Reconstruction of pedigrees in clonal plant populations

Markus Riester[*,a], Peter F. Stadler[a,b,c,d,e], Konstantin Klemm[a]

[a] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany*
[b] *Max-Planck-Insitute for Mathematics in Sciences (MPI-MIS), Inselstrasse 22, D-04103 Leipzig, Germany*
[c] *RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology (IZI), Deutscher Platz 5e, D-04103 Leipzig, Germany*
[d] *Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria*
[e] *The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico*

Abstract

We present a Bayesian method for the reconstruction of pedigrees in clonal populations using co-dominant genomic markers such as microsatellites and single nucleotide polymorphisms (SNPs). The accuracy of the algorithm is demonstrated for simulated data. We show that the joint estimation of parameters of interest such as the rate of self-fertilization is possible with high accuracy even with marker panels of moderate power. Classical methods can only assign a very limited number of statistically significant parentages in this case and would therefore fail. Statistical confidence is estimated by Markov Chain Monte Carlo (MCMC) sampling. The method is implemented in a fast and easy to use open source software that scales to large datasets with many thousand individuals.

Key words: pedigree, parentage, microsatellite, SNP, clonal, selfing, MCMC

1. Introduction

Molecular markers such as highly polymorphic microsatellites (Queller et al., 1993) and more recently also diallelic single nucleotide polymorphisms (SNPs) (Glaubitz et al., 2003; Anderson and Garza, 2006) are now routinely used to genotype individuals in natural populations. Pedigree reconstruction by means of molecular markers has a long history in flowering plant populations, see e.g. Ellstrand and Marshall (1985) and Meagher and Thompson (1987). It has been used mainly to find correlations between phenotypes and reproductive success, or to estimate pollen-mediated gene flow (Smouse and Meagher, 1994; Burczyk et al., 1996; Smouse et al., 1999; Meagher et al., 2003; Wright and Meagher, 2004). To a lesser extent, parentage inference and related methods are used to estimate recent rates of self-fertilization (selfing) in a population (Ritland and Jain, 1981; David et al., 2007; Wilson and Dawson, 2007; Jarne and David, 2008).

Pedigree reconstruction in clonal populations has received very little attention so far, although such an approach holds the promise to allow the direct inference of gene flow from a population's pedigree in particular in long-living clones with limited rate of sexual reproduction. It is a much harder problem than classical paternity or parentage inference for two main reasons: First, it is typically difficult to estimate the age of a clonal plant (Ally et al., 2008) and the absence of age data makes an *a priori* ordering of individuals in generations impossible. Such an ordering is assumed by traditional paternity inference software, see e.g. the works of Marshall et al. (1998); Cercueil et al. (2002); Gerber et al. (2003), and also dramatically restricts the pedigree search space. Second, while it is normally easy to estimate the number of individuals (*ramets*), N_r, in a clonal plant population over the occupied space, it is typically very hard to estimate the number of different genotypes (*genets*), N_g. The genotype number, N_g, usually is a required input parameter in most software for the estimation of the statistical significance of a parentage. Both restrictions can be overcome at least in principle, however. Our recently related pedigree reconstruction tool `FRANz` (Riester et al., 2009) is capable of handling partial or absent age information and can estimate N_g from the data if it is not provided as an input. In this contribution we describe an extension of the `FRANz` approach that specifically handles clonal plant populations.

[*] Corresponding author
Email addresses: `markus@bioinf.uni-leipzig.de` (Markus Riester), `studla@bioinf.uni-leipzig.de` (Peter F. Stadler), `klemm@bioinf.uni-leipzig.de` (Konstantin Klemm)

2. Methods

2.1. Pedigrees

The core of FRANz is a probabilistic model calculating the likelihood of a given *pedigree* of genotyped individuals. Let us start with the necessary definitions.

A pedigree $\mathscr{P} = (V, E)$ is an acyclic digraph with vertex set V and arc set E, where the vertices represent the individuals and the arcs the parent-offspring relationships. Thus V represents the set of all genotyped individuals in the sample. For an arc (u_i, v), we say that v is a *child* of u_i and u_i is a *parent* of v. The set of *parents* of v in $\mathscr{P}$ is denoted by $N^+(v) \subseteq V$; this set may contain two elements, $\{u_i, u_j\}$, one element, $\{u_i\}$, or none, $\emptyset$. In the latter case, v is called a *founder*. In selfing species, $u_i = u_j$ is allowed and $\mathscr{P}$ thus becomes a multigraph. With $N^-(u) \subseteq V$ we denote the *offspring* of u.

For a given individual i, we denote an observed single-locus genotype by g_i and its multi-locus genotype by G_i.

2.2. Likelihood Model

Using Bayes' Theorem, we calculate the posterior probability that the female F_i and the male M_j are the parents of O,

$$\Pr(G_{Fi}, G_{Mj} | G_O, G_F, G_M, A, N_m, N_f) = \frac{\mathrm{T}(G_O | G_{Fi}, G_{Mj}) \Pr(G_{Fi}, G_{Mj})}{\Pr(G_O)} \quad (1)$$

where G_O, G_F, and G_M are the offspring, candidate maternal, and paternal genotypes, A the population allele frequencies, N_m the total number of breeding males (alleged fathers) in the population. Correspondingly, N_f is the total number of candidate mothers. The symbol $\mathrm{T}(\cdot)$ denotes the Mendelian segregation probability, which is the probability that an offspring of F_i and M_j has the genotype G_O. For multi-locus genotypes, we assume here that the loci are *unlinked*, i.e., that the loci are inherited independently during meiosis. Explicit equations for $\mathrm{T}(\cdot)$ that tolerate genotyping errors are derived in Appendix A. $\Pr(G_{Fi}, G_{Mj})$ is the prior probability of F_i and M_j being parents of O. $\Pr(G_O)$ is the marginal probability of the offspring genotype. By assuming equal priors, it can be computed as

$$\Pr(G_O) = \frac{1}{N_m}\left[\sum_k^{n_m} \mathrm{T}(G_O | G_{Fi}, G_{Mk}) + (N_m - n_m)\, \mathrm{T}(G_O | G_{Fi}, A)\right] \quad (2)$$

provided the mother F_i is already known (Nielsen et al., 2001). The first term is the sum of segregation probabilities of all n_m sampled candidate paternal genotypes, the second term accounts for the unsampled ones. This second term is the probability, given the population's allele frequencies and assuming Hardy-Weinberg equilibrium, that a random genotype is the true father, weighted by the number of unsampled candidates. For the case that the mother is unknown, we may write

$$\Pr(G_O) = \frac{1}{N_f N_m}\left[\sum_k^{n_m}\sum_l^{n_f} \mathrm{T}(G_O | G_{Mk}, G_{Fl}) + (N_m - n_m)\sum_l^{n_f} \mathrm{T}(G_O | G_{Fl}, A) + (N_f - n_f)\sum_k^{n_m} \mathrm{T}(G_O | G_{Mk}, A) + (N_m - n_m)(N_f - n_f)\Pr(G_O | A)\right] \quad (3)$$

with n_f denoting the number of sampled female candidates. Here, the first term again collects the segregation probabilities of all sampled male and female genotypes. The following two sums are the cases that either the true father or the true mother are unsampled. The last term accounts for the case that both parents are unsampled, which is the probability of observing the offspring genotype in a population with allele frequencies A, weighted by the number of unsampled pairs.

