Pedigree reconstruction from SNP data: Parentage assignment, sibship clustering, and beyond

Jisca Huisman

August 26, 2016

Abstract

I present the R package *sequoia* for multi-generational pedigree reconstruction. It assigns parents, clusters half-siblings and assigns grandparents to half-sibships using potentially incomplete and partly erroneous data on several hundred Single Nucleotide Polymorphisms (SNPs). Assignments are made after consideration of the likelihoods of all possible alternative first, second and third degree relationships between the focal individuals, as well as the traditional alternative of being unrelated, thereby overcoming Thompson's sibling paradox. Distinction between the various categories of second degree relatives is possible when likelihoods are calculated conditional on at least on parent of each individual. Initial exclusion of pairs highly unlikely to be related is computationally fast, and based on opposite homozygosity for parentage and simplified pairwise likelihoods throughout. Parentage assignment is highly accurate down to 75 SNPs and takes a few minutes for thousands of individuals. The full algorithm converges to a pedigree with high global likelihood within at most 5–10 iterations in typically less than an hour, even for complex pedigrees involving inbreeding. Simulated datasets based on three different large pedigrees, with 40% of parents assumed non-genotyped, resulted in low error rates ($< 0.1\%$) and high assignment rates ($> 99\%$) when at least 200 SNPs were included.

Introduction

Pedigrees have many uses in a wide variety of fields, ranging from animal breeding and human genealogy to wildlife genetics and ethology. They can be used to derive estimates of pairwise relatedness coefficients r (Pemberton, 2008), required to estimate trait heritabilities. Even though r can now be estimated more precisely directly from genomic data than from a pedigree (Visscher *et al.*, 2006; Bérénos *et al.*, 2014), heritability estimates still require proper accounting for the similarity due to shared parents (Kruuk & Hadfield, 2007; Bérénos *et al.*, 2014). The relevant shared parent is unobservable in many marine species, den-sharing social mammals or seed-dispersing plants, and in such cases a pedigree is required to distinguish parents from full siblings and offspring, or between paternal and maternal half-siblings. Moreover, in natural populations pedigrees provide estimates of reproductive success, the key indicator of individual fitness, while in agriculture they enable estimation of (genomic) breeding values based on offspring performance. Thus, pedigree reconstruction remains useful in the current genomics era.

A plethora of methods have been developed to reconstruct pedigrees based on a dozen or so multi-allelic microsatellites (see Jones *et al.* (2010) for an overview). High resolution SNP data opens up new ways of pedigree reconstruction, amongst others via reliable distinction between different categories of relatives. Simultaneously, the smaller information content per SNP necessitates a large number of markers to obtain the same accuracy as with a dozen microsatellites. This puts a considerable strain on machinery intended to deal with variable number of alleles per marker, while the binary nature of typical SNPs allows some computational ‘short cuts’ to be taken, and therefore new tools are required.

Pedigree reconstruction not only entails parentage assignment, but when sampling of candidate parents is incomplete, also clustering of (half-)siblings sharing the same, non-genotyped parent. This can substantially increase the number of within-generation pedigree links (e.g. Walling *et al.*, 2010), but reconstructed sibships are typically unconnected to earlier parts of the pedigree. Assigning grandparents to sibship clusters would overcome this limitation, and involves highly similar comparisons to assigning half-siblings, but is to my knowledge not attempted in any existing method, although methods to assign grandparents to individuals have been described (e.g. VanRaden *et al.*, 2013).

Pedigree reconstruction methods Most pedigree reconstruction methods can broadly be grouped into three categories: Exclusion methods, relatedness based methods, and likelihood based methods, which are of increasing power, but have increasing computational cost as trade-off. The first simply excludes all candidate parents which do not share at least one allele with the focal individual at each marker locus. This method has been used with both microsatellites (see Thompson & Meagher, 1987) and SNP data (Hayes, 2011; Calus *et al.*, 2011). Often some genotyping errors or mutations are allowed for, and the main advantage is that it is very fast. When a very large number of SNPs is used, the number of opposing homozygotes can potentially also be used to differentiate full siblings and half siblings from unrelated pairs (Calus *et al.*, 2011). The major caveat is that when several candidate parents are non-excluded, it provides no way to differentiate between them.

Methods in the second category estimate pairwise relatedness or kinship coefficients between individuals, and use these to categorise the data into first degree relatives, second degree relatives, and unrelated (Glaubitz *et al.*, 2003). In systems with non-overlapping generations and no inbreeding, this may be sufficient to fully reconstruct a pedigree. When generations overlap, additional methods are required to differentiate between parent–offspring pairs and full siblings (both $r = 0.5$), and between half siblings, grandparents, and full aunts/uncles (all $r = 0.25$). Moreover, there is no way to differentiate second degree relatives from double third degree relatives, such as double full first cousins.

In comparison, likelihood methods (the third category) are considerably more powerful (Thompson, 1986; Hill *et al.*, 2008), although computationally notably slower. The likelihood of a particular pedigree configuration is the probability of observing the observed genotypes, conditional on the genotypes of the assigned parents (and the genotyping error rate), summed over all individuals and all loci. It makes use of heterozygous genotypes, which are ignored by exclusion methods, and can be calculated over many individuals jointly, whereas relatedness can only be calculated pairwise. The latter allows more powerful distinction between alternative candidate fathers when one can condition on the genotype of a known mother (Marshall *et al.*, 1998), and enables distinction between the three types of second degree relatives when conditioning on at least one parent each of a pair of individuals (see Appendix A), which is impossible in pairwise likelihoods (Epstein *et al.*, 2000).

Likelihood maximisation Maximising the total likelihood over all individuals is challenging, as the number of possible pedigree configurations increases disproportionally with the number of individuals. A common way to reduce computational cost is to consider only pairwise likelihoods, and find the most likely parent(s) for each individual in turn (e.g. Cervus, (Marshall *et al.*, 1998)). One caveat with this is that close relatives who are not parent and offspring (not PO), may have a higher pairwise likelihood to be PO than to be unrelated (U), and thus a positive log-likelihood ratio $\Lambda_{PO/U}$ (Thompson, 1986). Consequently, there is often considerable overlap in the distribution of $\Lambda_{PO/U}$ of true PO pairs and other types of relatives (Thompson, 1986; Marshall *et al.*, 1998), making it impossible to obtain both high confidence and high assignment rate. However, true full sibling (FS) pairs, for example, are always expected to have an even higher likelihood to be full siblings than to be parent and offspring (Thompson, 1986). Therefore, while $\Lambda_{PO/U}$ and $\Lambda_{PO/FS}$ are necessarily highly correlated, each provides information that the other does not (Thompson, 1986).

Thus, one solution to ensure that one indeed maximises the total likelihood, rather than finding a local maximum for a particular focal individual or pair, is to calculate for each set of candidate relatives the likelihoods under many possible alternative relationships. This is implicit to Kinship (Goodnight & Queller, 1999) and has been implemented in FRANz (Riester *et al.*, 2009), and is implemented even more comprehensively here. One reason for the limited implementation of this approach thus far is the large computational costs involved with calculating likelihoods under many relationship alternatives over a very large number of possible microsatellite genotypes. Moreover, with a typical number of 10-20 microsatellites it is nearly infeasible to distinguish reliably between the various relationship classes. In contrast, with a very large number of SNPs computation is relatively efficient and different relationships can be distinguished reliably.

An alternative solution that has been proposed is efficient partial-updates of the likelihood using a 'Sum-Product algorithm' (Anderson & Ng, 2016). This method deals explicitly with genotyping errors and unsampled individuals, but extension to deal with inbreeding loops is not straightforward (Anderson & Ng, 2016), nor is the post-processing of posterior probabilities to a single pedigree.

Inbred relationships, as well as double relatives, are often excluded from consideration to keep computations feasible and tractable (Wang, 2004; Goodnight & Queller, 1999;

Anderson & Ng, 2016). However, pedigree reconstruction in small populations is regularly performed with the specific aim to study the amount of inbreeding. Moreover, in a range of mammal species, female relatives live together and are therefore likely to mate with the same male (Stopher *et al.*, 2012, and references therein). The resulting offspring are related by more than 0.25, and can therefore easily be mis-classified as full siblings when full sibling, half sibling and unrelated are the only alternatives considered.

Sum up Here, I present an algorithm that compares likelihoods for seven different relationship alternatives, speeded up by exclusion steps based on opposite homozygosity for parentage, and approximate pairwise likelihood ratios of being related versus unrelated throughout. It (1) assigns parents, (2) clusters sibling groups across multiple cohorts, (3a) assigns grandparents to sibships and singletons, and (3b) identifies avuncular links between sibships (Figure 1, where grey symbols denote non-genotyped individuals). Performance is illustrated on three different pedigree structures, including an empirical red deer pedigree from an Scottish island population of red deer (*Cervus elaphus*), with extensively overlapping generations and numerous inbreeding events. I show that several hundred SNPs are sufficient for a high assignment rate ($> 99\%$) and low error rate ($< 0.1\%$).

Methods

Overview

The algorithm reconstructs the pedigree in an iterative, sequential fashion. During each step (described below), pairs of likely relatives are identified, and the likelihoods of seven different relationships between the pair ($H_0 - H_6$ in Table 1) are calculated. If the focal configuration (e.g. paternal siblings) has a higher likelihood than all alternative paternal relationships, by a fixed margin, an assignment is made. Maternal and paternal links are not considered alternatives to each other, as mother and offspring may also be paternal half-siblings, for example. Pairs may consist of two individuals, or an individual and a cluster of half-siblings with their unsampled 'dummy' parent, or two sibship clusters. Dummy individuals, like genotyped individuals, may subsequently be assigned as grandparent to other sibling clusters, providing pedigree links across generations.

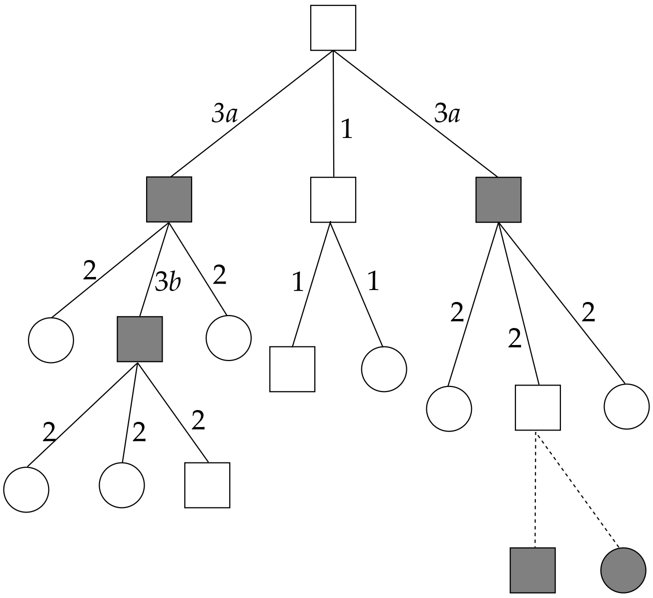


Figure 1: Example part-pedigree with only paternal links shown, and filled symbols indicating non-genotyped individuals. Letters indicate whether the link is inferred during (*1*) parentage assignment, (*2*) sibship clustering (assignment of a dummy parent), (*3a*) assignment of genotyped grandparents to sibships, (*3b*) assignment of dummy individuals as grandparents to other sibships, or (dashed lines) based on non-genetic data only (not by Sequoia). Note that links *3a* and *3b* are not inferred by other programs, which would result in four unconnected pedigree fragments.

Validity of the resulting pedigree is ensured by demanding that 1) an individual cannot be its own ancestor; 2) ancestors are born prior to their descendants, or either or both have an unknown birth year; and 3) the two parents of an individual are of opposite sex, or either one is of unknown sex (i.e., currently no hermaphrodites or asexual reproduction allowed). These simple rules are enforced by only calculating likelihoods over valid relationship alternatives.

Likelihood calculations

The quantity to be maximised is the total likelihood of the pedigree configuration $\boldsymbol{P}$ over all N genotyped individuals,

$$\mathcal{L}(\boldsymbol{P}) = \sum_{A=1}^{N} \mathcal{L}(A, D_A, S_A) = \sum_{A=1}^{N} \prod_{l} P(A_l = X | D_A, S_A) \ , \tag{1}$$

the probability to observe genotype X at locus l in individual A, conditional on its parents D_A and S_A in pedigree P. $\mathcal{L}(\boldsymbol{P})$ is is a simple summation over individuals and loci, as

Table 1: Genealogical relationships considered in this paper and their mean pairwise relatedness r, in absence of inbreeding or additional relationships between the pair of individuals.

	Relationship	Code	mean r
H_1	Parent-offspring	PO	1/2
H_2	Full siblings	FS	1/2
H_3	Half siblings	HS	1/4
	Maternal siblings (full or half)	MS	1/2 or 1/4
	Paternal siblings (full or half)	PS	1/2 or 1/4
H_4	Grandparent – grand-offspring	GG	1/4
H_5	Full aunt/uncle – niece/nephew	FA	1/4
H_{6a}	Half aunt/uncle – niece/nephew	HA	1/8
H_{6b}	Great-grandparent – great-grand-offspring	GGG	1/8
H_{6c}	Full cousins	CC	1/8
H_0	Unrelated	U	0

the probabilities of observing relatives' genotypes are independent conditional on their assigned parents, and it is assumed a set of SNPs can be found which are in low linkage disequilibrium.

The probability $P(A_l = X|D_A, S_A)$ can be broken down into a genotyping error term P_ϵ, a Mendelian inheritance term P_M (denoted transmission probability T in Meagher (1986) and Marshall *et al.* (1998)) and a parental genotype probability term P_P:

$$\mathcal{L}(A, D_A, S_A) = \prod_l \sum_x \sum_y \sum_z$$

$$P_\epsilon(A_l = X|A = x, \epsilon)\times \tag{2a}$$

$$P_M(A_l = x|D_{A_l} = y, S_{A_l} = z)\times \tag{2b}$$

$$P_P(D_{A_l} = y) P_P(S_{A_l} = z)\ . \tag{2c}$$

The first term (2a) is a function of A's actual genotype x and the genotyping error rate ϵ, which is assumed constant across loci. Details of the genotyping error model are given in Appendix A.

The second term (2b) is the probability that individual A inherited actual genotype x from its parents D_A and S_A, conditional on their actual genotypes. This probability can take values of 0, 1/4, 1/2 and 1. Since SNP genotypes can only take 3 possible values (0, 1 or 2 copies of the minor allele), the likelihood components P_ϵ and P_M can be calculated once at initiation and stored in look-up tables, for increased computational efficiency.

In contrast, the parental genotype probabilities P_P (2c) are continuously updated. They give the probability that A's parents carry actual genotypes y and z, and come in four different flavours. When the parent, say D_A is successfully genotyped at locus l, then, according to Bayes' theorem,

$$ {}^{g}P_P(D_A = y|D_A = Y) = P(D_A = Y|D_A = y)P(D_A = y)/P(D_A = Y) \ , \tag{3a} $$

where

$$ \begin{aligned} P(D_A = Y|D_A = y)P(D_A = y) = \mathcal{L}_{l,y}(D_A, D_{D_A}, S_{D_A}) \\ = P_\epsilon(D_A = Y|D_A = y, \epsilon) \sum_v \sum_w P_M(D_A = y|D_{D_A} = v, S_{D_A} = w) \times \\ P_P(D_{D_A} = v)P_P(S_{D_A} = w) \ , \end{aligned} \tag{3b} $$

and

$$ P(D_A = Y) = \sum_{y'} P(D_A = Y|D_A = y')P(D_A = y') \ . \tag{3c} $$

When D_A is not genotyped at locus l (but is genotyped at other loci), the term $P_\epsilon(D_A = Y|D_A = y, \epsilon)$ is omitted from Equation 3b to obtain ${}^{ng}P_P$. When neither parent of D_A is known, this reduces to ${}^{h}P_P(D_A = y|q_l)$, the parental genotype probability among founders. ${}^{h}P_P$ takes the standard values when assuming Hardy-Weinberg Equilibrium of q_l^2, $2q_l(1-q_l)$ and $(1-q_l)^2$.

