|\hat{\phi} - \phi|}{\phi},$$

where $\hat{\phi} = \frac{1}{4}\hat{w}_1 + \frac{1}{2}\hat{w}_2$ and $\phi = \frac{1}{4}w_1 + \frac{1}{2}w_2$.

We also used e and d to measure the pairwise error rate of CLAPPER, where the inferred pairwise relationships are those induced by the estimated pedigree, and the true pairwise relationships are those induced by the true pedigree. For PP and RP, which report all pedigrees with high likelihood scores, we computed the error rate by taking the average across all highest-scoring pedigrees. For PPP and RPP, which report a single best degree of relationship for each pair, we measured the error rate by e and d as defined above.

Results

Behavior of the composite likelihood

To examine the behavior of the composite likelihood, we simulated a nuclear family with two parents and their four children at 3,000 independent loci. We then computed the likelihood of the data under various pedigree configurations, ranging from the pedigree in which no one is related to the true pedigree. For each pedigree configuration, we computed the likelihood value with three different formulas: the full likelihood using MERLIN [44], composite likelihood A, given by Eq (2), and composite likelihood B, given by Eq (1).

The comparison of the three likelihood formulas are shown in S2 Fig. The x-axis is the distance of the test pedigree to the true pedigree, measured by the proportion of pairwise relationships that are correct in the test pedigree. As expected, the full likelihood increases as the test configuration becomes closer to the true pedigree. Both composite likelihood formulas preserve the ordering of the pedigrees induced by the full likelihood. That is, the order of pedigrees from the least likely to the most likely based on the full likelihood corresponds to the ordering based on the composite likelihood formulas. Although both composite likelihood formulas preserve this ordering, the likelihood surface given by Eq (2) is much flatter than the full likelihood, whereas the likelihood surface of Eq (1) is roughly on the same order of magnitude as the full likelihood.

Effects of linkage disequilibrium on pairwise relationship inference

As mentioned in the Methods section, we examined different thresholds for LD pruning. The appropriate level of pruning depends both on the genome length and the types of relationships we want to infer accurately. As shown in Fig 2, there is a trade-off between keeping enough markers to estimate distant relationships and removing markers to reduce false detection of relatives. For unrelated pairs, the most stringent LD pruning we tested (r^2 = .025) showed the best relationship prediction accuracy. For third cousin relationships, however, pruning the markers too severely caused too much information loss, leading to a decrease in prediction accuracy. A similar pattern is observed for the second cousin relationships. For our simulated and empirical data, we prune the markers at r^2 = .05, which according to our simulations, retained enough information to estimate second and third cousins while keeping the false positive rate (i.e. estimating unrelated pairs as related) relatively low. We note that finding optimal strategies for dealing with background LD when inferring relatedness is an important topic that merits further research.

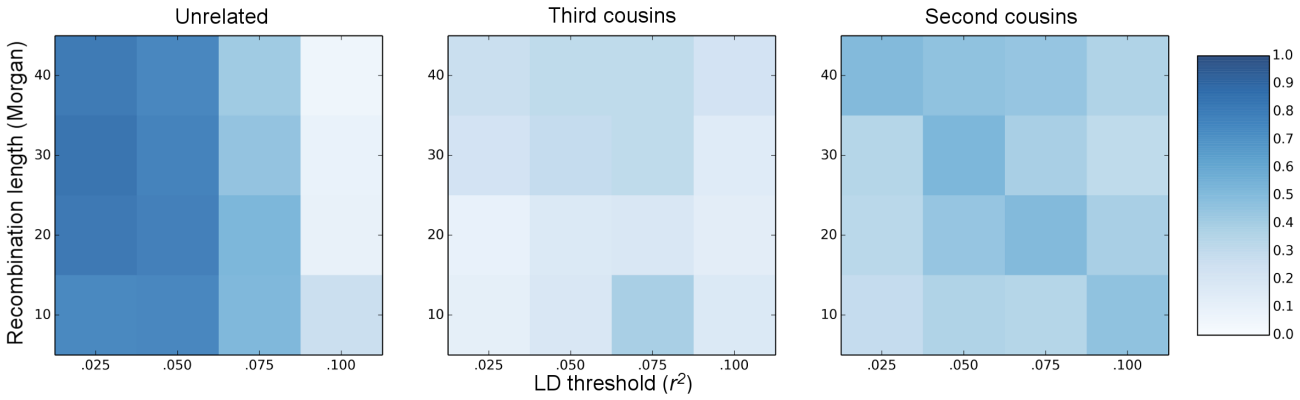

Fig 2. Effects of LD-pruning on pairwise prediction accuracy. The three panels show different true pairwise relationships: unrelated, third cousins, and second cousins. Each square in a panel corresponds to the relationship prediction accuracy for a particular genome length and LD-prune threshold. The color indicates the accuracy rate between 0 and 1.

https://doi.org/10.1371/journal.pgen.1006963.g002

Estimating simulated pedigrees

Table 1 summarizes the frequency of estimating the true pedigree and the average number of best pedigrees reported by each method. For simulation C, where all samples were connected by parent-offspring relationships, CLAPPER was able to find the true pedigree in all 100 experiments. This showed that when the sampled individuals are connected by very close relationships, CLAPPER can unambiguously find the correct pedigree. Similarly, RP inferred the true pedigree as the single best estimate in 96 out of 100 experiments. The remaining 4 experiments did not output any pedigrees because all likely pedigrees exceeded the maximum number of generations we imposed (5 generations). On the other hand, PP showed a lower accuracy rate than both CLAPPER and RP. Several experiments finished with errors due to too large a number of likely pedigrees to process, while some only produced inbred pedigrees. However, the true pedigree was estimated in the majority of the experiments that finished successfully.