In monoecious plant populations, where all individuals can mate with each other and with themselves, we finally have:

$$\Pr(G_O) = \frac{2}{N_g(N_g+1)}\left[\sum_{k=1}^{n_g}\sum_{l=k}^{n_g} \mathrm{T}(G_O | G_k, G_l) + (N_g - n_g)\sum_k^{n_g} \mathrm{T}(G_O | G_k, A) + \frac{(N_g - n_g)(N_g - n_g + 1)}{2}\Pr(G_O | A)\right] \quad (4)$$

with n_g denoting the number of sampled genets and G_k the multi-locus genotype of k-th genet; $G_k \neq G_O$. This assumes equal prior probabilities for selfed and outcrossed parentages.

2.3. Efficient Likelihood Calculation

If a single-locus genotype of an offspring does not share one allele with each of the candidate parents, we call this a *mismatch*. In true parent(s)-offspring pairs

or triples, we will observe mismatches only in case of genotyping error (i.e., the true genotype is different from the genotype stored in our dataset) or in case of somatic mutations. With a marker panel of sufficient power for parentage inference, therefore, most of the multi-locus segregation probabilities $T(\cdot)$ that appear in the marginal probabilities $\Pr(G_O)$ (Eq. 2, 3, and 4) will be 0 in the absence of typing errors and mutations. Thus only parentages without mismatching loci need to be considered, which reduces the pedigree search space and the parentage posterior calculation. The first is important for the mixing time of the Markov Chain Monte Carlo (MCMC) sampling, the latter can be a significant speedup especially when A, N and G are variables (see section 2.4).

When tolerating typing errors, however, all parentages have a non-zero probability. An exact computation, therefore, needs to take into account all pairs and triples. However, parentages with many mismatches will have a very small posterior probability and can be ignored. Our implementation uses simulations to generate mismatch distributions for parent(s)-offspring/unrelated relationships in order to determine an appropriate mismatch cut-offs.

For an offspring v, we denote the set of all plausible parents according to this cut-off by $\mathscr{H}_v$. It includes in particular also the cases that none or only one of the parents are sampled. Note that $\mathscr{H}_v \subset V \times V \cup V \cup \{\varnothing\}$. Apart from the number of mismatches, also *prior information* such as sex, age, and known mothers restrict $\mathscr{H}_v$. The posterior probability of a parentage x of offspring i can be expressed in the form:

$$\pi_i(x) = \frac{\Pr(O_i|x)}{\sum\limits_{y_j \in \mathscr{H}_i} \Pr(O_i|y_j)} \,. \tag{5}$$

With $\Pr(O_i|x)$ denoting the probability of parentage x as shown in Eq. 2 to 4. For Eq. 3 we have for example:

$$\begin{aligned}
\Pr(O_i|\{F_i, M_j\}) &= \mathrm{T}(G_{Oi}|G_{Fi}, G_{Mj})/(N_f N_m) \\
\Pr(O_i|\{F_i\}) &= \frac{(N_m - n_m)}{(N_f N_m)} \mathrm{T}(G_{Oi}|G_{Fi}, A) \\
\Pr(O_i|\{\varnothing\}) &= \frac{(N_m - n_m)(N_f - n_f)}{(N_f N_m)} \Pr(G_{Oi}|A)
\end{aligned}$$

Eq. 5 is thus an approximation of Eq. 1 because only plausible parentages are considered. In a second filter step, all parentages with negative LOD score (*e.g.* Meagher and Thompson, 1986) are ignored. These are all parentages which would decrease the pedigree likelihood if the corresponding arcs would be added to the pedigree. If N is estimated jointly with the pedigree, then these two filter steps can introduce a bias. We thus store the sum of the probabilities of all in the second step filtered offspring-candidate mother and offspring-candidate father pairs and all offspring-candidate parents triples. This sum is then added to the denominator of Eq. 5 and the two pair sums are weighted according Eq. 2 to 4.

The posterior probability of the genotype of individual i, conditioned on its parents in a given pedigree $\mathscr{P}$, may be written as $\pi_i(N^+(i))$. Under the assumption that founders are unrelated, the log-likelihood of $\mathscr{P}$ is now the sum of the log-transformed parentage posterior probabilities over all individuals in the pedigree:

$$\mathrm{LL}(\mathscr{P}) \quad = \quad \sum_{i \in V} \log \pi_i(N^+(i)) \tag{6}$$

In clonal populations, we can include the number of ramets for every genet in the posterior calculation:

$$\pi_i(x) = \frac{\Pr(O_i|x) n_r(x)}{\sum\limits_{y_j \in \mathscr{H}_i} \Pr(O_i|y_j) n_r(y_j)} \tag{7}$$

where $n_r(x)$ is the sum of the number of ramets of the parents in parentage x, and $n_r(x) = 1$ if $x = \{\varnothing\}$. This prior increases the likelihood of parentages with frequently observed genets.

2.4. Algorithm

If age data is not or only partially available, then an ordering of individuals in generations is not possible. Thus not all combinations of parentages may represent a valid pedigree of the sample as some of these combinations may introduce directed cycles into the pedigree. In such a "cyclic pedigree", some individuals would be their own ancestors. The MCMC and Simulated Annealing (SA) procedures now sample valid, cycle-free, pedigrees from the pedigree posterior distribution (Almudevar, 2007). We will later use these sampled pedigrees to estimate parameters and to estimate the statistical significance of a parentage. `FRANz` also supports the joint estimation of the population's allele frequencies, the number of unsampled candidate parents or missing data imputation (see below in section 2.5) in which cases we also need MCMC or SA. In these algorithms, one computes the likelihood of a given pedigree $\mathscr{P}_{i-1}$, randomly generates a new pedigree $\mathscr{P}_i$ and then accepts the change if either $\mathscr{P}_i$ has a higher likelihood or with probability $\exp([\mathrm{LL}(\mathscr{P}_i) - \mathrm{LL}(\mathscr{P}_{i-1})]/T)$. More precisely, in the following $\mathscr{P}_{i-1,j}(y)$ denotes a pedigree $\mathscr{P}_{i-1}$ where the parentage of offspring O_j has been changed from x to y; $x, y \in \mathscr{H}_j, x \neq y$. We

use a thread-safe Mersenne Twister (Matsumoto and Nishimura, 2000) to obtain random numbers.

We set $i = 1$ and repeat the following steps until convergence (SA) or until i is greater than a given maximum number of iterations (MCMC).