Sibships and dummy individuals When D_A is a dummy individual, ${}^{d}P_P$ is calculated from the probabilities of observing the genotypes of its assigned offspring if its genotype were 0, 1 or 2 (i.e., the likelihood of the sibship), multiplied by the probability of that genotype conditional on any sibship-grandparents, and scaled to sum to unity.

The likelihood of a sibship $\boldsymbol{A}$ is calculated as

$$\mathcal{L}(\boldsymbol{A}) = \prod_l \sum_x \sum_v \sum_w P_M(D_{\boldsymbol{A}} = x | GM_{\boldsymbol{A}} = v, GF_{\boldsymbol{A}} = w) P_P(GM_{\boldsymbol{A}} = v) P_P(GF_{\boldsymbol{A}} = w) \times$$
$$\prod_{i=1}^{n_A} \sum_{y_i} P_P(S_i = y_i) \prod_{j=1}^{m_{A,i}} \sum_z P_\epsilon(A_{i,j} = Z | A_{i,j} = z, \epsilon) P_M(A_{i,j} = z | D_{\boldsymbol{A}} = x, S_i = y_i) \ , \tag{4}$$

where $GM_{\boldsymbol{A}}$ and $GF_{\boldsymbol{A}}$ are the grandmother and grandfather of sibship $\boldsymbol{A}$, respectively, S_i is the parent of full siblings $A_{i,1} \ldots A_{i,m_{A,i}}$, S_i of opposite sex than $D_{\boldsymbol{A}}$, and sibship $\boldsymbol{A}$ consists of n_A full sib families. The parental probability is then calculated as

$$^dP_P(D_{\boldsymbol{A}} = x) = \frac{\mathcal{L}(\boldsymbol{A}|D_{\boldsymbol{A}} = x)}{\sum_{x'} \mathcal{L}(\boldsymbol{A}|D_{\boldsymbol{A}} = x')}. \tag{5}$$

Note that S_i may also be a dummy parent, in which case the parental probability $^dP_P(S_i = y_i)$ is calculated without the contribution of the joined offspring A_i, to avoid double counting. Most often, the joined likelihood over $\boldsymbol{A}$ and all directly connected sibships is calculated, as $P_P(S_i = y_i)$ will be a function of the presumed genotype of $D_{\boldsymbol{A}}$, and therefore the $P_P(S_i = y_i)$'s of different connected sibships are interdependent (detailed in Appendix B).

Likelihood equations The aforementioned components P_ϵ, P_M and P_P can be combined to calculate the likelihood of observing the genotypes of a group of individuals ($n \geq 1$) under any relationship configuration (equations given in Appendix B). Single locus likelihoods are illustrated in Figure 2 for the special case of two focal individuals, when neither individual has any parent yet assigned. In this case, second degree relatives (HS, GG and FA) can not be distinguished from each other. However, when one can condition on the genotype of a parent or dummy-parent of each individual, such a distinction can be made. Under GG, D_B and A are independent when conditioning on B, while under FA D_B is a grandparent of A, and under HS D_B contributes information about which allele B inherited from the unobserved parent S_B, for loci at which B is heterozygous. (details in Appendix C). In addition, age information may be used (Appendix D). Equations for some likely (inbred) combinations of relationships are given in Appendix E.

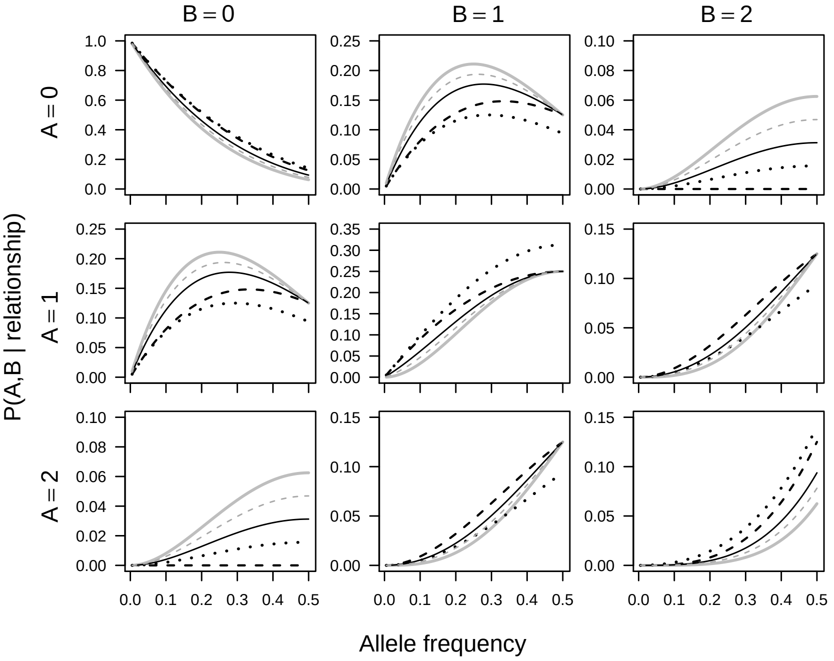


Figure 2: Single-locus probability of observing genotypes A and B (0, 1 or 2 copies of the minor allele) as a function of the minor allele frequency q, under the hypotheses U (solid grey line), PO (dashed black), FS (dotted black), HS, GG or FA (solid black, indistinguishable from each other), or HA or GGG (dashed grey). Formulae for the lines are given in Appendix B, Equations 7–16.

Age and sex information, and absence thereof

The age difference between individuals can be highly informative to distinguish between for example parents and full siblings, or between grandparents and half siblings. Sequoia makes use of an age-difference based prior, which in its simplest form is an indicator whether a given relationship is possible (1) or not (0) given the age difference between the two individuals. Optionally, after parentage assignment the empirical age distribution of fathers and mothers and between maternal and paternal siblings is used as prior to assist subsequent sibship clustering (Appendix D). It corrects for any age structure in the sample by using the fraction of all pairs of genotyped individuals born t years apart that are (say) mother-offspring pairs. For each hypothesised relationship, the genetic-based likelihoods are multiplied by these age-difference based prior probabilities, i.e. genotypes and age differences are treated as independent sources of information.

When the age difference between a candidate parent-offspring pair (say A and B) is unknown, an assignment can often still be made. When individual A already has a parent D_A assigned, of different sex than B, the likelihood of B being a joined parent of A will differ from B being an offspring of A. Alternatively, if A has no parents but does have offspring, B must be at least 2 years (time-units) older than A's offspring to be assigned as A's parent.

When B is of unknown sex, an assignment can similarly be made, as the likelihood to be joined parent with D_A will differ from the likelihood to replace D_A.

Algorithm steps

An overview of the program is given in Figure 3. First, the genotype file will be converted if necessary, and a file with parameter values will be created. A check for duplicate identities and genotypes is performed, to avoid downstream problems.

Then, several iterations of parentage assignment are performed, and results returned for user inspection. Parentage assignment is fast (see Results), and can therefore be easily incorporated as a step of data quality control. To this end, a list of identified parent-offspring pairs which could not be assigned due to absent or incompatible age or sex information is returned as well.

The speed is achieved by use of opposite homozygosity as a filtering step, a computationally fast method to dramatically reduce the number of potential parent-offspring pairs (Hayes, 2011; Hill *et al.*, 2008). By default a liberal threshold of $T_{OH} = 3 + \epsilon L$ is used to avoid exclusion of true PO pairs, which will non-exclude some pairs of non-PO close relatives, particularly FS pairs (see Calus *et al.*, 2011).

A second filtering step for parentage assignment, and the only filtering for the sibship clustering steps, consists of calculating the likelihood ratio $\Lambda^*_{R/U}$ between the focal relationship R and unrelated U, without conditioning on any already assigned parents or close relatives. A liberal, log-scale negative threshold is used to again avoid exclusion of true relatives.

For each individual in turn, from earliest born to last born, it then calculates the likelihood of the focal individual and candidate parent being U, PO, FS, HS, GG, FA, or HA/GGG ($\mathcal{L}_{H0}$–$\mathcal{L}_{H6}$, Table 1). If there are multiple candidate parents, these likelihoods are calculated conditional on none, one and both earlier assigned parents (details in Appendix F). The same likelihoods are calculated for the focal individual and the earlier assigned parents, conditional and unconditional on the new candidate parent. All results are scaled to reflect the joined likelihood over the up to four individuals, and par-

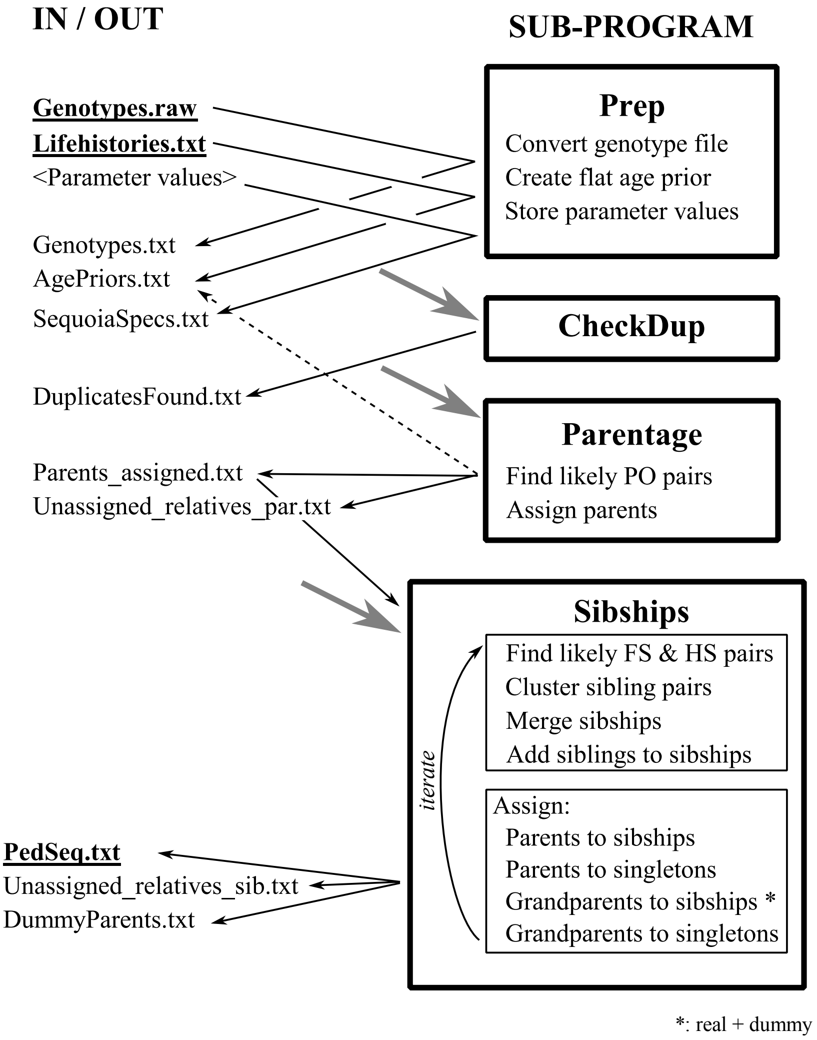


Figure 3: Overview of the various parts of the program.

ent assignment made according to the highest likelihood (which may include removal of earlier-assigned parents). This approach maximises assignment rate, and minimises the chance that for example double-grandparents are assigned as parents.

The same approach is followed to assign grandparent to sibships.

During each iteration of sibship clustering, first all pairs of likely HS and FS are identified (Figure 3), using $\Lambda^*_{HS/U}$ followed by calculation of $\mathcal{L}_{H0}$–$\mathcal{L}_{H6}$. Clustering is performed by for each candidate pair in turn either using it to found a new sibship, when neither individual had yet been assigned to a sibship of type k (maternal or paternal); adding it to an existing sibship, when one of the pair was already assigned to a sibship of type k in an earlier step, or using it to merge two existing sibships, when the pair members were previously assigned to different sibships of the same type. Each step is only followed through when $\mathcal{L}$(FS) or $\mathcal{L}$(HS) , calculated over the pair and all putative siblings, exceeds the likelihoods under all other relationships.

After clustering, all sibship pairs of the same type are considered for merging, using Λ^*(merged/separate) based on stored values of $\mathcal{L}(\boldsymbol{A}|D_A = x)$ and $\mathcal{L}(\boldsymbol{B}|D_B = y)$, followed by calculation of $\mathcal{L}_{H0}$–$\mathcal{L}_{H6}$ over the two sibships. The relationships between sibships $\boldsymbol{A}$ and $\boldsymbol{B}$ can be expressed as the relationship between A_1, A_2, $\ldots A_n$ and dummy parent ${}^dD_{\boldsymbol{B}}$, so that the considered relationships become identical to the relationships considered for pairs of individuals, as listed in Table 1.

Subsequently, all individuals who lack a parent of type k are considered for addition to each sibship of type k. This ensure that individuals who are ambivalent with respect to their relationship towards other lone individuals, may get clustered due to the combined 'evidence' provided by those in an existing sibship.

Grandparents Lastly in each iteration, grandparents are assigned to each sibship, in a process highly similar to parentage assignment. This step is omitted in the very first iteration, as this proved to reduce false assignments. From the second iteration onwards, also GG pairs are identified, and assigned as sibship clusters with one member. In subsequent iterations, additional siblings may be assigned to this sibship. It may not have been possible to assign these as siblings earlier, because for example without the assigned grandparent they were indistinguishable from an FA pair.

Assignment confidence

In the returned pedigree, a LLR(parent/not parent) is associated with each assigned parent and dummy parent, for each genotyped and dummy individual. These are calculated conditional on all other pedigree links, by, for each individual in turn, stripping its (dummy)parents, and re-calculating $\mathcal{L}_{H0}$–$\mathcal{L}_{H6}$ with the assigned parent or sibship. The LLR for the parent pair is calculated relative to the highest likelihood scenario with one or neither parent assigned. For dummy individuals, a similar approach is followed with respect to the sibship grandparents; the assigned offspring in the sibship are always all conditioned on.

Confidence probabilities could be obtained using a bootstrap approach, as is common in the inference of phylogenies since proposed by Felsenstein (1985). In each bootstrap iteration, L of the L loci of the genotype data are sampled within each individual, *with* *replacement*, followed by re-reconstruction of the pedigree. However, since this would further increase the dependence between (bootstrapped) markers, its accuracy is unclear.

A global estimate of the error rate over all individuals can be obtained using a simulation approach as in Cervus (Marshall *et al.*, 1998). This simulation generates founder genotypes according to population allele frequencies, and simulates the remaining genotypes based on the pedigree from observed data ($\boldsymbol{P}_{obs}$) (described below, 'Simulation of SNP data'). The average number of mismatches between pedigree reconstructed from simulated data ($\boldsymbol{P}_{sim}$) and $\boldsymbol{P}_{obs}$ then provides an estimate for the number of mismatches between $\boldsymbol{P}_{obs}$ and the true pedigree.