For simulation D, all methods had a lower accuracy rate for estimating the true pedigree compared to simulation C. Some of the samples in this scenario were connected only through an avuncular relationship, which made the inference more difficult than the pedigree given in simulation C. Nonetheless, CLAPPER showed a higher accuracy rate than both PP and RP even though we counted the estimated pedigree as correct if the true pedigree was found in any of the best reported pedigrees for RP and PP. Simulations A and B were omitted from our

Table 1. Accuracy for estimating the true pedigree.

Simulation	# Reported[a] (CLAPPER)	# Correct[b] (CLAPPER)	# Reported[a] (PP)	# Correct[b] (PP)	# Reported[a] (RP)	# Correct[b] (RP)
C	1	100/100	6	65/76 *	1	96/96 **
D	1	56/100	49	18/79 ***	3	20/100

[a] Number of highest scoring pedigrees reported.

[b] Numerator is the number of times the true pedigree was among the highest scoring pedigrees; denominator is the the number of successful experiments that produced at least one outbred pedigree.

*Excludes 6 runs that finished with errors and 18 runs that did not produce any outbred pedigrees.

**Excludes 4 runs that did not produce any pedigrees.

***Excludes 20 runs that finished with errors and 1 run that did not produce any outbred pedigrees.

https://doi.org/10.1371/journal.pgen.1006963.t001

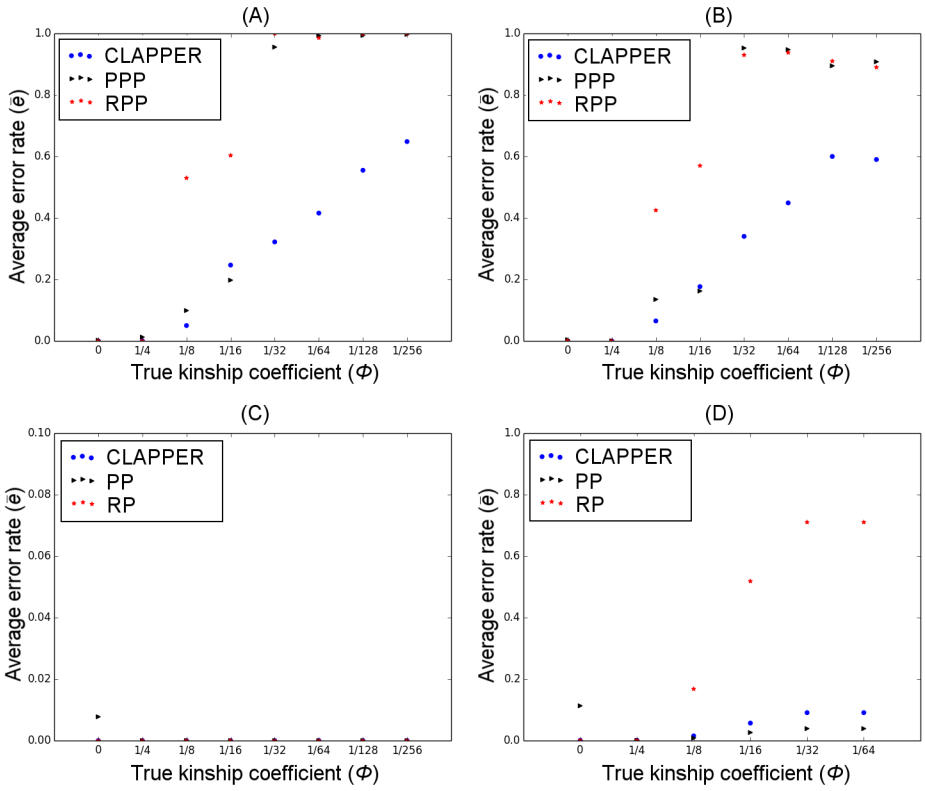

Fig 3. Comparison of prediction error rates. Each panel compares the average error rate between CLAPPER and competing methods for a particular simulation scenario: (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D. The x-axis shows different relationship categories measured by the kinship coefficient; the y-axis is the average error rate $\bar{e}$ (See Measuring the Error Rate). Analysis excludes all experiments that did not finish successfully or did not produce any outbred pedigrees.

https://doi.org/10.1371/journal.pgen.1006963.g003

analysis since they contained samples that were not connected by third degree relationships or closer, which made PP and RP inappropriate to use to estimate the full pedigree.

Figs 3 and 4 show the average pairwise error rate across all replicate experiments, categorized by different levels of true relatedness, ϕ. For simulation A, PPP did not finish successfully in 19 out of 100 experiments due to errors encountered in PRIMUS (e.g. too many likely pedigrees to process). Similarly, PPP did not finish successfully in 24 experiments for simulation B. Furthermore, PP and RP encountered errors or did not produce any outbred pedigrees in some experiments (Table 1). These experiments were removed from our analyses and are not reflected in Figs 3 and 4.