Pedigree Change Step:
We select a random offspring genotype O_j and select a new parentage for O_j from $\{q(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y))\}$ which is defined as

$$q(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y)) = \begin{cases} \frac{\pi_j(y)}{1-\pi_j(x)} & \text{if } y \neq x \\ 0 & \text{if } y = x \end{cases} \quad (8)$$

This proposal function thus selects a new parentage according their posterior probabilities. We then accept this change with the following probability:

$$\alpha(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y)) = \begin{cases} 0 & \text{if } \mathscr{P}_{i-1,j}(y) \text{ cyclic} \\ \exp\Big(\Big[\mathrm{LL}(\mathscr{P}_{i-1,j}(y)) - \mathrm{LL}(\mathscr{P}_{i-1}) + \\ \qquad \log \frac{q(\mathscr{P}_{i-1,j}(y), \mathscr{P}_{i-1})}{q(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y))}\Big]/T\Big) & \text{otherwise} \end{cases} \quad (9)$$

With (partially) missing age data, to ensure irreducibility of the MCMC sampler, it is necessary to perform swap steps in which the direction of a random arc (j, k) of the pedigree is reversed (Koch et al., 2008). Note that age data implies the direction of an pedigree arc, so a swap step would always return an invalid pedigree in the case that age data is available. We can therefore write down the probability to perform such a swap change in the following form

$$\Pr(\text{swap}) = \begin{cases} \frac{|A|}{|A|+n_o} & \text{if age data missing} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Otherwise we change the parentage of a random offspring as described above. With n_o we denote the number of sampled individuals in the offspring generation(s) ($n_o = n_g$ in the absence of age data) and $|A|$ is the number of arcs in the pedigree. In a swap change, the parentages of two individuals j and k are changed and this change is accepted with probability

$$\alpha(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j,k}(y,z)) = \begin{cases} 0 & \text{if } \mathscr{P}_{i-1,j,k}(y,z) \text{ cyclic} \\ 0 & \text{if } y \notin \mathscr{H}_j \vee z \notin \mathscr{H}_k \\ \exp([\mathrm{LL}(\mathscr{P}_{i-1,j,k}(y,z)) - \mathrm{LL}(\mathscr{P}_{i-1})]/T) & \\ \qquad \text{otherwise} & \end{cases} \quad (11)$$

The second case is necessary because a swap might generate an invalid parentage, for instance one with too many mismatches. In selfing parentages, both arcs are swapped as otherwise a swap would always produce a cycle.

Estimation of the size of the unsampled population:
If the number of unsampled candidates is not known within reasonable accuracy, it is possible to estimate this number together with the pedigree, either by sampling N every n_o steps from a uniform distribution in the interval $[n, N_{max}]$ (where N_{max} is specified by the user) or by treating N as a latent variable, estimated again every n_o steps from the indegree distribution of the pedigree (see Riester et al. (2009) for details).

Temperature schedule:
In the MCMC sampling, the temperature T is always kept at the constant value 1. To speedup mixing of the sampler, `FRANz` supports Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC) (Geyer, 1991). On computers with multiple CPU cores, we recommend starting a chain on every CPU core. The temperature of the i-th chain is set to $1/(1 + 0.5(i - 1))$. Then, the states of the chains i and j are swapped and accepted with probability $\exp([\mathrm{LL}(\mathscr{P}_i) - \mathrm{LL}(\mathscr{P}_j)]/T_j + [\mathrm{LL}(\mathscr{P}_j) - \mathrm{LL}(\mathscr{P}_i)]/T_i)$. Pedigrees are only sampled from the first, unheated chain. In the SA optimization, we use the temperature schedule as described in Aarts and Korst (1989) which has been shown to be efficient in pedigree reconstruction (Almudevar, 2003). For datasets with less than about 30 individuals, we do not use the SA heuristic. Instead, we find the maximum likelihood pedigree with the exact algorithm proposed by Silander and Myllymäki (2006) as described for the pedigree reconstruction problem in Cowell (2009).

2.5. Missing Values

A common problem in most datasets is that genotyping failed for a significant amount of loci. The problem of dealing with missing data has seen remarkably few attention in parentage analysis. `FRANz` offers two options for dealing with missing data. The first is imputation by a single-site Gibbs sampler. Here, the alleles of a random genotype g_{ij} with unobserved data of individual i at locus j are sampled proportional to the product of all affected segregation probabilities of the pedigree:

$$\Pr(g_{ij}|\mathscr{P}, A) = \mathrm{T}(G_i|N^+(i), A) \prod_{o \in N^-(i)} \mathrm{T}(G_o|N^+(o), A) \quad (12)$$

So the segregation probability of i and of all offspring of i need to be updated. The reason for sampling conditional to this product instead of the pedigree likelihood is that it is more sensitive to changes of a single allele.

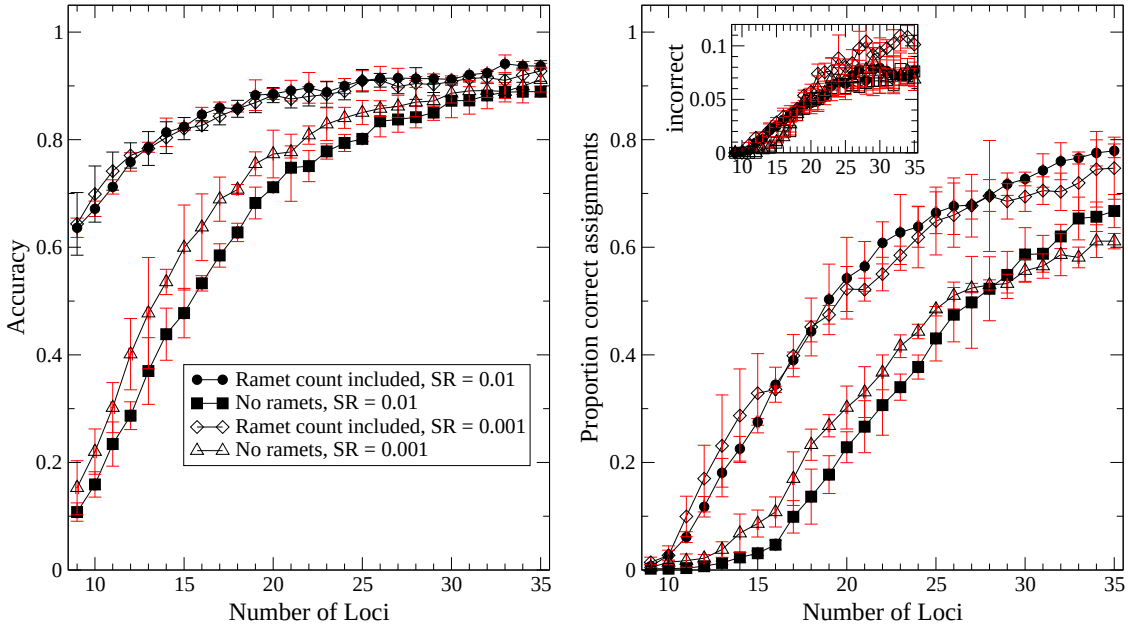

Figure 1: The accuracy of the reconstructed Maximum Likelihood pedigrees is plotted in Fig. 1a as a function of the number of loci. The values are the median accuracy of ten randomly generated pedigrees of size 10.000 genets with a sampling rate (SR) of 0.01 (filled symbols) and 35.000 genets with a sampling rate of 0.001 (unfilled symbols). The simulated datasets are reconstructed once with the standard parentage posterior probabilities (Eq. 5) and once with the number of ramets included as priors (Eq. 7). The error bars indicate the first and third quartile. Fig. 1b lists the proportion of correct and incorrect parentages with a posterior probability > 0.95.

