BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 16 2009, pages 2134–2139
doi:10.1093/bioinformatics/btp064

Genetics and population analysis

FRANz: reconstruction of wild multi-generation pedigrees

Markus Riester[1,*], Peter F. Stadler[1,2,3,4] and Konstantin Klemm[1]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, [2]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology (IZI), Perlickstraße 1, D-04103 Leipzig, Germany, [3]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria and [4]The Santa Fe Institute, 1399 Hyde Park Road., Santa Fe, New Mexico, USA

Received and revised on October 31, 2008; accepted on January 26, 2009

Advance Access publication February 8, 2009

Associate Editor: Trey Ideker

ABSTRACT

Summary: We present a software package for pedigree reconstruction in natural populations using co-dominant genomic markers such as microsatellites and single nucleotide polymorphisms (SNPs). If available, the algorithm makes use of prior information such as known relationships (sub-pedigrees) or the age and sex of individuals. Statistical confidence is estimated by Markov Chain Monte Carlo (MCMC) sampling. The accuracy of the algorithm is demonstrated for simulated data as well as an empirical dataset with known pedigree. The parentage inference is robust even in the presence of genotyping errors.

Availability: The C source code of FRANz can be obtained under the GPL from http://www.bioinf.uni-leipzig.de/Software/FRANz/.

Contact: markus@bioinf.uni-leipzig.de

1 INTRODUCTION

The reconstruction of genealogical relationships among diploid species has been an active field of research for more than three decades. A well-developed statistical theory of paternity inference has been published in series of articles by E. A. Thompson (e.g. Thompson, 1976). The study of parentage in natural populations was the topic of the pioneering papers by Meagher and Thompson (1986) and Marshall *et al.* (1998) and is recently reviewed in Blouin (2003); Jones and Ardren (2003); Pemberton (2008). The pedigree structure of a sample of individuals is important for a wide range of ecological, evolutionary and forensic studies. Applications include genealogy reconstruction (e.g. for wine grape cultivars Vouillamoz and Grando, 2006), the estimation of heritabilities in the wild (Thomas and Hill, 2000) and victim identification (Lin *et al.*, 2006).

In order to reconstruct the pedigree of a sample, the parents of each individual in the sample need to be determined. If one has a large amount of genomic data, the task of identifying first degree relationships, i.e. parent–offspring and full-sibs relations, is trivial. Unfortunately, many datasets in natural populations do not contain enough information to unambiguously determine the parents. Another problem is that datasets often contain only a subset of a population. Thus, one or both parents of an observed individual may be missing from the dataset. Furthermore, many datasets are not free of errors.

Most programs support only datasets comprising one or two generations. The approach to partial pedigree reconstruction in one generation datasets are sibship algorithms. Here, genotype data is used to infer full-sib and half-sib relationships (Berger-Wolf *et al.*, 2007; Thomas and Hill, 2002; Wang, 2004b). The parentage inference programs for two generations typically take an offspring list, if known their mothers, and a list of candidate parents or fathers as input and generate the possible parent combinations (Hadfield *et al.*, 2006; Kalinowski *et al.*, 2007). Much less attention (e.g. in Almudevar, 2003) has been given to multi-generation pedigrees in which the offspring and candidate parent sets are not necessarily non-overlapping. This is the case, for example, in the absence of age data. Then the ordering of genotypes into generations is not known a priori and has to be estimated from the genotype data only. Thus, at difference with parentage inference programs, the general case treated also here does not admit all possible parentage combinations as valid pedigrees. The task is therefore to find the parentage combinations that define the *maximum likelihood pedigree*. If the number of possible pedigrees is too large too enumerate, heuristics are necessary. So far, a flexible software package has not been available that allows the incorporation of prior information in addition to the genotypes and that is robust in the case of errors. It is the purpose of this contribution to fill this gap.

2 DEFINITIONS

A pedigree $\mathscr{P}=(V,A)$ is an acyclic digraph with vertex set V and arc set A. 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 (putative) *parents* of v is denoted by $N^+(v)\subseteq V$; it may have cardinality 2 $\{u_i,u_j\}$, 1 $\{u_i\}$ or 0 $\emptyset$. In the latter case, v is called a *founder*. In selfing species, $u_i=u_j$ is allowed and $\mathscr{P}$ is a multigraph. The set of all valid parent combinations of v is denoted by $\mathscr{H}(v)$. Again we include the cases that none or only one of the parents are present in V. Note that $\mathscr{H}(v)\subset V\times V\cup V\cup\{\varnothing\}$. The Mendelian laws of inheritance and *prior information* such as sex, age and known mothers restrict $\mathscr{H}(v)$.