For simulations A and B, all methods had a very low false positive rate (i.e. error rate for $\phi = 0$), and relatively low error rates for estimating close relationships (Fig 3A and 3B). For more distant relatives such as those beyond first cousins ($\phi \leq 1/32$), however, CLAPPER was able to estimate the relationships more accurately than both PPP and RPP. For simulation C, all methods had zero error rates for all relationship categories except PP, which showed a non-zero false positive rate (Fig 3C). For simulation D, CLAPPER outperformed RP across all relationship categories, but had a lower accuracy rate than PP in many relationship categories. However, PP showed a significantly higher false positive rate than CLAPPER (Fig 3D).

Furthermore, Fig 4 shows that even when the estimated relationship by CLAPPER is wrong, it is generally close to the true relationship. For example, the median error rate for $\phi = 1/128$ was 0.5, which is equivalent to estimating second cousins once removed as third cousins.

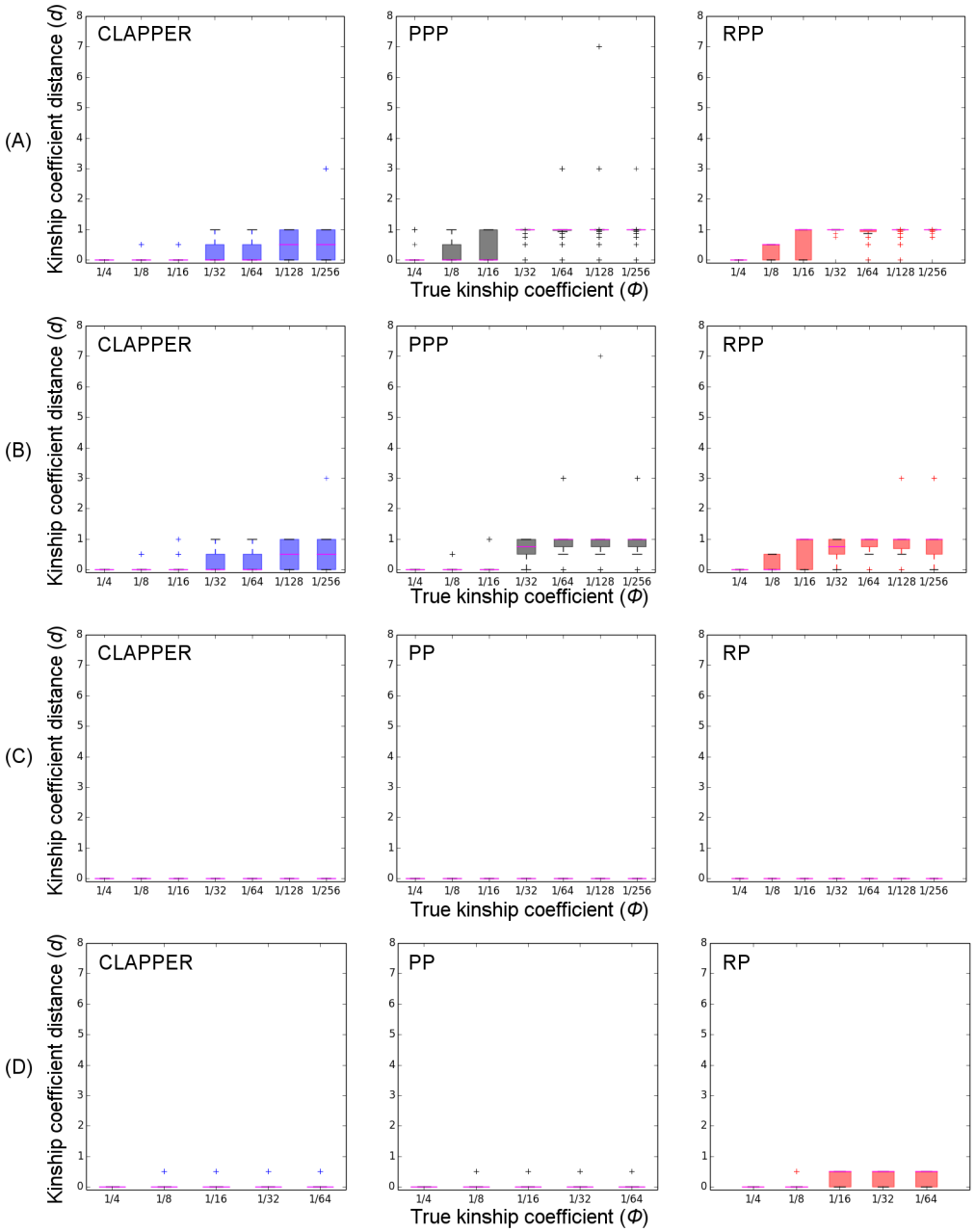

Fig 4. Absolute between the expected kinship coefficient under true and inferred relationships, normalized by the true kinship coefficient. (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D. The x-axis is the relationship category measured by the kinship coefficient; the y-axis is the distance *d* between the true relationship and the relationship estimated by our method (See Measuring the Error Rate in Materials and methods section). The magenta line indicates the median value for each box plot. Analysis excludes all experiments that did not finish successfully or did not produce any outbred pedigrees.

https://doi.org/10.1371/journal.pgen.1006963.g004

Overall, the median error rate of CLAPPER was equal to or lower than that of the competing methods across all relationship categories.

