FRANz: Fast reconstruction of wild pedigrees

Markus Riester[1], Peter F. Stadler[1,2,3,4], Konstantin Klemm[1]
[1]Bioinformatics Group, Department of Computer Science,
and Interdisciplinary Center for Bioinformatics,
University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany.
[2]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology (IZI),
Deutscher Platz 5e, D-04103 Leipzig, Germany
[3]Institute for Theoretical Chemistry, University of Vienna,
Währingerstrasse 17, A-1090 Vienna, Austria
[4]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico
{markus,studla,klemm}@bioinf.uni-leipzig.de

Abstract: We present a software package for fast pedigree reconstruction in natural populations using co-dominant genomic markers such as microsatellites and 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 a simulation of the sampling process. The parentage inference is robust even in the presence of genotyping errors.

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 developed in series of articles by E.A. Thompson, see e.g. [Tho76]. The study of parentage in natural populations was the topic of the pioneering papers by T.R. Meagher [MT86] and T.C. Marshall [MSKP98], recently reviewed in [Blo03, JA03, Pem08]. 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 [VG06]), the estimation of heritabilities in the wild [TH00], and victim identification [LMX06].

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 one or two generation datasets. 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 [TH02, Wan04, BWSD$^+$07]. The two generation parentage inference programs 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. Much less attention has been given to multi-generation pedigrees. The main difference to parentage inference programs is that in the general case not all possible parentage combinations are 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 become 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

We follow the formalism introduced in [SH06]. A pedigree $\mathscr{P}$ is an acyclic digraph, for which the vertex set V is the disjoint union of the subsets F, M and U (*'Female'*, *'Male'* and *'Unknown Sex'*) and for each vertex $v \in V$ satisfies the condition:

(P) if v has positive indegree then v has exactly two incoming arcs, say (u, v) and (u', v), where $u \in F \cup U$ and $u' \in M \cup U$, or v has one incoming arc.

In selfing species, $u = u'$ is allowed and $\mathscr{P}$ is a multigraph.

Condition (P) formalizes the requirement that the sex of the parents of an individual must be different if and only if both parents and both sexes are known.

For an arc (u, v) of $\mathscr{P}$ we say that v is a *child* of u and u is a *parent* of v. The set of (putative) *parents* of v is denoted by $N^+(v) \subseteq V$; it may have cardinality 2, 1 (only one parent sampled), or 0 if $N(v)^+ = \emptyset$. In this case, v is called a *founder*. 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.

3 Background

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 LOD score* (e.g. [MT86]):

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

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 $P(a_1, a_2) = 2pq$; for homozygotes, we have $P(a_1, a_1) = p^2$. The Mendelian transmission probability is denoted by $T(\cdot)$. 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 [MT86, KTM07].

For each dyad, we can calculate the probability that the two individuals 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 $P(g_A.g_B|\mathbb{R})$ uses the so-called *IBD coefficients* [Blo03]. For unlinked loci, which we assume in the following, the logarithms of these likelihoods and the LOD scores are additive over the loci.

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 [BBBE$^+$04] 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 [KTM07], and to corresponding modifications in the IBD likelihood calculations, see [BW98] for details.

4 Methods

4.1 Simulation of the sampling process

To estimate the power of the marker suite, our software performs several standard tests and calculations. This alone, however, will not be sufficient to estimate the accuracy of the pedigree reconstruction. A simulation of the sampling process is therefore necessary. Given the population's 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. To assess the degree of con-

fidence of the parent-offspring arcs in $\mathscr{P}$, we follow [MSKP98] in using ΔLOD as test statistic. ΔLOD is the difference of the LOD scores between the two most likely parent combinations (or fathers).

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

Full sibs can distinguished from parent-offspring pairs based on the log-likelihood differences $\Delta_{po} = P(G_i.G_j|\mathbb{FS}) - P(G_i.G_j|\mathbb{PO})$. The distribution of Δ_{po} for true full-sib dyads and for parent-offspring dyads. We later only consider dyads that exceed a critical value of Δ_{po} as full-sib candidates. If the intersection of their candidate parents includes at least one parent pair, we finally define this dyad as full-sibs. If not, then the dyad could still be a full-sib pair, but with unsampled parents. In this case, this dyad could also be a half-sib pair, so we use the distribution of the log-likelihood differences $\Delta_{hs} = P(G_i.G_j|\mathbb{FS}) - P(G_i.G_j|\mathbb{HS})$ to distinguish full-sibs from half-sibs. The values of Δ_{hs} are generated for true full-sib dyads and true half-sib dyads. Now, full-sib candidates without a common parent pair that exceed a critical value of Δ_{hs}, are defined as full-sibs.