For each individual, we have to choose one parent combination $N^+(v)\in\mathscr{H}(v)$. Not all such combinations of parents are possible, because this may introduce directed cycles into the pedigree. $\mathscr{T}$ denotes the set of all *valid pedigrees*.

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

[*]To whom correspondence should be addressed.

3 BACKGROUND

3.1 LOD scores

Consider a triplet of individuals (A, B, C) with single-locus genotypes g_A, g_B and g_C. In likelihood-based paternity analyses, one compares the likelihood of the hypothesis (H_1) that the three individuals are offspring, mother and father, with the likelihood of the alternative hypothesis (H_2) that the three individuals are unrelated. This comparison is usually expressed as a log-ratio, the *parent-pair log-odds ratio (LOD) score* (e.g. Meagher and Thompson, 1986):

$$\mathrm{LOD}(g_A,g_B,g_C)=\log\frac{\Pr(g_A,g_B,g_C|H_1)}{\Pr(g_A,g_B,g_C|H_2)}=\log\frac{\mathrm{T}(g_A|g_B,g_C)}{\Pr(g_A)}$$

The likelihood of (H_2) is the probability of observing the three genotypes when randomly drawn from 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 Mendelian transmission probability is denoted by $\mathrm{T}(\cdot)$. Variations of this equation can be derived for the cases where only one parent is sampled (*single-parent* LOD scores) and for triples where the relationship of two individuals A and B, typically mother and offspring, is known (Kalinowski *et al.*, 2007; Meagher and Thompson, 1986).

3.2 Statistical significance of a parentage

Different ways of assessing the confidence of the parentage with the largest LOD score have been proposed. Marshall *et al.* (1998) use ΔLOD as test statistic, which is the difference of the LOD scores between the two most likely parentages. The critical value of this test statistic is obtained by simulation. If not all individuals of the population are sampled, then the total number of breeding individuals N in the population must be estimated and incorporated in the simulation. Nielsen *et al.* (2001) proposed a Bayesian approach, extending the *fractional paternity* approach suggested by Devlin *et al.* (1988). The posterior probability that male F_i is the father of O can now be calculated for the case when the mother M is known as

$$\Pr(F_i|G_O,G_M,G_F,A,N)=\frac{\mathrm{T}(G_O|G_M,G_{Fi})}{\sum_j^n \mathrm{T}(G_O|G_M,G_{Fj})+(N-n)\,\mathrm{T}(G_O|G_M,A)}$$

where G_O, G_M and G_F are the offspring, maternal and paternal genotypes, A the population allele frequencies and n the number of sampled males. So $(N-n)$ weights the case that the true father is unsampled accordingly. Ignoring this weighting will give many false matches when the sampling rate and the amount of genomic information is low (Nielsen *et al.*, 2001). In the following, we shortly write $\Pr(N^+(v_i)|A,N)$ for the parentage posterior probability of vertex v_i.

For the case that the mother is unknown and assuming that the numbers of breeding males and females do not differ significantly, we have to add $(N-n)^2\ \Pr(G_O|A)$ to the denominator to weight the case that both parents are unsampled.

One important advantage of this Bayesian approach over the simulation approach is that for the case that N is not known with high confidence, it is possible to estimate this value simultaneously with the pedigree reconstruction.