Datasets

The algorithm was tested on simulated datasets generated from three different pedigree structures, described below, in order to give a general indication of performance. For each pedigree, after simulation of genotype data, a varying proportion of parental genotypes was discarded in order to assess sibship clustering. In addition, the algorithm was run on the large empirical SNP dataset from which Pedigree III was derived, and resulting pedigree relatedness estimates compared to genomic estimates of relatedness.

Pedigree I: Full sib families Pedigree I consists of 1157 genotyped individuals in a single generation, divided over 432 full sib families with 1–11 individuals each (mean 2.68, 143 singletons). It is identical to the pedigree structure used in Anderson & Ng (2016) to compare performance of Colony (Wang, 2012) and Fullsnplings (Anderson & Ng, 2016), and derived from an empirical salmon dataset.

Pedigree II: Multi-generational half-sib The second pedigree consisted of five non-overlapping generations in an small closed population, with full sib families nested within interconnected half sib clusters. Each female mated with 2 random males and each male with 3 random females, producing 4 full-sib offspring per mating. Each generation, 24 female and 16 male breeders were drawn at random from the 192 offspring born, and matings between full or half sibs were allowed. This simulated pedigree is provided with the R package. Setting a proportion (0–1) of parental genotypes to missing mimics a situation in which from each (potentially large) full-sib family only a few individuals are

Table 2: Total number of individuals in various categories for each Pedigree. Ped I consists of a single generation of full-sib families, Ped II of 5 discrete generations of full- and half-sib families, and Ped III is the empirical pedigree of the 17 most recent birth year cohorts of a wild Red deer population.

	Ped I	Ped II	Ped III
Total	2021	1000	1998
Mother known	1157	960	1642
Father known	1157	960	1202
Unique mothers	432	80	462
Unique fathers	432	120	193

genotyped, some of which may become parents of the next generation.

Pedigree III: Red deer Lastly, I used the pedigree from the study population of wild red deer (*Cervus elaphus*) on the Isle of Rum (Clutton-Brock *et al.*, 1982). This pedigree is characterised by extensively overlapping generations, immigration of males born elsewhere on the island, and numerous instances of close and moderate inbreeding. In addition, many females live in matrilinial groups, and consequently female relatives tend to mate with the same male more often than expected under random mating (Stopher *et al.*, 2012).

Recently, 2517 individuals were genotyped for nearly 40 000 polymorphic SNPs (Huisman *et al.*, 2016), and 440 SNPs with minor allele frequency above 0.4 and in low linkage disequilibrium with each other were selected for pedigree inference using Sequoia. The resulting pedigree was compared to and supplemented with parents from observational data, and from the previous microsatellite based pedigree (Walling *et al.*, 2010), which was reconstructed using MasterBayes and Colony. The full pedigree includes a total of 4439 individuals, born over a period of nearly 50 years. Out of time considerations, only the last 17 birth year cohorts and their parents were used as the basis for most simulated datasets, comprising 1998 individuals (Ped III). This pedigree was also used to investigate the joined effect of sample size and pedigree depth on performance, for which subsets consisting of the most recent 3, 8, 17 or 27 birth year cohorts and their parents were used, with respectively 503, 1003, 1998 and 3000 individuals (a few random individuals from the preceding cohort were added to create approximately round numbers).

Simulation of SNP data To simulate SNP genotypes, each pedigree was split into generations, where those in generation 1 had neither parent known, those in generation 2 had either both parents in generation 1, or one parent in generation 1 and one unknown parent, and those in generation g had parents in generations $< g$, and possibly an unknown parent. Founder genotypes ($i_l = 0$, 1 or 2) were simulated by drawing twice from a binomial distribution, with probabilities equal to the frequency q_l of the reference allele, in turn drawn from a uniform distribution between 0.35 and 0.5 (mimicking a selected subset of highly informative SNPs). Loci were simulated as unlinked, and in linkage equilibrium amongst founders. For subsequent generations, the parental inherited alleles were drawn from a binomial distribution with probability equal to half the parental genotype if known, and probability q_l otherwise. Data was made more realistic by setting 0.5% of individual locus genotypes as missing, and replacing 0.5% of genotypes by a random genotype, which may or may not be identical to the original one ($\epsilon \approx 0.005$, following Anderson & Garza (2006)). The function to simulate genotype data from a known pedigree is included in the R package.

Assignment and error rates The assignment rate (AR) and error rate (ER) were calculated as the number of individuals with a correctly respectively incorrectly inferred parent, divided by N_k, the number of individuals with a parent of sex k in the true pedigree, averaged over maternal and paternal links. A sibship parent, say dummy father, was deemed correct if the majority of inferred paternal HS were true paternal HS. For both erroneous merging and erroneous non-merging, the error count equals the size of the smaller of the two sibships. The partition distance (PD), the minimum number of elements (individuals) that must be deleted, so that the two induced partitions (the true and inferred sibships) are identical (Gusfield, 2002; Anderson & Ng, 2016) equals the sum of false positives and false negatives $PD = (ER + (1 - AR)) * N$.

Results

Distribution of LLR

As expected, the simulated distribution of the LLR(PO/most likely not PO) showed a clearer divide between true PO pairs and non-PO pairs than did LLR(PO/U) (Figure 4,

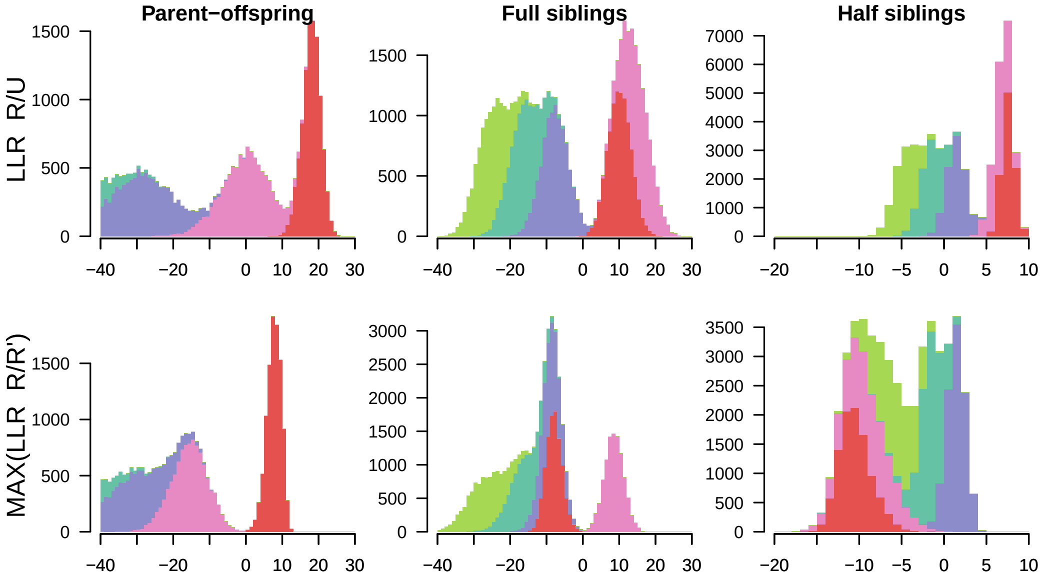


Figure 4: Distributions of the pairwise log10-likelihood ratios between a focal relationships (headers) versus unrelated (top row) and versus the most likely alternative (bottom), for pairs that are PO (red), FS (pink), HS (blue), HA (blue-green) or U (green), not conditioning on any parental genotypes. Based on 10.000 simulations of a simple pedigree with unrelated founders and 400 SNPs with MAF 0.3–0.5 and $\epsilon = 0.005$.

left panels). A similar pattern is apparent for FS (middle), but not for HS (right), where the distributions of HS and HA show considerable overlap at LLR > 0. However, when both parents of both individuals are unknown, as assumed in Figure 4, no HS assignments can be made as it is impossible to distinguish between maternal and paternal HS, and the likelihoods for HS, FA and GP are identical. When at least one parent of each individual is known, there is strong but imperfect segregation between the distributions of HS and HA (Figure 13).

The LLR threshold(s) for an optimal trade-off between AR and ER will depend on the proportions of different categories of relatives in the sample, which by definition are not known *a priori*. Therefore, results will be shown using the same thresholds across all simulations, of $T_{Filter} = -2$ for LLR(R/U) to filter out pairs which are highly unlikely to be relatives, and $T_{Assign} = +0.5$ for LLR(R/most likely not R) to make assignments. Use of different combinations of thresholds had minor effects on AR and ER, in different patterns across the three pedigrees (Figure 14).

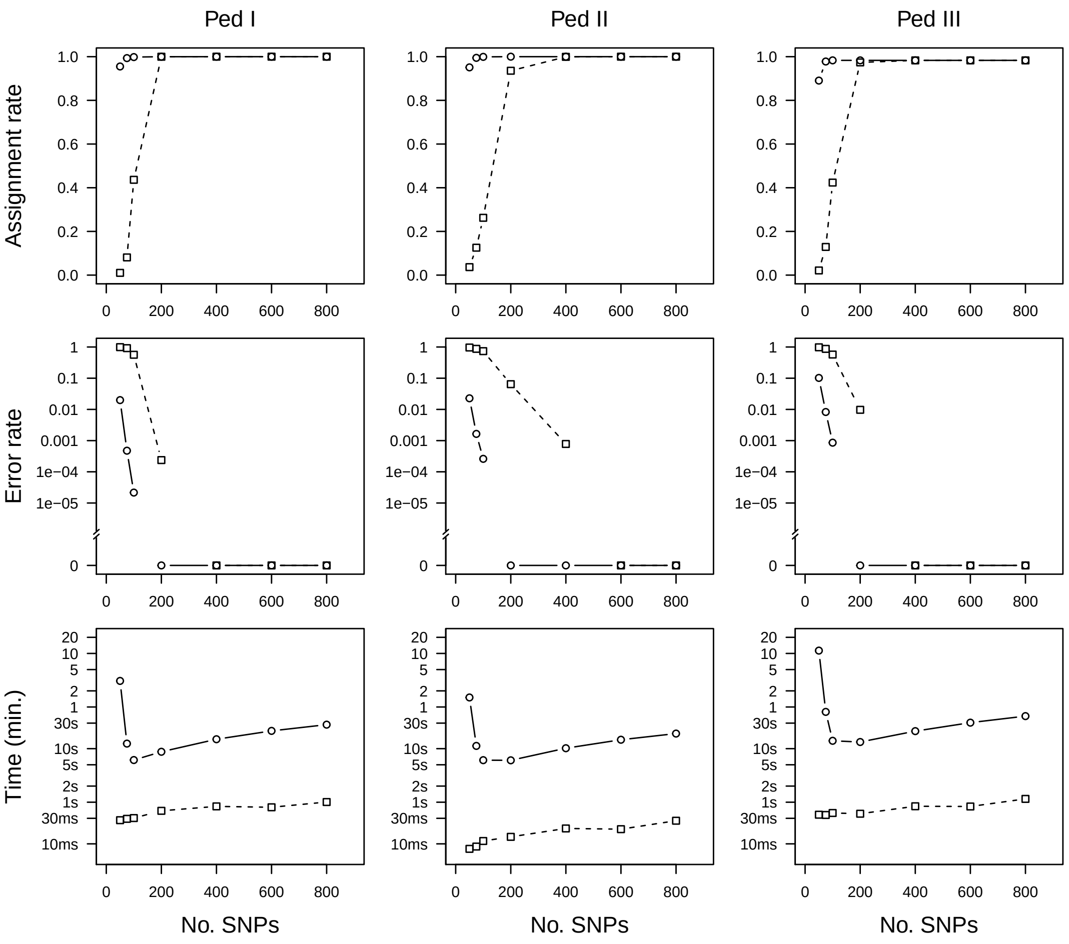


Figure 5: Effect of SNP number (x-axes) on parent assignment rate (top), error rate (middle) and computation time (bottom), when all parents are genotyped. Columns show Pedigree I, a single generation of full-sib families, Pedigree II, a five generation pedigree with full-sib families within interconnected half-sibships, and Pedigree III, based on en empirical red deer pedigree. Open circles indicate Sequoia (parentage assignment only), open squares exclusion; each point denotes the average over 10 independent simulations.

Parentage assignment

Assignment of genotyped parents to genotyped offspring was highly accurate, with assignment rate (AR) > 99.8% in Pedigrees I and II and error rate (ER) <0.1% when at least 100 SNPs are used and all parents are genotyped (Figure 5). AR is slightly lower (98.3%) in Pedigree III due to 12 parents with unknown birth year. Simply assigning the first non-excluded parent of each sex (opposite homozygotes at < 3 SNPs) performs as good as the combined likelihood approach when at least 400 SNPs are used, or at least 600 when SNPs the data contains many full siblings (Pedigree II), but performs extremely poorly when 100 SNPs or less are used. In contrast, Sequoia gives good ARs (89%–95%)

and reasonable ERs (2-10%) even when as few as 50 SNPs are used (Figure 5).

Computation time Exclusion is considerably faster than Sequoia at 0.1 – 1.2 seconds, compared to 6 – 30 seconds for Sequoia under most conditions, up to 11 minutes for 50 SNPs and Pedigree III (on a laptop with a quadcore intel i7 2.3 GHz processor and 8 GB RAM). Both are much faster than MCMC based approaches such as MasterBayes, which take many hours for large datasets with thousands of individuals and several hundred SNPs (C. Berenos, pers. comm.).

Sequoia's computational time has a minimum around 100–200 SNPs, both for parentage assignment (Figure 5) and in general (Figures 6 – 7). With limited information, many pairs cannot be excluded from being PO (or FS, HS), and the more intensive likelihood calculations over H0–H6 are performed for a larger number of pairs. Beyond this minimum, computation time increases approximately linearly with marker number.

Clustering of full sib families

Clustering of full sib families within a single generation, without any parental genotypes, gave high ARs ($> 98.4\%$) and low ER ($< 0.1\%$) when at least 200 SNPs were used (Figure 6). Even at high marker numbers some FS are erroneously inferred as HS (Figure 15), when the likelihood of the father(s) to be two FS was similar to the likelihood for it to be a single individual.

Using Colony with its monogamous breeding system option gave both higher AR and (much) lower ER (Figure 6), especially when the number of SNPs was less than 200, at a higher but acceptable computational cost. Assuming a polygamous breeding system, however, resulted in ER=0.0044 with 100 SNPs, and a running time of 79 minutes.

Combination of parentage assignment and sibship clustering

Assignment rate The combination of parentage assignment, sibship clustering and grandparent assignment resulted in reconstruction of 99% of parent-offspring links in Pedigree II when 40% of parental genotypes was treated as unknown (Figure 7, Table 3). AR was somewhat lower in Pedigree III, at 86%–89%, partly because for some identified likely half-sib pairs it could not be determined whether they were paternal or maternal half siblings, or a full avuncular pair. This is also the main reason of the low AR (56%)

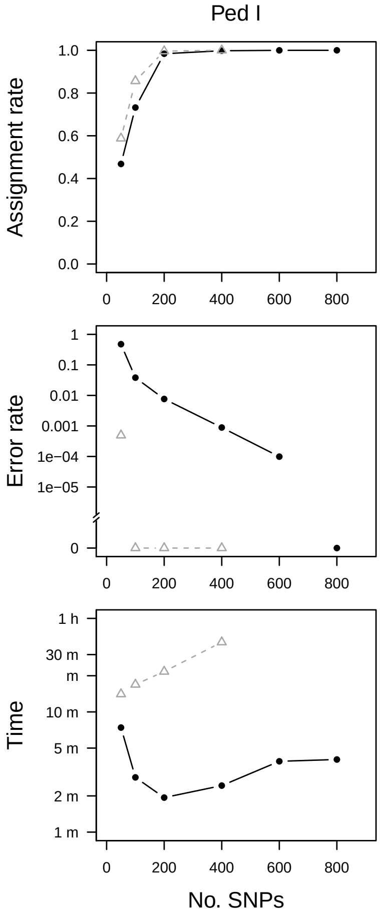


Figure 6: As Figure 5, for Pedigree I with no genotyped parents. Filled circles indicating results from Sequoia (parentage + sibship clustering) averaged over 10 simulations, and grey triangles from Colony (monogamous breeding system, FPLS) averaged over 3 simulations.

when only 20% of parents is genotyped in Pedigree III (Figure 8). Generally, however, sibship clustering and grandparent assignment ensured that the proportion of individuals with a (real or dummy) parent assigned was considerably higher (AR=82% – 99% for 20%–80% of parents genotyped) than would be feasible with parentage assignment alone (AR=17%–78%) (Figure 8).