A well-known problem of this approach is that the irreducibility condition is only guaranteed for diallelic loci (Thompson, 2000). This can be circumvented with non-zero segregation probabilities or for example with MCMCMC (Geyer, 1991) samplers (Sheehan and Thomas, 1993; Cannings and Sheehan, 2002). The implemented error model ensures the first with non-zero typing error rates and the latter is also available in our implementation. We make such a Gibbs sampling step after n_o pedigree changes.

The second option for dealing with missing data is to include only observed alleles in the likelihood calculations. Noteworthy, this is the standard method employed by most other parentage or paternity inference tools. `FRANz` also supports partially observed genotypes, see Appendix B for the corresponding segregation probabilities.

2.6. Rates of Self-fertilization

Given a pedigree $\mathscr{P}$, we can estimate the selfing rate r_s over the number of observed self-fertilizations S in $\mathscr{P}$:

$$r_s = \frac{2S}{\sum\limits_{i \in V} |N^+(i)|} \tag{13}$$

The normalization according the indegrees takes account for the fact that the probability of observing an outcrossed parentage is twice as high observing a selfed one, as in the latter case there is only one instead of two parents.

2.7. Rates of Clonality

Estimating the rate at which a population reproduces clonally is notoriously difficult (de Meeûs and Balloux, 2004). We assume in the following that we can estimate N_r, the total number of ramets in the population, within reasonable accuracy for example over the occupied space. We further assume a population in which N_r is constant across generations. In every generation, a ramet reproduces either clonally with rate c or sexually with rate $(1 - c)$. A random ramet is killed to keep N_r constant if necessary. Then we use pedigree reconstruction to estimate the total number of genets, N_g. We then

estimate the rate of clonal reproduction with the following equations.

By $h(s)$ we denote the expectation value of the number of genets with exactly s ramets, for integer $s \in \{0, \ldots, N_r\}$. We have $h(0) = 0$ by definition. Let us establish conditions for the stationary values of h. In a birth event, $h(s)$ changes by

$$\Delta^{\mathrm{b}}(s) = -c\frac{s}{N_r}h(s) + c\frac{s-1}{N_r}h(s-1) \tag{14}$$

for all $s \in \{2, \ldots, N_r\}$, whereas $h(1)$ is changed by

$$\Delta^{\mathrm{b}}(1) = (1-c) - \frac{c}{N_r}h(1)\,. \tag{15}$$

A death event causes a change

$$\Delta^{\mathrm{d}}(s) = -\frac{s}{N_r}h(s) + \frac{s+1}{N_r}h(s+1)\,. \tag{16}$$

for all $s \in \{1, \ldots, N\}$. At constant population size N_r, birth and death events occur equally often. Therefore

$$\Delta^{\mathrm{b}}(s) + \Delta^{\mathrm{d}}(s) = 0 \tag{17}$$

must be fulfilled in equilibrium for all $s\{1, \ldots, N_r\}$. We obtain the set of equations

$$\begin{aligned} h(0) &= 0 \\ h(1) &= N_r(1-c) \\ h(2) &= \frac{1}{2}\left[(1+c)h(1) - N_r(1-c)\right] \\ h(s+1) &= \frac{1}{s+1}\left[s(1+c)h(s) - c(s-1)h(s-1)\right] \end{aligned} \tag{18}$$

the last equation being valid for all $s \in \{2, \ldots, N_r\}$. Taken together, this is a second order linear difference equation for a given c and N_r. We solve it numerically and obtain the expected number of genotypes as

$$N_g = \sum_{s=1}^{N_r} h(s)\,. \tag{19}$$

For the inverse problem with given N_g, we obtain c by means of nested intervals.

3. Simulated Test Data

3.1. Growing Population

To test our algorithm and implementation, we first simulate data under the model of a growing population, where individuals do not die once they reproduced sexually. The data is simulated with allele frequencies with 8 alleles per locus. We use random ("broken stick") frequencies with a maximum frequency of 0.5. In every year, a ramet reproduces either sexually or clonally. For sexual reproduction we assume a fixed rate of self-fertilization or outcrossing with a random ramet. With a probability of 0.01, we replace one allele of a single-locus genotype with a random one to simulate genotyping errors.

We choose relatively high rates of selfing (0.1) and clonality (0.9). This results in an extremely difficult test dataset as individuals are closely related because old plants grow fast and thus mate very often. Exclusion probabilities (Jamieson and Taylor, 1997; Wang, 2007), which assume unrelatedness among candidate parents, thus overestimate the power of a marker suite. The exclusion probabilities of the simulated datasets are shown in Supplementary Fig. S1.

For the sampling rates, we choose a relatively high one of 0.01 of (1.000 of about 100.000 ramets) and a more realistic one of 0.001 (350 of about 350.000). We generate 10 datasets for each of the two sampling rates.

3.2. Constant Population Size

To show that the present approach is able to estimate the sampling rate of the genets, we next simulate data under the model of population with a constant number of ramets N_r. We start with a founder generation of 100 unrelated genets with 9 loci and use the same allele frequencies as before. Then in every generation, all ramets reproduce again either sexually or clonally. If sexually, then again with a selfing rate of 0.1. Outcrossing happens with a random living ramet or with an migrated ramet. Here we use a migration rate of 0.01. If after such a reproduction event there are more than N_r ramets, one random living ramet is killed. We stop the simulation after the birth of the 20000-th genet. Then we sample $n_r = 500$ living ramets. We generate again 10 datasets for every parameter combination. Here we vary the rate of clonality (0.5, 0.8, 0.9 and 0.95) and N_r (4.000 and 10.000). N_g is then estimated over the indegree distributions of the MCMC sampled pedigrees (Riester et al., 2009). The rate of clonality is then estimated with the model described in Sec. 2.7. N_r is assumed to be known *a priori*.