3.3 IBD coefficients

For each pair of individuals, we can calculate the probability that the two have a particular relationship $\mathbb{R}$: unrelated $\mathbb{U}$, parent–offspring $\mathbb{PO}$, full-sib $\mathbb{FS}$, half-sib $\mathbb{HS}$, etc. The usual way of calculating the likelihoods $\Pr(g_A.g_B|\mathbb{R})$ uses the so-called *IBD* (*identical by descent*) *coefficients* k_0, k_1 and k_2. Alleles are IBD if they are identical and are segregated from a recent common ancestor. A child, for example, shares with each parent exactly one allele that is IBD ($k_1=1$); monozygotic twins share two ($k_2=1$) whereas unrelated individuals share no alleles ($k_0=1$) IBD. For full-sibs, it is easy to show that the probability that they share one allele IBD is 0.5 and that they share no or two is in both cases 0.25 (so $k_0=0.25$, $k_1=0.5$ and $k_2=0.25$). Given the allele frequencies, the probabilities that the genotype pair $g_A.g_B$ shares 0, 1 or 2 alleles IBD, P_0, P_1 and P_2, are then calculated and are inserted in the final IBD likelihood formula (for details, see e.g. Blouin, 2003):

$$\Pr(g_A.g_B|\mathbb{R})=k_0P_0+k_1P_1+k_2P_2 \quad (k_0+k_1+k_2=1)$$

For unlinked loci, which we assume in the following, the logarithms of the IBD relationship likelihoods and the LOD scores are additive over the loci.

3.4 Genotyping errors

Even high quality datasets contain errors where at least one allele at a given locus does not match with what we expect from the Mendelian laws. Thus, it is unwise to exclude a parent immediately when observing such a mismatch. There are many reasons for such mismatches, see Bonin *et al.* (2004) for a review. Genotyping errors occur when the genotype determined by molecular analysis does not correspond to the real genotype. For instance, a common type of genotyping error in microsatellite datasets are null alleles, which are often the result of a mutation in the primer annealing site. Somatic mutations form another source of mismatches.

The model implemented here defines an error to be the replacement of the true genotype at a particular locus in an individual with a random genotype. This leads to a modification of the expressions for the LOD score, see Kalinowski *et al.* (2007), and to corresponding modifications in the IBD likelihood calculations, see Broman and Weber (1998) for details.

4 METHODS

4.1 Simulation

Given the population allele frequencies and the expected typing error rate, which are either estimated using the sample itself or provided by the user, we generate individuals with known relationships to determine various distributions.

One important characteristic is the distribution of the number of mismatching loci given the expected error rate for pairs (parent–offspring versus unrelated) as well for triples (offspring, mother and father versus offspring, mother and unrelated male). This knowledge allows us to speed up the algorithm, because we know when likelihood calculations can be terminated. We can furthermore omit the $O(n^3)$ triple calculation for pairs with more mismatches than maximally expected for a triple. These parameters are also important because too many allowed mismatches may lead to an increased number of false positive parent–offspring arcs.

Furthermore, we will later test the null hypothesis that a pair is a full-sib against the alternative hypotheses that they are unrelated, parent–offspring or half-sib. We calculate the P-values by generating following distributions for full-sibs and for pairs of the alternative hypothesis relationship:

$$\begin{aligned}\Delta_u &= \log\Pr(G_i.G_j|\mathbb{FS})-\log\Pr(G_i.G_j|\mathbb{U})\\ \Delta_{po} &= \log\Pr(G_i.G_j|\mathbb{FS})-\log\Pr(G_i.G_j|\mathbb{PO})\\ \Delta_{hs} &= \log\Pr(G_i.G_j|\mathbb{FS})-\log\Pr(G_i.G_j|\mathbb{HS})\end{aligned}$$

So for example Δ_{po} is generated for full-sibs and parent–offspring pairs to estimate the statistical significance of an observed positive Δ_{po} value. Note that $\mathbb{HS}$ are all second degree relationships (half-sib, grandparent–grandoffspring and avuncular), which has to be considered in the P-value calculation.

4.2 Calculation of the possible parent–offspring arcs

For every individual v, we calculate the LOD scores with all candidate parents u_i, individuals we cannot exclude a priori as parents, for example, because of their age. We discard pairs (u_i,v) or triples (u_i,u_j,v) with negative multi-locus LOD scores from our further analyses, because adding the corresponding arcs to the pedigree would decrease its likelihood. Hence, for every pair of individuals with positive single-parent LOD score, $(u_i,?)$ is included in the set of valid parent combinations $\mathscr{H}(v)$, just as well (u_i,u_j) for every triple with positive parent-pair LOD score. Unless we know that at least one parent of v is sampled, we include the empty parent pair $(?,?)$ in $\mathscr{H}(v)$.