Error rate Error rates were higher for the complete pedigree reconstruction than for parentage assignment, and therefore higher when a larger proportion of parents was non-genotyped (Figure 8). Nonetheless, ER was low (0.1%–0.3%) for both pedigrees when at least 200 SNPs were used and 60% of parents was genotyped (Figure 7), and was maximum 2.5%, when 20% of parental genotypes were simulated as known Pedigree II (Figure 8).

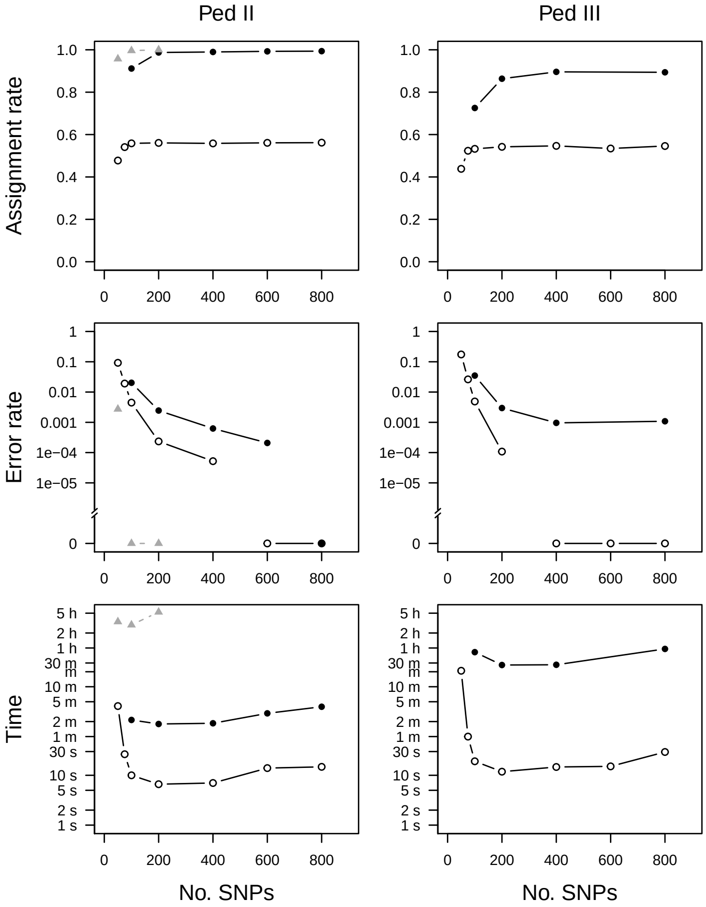


Figure 7: As Figures 5 and 6, with 60% of parents genotyped. Open circles indicate parentage assignment with Sequoia, filled circles full pedigree reconstruction with Sequoia; each point denotes the average over 10 independent simulations. Grey triangles show results from Colony (polygamy, FPLS) run separately on generations 2 and 5, and results averaged (and time multiplied by 5).

Similar to the reconstruction of full sib families, Colony resulted in higher AR than Sequioa when fewer than 200 SNPs were used with Pedigree II, and slightly lower ER throughout (Figure 7). However, computational times were considerably longer, at around an estimated 3 hours for Colony (based on separately running generations 1 and 5 only), compared to 2–4 minutes for Sequoia. Computational times for Pedigree III ranged from 30–60 minutes, due to the larger number of individuals as well as the considerably more complex pedigree structure.

Computation time As parentage assignment is much less computationally intensive than sibship clustering, computational time decreased with the fraction of genotyped parents (Figure 8). Computational time increased with the number of individuals in an approximately quadratic fashion, as one of the most computationally intensive steps is

Table 3: Counts of not inferred and erroneously inferred links per pedigree for a range of marker numbers. PD = Partition distance = not-inferred + erroneous.

Pedigree	SNPs	Total no. links	not assigned	wrong assigned	PD	Colony not ass.	Colony wrong ass.
Ped I	100.0	2028.0	420.8	77.4	498.2	291.3	0.0
	200.0	2028.0	5.9	15.5	21.4	8.0	0.0
	400.0	2028.0	0.2	1.8	2.0	0.0	0.0
	800.0	2028.0	0.0	0.0	0.0		
Ped II	100.0	1920.0	32.0	38.9	70.9	1.5	0.0
	200.0	1920.0	13.0	4.7	17.7	0.0	0.0
	400.0	1920.0	14.3	1.2	15.5		
	800.0	1920.0	12.3	0.0	12.3		
Ped III	100.0	2844.0	328.5	96.1	424.6		
	200.0	2844.0	213.4	8.6	222.0		
	400.0	2844.0	158.9	2.6	161.5		
	800.0	2844.0	167.6	2.9	170.5		

the identification of sibling pairs among the approximately $N^2/2$ pairs of individuals. Another strong determinant of the computational time is the number of sibships S, as the merging step will consider around $(S/2)^2$ combinations, and the subsequent step adding individuals to existing sibships will consider in the order of SN combinations.

Empirical red deer dataset

For the red deer data, as for most empirical data, many mothers are known from observations with high accuracy, while the true fathers are unknown or uncertain. As a proxy for the true pedigree relatedness between pairs of individuals, I used the genomic relatedness estimated by GCTA (Yang *et al.*, 2011) from all 40.000 SNPs (y-axes in Figure 9), and compared this to the relatedness estimated from the pedigree reconstructed using 440 selected SNPs (Figure 9b). The correlation was higher than in the same comparison using the previous previous microsatellite-based pedigree (Figure 9a), although note that this pedigree does not include individuals born in 2012–2015. It was also higher than the correlation with the genomic relatedness estimated from the 440 selected SNPs (Figure

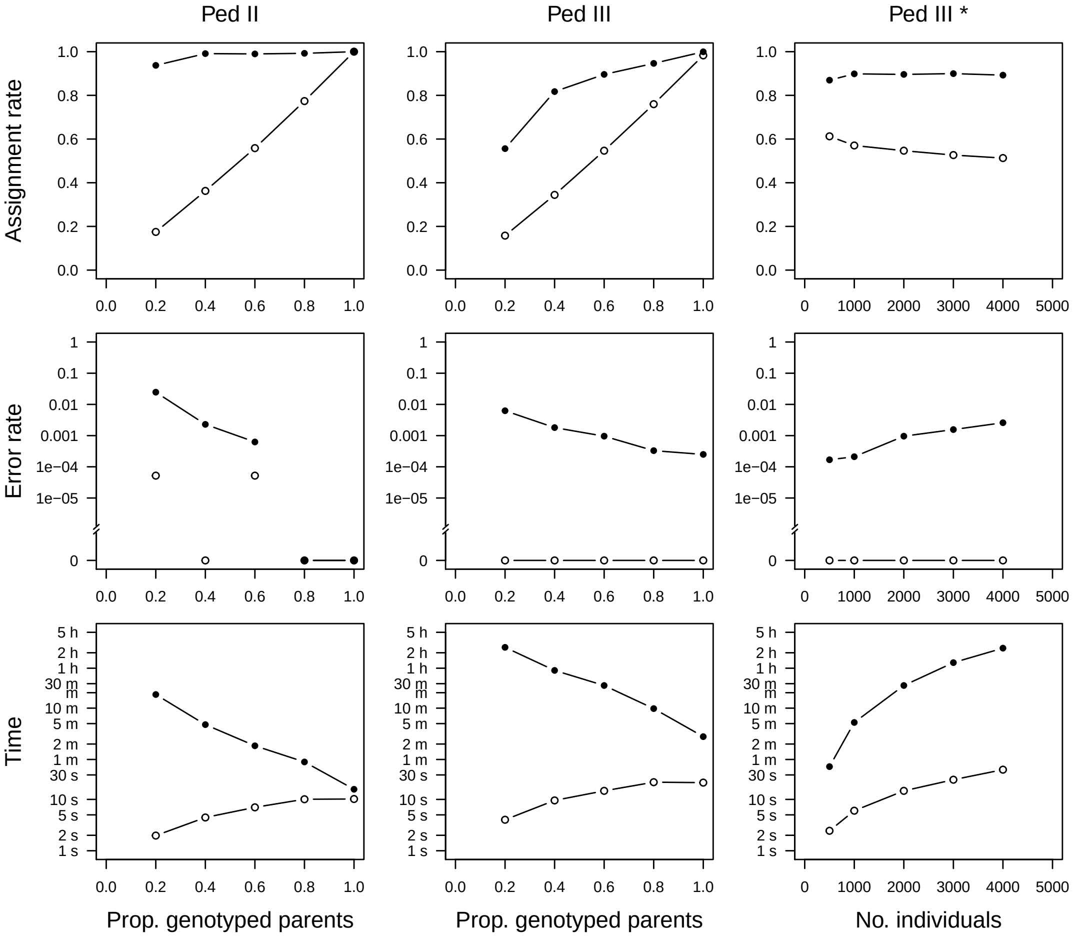
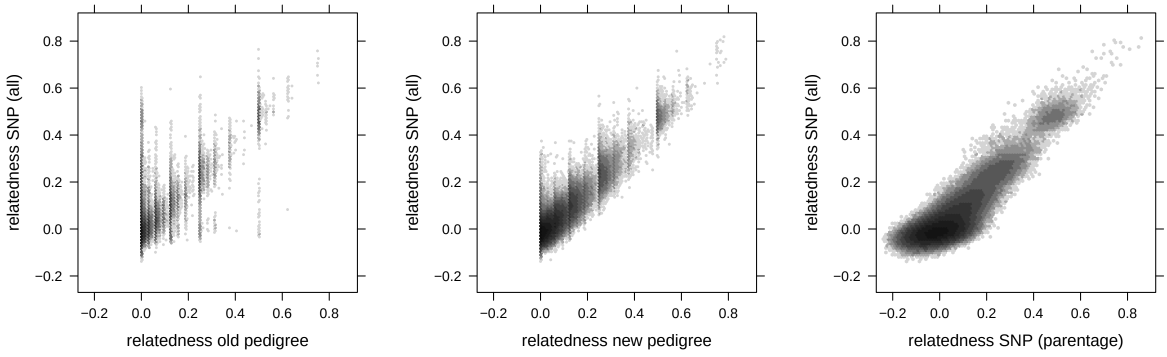


Figure 8: Effect of the proportion of genotyped parents in Pedigrees II and III, as well as the effect of pedigree size (more individuals confounded with a deeper pedigree; 60% of parents genotyped) on AR, ER and computational time; averages over 10 replicates are shown. Symbols as in Figure 7.

9c), which may partly be an artefact of the different average allele frequencies in the two sets of markers, but is largely due to Mendelian noise.

Discussion

Using Sequoia, reliable pedigree inference is possible even with complex mating structures, extensively overlapping generations, and inbreeding. Parentage assignment performs very well down to about 100 independent SNPs, while for sibship clustering a few hundred SNPs are required. For these marker numbers, false positive rates are low (< 0.1%) and assignment rates high (> 99%), for all three simulated pedigree structures considered. Due to its speed, pedigree reconstruction using Sequoia can be used as a part of data quality control even for very large datasets, to ensure samples of supposed relatives are



(a) $r = 0.66$, $n = 1.4 \times 10^6$ (b) $r = 0.81$, $n = 3.7 \times 10^6$ (c) $r = 0.68$, $n = 3.8 \times 10^6$

Figure 9: Pairwise relatedness in an empirical red deer dataset, as estimated from 40,000 polymorphic SNPs using GCTA (y-axes), and (a) a previous microsatellite based pedigree, (b) from the pedigree inferred using Sequoia on 440 SNPs with high MAF and in low LD, or (c) from these same 440 SNPs using GCTA. n denotes the number of pairwise relationships, related to the number of unique individuals i as $n = i \times (i - 1)/2$

indeed assigned as such.

Comparison to MCMC methods The main advantage of Sequoia over existing methods is that through explicit consideration of all possible relationship alternatives before making an assignment, including inbred configurations, both high assignment rate and high accuracy can be obtained for complex pedigrees. The time needed for each iteration is much longer than in a typical MCMC iteration, but since the likelihood asymptotes within only a few iterations, total computation time is similar to Colony for simple pedigrees, and orders of magnitude faster for complex, interconnected pedigrees (Figures 6–7).

On the flip side, Sequoia does not benefit from the strengths of MCMC based methods, such as its reliable estimation of assignment confidence. The likelihood ratio between the focal relationship and all other relationships is a good indicator whether the focal relationship is the true one when the number of independent markers is large, as shown by the low error rates found. However, when the number of markers is limited, Mendelian noise can be substantial, such that the true configuration may not result in the highest partial likelihood. Well-written MCMC algorithms will not get stuck on such local maxima, while our current algorithm can only go 'uphill' when clustering siblings. Consequently, for low marker numbers Sequoia is out-competed by MCMC based approaches such as Fullsnplings Anderson & Ng (2016) and Colony (Wang, 2012), which can reconstruct noninterconnected half-sib families with high accuracy ($> 95\%$) based on just a few dozen SNP markers (Wang, 2012). However, with the advance of SNP genotyping technology and steadily decreasing costs, I expect that in many situations an abundance of markers will be available, although for species with small genomes the number of semi-independent markers will be restricted.

Parentage assignment When interest is solely in parentage assignment, the approach presented here can dramatically decrease the false assignment rate in the presence of close relatives of the true parent, compared to methods based on the number of opposite homozygous loci alone (Figure 5). Consequently, it reduces the number of markers required to obtain a given accuracy. In a study across various livestock breeds, 200–400 SNPs with high MAF were required to fully separate parent-offspring pairs from other pairs based on the number of opposing homozygous loci only (Strucken *et al.*, 2015), while Sequoia requires around 75 SNPs (Figure 5), and the likelihood-based MCMC method 'snpSumPed' requires 60-100 SNPs (Anderson & Garza, 2006).

One additional benefit of Sequoia is that it does not require the user to create per-cohort lists of candidate parents, as Colony (Wang, 2004) and MasterBayes (Hadfield *et al.*, 2006) do. Creating such lists can be time consuming, and the true parent may erroneously be left out. In the empirical red deer dataset, I identified several males which were genotyped at birth, emigrated from the study area as sub-adults and returned many years later to sire offspring, but were not recognised upon return. Consequently, these genotyped males were never considered as candidate fathers, but instead existed with a dual identity.