CLAPPER also performed considerably better than the pairwise inference method. The likelihoods in the pairwise prediction were weighted so that its false positive rate roughly matched that of our method. Fig 5 show that at similar false positive rates, our method

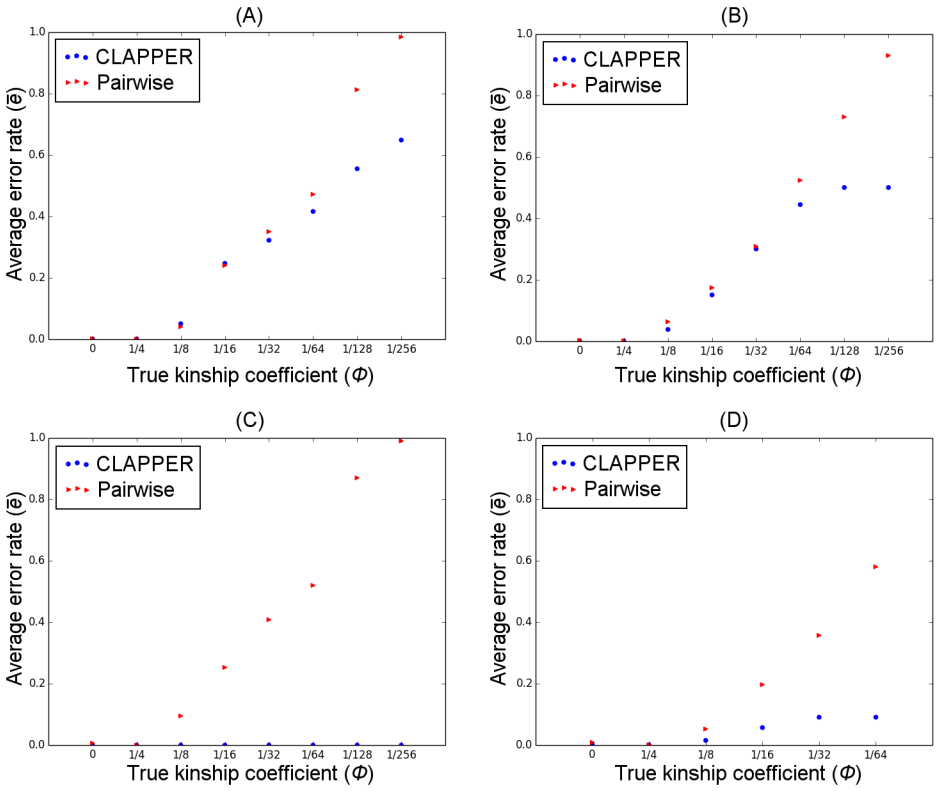

Fig 5. Comparison of prediction error rates between CLAPPER and pairwise inference. Each panel compares the average error rate between the pairwise method and CLAPPER for a particular simulation scenario: (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D.

https://doi.org/10.1371/journal.pgen.1006963.g005

estimated pairwise relationships with a greater accuracy than the pairwise method across almost all relationship categories. Fig 6 further demonstrates that our method has a significant advantage over the pairwise prediction method in detecting relatives. If the purpose of relationship inference is to find relatives–to discover the number of family clusters present in the

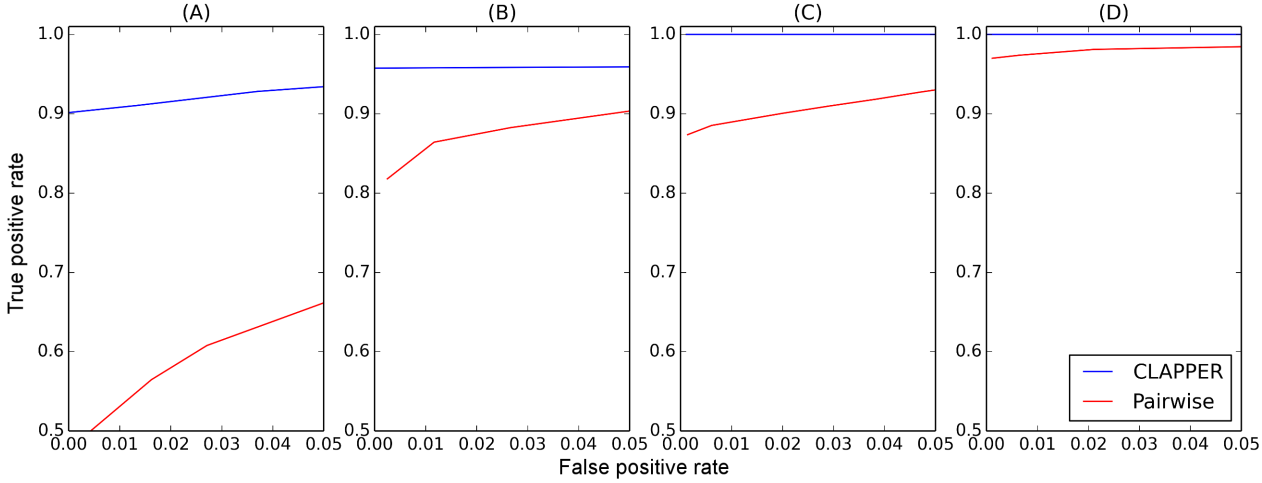

Fig 6. ROC curve for detecting relatives in a sample: Pairwise vs. CLAPPER. (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D.

https://doi.org/10.1371/journal.pgen.1006963.g006

data, for example–Fig 6 demonstrates that our method is able to detect relatives far more accurately than the pairwise method. These figures show that even though our method and the pairwise inference method both use the same pairwise likelihood values to estimate relationships, leveraging information from all pairs of relationships improves the inference significantly compared to considering each pair in isolation.