The parentage likelihood calculation is the most important step in the pedigree reconstruction procedure as these likelihoods define the set of all possible arcs in the pedigree. However, as described in detail in Thompson and Meagher (1987), if we cannot exclude two full-sibs, v_i and v_j, as parent and offspring, they in general give a higher likelihood than do true parents. Thus, for highly probable full-sibs, a reasonable strategy is to use only the intersection of the valid parent combinations: $\mathscr{H}(v_i)=\mathscr{H}(v_j)=\mathscr{H}(v_i)\cap\mathscr{H}(v_j)$. The critical values of Δ_{po} and Δ_{hs} that a full-sib pair must exceed should be high enough to prevent false positives, which may result in an exclusion of the true parents in the next step, the pedigree reconstruction. Note that if the intersection contains a parent pair, this is an additional hint that v_i and v_j are full-sibs. Modeling this in the P-value calculation is difficult, we could use however a less conservative critical α value in this case. As default values for α, we use 0.01 and 0.05, respectively. The observed P-values are adjusted for multiple testing (Benjamini and Hochberg, 1995).

4.3 Pedigree likelihood

The log-likelihood of a pedigree $\mathscr{P}$ is now computed as the sum of the logarithms of the N_I parentage posterior probabilities given this pedigree:

$$\max_{\mathscr{P}\in\mathscr{T}} \mathrm{LL}(\mathscr{P}|A,N) = \sum_{i=1}^{N_I}\log\Pr(N^+(v_i)|A,N)$$

We use simulated annealing (Kirkpatrick *et al.*, 1983) for the pedigree reconstruction as described in Almudevar (2003) to find the maximum likelihood pedigree. If necessary, then every N_I+2 iterations a random missing value is estimated by Gibbs sampling.

4.4 Incomplete sampling

As already stated in Section 3.2, if not all candidate parents are sampled, it is important to estimate the number of unsampled candidates. This number could be either estimated by additional experiments, for example capture–recapture surveys or by using the data alone. The pedigree structure itself contains information about the sampling rate in the ratio of the number vertices with indegree 1 and with indegree 2, d_1 and d_2:

$$r=\frac{1}{(d_1/2d_2)+1} \quad \text{and} \quad N\approx\frac{n}{r}\cdot x \quad \text{for} \quad x\geq r.$$

For larger samples, setting $x=1$ should give a good point estimate of N when we assume that r and x are constant across sampled generations. Again every N_I+2 iterations, we draw a new value of x from a flat distribution $\mathcal{U}(r,x_{\max})$ and accept the change with the simulated annealing acceptance probability. A value of 4 for $x_{\max}$ showed a very robust performance in our tests. Depending on the data, it might be also necessary to specify a $N_{\max}$ (Nielsen *et al.*, 2001). In the absence of age data, it is not known a priori which sampled individuals are candidate parents. So it might also be necessary here to specify n and to exclude at least the direct descendants in the parentage posterior calculation.

4.5 MCMC

When $\mathscr{T}$ does not allow all parentage combinations, the parentage posterior probabilities $\Pr(N^+(v_i)|A,N)$ (Section 3.2) must be corrected accordingly. FRANz samples from the pedigree posterior distribution $\Pr(\mathscr{P})$ by Markov Chain Monte Carlo (MCMC) and redefines $\Pr(N^+(v_i))$ as the probability of observing the parentage $N^+(v_i)$ when drawing from $\Pr(\mathscr{P})$. Another benefit of MCMC sampling is that it allows to incorporate the uncertainty of the pedigree reconstruction when estimating parameters from the pedigrees (Hadfield *et al.*, 2006).

To speed up mixing, FRANz automatically uses parallel *Metropolis coupled Markov chain Monte Carlo* (MCMCMC; Huelsenbeck and Ronquist, 2001), implemented in a shared memory programming model, when run on computers with multiple CPU cores. In short, in addition to the normal, unheated chain, $n-1$ heated chains are started on the CPU cores $2,\dots,n$ and states are attempted to swap with a given probability. Swaps are then accepted with normal Metropolis–Hasting acceptance probability. Pedigrees are only sampled from the unheated chain.