4. Results

4.1. Growing Population

The accuracy of the pedigree reconstructions, defined as the proportion of correct parentages in the SA Maximum Likelihood pedigree, are shown in Fig. 1a. The incorporation of the number of sampled ramets per genet

(using Eq. 7 instead of Eq. 5) improves the reconstruction significantly. A reason for that improvement is that the number of ramets is an approximation of the age of genets; without age data it is sometimes not possible to identify parent and offspring in a parent-offspring pair, which is also the reason why the accuracy does not reach 100%. The plot also shows that the sampling rate has surprisingly little influence on the accuracy, the amount of available genomic information is the crucial factor here.

Fig. 1b shows the fraction of correct and incorrect parentages with a pedigree posterior probability of > 0.95. These are parentages that are observed in at least 95% of all sampled pedigrees. Although these probabilities are not exactly comparable to the thresholds of classical, simulation based paternity inference tools such as CERVUS (Marshall et al., 1998) due to the very different approaches, the plot nevertheless shows that even with powerful marker panels, i.e., when there is a high amount of genomic information in the data, we can only assign relatively small numbers of parentages, which is especially a problem in low sampling rate datasets. It is therefore crucial to use an approach that uses the complete data for the estimation of parameters of interest, not only highly significant parentages. The high rate of incorrect assignments, especially in the 0.001 dataset, is explained by the large violation of the assumption that candidate parents are unrelated.

We then use the MCMC sampled pedigrees to estimate the parameters of interest. As an example, we plot in Fig. 2 the estimated rates of self-fertilization (Eq. 13) for both sampling rates (0.01 and 0.001, using Eq. 7) as a function of the number of loci. As to expect, the accuracy of the test dataset with high sampling rate (Fig. 2a) is higher than the one with lower sampling rate (Fig. 2b) because the number of observed parentages is much higher. The estimates are fairly independent of the number of loci and already accurate with very low amount of genomic information. Other parameters such as male fertilities (Morgan and Conner, 2001) could be calculated analogously. In Fig. 2c we show the good mixing properties of our sampler with trace plots of the pedigree log-likelihood and the selfing rate. Sampling of 20.000 pedigrees (9 loci, 0.01 dataset) takes less than a second on a modern[1] computer and we observe a Metropolis-Hastings acceptance rate of 0.33 (data not shown).

We also compare our selfing rate estimates with the Maximum Likelihood estimates of the RMES software (David et al., 2007) in Table 1. RMES estimates selfing rates over observed multi-locus heterozygosity deficiencies and does not require parent-offspring relationships. These allele frequency approaches are therefore in principle capable of estimating *long-term* selfing rates but are inherently less robust with respect to violations of the assumptions. `FRANz`, on the other hand, provides quite accurate estimates of the *recent* selfing rates for the 10 datasets. This pedigree reconstruction approach works in the model of a growing population even with very low sampling rates extremely well because old founder plants are sampled with probability close to one and these plants have many offspring. This is the reason why we observe enough parentages for reliable parameter estimation.

S.R.[a]	Loci	`RMES`		`FRANz`	
0.01	9	0.154	±0.044	0.100	±0.012
0.01	35	0.176	±0.026	0.096	±0.023
0.001	9	0.148	±0.055	0.099	±0.027
0.001	35	0.175	±0.025	0.103	±0.034

Table 1: Comparison of the estimated rates of self-fertilization. This table lists the means and standard deviations of the selfing rate estimates from the RMES software (David et al., 2007) and the present approach for simulated datasets (Sec. 3.1) with a true selfing rate of 0.1. [a]S.R. sampling rate (ramets)

4.2. Constant Population Size

In Table 2 we present the results of the pedigree reconstruction of the datasets with constant population size (Sec. 3.2). Our approach significantly underestimates the true N_g. This is partly explained by the fact that the probability of sampling old and big plants, i.e., ones with many ramets is higher than sampling small genets. And as old plants have in general more offspring than young ones, we observe more parentages as we would expect by assuming equal sampling probabilities of parents for all individuals, which we do. More observed parentages result in a higher sampling rate and therefore a smaller N_g. Our model (Sec. 2.7) also slightly underestimates the true N_g. Nevertheless, with relatively high sampling rates of 0.125, we can observe fairly accurate estimates. As the number of genets increases with decreasing rate of clonal reproduction, we observe less parentages if these rates are low. This explains the high variances in the datasets with a clonal rate of 0.5. At clonal rates smaller than 0.9, we see that a sampling rate of 0.05 is not high enough for a reliable