Each experiment was run 3 times with different random number seeds, where each run consisted of 2 million iterations. The runtime of our method depends on many factors, including the number of individuals, the hidden pedigree structure, the number of missing individuals, and the annealing schedule in the simulated annealing algorithm. That said, each run on our simulated data, excluding the pre-computation time for calculating the pairwise likelihoods, took about 9 seconds on 2.5 GHz Intel Core i5 processor.

Estimating the greenlandic inuit pedigrees

To demonstrate our method's ability to infer pedigrees in practical applications, we estimated the previously unreported pedigrees of 100 individuals from Tasiilaq villages in Greenland. Because the Greenlandic Inuit population has high levels of LD, only 1868 SNPs remained after pruning the markers at $r^2 = .05$. Our simulation study showed that at this number of SNPs, regularization with $Poi(n)$ caused the error rate for estimating distant relatives ($\phi < 1/32$) to be very high; but using no regularization at all led to a high false positive rate (S3 Fig). So we chose to use $Poi(n/2)$ as our regularization, which still produced a lower false positive rate, yet performed better in inferring distant relatives on simulated data.

We ran our algorithm 5 times with different random number seeds, resulting in 5 pedigrees estimates. The top three estimates with the highest composite likelihood scores were within 1.2 likelihood units of each other. The other two estimates were both about 20 likelihood units away from the top three. The inconsistency of the multiple runs was likely caused by the existence of multiple local peaks on the likelihood surface, which makes finding the global optimum difficult in our heuristic optimization. Each run, which consisted of 80 million iterations, finished in about 24 minutes on 2.5 GHz Intel Core i5 processor. S4 Fig shows the estimated pedigree drawn by PhenoTips [45]. The reconstructed pedigree consisted of 38 singletons and 8 non-singleton family clusters. Many of these clusters consisted of close relationships such as parent-offspring, full siblings, half-siblings, and avuncular relationships. Based on our simulations, we expect more than 90 percent of the estimated relationships in these categories to be correct.

Discussion

In this report, we have shown that the use of composite likelihood allows us to analyze pedigrees containing many individuals at many loci, where computing the full likelihood would be prohibitively slow. Our method can estimate pedigrees when the number of possible pedigrees is too large to enumerate, which is true even for tens of individuals in a multi-generational pedigree. Our method is also one of the very few methods that can support complex pedigree structures such as polygamy, multigenerational pedigrees (up to 5 generations), and missing individuals. In addition, we can incorporate information about sex, age, and the number of generations spanned by the sample to better estimate the pedigree.

We have shown that our method has a significant advantage over the pairwise inference method. It can better estimate relationships beyond first cousins (Fig 5) and is able to detect relatives much more accurately (Fig 6). The composite likelihood considers all pairwise likelihoods jointly, which in turn can help resolve uncertain relationships in the context of other pairwise relationships. Therefore, even for pairwise relationship inference, where estimating

the entire pedigrees may not necessarily be of interest, our method can be used to estimate the relationships more accurately.

Our method also showed an improvement over PRIMUS (PP and RP) and PADRE (PPP and RPP). PRIMUS's reconstruction algorithm relies on accurate pairwise relationship assignments based on IBD estimates. If the sample consists mostly of distant relatives, however, relationship assignment becomes uncertain due to high variance in IBD sharing, which often leads to incorrect pedigree reconstruction. Although our method also relies on pairwise information, we showed that working directly with pairwise likelihood values rather than IBD-based relationships assignments improved the power significantly. Furthermore, PRIMUS's enumeration of possible pedigrees becomes computationally cumbersome as the number of likely pedigrees increases rapidly for a set of distantly related samples. If the data contains many close relationships, however, PRIMUS can reconstruct all likely pedigrees very fast, whereas our method produces a single best pedigree, which may be close but not exactly correct. Thus the performance of each method depends on the sample structure and a suitable method must be chosen accordingly. Similar to PRIMUS, the performance of PADRE depends crucially on accurate estimates of IBD proportions and segments, and poor estimates of either parameter can lead to biases in the relationship inference. We note that IBD estimation is a difficult problem and better estimates of IBD would improve the performance of both PRIMUS and PADRE.

We applied our method on the Greenlandic Inuit dataset to demonstrate its ability infer previously unknown pedigrees from genetic data. Although the estimates of distant relationships are uncertain, we can still get a general sense of pedigree structures hidden in the data and take appropriate actions for downstream analyses. For example, the inferred pedigree can be used to filter out close relatives or model relatedness among samples in association studies. Furthermore, we can validate or improve the estimated pedigree with other evidence such as age.