4.6 Allele frequencies

The population allele frequencies are often unknown. If the sample size is large and family sizes are small, it is reasonable to assume that individuals are unrelated and then to use all genotypes for the estimation. If not, however, then this strategy will overestimate the frequency of rare alleles in large families. FRANz therefore updates the allele frequencies during SA optimization or MCMC sampling. This is computationally extensive, but it is not necessary to update after every change of the pedigree (Thomas and Hill, 2000).

5 RESULTS

5.1 Real microsatellite data

Our first dataset is a microsatellite dataset of the black tiger shrimp *Penaeus monodon* (Jerry *et al.*, 2006). The true pedigree is known from direct observation. The dataset consists of 13 families with a total number of 85 individuals (of which 59 offspring), genotyped at seven highly polymorphic loci. For 10 individuals, alleles are missing at one locus. The error rate is very low, with only one observed mismatch. Figure 1 shows the best pedigrees with and without full-sib calculation (Section 4.2). Full-sibs tend to have higher parentage likelihoods, but large full-sib groups greatly enhance the performance of our algorithm such that the accuracy of the reconstructed pedigree increases from 82.8 to 97.1%. A recent publication (Berger-Wolf *et al.*, 2007) listed an accuracy rate of several sibling reconstruction methods ranging from 67.8 to 78.0% percent on the same dataset. Classic parentage inference programs such as CERVUS (Marshall *et al.*, 1998), where the absence of age data violates main assumptions, assign statistical significant parentages to the parental genotypes even when the correct parameters (sampling rate, fraction of relatives in the candidate parents) are provided.

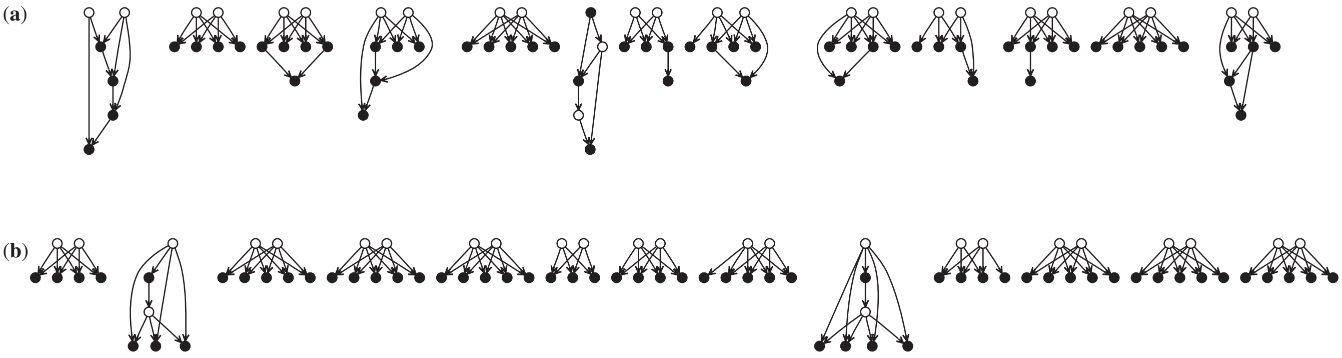

Fig. 1. Reconstructed *Penaeus monodon* pedigree (Section 5.1). The white vertices are the parental genotypes, black the offspring genotypes. (**a**) without full-sib calculation. (**b**) with full-sib calculation.

5.2 Simulated data

We artificially generate population datasets as follows. A population of 100 unrelated founders is created by drawing genotypes independently with allele frequencies of 64 human microsatellites (Jin *et al.*, 2000). Then we let individuals die, mate or marry according to rates extracted from the statistics of the German population (Federal Statistical Office, 2007). As mating partners or husbands, we only allow unrelated individuals. Married couples only mate with each other. We stop when the desired number of individuals is reached. In order to simulate typing errors, we replace the true allele with a random one. Null alleles are simulated in heterozygote genotypes by replacing the null allele with the other allele ($a_i.a_n$ becomes $a_i.a_i$). Homozygote genotypes are marked as missing.

We analyze the accuracy of the pedigree reconstruction as a function of the number of available loci, see Figure 2. In all cases where the accuracy is below 1, the optimal pedigree from our algorithm has an even larger likelihood than the true one. Thus without exceptions, our algorithm finds a pedigree with at least the log-likelihood of the true pedigree (data not shown). The plots show that the reconstruction is robust even when the upper limit of the total number of breeding individuals per generation in the population $N_{\max}$ was largely overestimated (164 versus 1000).