[1]Intel® Xeon® CPU, 2.33GHz, 8 cores, 16GB RAM

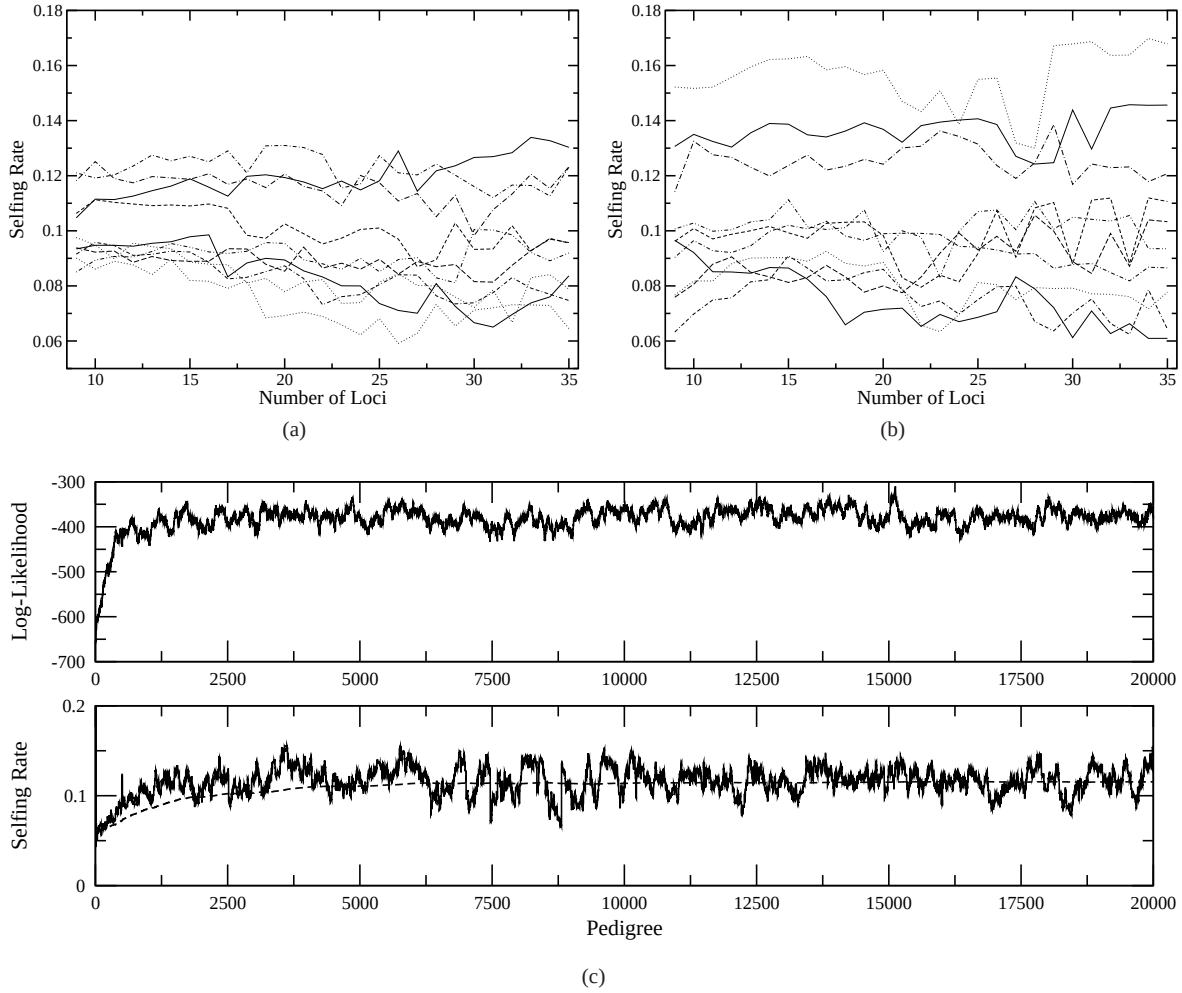

Figure 2: The estimated rates of self-fertilization of the simulated datasets (Sec. 3.1). A line visualizes the selfing rate of one of the 10 datasets for different amounts of genomic information. Fig. 2a are the rates from the 0.01 sampling rate datasets, Fig. 2b from the 0.001 one. In Fig. 2c we show a trace plot of the MCMCMC sampling of one simulated dataset (9 loci, sampling rate 0.01, 8 parallel chains). The dashed curve visualizes the mean of the selfing rate.

parameter estimation: the selfing rates are in these cases also significantly underestimated.

5. Discussion

We have presented a novel likelihood model for the reconstruction of pedigrees in monoecious clonal plant populations and demonstrated the accuracy and good MCMC(MC) mixing properties of our open source implementation on simulated data. Our efficient likelihood calculations allows parentage analysis on huge datasets with thousands of individuals. We have shown that the joint estimation of parameters of interest such as the rate of self-fertilization is possible with high accuracy even with marker panels of moderate power. Classical methods can only assign a very limited number of statistically significant parentages in this case and would therefore fail, especially if sampling rates are low which is still a problem in most parentage studies. We have also shown that our likelihood model is surprisingly robust for violations of assumptions such as unrelatedness of candidate parents and constant effective population size. With relatively high sampling rates, we were further able to give fairly accurate estimates of the rates of clonal reproduction in simulated populations with constant size.

As mating success drops off with distance between mates, several authors suggested likelihood models that

S.R.[a]	N_g					Rate of Clonality			Selfing Rate	
	FRANz		True		Model	FRANz		True	FRANz	
0.125	2574.95	±614.77	2834.00	±20.18	2772.59	0.541	±0.19	0.5	0.079	±0.04
0.05	5832.13	±3508.20	6955.67	±43.36	6931.47	0.611	±0.32	0.5	0.037	±0.03
0.125	1460.90	±228.39	1711.20	±18.88	1609.44	0.826	±0.05	0.8	0.082	±0.02
0.05	2447.02	±522.00	4000.80	±31.02	4023.59	0.905	±0.03	0.8	0.068	±0.02
0.125	876.26	±80.41	1121.90	±18.89	1023.37	0.920	±0.01	0.9	0.095	±0.02
0.05	1814.28	±243.40	2792.70	±37.12	2558.43	0.939	±0.01	0.9	0.066	±0.03
0.125	555.72	±45.25	720.10	±21.70	630.68	0.958	±0.01	0.95	0.094	±0.03
0.05	1218.60	±121.70	1847.80	±31.42	1576.70	0.965	±0.01	0.95	0.081	±0.03

Table 2: Estimated rates of N_g, clonality and self-fertilizations. This table lists means and standard deviations of the in FRANz estimated N_g for ten simulated datasets (Sec. 3.2) for each of the 8 parameter combinations. It further lists the true N_g and the ones estimated of our model (Sec. 2.7). Then the estimated rates of clonality are presented. Finally, the mean and standard deviations of estimated rates of self-fertilization are shown (with a true selfing rate of 0.1). [a]S.R. sampling rate (ramets)

include the sampling location of the genotypes (*e.g.* Adams et al., 1992; Burczyk et al., 1996; Smouse et al., 1999; Hadfield et al., 2006). In principle, it is possible to add the corresponding prior probability distributions in our model. To calculate the distance between two clones A and B, δ_{AB}, one can use the locations of all sampled ramets as an approximation for the real distance between the (maybe unsampled) mating ramets. An obvious strategy for the calculation of δ_{AB} would be the average distance between all sampled ramets of A and B. Another possibility would be to use the minimum distance.

It should be noted that other methods for the estimation of recent selfing rates exist which do not necessarily require that parental genotypes are sampled. For example if it is possible to obtain progeny arrays, the known family structure in the data can be used to reconstruct maternal genotypes. Selfing rates are then estimated by comparison of maternal with offspring genotypes (*e.g.* Jarne and David, 2008, for a review). If neither such a family structure nor parental genotypes are known, then reconstruction of the genotypes of the previous generations might be possible by MCMC sampling (Wilson and Dawson, 2007). However, this assumes that the model used in MCMC sampling fits the population under investigation.

We assumed in this article that all ramets have the same genotype. However, especially in long-living plant populations with high rates of clonality, somatic mutations may lead to clones with different genotypes. In this case it could be necessary to extend the model to allow multiple genotypes per genet and include them in the segregation probability calculation (see Appendix A). Our implementation FRANz supports partially genotyped loci where only one of the two alleles are known and this feature could be used in these cases to mark an observed mutation as unknown without loosing much information.

Availability of the implementation

An open source implementation of FRANz is available under http://www.bioinf.uni-leipzig.de/Software/FRANz. For a simple interaction and comparison with other tools, we provide a user-friendly Web 2.0 input file generator on the FRANz website. Furthermore, it is now possible to convert FRANz input files into several other formats (currently supported are CERVUS (Marshall et al., 1998; Kalinowski et al., 2007), PARENTE (Cercueil et al., 2002), GENEPOP (Roussett, 2008), and RMES (David et al., 2007)).

Acknowledgements

We are grateful to Thomas R. Meagher for his literature recommendations and the referees for insightful comments. The authors further want to thank Gunnar Boldhaus and Florian Greil for comments on the manuscript. This paper also benefited from discussions with the people of the Leipzig Network Seminar and from inspiration from Hal Incandenza. This work was supported by the 6th Framework Programme of the European Union, project 043251 “EDEN”.

References

Aarts, E., Korst, J., 1989. Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing. Wiley, Chichester.