Pedigree inference based on our composite likelihood is heavily influenced by how well we can compute the pairwise likelihoods. An important factor that affects the pairwise likelihood computation is LD, which often leads to overestimation of relatedness. Although the HMM by [37] conditions on nearby markers, it does not remove the effects of LD completely and necessitates LD-pruning. Unfortunately, there is no consensus on how best to prune markers while still retaining enough information to infer distant relatives. Although we carried out a simple simulation study to get a rough sense of appropriate level of pruning, it is by no means a complete solution. More work is needed on the effects of LD on relatedness inferences and how to remedy the problem, whether it be by more extensive simulations studies, or by modeling LD in the likelihood computation. Furthermore, care must be taken to use appropriate allele frequencies in likelihood computation to account for other potentially confounding factors such as population substructure [46, 47] and admixture [48, 49]. As better methods for estimating pairwise likelihoods become available, our method for estimating pedigrees should also improve.

There are limitations to our method that require further work. Our method assumes that all individuals are outbred, which may not be true of many systems including some human populations [50, 51]. It currently does not support pedigrees with cycles caused by inbreeding or complex cyclic relationships such as double first cousins. When inbreeding is present, CLAPPER infers pedigrees that are close to the underlying truth under the assumption that there is no inbreeding (S3 Text). Pedigree non-identifiability also poses a challenge to pedigree estimation. Donnelley [52] remarked that two pairs of cousin-type pedigrees that have equal numbers of meioses are not identifiable (e.g. half cousins vs. great half avuncular) no matter how much genetic data are available. Furthermore, Kirkpatrick [53] gave examples of non-identifiable

3-person pedigrees where no likelihood-based methods, including the full likelihood, can find the correct pedigree for certain. Another limitation of our method is that it does not provide an uncertainty measure on the estimated pedigree. This could be solved in two ways: by block-bootstrapping the data and repeating the inference, which would be slow; or using a Bayesian approach by assigning a prior to pedigrees and attempting to sample from the posterior distribution. Furthermore, while computationally efficient compared to full likelihood methods, our method is still based on calculation of pairwise relationships and does, therefore, not scale up to GWAS data sets with hundreds of thousands of individuals. However, it may be possible to use a divide-and-conquer approach in which individuals are first divided into clusters using methods such as [54], then estimating the pedigree of each cluster separately, and finally estimating more distant relationships among clusters.

Overall, our method provides a computationally efficient way to estimate pedigrees of seemingly unrelated individuals. It improves our ability to validate and discover pedigrees in realistic genetic datasets where we expect a high level of missing data. The ability to estimate pedigrees more accurately opens up possibilities to develop and improve numerous pedigree-based or pedigree-aware studies, from correcting cryptic relatedness in GWAS to estimating demographic parameters of the very recent past. However, as noted in Introduction, the naive use of estimated pedigrees in downstream analyses may not be justified when there is significant statistical uncertainty in the estimation of the pedigree. Such analyses would need to take the statistical uncertainty in pedigree estimation into account, a topic of potential future research.

Our software is available for download at https://github.com/amyko/clapper.

Supporting information

S1 Table. Summary of possible pairwise relationships in a 5-generation pedigree.
(PDF)

S2 Table. Composite likelihood convergence. Composite likelihood score at various stopping times given by I for a particular instance of simulation B.
(TIFF)

S1 Text. Description of transitions between pedigree graphs.
(PDF)

S2 Text. Command lines Used for running external softwares.
(PDF)

S3 Text. Effects of inbreeding on CLAPPER. Summarizes a simulation study that shows how CLAPPER performs on inbred pedigrees.
(PDF)

S1 Fig. Effects of LD on relatedness estimation. The figure shows the histogram of the log likelihood difference, L(unrelated) − L(third cousins), when the true relationship is unrelated. Unrelated pairs often have higher likelihoods for being third cousins when LD is present in the data, as shown by the histogram corresponding to linked markers. The data were simulated with msprime and the likelihoods were computed using RELATE.
(TIF)

S2 Fig. Comparison of various likelihood formulas on simulated data. The x-axis measures how close the test pedigree is to the true pedigree; the test pedigree becomes closer to the truth from left to right. In this simulation, the composite likelihood given by Eq (1) approximates

the full likelihood more closely than Eq (2).
(TIFF)

S3 Fig. Effects of regularization term. Accuracy of simulated annealing method on simulated data at 2000 markers under different levels of regularization.
(TIFF)

S4 Fig. Estimated pedigree of 100 Tasiilaq individuals in the Greenlandic Inuit dataset. Shaded nodes indicate sampled individuals; unshaded for unsampled; squares for male; circles female; diamonds for unknown sex.
(PDF)

S5 Fig. Likelihood convergence for the Greenlandic Inuit pedigree estimation.
(TIFF)

Acknowledgments

We would like to thank Jeffrey Spence and Peter Wilton for helpful discussion and feedback, and Matteo Fumagalli for providing the Greenlandic dataset.

Author Contributions

Conceptualization: Amy Ko, Rasmus Nielsen.

Data curation: Amy Ko.

Formal analysis: Amy Ko.

Funding acquisition: Rasmus Nielsen.

Investigation: Amy Ko.

Methodology: Amy Ko, Rasmus Nielsen.