Age data is clearly the most informative prior knowledge. Knowledge about the sex rarely helps to exclude a false parentage mainly because mothers are sampled like all individuals with a rate of 0.5 and sex requires candidate parent pairs for exclusion. Thus, the knowledge of the sex does not resolve the difficult cases where the true parents are unsampled but a close relative (e.g. aunt or uncle) is sampled.

Without age data, the direction of a large fraction of parent–offspring arcs cannot be determined, which explains the plateaus in the plots. These parentages are easily identified by their posterior probability which is typically near 0.5. In Nielsen *et al.* (2001), a parentage was assigned when the posterior probability was higher than 0.95. Figure 2 visualizes the proportion of correct and incorrect assignments. In almost all cases, the proportion of wrongly assigned parentages was smaller than 0.01. These parentages are mainly the difficult cases mentioned above or false positives of the sibling calculation, whose sensitivity and specificity is plotted in Figure 2c.

6 DISCUSSION

We have presented a new algorithm for the multi-generation pedigree reconstruction problem. The publicly available implementation is written in the C programming language and is platform-independent. The genealogy of datasets with thousands of individuals is typically reconstructed in a few minutes. Our implementation is flexible in incorporating additional data like age, sex, sampling locations, sub-pedigrees and allele frequencies. This was suggested in Almudevar (2003) but not previously implemented in a publicly available software package. The reconstruction of large and deep pedigrees is highly accurate with only 10–15 polymorphic microsatellite loci. Our approach is to our knowledge the first one that combines paternity inference and sibship reconstruction.

In Almudevar (2003), some remaining challenges in the pedigree reconstruction problem were listed. These are the assumption that founders are unrelated, a better estimation of allele frequencies, linkage, support for typing errors or mutation and estimation of the error of the reconstruction procedure. FRANz makes significant progress in the latter two tasks by combining the error model described in Kalinowski *et al.* (2007) with an MCMC sampling.

The error model, however, was criticized in the literature because of its simplicity. Other programs explicitly model special kinds of errors, for example null alleles and sample the true genotypes with an individual-by-individual Gibbs sampling (Hadfield *et al.*, 2006; Wang, 2004b). For multi-generation pedigrees, one has to sample over the family to ensure irreducibility of the Markov chain (Sheehan, 2000). For large pedigrees, this becomes very fast computationally infeasible and the gain is questionable. Extending the likelihood formulas in (Kalinowski *et al.*, 2007) to model null alleles, however, could be a valuable extension if they occur at higher rates. Now, FRANz estimates the null allele frequency (Kalinowski and Taper, 2006) and warns the user when null alleles are likely to be present in the data.

Extensions of the LOD scores for linked loci when the linkage phase is known are proposed in Devlin *et al.* (1988). If the linkage phase and recombination rates are known with high accuracy, the incorporation of this prior information can significantly enhance the performance of the parentage assignments (Devlin *et al.*, 1988). However, in most cases the linkage phase is unknown and has to be estimated jointly. Loose linkage of a small fraction of markers should not seriously bias multi-locus likelihood calculations