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 multilocus LOD scores from our further analyses. 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 $\mathscr{H}(v)$.

These parentage likelihoods are the most important step in the pedigree reconstruction procedure as they define the set of all possible arcs in the pedigree. However, as described in detail by Meagher and Thompson [TM87], 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 candidate parents: $\mathscr{H}(v_i) = \mathscr{H}(v_j) = \mathscr{H}(v_i) \cap \mathscr{H}(v_j)$. The critical values of Δ_{po} and Δ_{hs} that a full-sib dyad 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.

4.3 Pedigree Reconstruction

The likelihood of a pedigree $\mathscr{P}$ is computed as the probability of the genotypes given this pedigree. So the goal is to find the pedigree which maximizes the log-likelihood:

$$\max_{\mathscr{P} \in \mathscr{T}} L(\mathscr{P}) \quad = \quad \sum_{i=1}^{N_I} \log P(G_i | N^+(v_i))$$

Here, $P(\cdot)$ is the probability of observing the multilocus genotype G_i given the parents $N^+(v_i)$. For founders ($N^+ = \emptyset$), $\log P(\cdot)$ equals the denominator of the multilocus LOD score. This is equivalent to the assumption that all founders are unrelated. For the offspring, these probabilities are the multilocus Mendelian transition probabilities in our error model. So for vertices where $|N^+| = 1$, $\log P(\cdot)$ is the *single-parent*, when $|N^+| = 2$ the *parent-pair* LOD enumerator.

For each individual, we now sort the possible parent combinations by their probability. The maximal possible score is simply the sum of all most likely parent combinations. Our greedy algorithm works by selecting one vertex v and then adding the arcs corresponding to the most likely parent combination $N^+ \in \mathscr{H}(v)$. If the arcs introduce a directed cycle in $\mathscr{P}$, we try the second most likely parent combination and so on. If no parent-offspring relationships are known, this algorithm produces a valid pedigree, because the 'empty' parent combination (v is a founder) is always in $\mathscr{H}(v)$, which can never introduce a cycle. We proceed until all vertices are added.

For vertices with known parents, every parent combination adds at least one arc. A simple strategy is now to start with vertices where $|\mathscr{H}(v)| = 1$. Unless the "known" parent-offspring relationships are wrong, this introduces no directed cycles. Then we proceed with the remaining vertices with known parents. If this succeeds, we add the remaining vertices without known parents as described above. If not, or if the final score is not the maximal score, we use Simulated Annealing [KGV83] for the pedigree reconstruction as described in [Alm03].

5 Results

Black Tiger Shrimp *Penaeus monodon*. Our first dataset is a microsatellite dataset of the black tiger shrimp *Penaeus monodon* [JBMW06]. 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 ten individuals, alleles are missing at one locus. The error rate is very low, with only one observed mismatch. Figure 1 are the best pedigrees with and without full-sib heuristic (assumed typing error rate of 0.01) and shows that large full-sib groups greatly enhance the performance of our algorithm. The accuracy of the complete pedigree without full-sib heuristic is 82.0% in comparison to 99.58% with this heuristic. A recent publication [BWSD$^+$07] listed an accuracy rate of several sibling reconstruction methods ranging from 67.8 to 77.97 percent

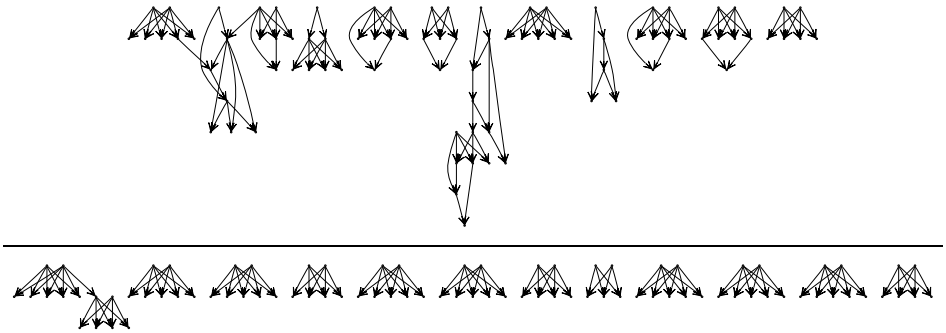

Figure 1: Reconstructed *penaeus monodon* pedigree. Without (top) and with (below) full-sib calculation.

on the same dataset.