Project administration: Rasmus Nielsen.

Resources: Rasmus Nielsen.

Software: Amy Ko.

Supervision: Rasmus Nielsen.

Validation: Amy Ko.

Visualization: Amy Ko.

Writing – original draft: Amy Ko.

Writing – review & editing: Rasmus Nielsen.

References

1. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. Nat Rev Genet. 2011; 12(7):465–474. https://doi.org/10.1038/nrg2989 PMID: 21629274
2. Livne OE, Han L, Alkorta-Aranburu G, Wentworth-Sheilds W, Abney M, Ober C, et al. PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population. PLoS Comput Biol. 2015; 11(3). https://doi.org/10.1371/journal.pcbi.1004139 PMID: 25735005
3. Vinkhuyzen AAE, Wray NR, Yang J, Goddard ME, Visscher PM. Estimation and partition of heritability in human populations using whole-genome analysis methods. Annu Rev Genet. 2013; 47:75–95. https://doi.org/10.1146/annurev-genet-111212-133258 PMID: 23988118

4. Blouin MS. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol. 2003; 18(10):503–511. https://doi.org/10.1016/S0169-5347(03)00225-8
5. Kingman JFC. The coalescent. Stochastic processes and their applications. 1982; 13(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4
6. Kingman J. Exchangeability and the evolution of large populations. 1982;.
7. Kingman JF. On the genealogy of large populations. Journal of Applied Probability. 1982; p. 27–43. https://doi.org/10.1017/S0021900200034446
8. Wakeley J, King L, Low BS, Ramachandran S. Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent. Genetics. 2012; 190(4):1433–1445. https://doi.org/10.1534/genetics.111.135574 PMID: 22234858
9. Wakeley J, King L, Wilton PR. Effects of the population pedigree on genetic signatures of historical demographic events. Proceedings of the National Academy of Sciences. 2016; 113(29):7994–8001. https://doi.org/10.1073/pnas.1601080113
10. Moltke I, Fumagalli M, Korneliussen TS, Crawford JE, Bjerregaard P, Jørgensen ME, et al. Uncovering the genetic history of the present-day Greenlandic population. Am J Hum Genet. 2015; 96(1):54–69. https://doi.org/10.1016/j.ajhg.2014.11.012 PMID: 25557782
11. Pemberton TJ, Wang C, Li JZ, Rosenberg NA. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. The American Journal of Human Genetics. 2010; 87(4):457–464. https://doi.org/10.1016/j.ajhg.2010.08.014 PMID: 20869033
12. Sun L, Dimitromanolakis A. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. In: BMC proceedings. vol. 8. BioMed Central; 2014. p. S23.
13. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 2005; 1(3):e32. https://doi.org/10.1371/journal.pgen.0010032 PMID: 16151517
14. Eu-ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SMB, Blackwell JM, Cordell HJ, et al. Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. PLoS Genet. 2014; 10(7). https://doi.org/10.1371/journal.pgen.1004445 PMID: 25033443
15. Almudevar A. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theor Popul Biol. 2003; 63(2):63–75. https://doi.org/10.1016/S0040-5809(02)00048-5 PMID: 12615491
16. Smith BR, Herbinger CM, Merry HR. Accurate partition of individuals into full-sib families from genetic data without parental information. Genetics. 2001; 158(3):1329–1338. PMID: 11454779
17. Thomas SC, Hill WG. Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics. 2000; 155(4):1961–1972. PMID: 10924488
18. Wang JL. Sibship reconstruction from genetic data with typing errors. Genetics. 2004; 166(4):1963–1979. https://doi.org/10.1534/genetics.166.4.1963 PMID: 15126412
19. Hadfield JD, Richardson DS, Burke T. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Mol Ecol. 2006; 15(12):3715–30. https://doi.org/10.1111/j.1365-294X.2006.03050.x PMID: 17032269
20. Wang J, Santure AW. Parentage and Sibship Inference From Multilocus Genotype Data Under Polygamy. Genetics. 2009; 181(4):1579–1594. https://doi.org/10.1534/genetics.108.100214 PMID: 19221199
21. Wang J. Computationally Efficient Sibship and Parentage Assignment from Multilocus Marker Data. Genetics. 2012; 191(1):183–194. https://doi.org/10.1534/genetics.111.138149 PMID: 22367033
22. Almudevar A, Anderson EC. A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. Mol Ecol Resour. 2012; 12(1):164–178. https://doi.org/10.1111/j.1755-0998.2011.03061.x PMID: 21883980
23. Anderson EC, Ng TC. Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation. Theor Popul Biol. 2016; 107:39–51. https://doi.org/10.1016/j.tpb.2015.09.005 PMID: 26450523
24. Cowell RG. Efficient maximum likelihood pedigree reconstruction. Theor Popul Biol. 2009; 76(4):285–91. https://doi.org/10.1016/j.tpb.2009.09.002 PMID: 19781561
25. Cowell RG. A simple greedy algorithm for reconstructing pedigrees. Theor Popul Biol. 2013; 83:55–63. https://doi.org/10.1016/j.tpb.2012.11.002 PMID: 23164633
26. He D, Wang Z, Han B, Parida L, Eskin E. IPED: Inheritance Path-based Pedigree Reconstruction Algorithm Using Genotype Data. J Comput Biol. 2013; 20(10):780–791. https://doi.org/10.1089/cmb.2013.0080 PMID: 24093229
27. Kirkpatrick B, Li SC, Karp RM, Halperin E. Pedigree Reconstruction Using Identity by Descent. J Comput Biol. 2011; 18(11):1481–1493. https://doi.org/10.1089/cmb.2011.0156 PMID: 22035331