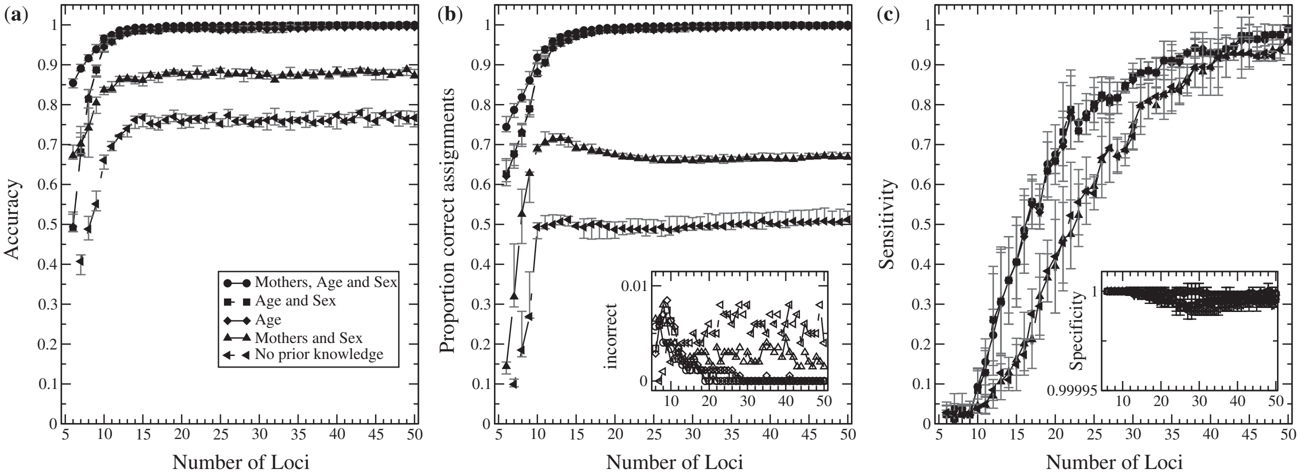

Fig. 2. These plots visualize the results of the reconstruction of simulated pedigrees (Section 5.2). The various measurement are plotted as a function of the number of loci. The values are the median of 10 randomly generated pedigrees of size 1000, reconstructed with different combinations of available prior knowledge. The error bars indicate the first and third quartile. The dataset has a sampling rate of 0.5 (1000 of 2000 individuals sampled) and has an overall typing error rate of 0.01. In addition, the first locus comprises one null allele ($p_n = 0.05$). The pedigree depth ranges from 5 to 9 and the mean number of sampled candidate parents is 82. $N_{\max}$ (see Section 4.4) was largely overestimated set to 1000. (**a**) The accuracy of the maximum likelihood pedigree. (**b**) The proportion of incorrect (unfilled symbols) and correct parentages with a *posterior probability* > 0.95. (**c**) The sensitivity and specificity of the sibling calculation.

(Meagher, 1991). Tightly linked loci in contrast, such as neighboring single nucleotide polymorphisms (SNPs), can be combined and treated as one single *pseudolocus*. In general, linked loci are less informative than unlinked ones and therefore the calculated LOD scores are too large. The best advice now is probably to avoid medium linked loci (Jones and Ardren, 2003).

The framework we have presented in this article may easily be extended to incorporate prior knowledge in the likelihood calculation (Neff *et al.*, 2001). Currently, prior knowledge is only used to reduce the search space. For parentages, sampling locations and behavioural data have been successfully used to increase the parentage assignments in Hadfield *et al.* (2006). Priors about the pedigree structure (the expected inbreeding rates, number of offspring, etc.) might further improve the performance (Sheehan and Egeland, 2007). Information of this kind is oftentimes unknown a priori, however. In fact, these are parameters that one typically would like to infer from the reconstructed pedigrees.

Our implementation currently only allows co-dominant markers. In Gerber *et al.* (2000), the original LOD scores for co-dominant markers (Meagher and Thompson, 1986) were modified for dominant markers, such as *amplified fragment length polymorphisms*. Statistics for estimating pairwise relationships with dominant markers were proposed e.g. in Wang (2004a).

Our incorporation of full-sib probabilities is a reaction to the concern expressed in Meagher and Thompson (1986) that non-excluded full-sibs of the offspring have on average a higher LOD score than the true father. To keep the pedigree likelihood function simple and efficient to calculate, we use only highly significant full-sibs to reduce the pedigree space. It seems possible to include more siblings than just the highly significant ones into the pedigree likelihood calculation without the risk of excluding the true parents. Since such ‘local’ factors in the pedigree likelihood are also not very computationally intensive, we plan to explore this avenue in future work.

With the rapid progress and decay of cost in high-throughput sequencing techniques, it is just a matter of time until there are whole genomes of complete populations available. Large amounts of SNP data with high quality genetic maps will be therefore available, at least for some model organisms. The identification of parents with such an amount of data is a trivial task and the methods are well known (Boehnke and Cox, 1997). A challenging question is then how many unobserved generations we can reconstruct back in time [see Steel and Hein (2006) and Thatte and Steel (2007) for first results]. As we cannot expect an elegant solution to this problem, MCMC heuristics are promising tools for throwing some light on a population’s immediate past.

ACKNOWLEDGEMENTS

We would like to thank Dean Jerry for the *P.monodon* dataset, the anonymous reviewers for many helpful comments and Elizabeth Thompson for elaborately answering our questions.