Sibship clustering It has been observed that likelihood scores tend to favour more complex explanations (Thomas & Hill, 2002; Almudevar, 2007), resulting in splitting true sibling groups (Almudevar, 2007) as well as creation of spurious sibling groups (Anderson & Ng, 2016). With Sequoia, the pairwise error rate due to inference of spurious siblings was orders of magnitude lower than non-assignment of true siblings (Figure 15). Non-assignment as full siblings was predominantly due to a limited likelihood difference for true full siblings to be full siblings versus maternal half-siblings and paternal full cousins ($r = 1/4 + 2/16$), a configuration which will be rare in many systems. Possibly a-priori estimates of the fraction of pairs in each type of relationship (PO, FS, HS, CC, etc.) could lessen this problem, although sensitivity to this additional parameter should be carefully considered.

In the empirical red deer dataset, Sequoia identified more siblings links than were identified previously using Colony (Walling *et al.*, 2010). Many of these were the consequence of being able to analyse all birth year cohorts jointly, such that the offspring of non-sampled males siring one or two offspring per year were now clustered together. With Colony each cohort is analysed separately, and while several birth years may be analysed together using a sliding-window approach, combining the results into a single pedigree is hindered by the presence of erroneous sibship clusters, and the lack of concordance between a sibship's posterior probability and it's correctness (Anderson & Ng, 2016). Moreover, Colony's computational time increases more than linearly with sample size, such that joined analysis of more than four birth years of the red deer dataset was impractical due to computational times of several days.

Grandparents In the red deer dataset, grandparent assignment identified parents for several non-genotyped males, and in many cases the dummy male could be linked to an observed male based on association with the mothers during the preceding mating season (following Walling *et al.*, 2010). Various paternal links between males not born in the study area were identified, so that analyses based on the previous microsatellite-based pedigree (Walling *et al.*, 2010) violated the assumption that all founders and immigrants are unrelated. Nonetheless, the effect of these pedigree improvements on for example heritability estimates is expected to be generally small, as the existing pedigree was already very extensive, and the relatedness between immigrants a minor source of error. However, similar effects are expected to be more substantial when the proportion of sampled parents is lower, or when the pedigree is shallower.

Future directions One problem with pedigree reconstruction is the differentiation between maternal and paternal relatives. When the sex of a candidate parent is unknown, it may be deduced from the sex of the other parent when assigned as part of a parent-pair. Similarly, half-sibling pairs can be identified as being maternal or paternal if at least one has a parent of known sex assigned. However, in many cases distinction is not possible, and for example for a full sib family ($n \geq 1$) where both parents are non-genotyped, it is impossible to determine whether a grandparent is a parent of the mother or the father.

One solution would be to incorporate data on sex linked markers, which are inherited differently via the maternal than paternal line.

In addition, one could incorporate prior information on observation-based parents during sibship clustering, as is done in Colony (Wang, 2012) and MasterBayes (Hadfield *et al.*, 2006). An advantage of the current approach, however, is that it allows the user to spot mismatches between genetic and observational or age data, as they are not combined during parentage assignment. Moreover, it is not directly obvious how to extend the current algorithm to incorporate this type of data.

An alternative way to increase performance when a large number of markers is available, is to base pedigree inference on the length and distribution of genome segments shared between individuals, which theoretically is a more powerful approach than methods based on single markers (Hill & White, 2013). It enables for example distinction between half-sibs and avuncular pairs, without the need for genotyped close relatives or age information. Nonetheless, it is expected that for non-model organisms, genotyping anonymous SNPs using, for example, RAD-sequencing will be more affordable and common place than whole-genome sequencing and genome assembly. Methods to construct linkage maps are currently only available for large full-sib families (Rastas *et al.*, 2013), and would always require an pedigree as input, although methods to jointly estimate a pedigree and map positions are theoretically conceivable.

Acknowledgements

I wish to thank J Hickey for motivating me to re-learn Fortran, J Pemberton for the liberty to pursue this side-project and use of the red deer data, J Pemberton, T Ellis and J Hadfield for comments on the manuscript, the International Centre for Mathematical Sciences for organising an inspirational workshop on pedigree reconstruction, and NERC for funding.

References

Almudevar A (2007) A graphical approach to relatedness inference. *Theoretical population biology*, **71**, 213–229.

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Anderson EC, Ng TC (2016) Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation. *Theoretical population biology*, **107**, 39–51.

Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Molecular ecology*, **23**, 3434–3451.

Calus MPL, Mulder HA, Bastiaansen JWM (2011) Identification of mendelian inconsistencies between snp and pedigree information of sibs. *Genet. Sel. Evol*, **43**, 34.

Clutton-Brock TH, Guinness FE, Albon SD (1982) *Red deer: behaviour and ecology of two sexes.* University of Chicago Press.

Epstein MP, Duren WL, Boehnke M (2000) Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, **67**, 1219–1231.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, S. 783–791.

Glaubitz JC, Rhodes OE, DeWoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039–1047.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**, 1231–1234.

Gusfield D (2002) Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, **82**, 159–164.

Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a bayesian framework. *Molecular Ecology*, **15**, 3715–3730.

Hayes BJ (2011) *Technical note:* efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *Journal of dairy science*, **94**, 2114–2117.

Hill WG, Salisbury BA, Webb AJ (2008) Parentage identification using snp genotypes: application to product tracing. *Journal of animal science.*

Hill WG, White IMS (2013) Identification of pedigree relationship from genome sharing. *G3: Genes— Genomes— Genetics*, S. g3–113.

Huisman J, Kruuk LEB, Ellis P, Clutton-Brock TH, Pemberton JM (2016) Inbreeding depression across the lifespan in a wild mammal population. *PNAS.*

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular ecology resources*, **10**, 6–30.

Kruuk LEB, Hadfield JD (2007) How to separate genetic and environmental causes of similarity between relatives. *Journal of evolutionary biology*, **20**, 1890–1903.

Marshall TC, Slate JBKE, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular ecology*, **7**, 639–655.

Meagher TR (1986) Analysis of paternity within a natural population of chamaelirium luteum. 1. identification of most-likely male parents. *American Naturalist*, S. 199–215.

Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 613–621.

Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P (2013) Lep-map: fast and accurate linkage map construction for large snp datasets. *Bioinformatics*, S. btt563.

Riester M, Stadler PF, Klemm K (2009) Franz: reconstruction of wild multi-generation pedigrees. *Bioinformatics*, **25**, 2134–2139.

Stopher KV, Nussey DH, Clutton-Brock TH, Guinness F, Morris A, Pemberton JM (2012) Re-mating across years and intralineage polygyny are associated with greater than expected levels of inbreeding in wild red deer. *Journal of Evolutionary Biology.*

Strucken EM, Lee SH, Lee HK, Song KD, Gibson JP, Gondro C (2015) How many markers are enough? factors influencing parentage testing in different livestock populations. *Journal of Animal Breeding and Genetics.*

Thomas SC, Hill WG (2002) Sibship reconstruction in hierarchical population structures using markov chain monte carlo techniques. *Genetical research*, **79**, 227–234.

Thompson EA (1986) *Pedigree analysis in human genetics.* Johns Hopkins University Press Baltimore.

Thompson EA, Meagher TR (1987) Parental and sib likelihoods in genealogy reconstruction. *Biometrics*, S. 585–600.

VanRaden PM, Cooper TA, Wiggans GR, OConnell JR, Bacheller LR (2013) Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle. *Journal of dairy science*, **96**, 1874–1879.

Visscher PM, Medland SE, Ferreira MAR, *et al.* (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, **2**, e41.

Walling CA, Pemberton JM, Hadfield JD, Kruuk LEB (2010) Comparing parentage inference software: reanalysis of a red deer pedigree. *Molecular Ecology*, **19**, 1914–1928.

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

Wang J (2012) Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*, **191**, 183–194.

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88**, 76–82.

Appendix A: Genotyping error model

To account for genotyping errors, I distinguish between the observed genotype X, and the actual genotype x. I assume the genotyping error rate ϵ is constant across loci, that errors occur independently of each other, and that there are no (heritable) mutations.

In the default error structure used, the chance of observing a true minor homozygote as major homozygote, or vice versa, is assumed negligible (Table 4), based on own observations when scoring SNP array data in Illumina's GenomeStudio. True heterozygotes are observed as either homozygote with probability $\epsilon/2$, and true homozygotes are observed as heterozygous with probability ϵ. Different error structures can be easily implemented by changing the analogue of Table 4 in the source code.

Table 4: Default probabilities of observing genotype X, conditional on actual genotype x.

	X		
x	0	1	2
0	1-ϵ	ϵ	0
1	$\epsilon/2$	1-ϵ	$\epsilon/2$
2	0	ϵ	1-ϵ

The probability that a given observed genotype is erroneous depends slightly on the allele frequency, similar to in e.g. Wang (2004). Intuitively, the reason is that an observed homozygote for a very rare allele may be less likely to be an actual homozygote (with probability q^2) than to be an actual heterozygote ($\epsilon/2 \cdot 2q(1-q)$). Note however that in practice, q will typically exceed ϵ by several orders of magnitude.

More formally, the joint distribution P(Observed, Actual) is a function of the frequencies of the actual genotypes, which are assumed to be according to HWE, and the conditional probabilities P(Actual | Observed) are calculated as

$$P(\text{Actual}|\text{Observed}) = \frac{P(\text{Actual, Observed})}{P(\text{Observed})} = \frac{P(\text{Observed}|\text{Actual})P(\text{Actual})}{\sum_{x=0}^{2} P(\text{Observed}|\text{Actual} = x)}, \tag{6}$$

where P(Actual) assumes HWE when both parents of the focal individual are unknown, and is otherwise dependent on the parental genotype(s) (see Equation 3). The error model used assumes a slightly inflated number of observed heterozygotes (of $2q(1-q)+\epsilon$), which will bias the estimated allele frequencies $\hat{q}$ upwards by $1/2\epsilon(1-2q)$. This bias is corrected for, but is negligible for typical error rates ($\epsilon = 0.005$, (Anderson & Garza, 2006)) and the most informative allele frequencies (q =0.3–0.5).

Note that null alleles are currently not explicitly incorporated. They are much less of a problem than with microsatellite markers, but occasionally a third allele may occur at a SNP locus, which binds to neither probe of the SNP array. Heterozygotes for such an allele may either be scored as homozygote, or not at all. Typically, however, such tri-allelic SNPs are excluded during quality control, as they either have low call rate, or show considerable deviation from HWE.

Appendix B: Likelihood equations

As an example, all pairwise likelihood equations are shown for a pair of individuals A and B, for a maternal focal relationship, and B being older than A (e.g. the question whether B is the mother of A). For paternal focal relationships, one can simply swap the interpretation of the parental symbols below, currently meaning D_A for A's mother, and S_A for A's father.

Each equation can be generalised to the situation where either or both A and B are sibships, rather than individuals, in which case one multiplies over all members of the sibship(s), analogous to Equation 4.

H_0**: Unrelated** Firstly, the likelihood under the hypothesis that the pair is conditionally unrelated is

$$\mathcal{L}(\mathrm{U}|A, B, \ldots) = \mathcal{L}(A, D_A, S_A)\mathcal{L}(B, D_B, S_B) \tag{7}$$

where $\mathcal{L}(A, D_A, S_A)$ is defined in Equation 2, and . . . denotes the parents of A and B when known.

H_1–H_2**: First degree relatives** The first alternative relationship (H_1) considered is parent-offspring (PO), with say B being the candidate mother of A

$$\mathcal{L}(\mathrm{PO}|A, B, \ldots) = \prod_{l=1}^{L}\prod_{l}\sum_{y}\sum_{z}\sum_{v}\sum_{w} P_{M\epsilon}(A = X|B = y, S_A = z)P_\epsilon(B = Y|B = y, \epsilon)P_P(S_A = z)\times$$

$$P_M(B = y|D_B = v, S_B = w)P_P(D_B = v)P_P(S_B = w)\ , \tag{8}$$

dropping subscripts l for brevity, and using the shorthand

$$P_{M\epsilon}(A = X|B = y, S_A = z) = \sum_x P_\epsilon(A = X|A = x, \epsilon)P_M(A = x|B = y, S_A = z) \tag{9}$$

If A and B do not have a different mother or a different father assigned, secondly (H_2) the likelihood of being full siblings (FS) is calculated,

$$\begin{aligned}\mathcal{L}(\text{FS}|A, B, \ldots) = \prod_{l=1}^{L}\prod_{l}\sum_{u}\sum_{z} & P_{M\epsilon}(A = X|D_{AB} = u, S_{AB} = z)\times \\ & P_{M\epsilon}(B = Y|D_{AB} = u, S_{AB} = z)P_P(D_{AB} = u)P_P(S_{AB} = z)\ ,\end{aligned} \tag{10}$$

where D_{AB} and S_{AB} are the shared parents of A and B.

H_3–H_5**: Second degree relatives** The likelihood that A and B are maternal half-siblings is given by

$$\begin{aligned}\mathcal{L}(\text{HS}|A, B, \ldots) = \prod_{l=1}^{L}\prod_{l}\sum_{u}\sum_{z}\sum_{w} & P_{M\epsilon}(A = X|D_{AB} = u, S_A = z)\times \\ & P_{M\epsilon}(B = Y|D_{AB} = u, S_B = w)P_P(D_{AB} = u)P_P(S_A = z)P_P(S_B = w)\ ,\end{aligned} \tag{11}$$

and that they are grandparent and grand-offspring by (here via D_A; via S_A is considered too)

$$\begin{aligned}\mathcal{L}(\text{GG}|A, B, \ldots) = \prod_{l=1}^{L}\prod_{l}\sum_{u}\sum_{y}\sum_{z}\sum_{v}\sum_{w}\sum_{t} & P_{M\epsilon}(A = X|D_A = u, S_A = z)P_P(S_A = z)\times \\ & P_M(D_A = u|B = y, MGF_A = t)P_M(B = y|D_B = v, S_B = w)\times \\ & P_P(D_B = v)P_P(S_B = v)P_P(MGF_A = t)P_{P^*}(D_A = u)\ ,\end{aligned} \tag{12}$$

where MGF_A is the maternal grandfather of A, and $P_{P^*}(D_A = u) = P_\epsilon(D_A = U|D_A = u)$ for D_A known and genotyped, $P_{P^*}(D_A = u) = 1$ for D_A unknown, and when D_A is a dummy parent calculated from $\mathcal{L}(\boldsymbol{A})$ without the contributions of either grandparent or A.

The fifth alternative is full avuncular (H_5), i.e. either parent of A (here D_A) is a full sibling of B,

$$\begin{aligned}\mathcal{L}(\text{FA}|A, B, \ldots) = \prod_{l=1}^{L}\prod_{l}\sum_{u}\sum_{z}\sum_{v}\sum_{w} & P_{M\epsilon}(A = X|D_B = u, S_A = z)P_{M\epsilon}(B = Y|D_B = v, S_B = z)\times \\ & P_M(D_A = u|D_B = v, S_B = z)P_P(S_A = z)P_P(D_B = v)P_P(S_B = w)P_{P^*}(D_A = u),\end{aligned} \tag{13}$$

where S_A also might be a FS of B, or either parent of B a FS of A.

As mentioned in the Introduction, $\mathcal{L}$(HS | A, B) $=\mathcal{L}$(GG | A, B) $=\mathcal{L}$(FA | A, B) when neither A nor B has a (dummy)parent assigned. Distinction between these three relationship types can be made when either parent of B is known (Appendix C), or when the age difference between A and B excludes some of the configurations (Appendix D).

H_6**: Third degree relatives** Lastly, A and B may be third degree relatives. These are considered to prevent false positive assignments, as some third degree relatives may have a higher likelihood to be second degree relatives than unrelated - but will have an even higher expected likelihood to be third degree relatives. Assigning third degree relatives is not attempted, as the distinction with fourth degree relatives is difficult. Moreover, even if B were known to be a great-grandparent of A, it would be unclear which of the 8 great-grandparents of A it was, without knowledge on first and second degree relatives of A and B.