Adams, W. T., Griffin, A. R., Moran, G. F., Nov 1992. Using paternity analysis to measure effective pollen dispersal in plant populations. Am Nat 140 (5), 762–780.

Ally, D., Ritland, K., Otto, S. P., Nov 2008. Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in *Populus tremuloides*. Mol Ecol 17 (22), 4897–4911.

Almudevar, A., Mar 2003. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theor Popul Biol 63, 63–75.

Almudevar, A., Mar 2007. A graphical approach to relatedness inference. Theor Popul Biol 71 (2), 213–229.

Anderson, E. C., Garza, J. C., Apr 2006. The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics 172 (4), 2567–2582.

Burczyk, J., Adams, W. T., Shimizu, J. Y., 1996. Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuate Lemmon.*) stand. Heredity 77, 251–260.

Cannings, C., Sheehan, N. A., Oct 2002. On a misconception about irreducibility of the single-site gibbs sampler in a pedigree application. Genetics 162 (2), 993–996.

Cercueil, A., Bellemain, E., Manel, S., Nov-Dec 2002. Parente: computer program for parentage analysis. J Hered 93 (6), 458–459.

Cowell, R. G., Dec 2009. Efficient maximum likelihood pedigree reconstruction. Theor Popul Biol 76 (4), 285–291.

David, P., Pujol, B., Viard, F., Castella, V., Goudet, J., Jun 2007. Reliable selfing rate estimates from imperfect population genetic data. Mol Ecol 16 (12), 2474–2487.

de Meeûs, T., Balloux, F., Dec 2004. Clonal reproduction and linkage disequilibrium in diploids: a simulation study. Infect Genet Evol 4 (4), 345–351.

Ellstrand, N. C., Marshall, D. L., 1985. Interpopulation gene flow in *Raphanus sativus*. American Naturalist 126, 606–616.

Gerber, S., Chabrier, P., Kremer, A., 2003. FaMoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. Molecular Ecology Notes 3 (3), 479–481.

Geyer, C. J., 1991. Markov chain monte carlo maximum likelihood. In: Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation. Fairfax Station, pp. 156–163.

Glaubitz, J. C., Rhodes, O. E., Dewoody, J. A., Apr 2003. Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Mol Ecol 12 (4), 1039–1047.

Hadfield, J. D., Richardson, D. S., Burke, T., Oct 2006. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Mol. Ecol. 15, 3715–3730.

Jamieson, A., Taylor, S., Dec 1997. Comparisons of three probability formulae for parentage exclusion. Anim. Genet. 28, 397–400.

Jarne, P., David, P., Apr 2008. Quantifying inbreeding in natural populations of hermaphroditic organisms. Heredity 100 (4), 431–439.

Kalinowski, S. T., Taper, M. L., Marshall, T. C., Mar 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16, 1099–1106.

Koch, M., Hadfield, J. D., Sefc, K. M., Sturmbauer, C., Oct 2008. Pedigree reconstruction in wild cichlid fish populations. Mol Ecol 17 (20), 4500–4511.

Marshall, T. C., Slate, J., Kruuk, L. E., Pemberton, J. M., May 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639–655.

Matsumoto, M., Nishimura, T., 2000. Dynamic creation of pseudorandom number generators. In: Monte Carlo and Quasi-Monte Carlo Methods 1998: Proceedings of a Conference, Held at Claremont Graduate University, Claremont, California, USA. Springer, pp. 56–69.

Meagher, T. R., 1991. Analysis of paternity within a natural population of *Chamaelirium luteum*. II. Patterns of male reproductive success. The American Naturalist 137 (6), 738–752.

Meagher, T. R., Belanger, F. C., Day, P. R., Jun 2003. Using empirical data to model transgene dispersal. Philos Trans R Soc Lond B Biol Sci 358 (1434), 1157–1162.

Meagher, T. R., Thompson, E. A., February 1986. The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theoretical Population Biology 29 (1), 87–106.

Meagher, T. R., Thompson, E. A., 1987. Analysis of parentage for naturally established seedlings of *Chamelirium Luteum* (Liliaceae). Ecology 68 (4), 803–812.

Morgan, M. T., Conner, J. K., Feb 2001. Using genetic markers to directly estimate male selection gradients. Evolution 55 (2), 272–281.

Nielsen, R., Mattila, D., Clapham, P., Palsbøll, P., Apr 2001. Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. Genetics 157, 1673–1682.

Queller, D. C., Strassmann, J. E., Hughes, C. R., 1993. Microsatellites and kinship. Trends in Ecology & Evolution 8 (8), 285 – 288.

Riester, M., Stadler, P. F., Klemm, K., Aug 2009. FRANz: Reconstruction of wild multi-generation pedigrees. Bioinformatics 25 (16), 2134–2139.

Ritland, K., Jain, S., 1981. A model for the estimation of outcrossing rate and gene frequencies using n independent loci. Heredity 47 (1), 35–52.

Roussett, F., 2008. GENEPOP'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources 8 (1), 103–106.

Sheehan, N., Thomas, A., Mar 1993. On the irreducibility of a markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49 (1), 163–175.

Silander, T., Myllymäki, P., 2006. A simple approach for finding the globally optimal bayesian network structure. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). AUAI Press, Arlington, Virginia, pp. 445–452.

Smouse, P. E., Meagher, T. R., Jan 1994. Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae). Genetics 136 (1), 313–322.

Smouse, P. E., Meagher, T. R., Kobak, C. J., 1999. Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): why do some males have higher reproductive contributions? Journal of Evolutionary Biology 12 (6), 1069–1077.

Thompson, E. A., 2000. Statistical inference from genetic data on pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, IMS, Beachwood, OH.

Wang, J., Aug 2007. Parentage and sibship exclusions: higher statistical power with more family members. Heredity 99 (2), 205–217.

Wilson, I. J., Dawson, K. J., Nov 2007. A markov chain monte carlo strategy for sampling from the joint posterior distribution of pedigrees and population parameters under a fisher-wright model with partial selfing. Theor Popul Biol 72 (3), 436–458.

Wright, J. W., Meagher, T. R., Mar 2004. Selection on floral characters in natural spanish populations of *Silene latifolia*. J Evol Biol 17 (2), 382–395.