Funding: European Commission NEST Pathfinder [initiative on Complexity through project EDEN (Contract 043251)].

Conflict of Interest: none declared.

REFERENCES

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

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)*, **57**, 289–300.

Berger-Wolf,T. *et al.* (2007) Reconstructing sibling relationships in wild populations. *Bioinformatics*, **23**, 49–56.

Blouin,M.S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol.*, **18**, 503–511.

Boehnke,M. and Cox,N. (1997) Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.*, **61**, 423–429.

Bonin,A. *et al*. (2004) How to track and assess genotyping errors in population genetics studies. *Mol. Ecol.*, **13**, 3261–3273.

Broman,K. and Weber,J. (1998) Estimation of pairwise relationships in the presence of genotyping errors. *Am. J. Hum. Genet.*, **63**, 1563–1564.

Devlin,B. *et al*. (1988) Fractional paternity assignment: theoretical development and comparison to other methods. *Theor. Appl. Genet.*, **76**, 369–380.

Federal Statistical Office (2007) *Statistical Yearbook 2007 for the Federal Republic of Germany*. Federal Statistical Office, Wiesbaden.

Gerber,S. *et al*. (2000) Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Mol. Ecol.*, **9**, 1037–1048.

Hadfield,J. *et al*. (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol. Ecol.*, **15**, 3715–3730.

Huelsenbeck,J. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Jerry,D. *et al*. (2006) Development of a microsatellite DNA parentage marker suite for black tiger shrimp penaeus monodon. *Aquaculture*, **255**, 542–547.

Jin,L. *et al*. (2000) Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. *Ann. Hum. Genet.*, **64**, 117–134.

Jones,A. and Ardren,W. (2003) Methods of parentage analysis in natural populations. *Mol. Ecol.*, **12**, 2511–2523.

Kalinowski,S. and Taper,M.L. (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics*, **7**, 991–995.

Kalinowski,S. *et al*. (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.*, **16**, 1099–1106.

Kirkpatrick,S. *et al*. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Lin,T. *et al*. (2006) Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. *Bioinformatics*, **22**, 298–306.

Marshall,T. *et al*. (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, **7**, 639–655.

Meagher,T.R. (1991) Analysis of paternity within a natural population of chamaelirium luteum. ii. patterns of male reproductive success. *Am. Nat.*, **137**, 738–752.

Meagher,T.R. and Thompson,E. (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theor. Popul. Biol.*, **29**, 87–106.

Neff,B. *et al*. (2001) A Bayesian framework for parentage analysis: the value of genetic and other biological data. *Theor. Popul. Biol.*, **59**, 315–331.

Nielsen,R. *et al*. (2001) Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics*, **157**, 1673–1682.

Pemberton,J. (2008) Wild pedigrees: the way forward. *Proc. Biol. Sci.*, **275**, 613–621.

Sheehan,N. (2000) On the application of markov chain monte carlo methods to genetic analyses on complex pedigrees. *Int. Stat. Rev.*, **68**, 83–110.

Sheehan,N. and Egeland,T. (2007) Structured incorporation of prior information in relationship identification problems. *Ann. Hum. Genet.*, **71**, 501–518.

Steel,M. and Hein,J. (2006) Reconstructing pedigrees: a combinatorial perspective. *J. Theor. Biol.*, **240**, 360–367.

Thatte,B. and Steel,M. (2007) Reconstructing pedigrees: a stochastic perspective. *J. Theor. Biol.*

Thomas,S. and Hill,W. (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, **155**, 1961–1972.

Thomas,S. and Hill,W. (2002) Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genet. Res.*, **79**, 227–234.

Thompson,E. (1976). Inference of genealogical structure. *Soc. Sci. Inform.*, **15**.

Thompson,E. and Meagher,T. (1987) Parental and sib likelihoods in genealogy reconstruction. *Biometrics*, **43**, 585–600.

Vouillamoz,J. and Grando,M. (2006) Genealogy of wine grape cultivars: ‘Pinot’ is related to ‘Syrah’. *Heredity*, **97**, 102–110.

Wang,J. (2004a) Estimating pairwise relatedness from dominant genetic markers. *Mol. Ecol.*, **13**, 3169–3178.

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