28. Riester M, Stadler PF, Klemm K. FRANz: reconstruction of wild multi-generation pedigrees. Bioinformatics. 2009; 25(16):2134–9. https://doi.org/10.1093/bioinformatics/btp064 PMID: 19202194

29. Cussens J, Bartlett M, Jones EM, Sheehan NA. Maximum likelihood pedigree reconstruction using integer linear programming. Genet Epidemiol. 2013; 37(1):69–83. https://doi.org/10.1002/gepi.21686 PMID: 23034892

30. Staples J, Qiao D, Cho MH, Silverman EK, Nickerson DA, Below JE, et al. PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent. Am J Hum Genet. 2014; 95 (5):553–564. https://doi.org/10.1016/j.ajhg.2014.10.005 PMID: 25439724

31. Staples J, Witherspoon DJ, Jorde LB, Nickerson DA, Below JE, Huff CD, et al. PADRE: Pedigree-Aware Distant-Relationship Estimation. The American Journal of Human Genetics. 2016; 99(1):154–162. https://doi.org/10.1016/j.ajhg.2016.05.020 PMID: 27374771

32. Steel M, Hein J. Reconstructing pedigrees: a combinatorial perspective. Journal of theoretical biology. 2006; 240(3):360–367. https://doi.org/10.1016/j.jtbi.2005.09.026 PMID: 16325207

33. Thatte BD, Steel M. Reconstructing pedigrees: a stochastic perspective. J Theor Biol. 2008; 251 (3):440–9. https://doi.org/10.1016/j.jtbi.2007.12.004 PMID: 18249415

34. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered. 1971; 21 (6):523–42. https://doi.org/10.1159/000152448 PMID: 5149961

35. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987; 84(8):2363–7. https://doi.org/10.1073/pnas.84.8.2363 PMID: 3470801

36. Hill WG, Weir BS. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb). 2011; 93(1):47–64. https://doi.org/10.1017/S0016672310000480

37. Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet Epidemiol. 2009; 33(3):266–74. https://doi.org/10.1002/gepi.20378 PMID: 19025785

38. Kirkpatrick S VM Gelatt CD. Optimization by Simulated Annealing. Science. 1983; 220:671–680. https://doi.org/10.1126/science.220.4598.671

39. Sun M, Jobling MA, Taliun D, Pramstaller PP, Egeland T, Sheehan NA. On the use of dense SNP marker data for the identification of distant relative pairs. Theor Popul Biol. 2016; 107:14–25. https://doi.org/10.1016/j.tpb.2015.10.002 PMID: 26474828

40. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol. 2016; 12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842 PMID: 27145223

41. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

42. Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research. 2011; 21(5):768–774. https://doi.org/10.1101/gr.115972.110 PMID: 21324875

43. Jacquard A. The genetic structure of populations. vol. v. 5 of Biomathematics. Berlin: Springer-Verlag; 1974.

44. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30(1):97–101. https://doi.org/10.1038/ng786 PMID: 11731797

45. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: patient phenotyping software for clinical and research use. Human mutation. 2013; 34(8):1057–1065. https://doi.org/10.1002/humu.22347 PMID: 23636887

46. Anderson AD, Weir BS. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics. 2007; 176(1):421–440. https://doi.org/10.1534/genetics.106.063149 PMID: 17339212

47. Wang J. Unbiased relatedness estimation in structured populations. Genetics. 2011; 187(3):887–901. https://doi.org/10.1534/genetics.110.124438 PMID: 21212234

48. Rohlfs RV, Fullerton SM, Weir BS. Familial identification: population structure and relationship distinguishability. PLoS Genet. 2012; 8(2):e1002469. https://doi.org/10.1371/journal.pgen.1002469 PMID: 22346758

49. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. The American Journal of Human Genetics. 2012; 91(1):122–138. https://doi.org/10.1016/j.ajhg.2012.05.024 PMID: 22748210

50. Leutenegger AL, Sahbatou M, Gazal S, Cann H, Génin E. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us&quest. European Journal of Human Genetics. 2011; 19(5):583–587. https://doi.org/10.1038/ejhg.2010.205 PMID: 21364699

51. Gazal S, Sahbatou M, Babron MC, Génin E, Leutenegger AL. High level of inbreeding in final phase of 1000 Genomes Project. Scientific reports. 2015; 5. https://doi.org/10.1038/srep17453 PMID: 26625947

52. Donnelly KP. The probability that related individuals share some section of genome identical by descent. Theoretical population biology. 1983; 23(1):34–63. https://doi.org/10.1016/0040-5809(83)90004-7 PMID: 6857549

53. Kirkpatrick B. Non-identifiable pedigrees and a bayesian solution. In: International Symposium on Bioinformatics Research and Applications. Springer; 2012. p. 139–152.

54. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010; 26(22):2867–2873. https://doi.org/10.1093/bioinformatics/btq559 PMID: 20926424