The likelihood to be third degree relatives is taken as the most likely scenario of half-avuncular (HA), great-grand-parental (GGG), or full first cousins (CC):

$$\begin{aligned}
\mathcal{L}(\text{HA}|A,B) = \prod_l \sum_u \sum_z \sum_v \sum_w \sum_t & P_{M\epsilon}(A=X|D_A=u,S_A=z) P_{M\epsilon}(B=Y|D_B=v,S_B=w)\times \\
& P_M(D_A=u|D_B=v,MGF_A=t)P_P(S_A=z)P_P(D_B=v)P_P(S_B=w)\times \\
& P_P(MGF_A=t)P_{P^*}(D_A=u) \qquad (14) \\
\mathcal{L}(\text{GGG}|A,B) = \prod_l \sum_y \sum_u \sum_z \sum_v \sum_w \sum_s \sum_t & P_{M\epsilon}(A=X|D_A=u,S_A=z)P_P(S_A=z)\times \\
& P_M(D_A=u|s,MGF_A=t)P_{Mh}(s|B=y,q_l)P_M(B=y|D_B=v,S_B=w)\times \\
& P_P(D_B=v)P_P(S_B=v)P_P(MGF_A=t)P_{P^*}(D_A=u) \ , \qquad (15) \\
\mathcal{L}(\text{CC}|A,B) = \prod_l \sum_u \sum_z \sum_v \sum_w \sum_s \sum_t & P_{M\epsilon}(A=X|D_A=u,S_A=z)P_P(S_A=z)\times \\
& P_{M\epsilon}(B=Y|D_B=v,S_B=w)P_P(S_B=w)P_M(D_A=u|MGF_{AB}=t,MGM_{AB}=s)\times \\
& P_M(D_B=v|MGF_{AB}=t,MGM_{AB}=s)P_P(MGF_{AB}=t)P_P(MGM_{AB}=s), \\
& \qquad (16)
\end{aligned}$$

where $P_{Mh}(s|B=y,q_l)$ is the inheritance probability from a single parent (here B), and MGF_A and MGM_A are the maternal grandmother and maternal grandfather of A, respectively. Under HA, when the focal hypothesis is that A and B are relatives of type

k, we consider the possibilities that parent k of A is a paternal or maternal half-sibling of B, or that parent k B is a paternal or maternal half-sibling of A.

It can be shown that similar to 2nd degree relatives, all 3rd degree relatives have the same likelihood function when not conditioning on any parental or sibling genotypes. Therefore, full great-uncle – great-nephew pairs, which would require summation over four (unobserved) relatives, are currently not explicitly considered, as they either have a similar likelihood as HA or GGG, or one of the 'intermediate' individuals is known making A and B conditionally unrelated.

Note that although there are up to 6 or 7 summations in each likelihood equation, many shortcuts can be taken by the use of look-up tables. Moreover, as there are only 3 different possible states per individual, this constitutes only $3^6 = 729$ different joined states - identical to the number of possible states for a trio on a microsatellite locus with 9 alleles, a very typical number.

Interconnected sibships

When considering various hypothesised relationships between sibship $\boldsymbol{A}$ and individual B, or between sibships $\boldsymbol{A}$ and $\boldsymbol{B}$, likelihood calculations are mostly performed over the sibship cluster itself, and all sibships directly linked to it (e.g. for a maternal sibship, all paternal sibships of the males with whom the sibship mother mated). This is especially useful when there are multiple opposite-sex dummy parents, as using ${}^{d}P_P(S_1 = y_1|x = 0)$ and ${}^{d}P_P(S_2 = y_2|x = 0)$, will give different results from using $\sum_{x'} {}^{d}P_P(S_i = y_i|D_A = x')^{d}P_P(D_A = x')$ for $i = 1, 2$. Especially when $\boldsymbol{A}$ is small may ${}^{d}P_P(S_1)$ and ${}^{d}P_P(S_2)$ depend strongly on each other; for example, if at a locus A_1 and A_2 are both heterozygous, if $D_{\boldsymbol{A}} = 2$, most likely $S_1 = S_2 = 0$, while when $D_{\boldsymbol{A}} = 0$ most likely $S_1 = S_2 = 2$.

In some simpler scenarios, such as addition of a half-sibling with no current parents, changes in the connected sibships are presumed negligible. The more distantly, indirectly connected sibships are always treated as temporarily fixed. The latter approximation is necessary as the number of interconnected sibship may become very large, and calculations over such large webs are computational intensive, while their contribution to changes in the likelihood is much smaller than the contributions of the focal and directly-connected sibships.

To enable comparison of likelihoods calculated under the different hypotheses, only the

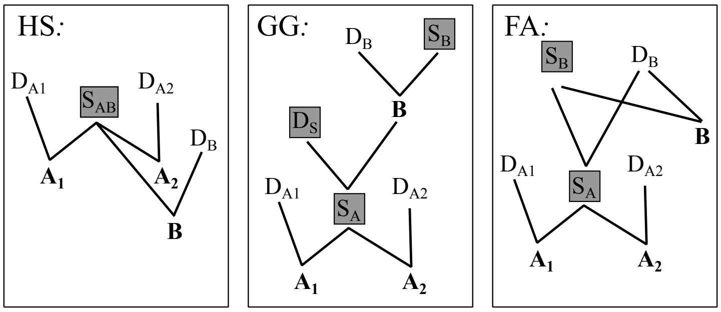


Figure 10: Schematic pedigrees of the three types of second degree relationships possible between individual B and sibship A with known siblings A_1 and A_2. The genotypes of the fathers (S_A, S_B and S_{AB}) are assumed unknown (denoted by grey square boxes), and the genotypes of the mothers (D_{Ai} and D_B) may or may not be known.

likelihood over the focal individuals should be returned. Therefore, in parallel the likelihood over all individuals *except* $\boldsymbol{A}$ and the other focal individual or sibship is calculated, and the required likelihood is taken as the difference between the two.

Appendix C: Differentiating between types of second degree relatives

A long standing problem in pedigree reconstruction is the differentiation between half-sibling (HS), grandparent–grand-offspring (GG) and avuncular (FA) pairs (see e.g. Epstein *et al.*, 2000), which all have an pedigree relatedness of $r = 0.25$. One remedy is to make use of the age difference of the pair (Appendix D), but this provides no conclusive distinction in species where the maximum reproductive lifespan is several times longer than the minimum generation time. Therefore, I (additionally) condition on the genotype of the (dummy)parents of the pair.

It is common practice to condition on the maternal genotype when inferring paternities (e.g. Marshall *et al.*, 1998), because if an heterozygous individual has a major homozygote as mother, it must have inherited the minor allele from its father (bare genotyping errors). Similarly, this approach has been used to distinguish between full sibs, half sibs and unrelated individuals within a single cohort (Wang, 2004). To our knowledge, however, such an approach has not been widely used to distinguish between types of second degree relatives (but see Anderson & Garza (2006)).

To illustrate, presume $B = 1$ (heterozygous) and $D_B = 2$ (homozygous for the rare

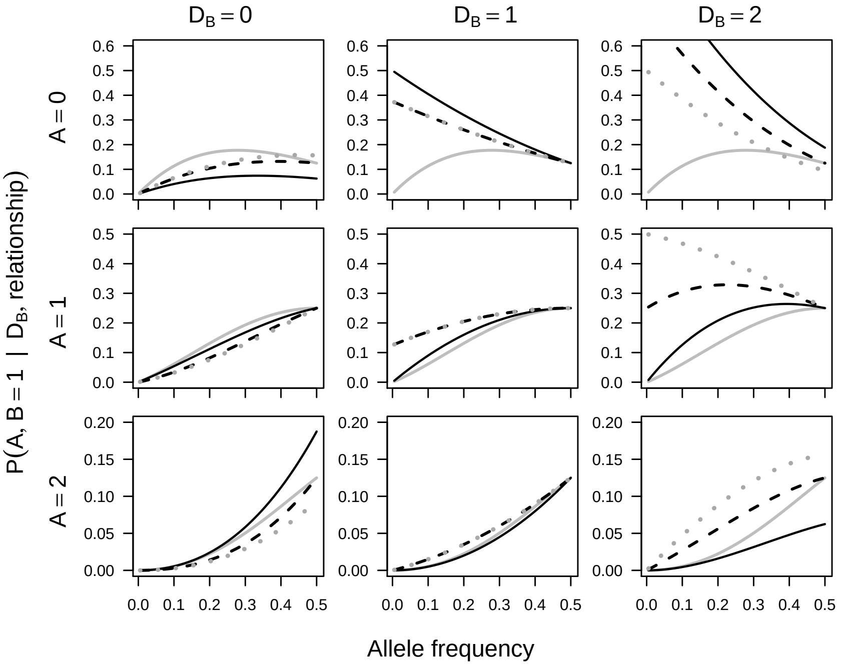


Figure 11: As figure 2, illustrating how knowledge of the genotype of B's parent D_B allows differentiation between HS (solid black lines), GG (dashed black) or FAU (dotted grey), while these are indistinguishable when D_B is unknown (solid grey line). Formulae are given in Equations 8–10 ($i = 1$).

allele, as in the right column in Figure 11). Under scenario HS (see Figure 10), S_A (=S_{AB}) must be a carrier of the common allele, and

$$\begin{aligned}
P(S_A = 0|\text{HS}, B = 1, D_B = 2) &= \frac{q^2}{q^2 + 2q(1-q)} = \frac{q}{2-q} \\
P(S_A = 1|\text{HS}, B = 1, D_B = 2) &= \frac{2q(1-q)}{q^2 + 2q(1-q)} = \frac{2(1-q)}{2-q} \\
P(S_A = 2|\text{HS}, B = 1, D_B = 2) &= 0, \qquad (17)
\end{aligned}$$

while under scenario GG, S_A only depends on B, and is conditionally independent of D_B, and the probabilities are given by:

$$\begin{aligned}
P(S_A = 0|\text{GG}, B = 1) &= \frac{1-q}{2} \\
P(S_A = 1|\text{GG}, B = 1) &= \frac{1}{2} \\
P(S_A = 2|\text{GG}, B = 1) &= \frac{q}{2}. \qquad (18)
\end{aligned}$$

Lastly, under scenario FA, S_A only depends on D_B, and is conditionally independent of

B:

$$\begin{aligned}
P(S_A = 0|\mathrm{FA}, D_B = 2) &= 0 \\
P(S_A = 1|\mathrm{FA}, D_B = 2) &= 1 - q \\
P(S_A = 2|\mathrm{FA}, D_B = 2) &= q.
\end{aligned} \tag{19}$$

These different probabilities for the possible genotypes of the unobserved S_A, result in different probabilities for the observed genotypes A and B, for all possible genotypes of D_B, as illustrated in Figure 11, and thus different likelihoods for three alternative relationships. Knowledge on D_A does not help in the differentiation, as it does not affect either the probability that S_A inherits an allele from B, nor the reverse probability. However, when generations overlap or the age difference between A and B is unknown, both D_A and D_B are required, as it cannot be determined with certainty whether B might be a full aunt or uncle of A, or instead A an aunt/uncle of B.

Appendix D: Age-difference based priors

The age difference between individuals can be very informative in pedigree reconstruction, as grandparents will on average always be older than siblings, and in many species the two age distributions may show little overlap (see Figure 12 for an example in red deer). Ideally, the effect of age difference on relationships is estimated jointly with the effect of genotypes in a Bayesian MCMC framework (as e.g. in MasterBayes, (Hadfield *et al.*, 2006)). However, this approach can be very time-consuming when the numbers of individuals and markers are large. As an heuristic approximation, we assume that the distribution of maternal and paternal ages amongst assigned parents is identical to that amongst ungenotyped parents. When a sufficient number of individuals has been assigned a sampled parents (by default a threshold of 25% is used), it is possible to estimate from these parental age distributions the empirical age-difference distributions for maternal and paternal siblings, maternal grandmothers, paternal grandfathers, paternal grandmothers and maternal grandfathers, and avuncular pairs. After extending the tails of these distributions, to allow for biologically plausible but un-observed values, and optional smoothening, use of these age-based priors aids in the distinction between various relationships in cases where the genetic data is inconclusive. Implicitly it is assumed that

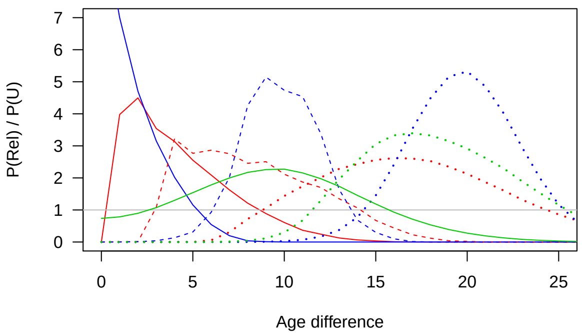


Figure 12: Empirical age-difference distribution for various classes of relatives in a Red deer population (described in main text), used as prior probabilities in sibship clustering. Shown is the proportion of pairs in a given relationship that has an age difference of t, relative to the proportion of all possible pairs in the dataset that have an age difference of t. Dashed red: Mother-O, dashed blue: Father-O, solid red: MS(no twins in this population), solid blue: PS, solid green: FA+HA, dotted blue: pat. grandfather, dotted red: mat. grandmother, dotted green: pat. grandmother & mat. grandfather.

the age distribution, and distribution of female and male age at reproduction, remain constant throughout the sampling period. This assumption can be relaxed in future versions if needed, by using birth years rather than, or in addition to, age differences.

We account for the fact that the distribution of absolute age differences within the sample is non-uniform, due to the finite time period in which samples are necessarily collected. For example, in a 10 year study period, sampling each individual at birth, many more sampled individuals will have been born 2 years apart than 10 years apart. We first calculate the proportion of all pairs of individuals which were born t years apart, $P_{U,t}$, with $\sum_{t=0}^{t_{max}} P_{U,t} = 1$. We similarly calculate the number of mother-offspring pairs born t years apart, as a fraction $P_{D,t}$ of the total number of assigned mother-offspring pairs, as well as for father-offspring pairs, $P_{S,t}$, maternal siblings $P_{MS,t}$ and paternal siblings $P_{PS,t}$. From this we calculate the age-difference probability ratio

$$APR_{\cdot,t} = \frac{P_{\cdot,t}}{P_{U,t}} \ , \tag{20}$$

which is stored in a user-editable text file, rounded to 3 decimal places.

The age distribution of maternal grandmother - grandoffspring $P_{MGM,t}$ is obtained as

$$\begin{aligned} X_{MGM,u,v} &= \sum_{u=0}^{t_{max}} \sum_{v=0}^{t_{max}} P_{D,u} P_{D,v} \\ P_{MGM,t} &= \sum_{u+v=t} X_{MGM,u,v} \ , \end{aligned} \tag{21}$$

and analogous for paternal grandfathers (PGF) and maternal grandfathers (MGF) / paternal grandmothers (PGM). As a parsimoneous approximation, we assume the same age distribution for paternal and maternal aunts and uncles, as well as for full and half aunts and uncles. It is calculated from the grandparental and parental age distributions as

$$X_{AU,u,v} = \frac{1}{4} \sum_{u=0}^{t_{max}} \sum_{v=0}^{t_{max}} P_{D,t} P_{MGM,t} + P_{D,t} P_{PGM,t} + P_{S,t} P_{MGF,t} + P_{S,t} P_{PGF,t}$$

$$P_{AU,t} = \sum_{u+v=t} X_{AU,u,v} \; . \tag{22}$$

Alternatively, a discrete prior can be used, where '1' signals an age difference is possible for a specific type of relationship (and that given this age difference, this relationship is as likely as being unrelated), and '0' signals it is not possible, given the minimum and/or maximum age of reproduction of the species.