A. CERVUS Likelihood formulas

In the following, we present the likelihood formulas for paternity or parentage inference of the typing error model described in Kalinowski et al. (2007). We corrected some typos in the original version presented in the appendix of Kalinowski et al. (2007) and also simplified the formulas where possible. $L(H_1)$ is the likelihood of the hypothesis H_1 that the alleged parent is the true parent; the alternative hypothesis H_2 is that the alleged parent is unrelated. We follow here the notation of the original paper instead of ours: the single locus genotypes of mother, alleged father and offspring are denoted with g_m, g_a and g_o (corresponding to g_f, g_m and g_o in our notation). $\Pr(g)$ is the probability of observing the genotype g in a population in Hardy-Weinberg equilibrium. For diploid heterozygotes, the probability of a genotype with the alleles a_1 and a_2 and with the allele frequencies p and q is $\Pr(a_1, a_2) = 2pq$; for homozygotes, we have $\Pr(a_1, a_1) = p^2$. The estimated typing error rate is the probability that one or both alleles of an genotype are not correctly observed and is denoted as ϵ. Finally, $\mathrm{T}(\cdot)$ denotes the Mendelian segregation probabilities (see for example Meagher (1991)) and $\mathrm{T}_e(\cdot)$ is the variant of the error model (Kalinowski et al., 2007). For details see Marshall et al. (1998); Kalinowski et al. (2007). The likelihoods for paternity when the mother is unknown are:

$$
\begin{aligned}
L(H_1) &= \Pr(g_a)\mathrm{T}_e(g_o|g_a, \epsilon) \\
L(H_1) &= \Pr(g_a)\{(1-\epsilon)^2\mathrm{T}(g_o|g_a) + \epsilon(1-\epsilon)2\Pr(g_o) + \epsilon^2\Pr(g_o)\} \\
&= \Pr(g_a)\{(1-\epsilon)^2\mathrm{T}(g_o|g_a) + \epsilon(2-\epsilon)\Pr(g_0)\} \\
L(H_2) &= \Pr(g_a)\{\Pr(g_o)\}
\end{aligned}
$$

The likelihoods for paternity and maternity jointly are

$$
\begin{aligned}
L(H_1) &= \Pr(g_m)\Pr(g_a)\mathrm{T}_e(g_o|g_m, g_a, \epsilon) \\
L(H_1) &= \Pr(g_m)\Pr(g_a)\{(1-\epsilon)^3\mathrm{T}(g_o|g_m, g_a) + \epsilon(1-\epsilon)^2 \\
&\quad [\mathrm{T}(g_o|g_m) + \mathrm{T}(g_o|g_a) + \Pr(g_o)] + \epsilon^2(3-2\epsilon)\Pr(g_o)\} \\
L(H_2) &= \Pr(g_m)\Pr(g_a)\{\Pr(g_o)\}
\end{aligned}
$$

The likelihood of the alternative hypothesis H_2 for paternity when the mother is known is:

$$
\begin{aligned}
L(H_2) = \Pr(g_m)\Pr(g_a)\{&(1-\epsilon)^3\mathrm{T}(g_o|g_m) + \epsilon(1-\epsilon)^2 \\
&[\mathrm{T}(g_o|g_m) + 2\Pr(g_0)] + \epsilon^2(3-2\epsilon)\Pr(g_o)\}
\end{aligned}
$$

$L(H_1)$ is the same as for the parentage inference case.

B. Missing Values

The following equations are the segregation probabilities with (partially) missing data for genotypes, pairs and triples.

B.1. Genotype probabilities

If one allele of a single locus genotype is missing, then all alleles are considered and we have $\Pr(?) = 1$, where the question mark codes a missing allele. The genotype probabilities are thus:

$$
\Pr(?.?) = 1,\ \Pr(a_i.?) = \Pr(a_i) \tag{20}
$$

B.2. Pairs

$$
\delta(a_o, a_p) = \begin{cases} 1 & \text{if } a_o = a_p \vee a_o = ? \\ \Pr(a_o) & \text{if } a_p = ? \\ 0 & \text{otherwise} \end{cases}
$$

Case 1: For parent-offspring pairs, we have with both parental alleles missing no additional information and thus have the genotype probability: $\mathrm{T}(a_i.a_j|?.?) = \Pr(a_i.a_j)$.

Case 2: One offspring allele missing

$$
\mathrm{T}(a_i.?|a_j.a_k) = 0.5\Pr(a_i) + 0.25\left[\delta(a_i, a_j) + \delta(a_i, a_k)\right] \tag{21}
$$

Case 3: One parental allele missing

$$
\begin{aligned}
\mathrm{T}(a_i.a_i|a_j.?) &= 0.5\left[\delta(a_i, a_j)\Pr(a_i) + \Pr(a_i.a_i)\right] \\
\mathrm{T}(a_i.a_j|a_k.?) &= 0.5\left[\delta(a_i, a_k)\Pr(a_j) + \delta(a_j, a_k)\Pr(a_i)\right] \\
&\quad + 0.5\Pr(a_i.a_j)
\end{aligned} \tag{22}
$$

B.3. Triples

Case 1: Both maternal or paternal alleles missing:

$$
\mathrm{T}(a_i.a_j|a_k.a_l, ?.?) = \mathrm{T}(a_i.a_j|a_k.a_l) \tag{23}
$$

Case 2: One offspring allele missing:

$$
\mathrm{T}(a_i.?|a_{j1}.a_{j2}, a_{j3}.a_{j4}) = \frac{1}{4}\sum_{k=1}^{4}\delta(a_i, a_{jk}) \tag{24}
$$

Case 3: One maternal and/or paternal allele missing:

$$
\delta(a_{o1}.a_{o2}, a_{p1}.a_{p2}) = \begin{cases} 1 & \text{if } a_{o1}.a_{o2} = a_{p1}.a_{p2} \\ \Pr(a_{o1}.a_{o2}) & \text{if } a_{p1} = ? \wedge a_{p2} = ? \\ \Pr(a_{o1}) & \text{if } a_{o2} = a_{p1} \wedge a_{p2} = ? \\ \Pr(a_{o2}) & \text{if } a_{o1} = a_{p1} \wedge a_{p2} = ? \\ 0 & \text{otherwise} \end{cases}
$$

$$
\mathrm{T}(a_{o1}.a_{o2}|a_{m1}.a_{m2}, a_{f1}.a_{f2}) = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\delta(a_{o1}.a_{o2}, a_{mi}.a_{fj}) \tag{25}
$$

C. Supplementary Figures

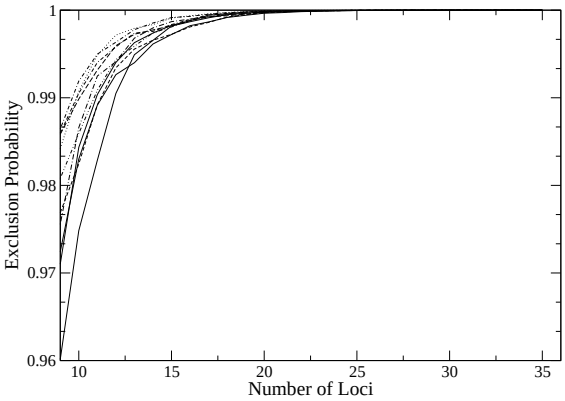

Figure S1: The exclusion probabilities (Jamieson and Taylor, 1997) of 10 randomly generated datasets (3). In our data, the average probability that a random individual has a genotype that is compatible with an offspring genotype is $< 1 \times 10^{-5}$ for more than 25 loci.