Appendix E: Extensions for close inbreeding and double relatives

The explicit consideration of inbred configurations is not required to detect those, as they will typically come about as by-products of parentage assignment and sibship clustering. However, when exploring the alternative hypothetical relationships, inbreeding loops need to be either considered carefully and explicitly, or declared impossible, to avoid erroneous likelihoods and false assignments. For example, if a mother and daughter mate with the same male (as common in a.o. red deer, (Stopher *et al.*, 2012)), their offspring are related by $1/4 + 1/8 = 0.375$ and can easily be mis-identified as full siblings if this double-relatedness is not explicitly considered.

Additional pairwise likelihoods When the focal individual is the result of a parent-offspring mating, the assumption in Equation 8 that the opposite-sex parent is a random draw from the population is severely violated. When considering say candidate father B, and the focal individual A has no mother D_A yet assigned, we therefore also consider the possibility that B might be the parent of both A and its mother D_A,

$$\begin{aligned}\mathcal{L}(\mathrm{PO}^{+}|A,B) = \prod_{l=1}^{L}\prod_{l}\sum_{y}\sum_{z}\sum_{v}\sum_{w}\sum_{t} & P_{M\epsilon}(A=X|B=y,D_A=z)\times \\ & P_M(D_A=z|B=y,MGM_A=t)P_{P^*}(D_A=z)P_P(MGM_A=t)\times \\ & P_\epsilon(B=Y|B=y,\epsilon)P_M(B=y|D_B=v,S_B=w)P_P(D_B=v)P_P(S_B=w)\ ,\end{aligned} \tag{23}$$

as well as the possibility that A and B may share the same mother D_{AB},

$$\begin{aligned}\mathcal{L}(\mathrm{PO}^{+'}|A,B) = \prod_{l=1}^{L}\prod_{l}\sum_{y}\sum_{u}\sum_{w} & P_{M\epsilon}(A=X|B=y,D_{AB}=z)\times \\ & P_M(B=y|D_{AB}=z,S_B=w)P_\epsilon(B=Y|B=y,\epsilon)P_P(S_B=w)P_P(D_A=z)\end{aligned} \tag{24}$$

and the maximum of $\mathcal{L}(\mathrm{PO})$, $\mathcal{L}(\mathrm{PO}^{+})$ and $\mathcal{L}(\mathrm{PO}^{+'})$ is used in comparison with the likelihoods of the alternative scenarios.

When during sibship clustering a pair has a higher likelihood to be FS than any other of the single-relationships considered, we additionally consider the possibility that they may be (e.g. paternal) HS with closely related opposite-sex parents (mothers) (currently considering PO and GG). Erroneous assignment as a FS pair can have considerable downstream consequences, by providing an erroneous 'core' from which a sibship may grow. Most similar to FS are pairs where the opposite-sex parents are parent and offspring themselves, for which the likelihood equation for half-siblings (Equation 11) is adapted to

$$\begin{aligned}\mathcal{L}(\mathrm{HS}^{+}|A,B) = \prod_{l=1}^{L}\prod_{l}\sum_{u}\sum_{z}\sum_{w}\sum_{t} & P_{M\epsilon}(A=X|D_A=u,S_{AB}=z)\times \\ & P_{M\epsilon}(B=Y|D_B=v,S_{AB}=z)P_M(D_A=u|D_B=v,MGF_A=t)\times \\ & P_{P*}(D_A=u)P_P(S_{AB}=z)P_P(S_B=w)P_P(MGF_A=t)\ ,\end{aligned} \tag{25}$$

and if this likelihood exceeds $\mathcal{L}(\mathrm{FS})$, $\mathcal{L}(\mathrm{FS})$ is set to missing, as in absence of any assigned parents, it cannot be determined whether A and B are paternal HS with related mothers, or maternal HS with related fathers. When $\mathcal{L}(\mathrm{HS}+)$ exceeds $\mathcal{L}(\mathrm{HS})$, $\mathcal{L}(\mathrm{HS}+)$ is used in further comparison with the other relationships.

The modular structure of the source code allows additional types of relationships to be added quite easily. This may be required if they are common in the population of interest, and provide a large source of false positives. These may be inbred relationships, or non-inbred double relationships, such as double first cousins, which are currently not explicitly considered.

Inbreeding within sibship cluster Within a sibship, a grandparent may also be an opposite-sex parent of one of the members, or one sibship member may be the opposite-sex parent of another member. To incorporate these possibilities, Equation 4 is generalised to

$$\begin{aligned}
\mathcal{L}(\boldsymbol{A}) = \prod_l \sum_x \sum_v \sum_w & P_M(D_{\boldsymbol{A}} = x | GM_{\boldsymbol{A}} = v, GF_{\boldsymbol{A}} = w) P_P(GM_{\boldsymbol{A}} = v) P_{P'}(GF_{\boldsymbol{A}} = w) \times \\
& \prod_{i=1}^{n_A} I(S_i = GF_{\boldsymbol{A}}) \prod_{j=1}^{m_{A,i}} P_{M\epsilon}(A_{i,j} = Z | D_{\boldsymbol{A}} = x, GF_{\boldsymbol{A}} = w) \times \\
& \quad I(S_i \in \boldsymbol{A}) \sum_u P_{M\epsilon}(S_i = y_i | D_A = x, S_{Si} = u) P_P(S_{Si} = u) P_{M\epsilon} \prod_{j=1}^{m_{A,i}} (A_{i,j} = Z | D_{\boldsymbol{A}} = x, S_i = y_i) \times \\
& \quad I(S_i \neq GF_{\boldsymbol{A}}) I(S_i \notin \boldsymbol{A}) \sum_{y_i} P_P(S_i = y_i) \prod_{j=1}^{m_{A,i}} P_{M\epsilon}(A_{i,j} = Z | D_{\boldsymbol{A}} = x, S_i = y_i) \ , \qquad (26)
\end{aligned}$$

where $P_{P'}(GF_{\boldsymbol{A}} = w)$ is calculated without the contribution of its shared offspring with D_A, and I are indicator variables taking the value 1 when true and 0 when false.

Appendix F: Parentage assignment

Parentage assignment is done by calculating the pairwise likelihood between the focal individual A and candidate parent B conditional on the earlier assigned parent(s) S_A and D_A ($H_{1,\,1,\,0-6}$), as well as under the hypotheses that S_A is unrelated ($H_{0,\,1,\,0-6}$) or D_A is unrelated ($H_{1,\,0,\,0-6}$) (columns in Table 5). To these pairwise likelihoods over A and B the likelihoods of D_A and S_A is added, to obtain the total likelihood over all 2–4 individuals involved. Similarly, the pairwise likelihoods over A and current parent S_A are calculated under the condition of B being the parent ($H_{0-6,\,,1,\,1}$, top row in Table 5) or unrelated ($H_{0-6,\,1,\,0}$, bottom row), after addition of the likelihoods of B and D_A, and analogously for D_A.

Calculation of the likelihoods under all $2 \times 7 \times 7 = 98$ possible quartet scenarios appears redundant; if for example candidate B truly were a grandmother of A, and S_A truly a full sibling, the likelihoods under the hypothesis (U + FS) and/or (GG + U) would still

		Currently assigned father (S_A)						
		PO	**FS**	**HS**	**GG**	**FA**	**HA**	**U**
Candidate parent (B)	**PO**	$B + S_A$	B	B	B	B	B	B
	FS	$D_A + S_A$						D_A
	HS	$D_A + S_A$						D_A
	GG	$D_A + S_A$						D_A
	FA	$D_A + S_A$						D_A
	HA	$D_A + S_A$						D_A
	U	$D_A + S_A$	D_A	D_A	D_A	D_A	D_A	D_A

		Currently assigned mother (D_A)						
		PO	**FS**	**HS**	**GG**	**FA**	**HA**	**U**
Candidate parent (B)	**PO**	N/A	$B + S_A$	$B + S_A$	$B + S_A$	$B + S_A$	$B + S_A$	$B + S_A$
	FS	$D_A + S_A$						S_A
	HS	$D_A + S_A$						S_A
	GG	$D_A + S_A$						S_A
	FA	$D_A + S_A$						S_A
	HA	$D_A + S_A$						S_A
	U	$D_A + S_A$	S_A	S_A	S_A	S_A	S_A	S_A

Table 5: Scheme of quartet relationships considered between a focal individual A, a candidate parent (B, here assumed female) (rows), its previously assigned father S_A (top) and mother D_A (bottom). Bold abbreviations as in Table 1. Values in middle cells indicate which individuals will be assigned as parents when that particular combination of pairwise relationships has the highest likelihood; blank cells are not considered. Note that consideration of candidate mother B may result in joint assignment with current father S_A (when $\mathcal{L}$(PO + PO) in upper matrix exceeds all others), but also in loss of a currently assigned parent.

exceed the likelihoods under the hypothesis (PO + FS) and (GG + PO), and the correct assignment made.

During initial parentage assignment, A, B, D_A and S_A are always real genotyped individuals. During later parentage assignment, D_A and S_A may also be dummy parents. In the current algorithm, assignment of dummy parents to A, i.e. sibship clustering, is performed in a separate step from assignment of real parents. In contrast, when assigning grandparents to sibship clusters $\boldsymbol{A}$, this step considers jointly all possible candidates, both genotyped individuals B and dummy individuals ${}^dB$.

Supplementary Figures

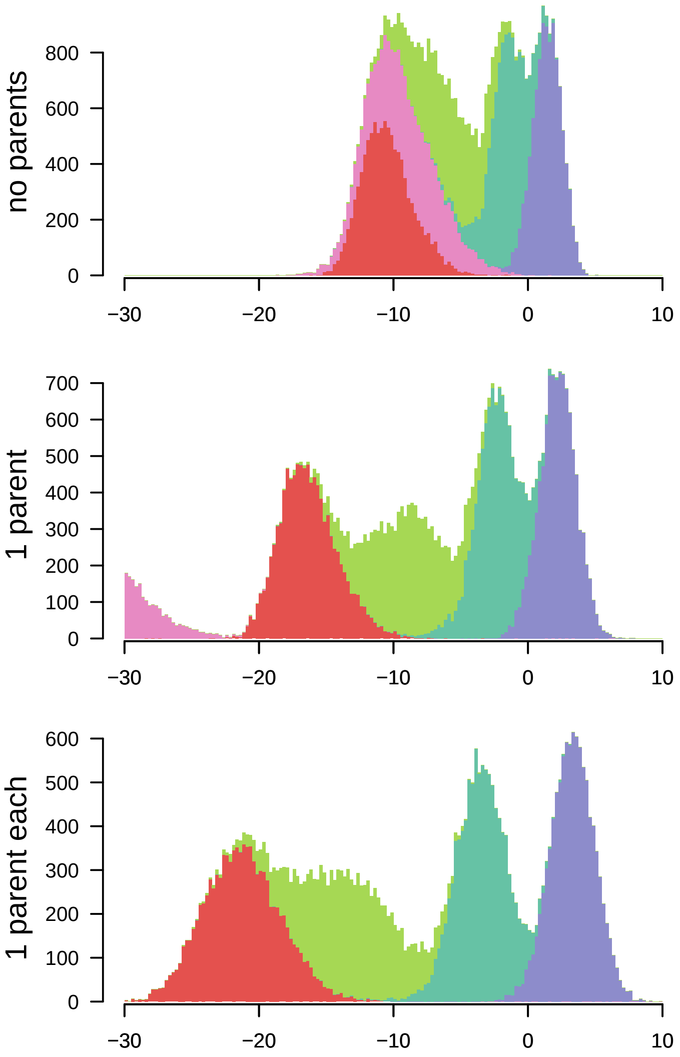


Figure 13: Distributions of the pairwise log10-likelihood ratios $\Lambda_{HS/U}$ versus the most likely alternative for pairs that are PO (red), FS (pink), HS (purple), HA (blue-green) or U (green), conditioning on no parents, one parent or one parent each. Based on 10.000 simulations of a simple pedigree with unrelated founders and 400 SNPs with MAF 0.3–0.5 and $\epsilon = 0.005$.

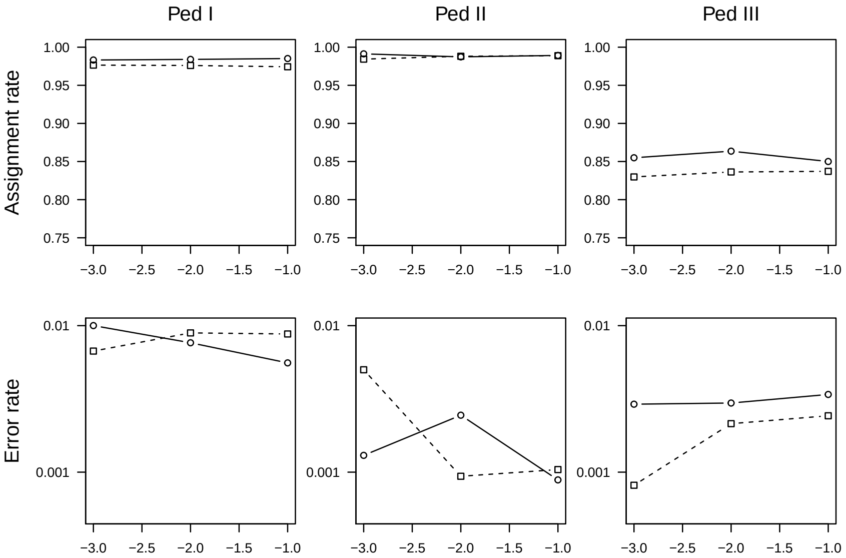


Figure 14: Assignment and error rates for Pedigree I–III, for three different thresholds to exclude unlikely relatives (x-axes) and an assignment threshold of 0.5 (solid lines) or 1.0 (dashed lines). Simulated datasets included 200 SNPs and 0% (Ped I) or 60% (Ped II–III) of parents genotyped.

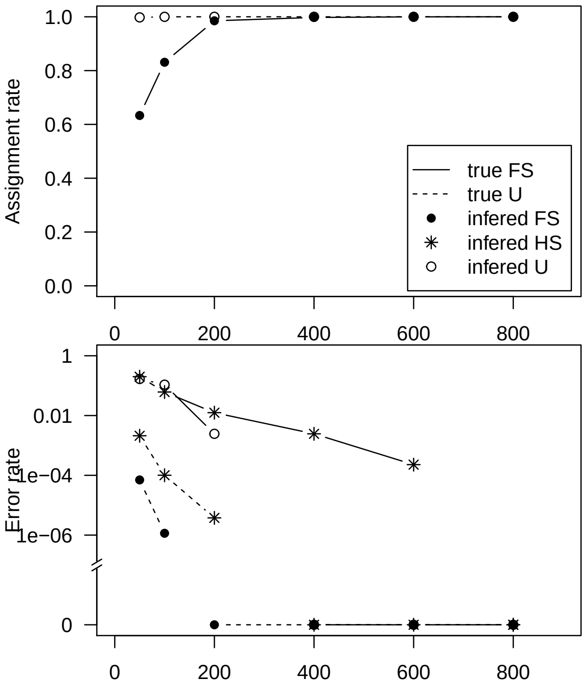


Figure 15: Proportion of pairs (A) correctly assigned or (B) misclassified in Pedigree I among FS (solid lines) and U (dashed) pairs. Filled points denote classification as FS, asterisks as HS, and open points as U; averages over 10 replicate runs are shown. Note that pedigree I consists of 1983 FS pairs and 6.6×10^5 U pairs, such that erroneous assignment of one individual to a full sib family with, say, 2 members results in an pairwise error rate $P(\text{FS}|\text{U}) = 3.03 \times 10^{-6}$.

sequoia

Jisca Huisman
jisca.huisman@gmail.com

June 29, 2016

Sequoia provides a method to reconstruct multi-generational pedigrees based on SNP data, as described in the manuscript “Pedigree reconstruction using SNP data: parentage assignment, sibship clustering, and beyond”. The bulk of the algorithm is written in Fortran, to minimise computation times.

An example R script for pedigree reconstruction is given below, followed by detailed description of each step.

Example

An example pedigree and associated life history file are provided with the package, which can be used to try out the steps detailed in this vignette. This pedigree consists of 5 generations with interconnected half-sib clusters (Pedigree II in the manuscript).

```
# install the package (only required once)
install.packages("sequoia")

# set the working directory
setwd("E:/Sequoia/test")

# load the package
library(sequoia)

# copy the example pedigree and associated life history file to the working
# directory.
file.copy(system.file("Ped_HSg5.txt", package="sequoia"), getwd())
file.copy(system.file("LH_HSg5.txt", package="sequoia"), getwd())

# simulate genotype data for 200 SNPs, and use otherwise default values
SimGeno(PedFile = "Ped_HSg5.txt", nSnp = 200)

# run the preparation step, duplicate checking and parentage assignment,
# but not yet the slower sibship clustering. Iterate as necessary,
# weeding out duplicated and erroneous samples (using PLINK's toolkit),
# and adding estimated birth years to the life history file.
sequoia(GenoFile = "SimGeno.txt", LifeHistFile = "LH_HSg5.txt", Sibships = FALSE)

# compare the assigned parents to those in the true pedigree
PedCompare(PedIN = "Ped_HSg5.txt", PedOUT = "Parents_assigned.txt")

# run sibship clustering
sequoia(Prep = FALSE, CheckDup = FALSE, Parentage = FALSE, Sibships = TRUE)

# compare the assigned real and dummy parents to the true pedigree
PedCompare(PedIN = "Ped_HSg5.txt", PedOUT = "PedSeq.txt")
```

1 Input

Two files are required as input. The first one contains the SNP data, with one line per individual, and one column for IDs followed by one column per SNP, where each SNP is coded as 0, 1, 2 copies of the reference allele, or missing (-9). This file format can for example be obtained using PLINK [Purcell et al., 2007] in combination with `sequoia`'s 'Prep' step, as described below.

The other file contains three columns: individual ID, sex (1 = female, 2 = male, other numbers = unknown), and birth year. Ideally all genotyped individuals are included in this life history file with sex and (estimated) birth year information, but this is not necessary. The life history file may include many more individuals than the genotype file, or in a different order. In species with more than one generations per year, a finer time scale than year of birth ought to be used, ensuring that parents are never born within the same time unit as their offspring.

Selection of SNP markers Using tens of thousands of SNP markers for pedigree reconstruction is unnecessary, will slow down computation, and may even hamper inferences by their non-independence. Rather, a subset of SNPs in low linkage disequilibrium (LD) with each other, and with high minor allele frequencies (MAF > 0.3), ought to be selected first. The calculations assume independence of markers, or absence of LD in the founder population. While low (background) levels of LD are unlikely to interfere with pedigree reconstruction, high levels may give spurious results. Markers with a high MAF provide the most information, as although rare allele provide strong evidence when they are inherited, this does not balance out the rarity of such events.

Creating a subset of SNPs can be done conveniently using PLINK (`http://pngu.mgh.harvard.edu/~purcell/plink/`), using for example the command

```
plink --file mydata --maf 0.4 --indep 50 5 2
```

which will create a list of SNPs with a minor allele frequency of at least 0.4, and which

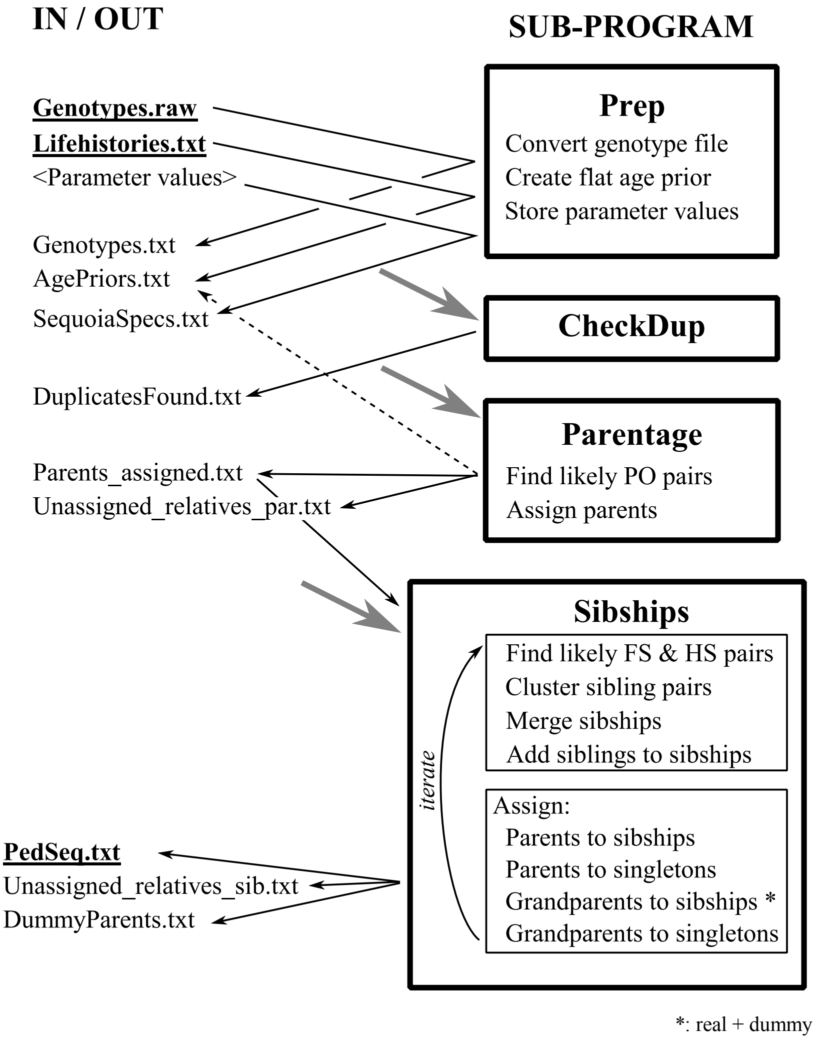


Figure 1: Overview of the input files required and output files generated by the various sub-programs of Sequoia. 'CheckDup', 'Parentage' and 'Sibship' each require a genotype file (Genotypes.txt), a lifehistory file (Lifehistories.txt), the age-difference based prior probabilities (AgePriors.txt) and the parameter values for the run (SequoiaSpecs.txt), symbolised by the grey incoming arrows.

in a window of 50 SNPs, sliding by 5 SNPs per step, have a VIF of maximum 2. VIF, or variance inflation factor, is $1/(1-r^2)$. It is advised to 'tweak' the parameter values until a set with a few hundred SNPs (300-700) is created. For further details, see `http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml#prune`.

The resulting list ('plink.prune.in') can be used to create the genotype file used as input for Sequoia, with SNPs codes as 0, 1, 2, or NA, with the command

```
plink --file mydata --extract plink.prune.in --recodeA --out inputfile_for_sequoia
```

which will create a file with the extension .RAW.

Simulating SNP data When SNP data is not (yet) available, but an approximate pedigree is, it is possible to test `sequoia` on a simulated dataset. This may be useful to for example explore the number of markers required to reliably infer a pedigree of that particular structure.

The function `SimGeno()` can be used for this, which lets the user specify the average proportion of missing genotypes per individual, the genotyping error rate, and the fraction of known parents (in the supposed 'true' pedigree) which have not been genotyped. Moreover, the data can be made to contain a fraction of low-quality samples, to assess whether inclusion of samples which did not pass stringent quality control would improve or hamper pedigree reconstruction.

2 Preparation

The program `sequoia` consists of various sub-programs, which are all called via

```
sequoia(RawFile = NULL, GenoFile = NULL, LifeHistFile = NULL,
Prep = TRUE, CheckDup = TRUE, Parentage = TRUE, Sibships = TRUE)
```

details of the various optional parameters can be found using `?sequoia`, and the sub-programs 'Prep' (prepare), 'CheckDup' (check for duplicates), 'Parentage' and 'Sibships' (sibship clustering and grandparent assignment) are described below. `RawFile` denotes the genotype file created by PLINK, and `LifeHistFile` the lifehistory file with invidual ID, sex and birth year.

The preparation step amongst others slightly reformats the genotype file, by removing the header row, removing columns 2–6, which in Plink are intended for family ID, sex and phenotypic data, and replacing 'NA' by '-9' for missing values. Genotype files which are already in this format, such as simulated genotype files, can be specified as `GenoFile`.

Setting parameter values When the program is called, with `Prep = TRUE`, it takes the parameter values supplied, or the default values, and writes those to the file 'SequoiaSpecs.txt' in the current working directory, or the specified directory (option `Dir`). This file is used by the Fortran part of the program, which does all the heavy lifting. Please do not alter the name of this file, or the number of rows in it. Do feel free to change parameter values within this file, provided the number of SNPs or individuals does not exceed the total of SNPs or individuals in the genotype file.

Check for duplicates The data may contain positive controls, as well as other intentional and unintentional duplicated samples, which ought to be removed prior to parentage assignment. Sequoia includes a function to quickly search the data for identical genotypes, called with the option `CheckDup = TRUE`. It allows a few mismatches between the genotypes (depending on the assumed genotyping error rate), with or without the same individual ID. Note that when the number of SNPs is limited, very inbred individuals may be nearly indistinguishable from their parent(s); such individuals should not be excluded.

This function additionally searches the life history file for duplicated entries, and will also print a list of individuals included in the genotype file, but not in the life history file. This latter list is merely a service to the user; individuals without life history information can often be successfully included in the pedigree.

Age based prior During preparation the file ‘AgePriors.txt’ is created, which contains 8 columns, and as many rows as the birth year range detected in the life history data input file. It initially only indicates whether a given relationship is biologically possible (1) or not (0) for a given age difference between individuals, where the first row is for individuals born in the same year, the second row for individuals born one year apart, etc. The columns are labelled for various relationship categories, with M = mother, P = father, MS = maternal sibling, PS = paternal sibling, MGM = maternal grandmother, PGF = paternal grandfather, MGF = maternal grandfather and paternal grandmother, and AU = avuncular.

For example, the first value in the column ’MS’ can be interpreted as ‘if I were to pick two individuals born in the same year, and two individuals from my sample at random, how much more likely are the first pair to be maternal siblings, compared to the second pair?’ Values below 1 indicate less likely, and values above 1 more likely. For MS, PS and AU absolute age differences are used (with overlapping generations, nephews may be older than their aunts), while parents and grandparents are necessarily older than their (grand-)offspring (categories M, P, MGM, PGF and MGF).

These age-difference based priors are by default automatically updated after parentage assignment and prior to sibship clustering, based on the empirical distribution of age differences between individuals and their assigned fathers and mothers. This update can be prevented with the option `Agepriors="old"`, or enforced later with the option `AgePriors="par"`.

As with the parameter specification file, feel free to alter the values in the AgePriors file to match the biological characteristics of the species, but please do not alter the number of columns. The number of rows may be increased (but not decreased below the age range amongst the genotyped individuals), in which case the entry ‘nAgeClasses’ in the file ‘SequoiaSpecs.txt’ should be updated to match the new number of rows.

3 Parentage assignment

Parentage assignment is performed with the option `Parentage = TRUE`, which will use the parameter settings in the file ‘SequoiaSpecs.txt’, created by `Prep = TRUE`. Parentage assignment is quick, and takes about 10–20 seconds for a dataset with 2,500 genotyped individuals on a laptop with an intel i7 2.3 GHz CPU and 8GB RAM.

Output The assigned parents are written to a text file (by default ‘Parents_assigned.txt’), rather than returned within R. This file contains columns with

- ID of the individual, its assigned mother and assigned father;
- The log10 likelihood ratio (LLR) of the mother, father and the parent pair; this is the ratio between the likelihood of the assigned parent being the parent, versus the most likely alternative type of relative (e.g. full sibling or grandparent) or unrelated, to the focal individual (999.0 = missing value);
- The number of loci at which the offspring and the mother or father are opposite homozygotes (-9 = missing value);
- The row number in the genotype file of the offspring, mother and father; used when reading in this file for subsequent sibship clustering.

Some parents may have a very small or even negative single-parent LLR, but the LLR of the parent pair will always be positive, and is relative to the most likely assignment of a single parent. Note that the reported LLR differs from for example Cervus [Marshall et al., 1998], which returns the natural log of the ratio between the probability that the assigned parent is the parent, and that the next most likely candidate is the parent.

Non-assigned parent-offspring pairs In addition, the file 'Unassigned_relatives_par.txt' may be created, with identified but non-assigned parent-offspring pairs. These are non-assigned either because it was not possible to tell which of the two was the parent and which was the offspring, due to either or both individuals having an unknown birth year, or because it was not possible to tell whether the candidate parent was the mother or the father. These situations can be remedied by providing estimated birth years, or guessed genders, for the individuals involved in the life history input file, and re-running `sequoia`. The short computational time of parentage assignment means that this step can be repeated a number of times as a part of data quality control.

4 Sibship clustering

Sibship clustering amongst those individuals which have not been assigned a genotyped parent of each sex is performed with the option `Sibships = TRUE`, which will again use the parameter settings in the file 'SequoiaSpecs.txt'. Sibship clustering is considerably slower than parentage assignment, and may take from a few minutes to a few hours, depending on the number of individuals without a parent, the number of sibships that is being clustered, and their degree of interconnection.

Convergence is typically reached within five iterations, even for complex pedigrees. When convergence is reached before the user-set maximum number of iterations, a final iteration with stronger dependence on the age prior is ran, (dummy)parental likelihoods are calculated, and the algorithm is terminated.

Dummy names By default, dummy parents are denoted by increasing numbers, with prefix 'F' for females and 'M' for males. The prefixes can be altered in the 'SequoiaSpecs.txt' file, for example to avoid confusion with IDs of real individuals.

Output The output of the sibship clustering is by default written to a text file called 'PedSeq.txt', and which is highly similar to the output of the parentage assignment. This file does contain all assigned real parents, as well as the dummy parents assigned during sibship clustering. Dummy individuals are appended at the bottom of this pedigree file, with their assigned parents, i.e. the sibship's assigned grandparents.

In addition, a file with non-assigned pairs of relatives may be created, containing for example half-siblings were it could not be determined whether they are maternal or paternal half-siblings, and grandparents of singletons.

5 Comparison with previous pedigree

Often times, a (part) pedigree is already available to which one wants to compare the results. This pedigree may consist only of maternal links, deduced from observations in the field. The R function `PedCompare()` performs such comparisons, and takes as only arguments the file names of the 'true' pedigree and of the inferred pedigree.

References

TC Marshall, JBKE Slate, LEB Kruuk, and JM Pemberton. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular ecology*, 7(5):639–655, 1998.

S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, J Maller, P Sklar, PIW De Bakker, MJ Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.