Simulated Data. We use the statistics of the German population [Off07] to calculate the probabilities of death, (multiple) birth and marriage at a given age for males and females. As initial population we generate 100 unrelated individuals. For the genotypes, we use the allele frequencies of 64 human microsatellites [JBCS$^+$00]. In every year, we let all individuals die, mate or marry according the corresponding probabilities. As mating partners or husbands, we only allow unrelated individuals. Married couples only mate with each other. We stop when the desired number of individuals is reached. In order to simulate typing errors, we replace the true allele with a random one. Null alleles are simulated in heterozygote genotypes by replacing the null allele with the other allele ($a_i.a_n$ becomes $a_i.a_i$). Homozygote genotypes are marked as missing, i.e., $a_n.a_n$ becomes $?.?$.

We analyzed the accuracy of our algorithm with different subsets of the simulated data, see Figure 2. If the accuracy is not 100%, then either the algorithm failed to find the maximum likelihood pedigree or there exists a valid pedigree that has a higher likelihood than the true one. Without exceptions, our optimization algorithm found a pedigree with at least the log-likelihood of the true pedigree (data not shown).

We also evaluated the performance of our full-sib heuristic directly. As we use this heuristic to reduce the pedigree space, we require a very small false positive rate. The sensitivity and specificity is plotted in Figure 2.

6 Discussion

We have presented a fast algorithm for the pedigree reconstruction problem. The publicly available implementation is written in the C programming language and is platform-independent. It can be obtained under the GPL[1]. The genealogy of datasets with thousands

[1] `http://www.bioinf.uni-leipzig.de/Software/FRANz/`

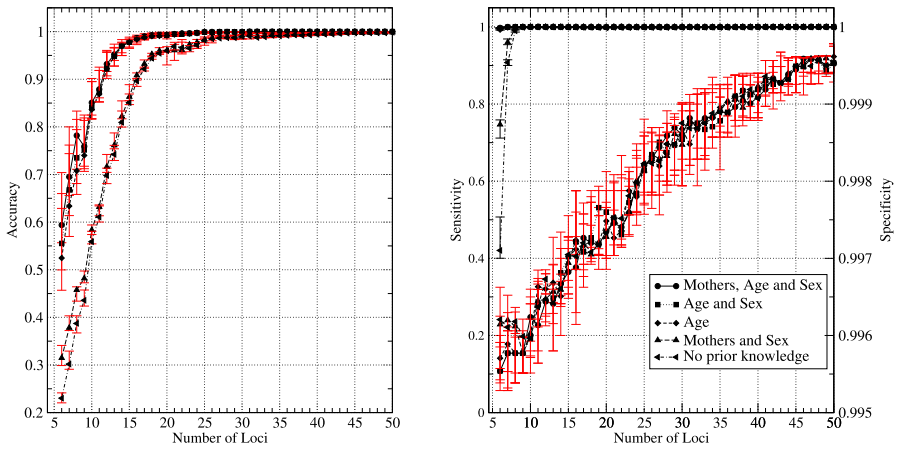

Figure 2: (Left) The accuracy of the reconstructed pedigrees is plotted as a function of the number of loci. The values are the median accuracy of ten randomly generated pedigrees of size 1000, reconstructed with different combinations of available prior knowledge. The error bars indicate the first and 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 ($p_n = 0.05$).
(Right) The sensitivity and specificity of the sibling calculation plotted again as a function of the number of loci.

of individuals is typically reconstructed in a few minutes. Due to the space constraints of this paper, we can only describe the core functionality of the software. Our implementation is flexible in incorporating additional data like age, sex, sampling locations, sub-pedigrees and allele frequencies. This was suggested in [Alm03] but not previously implemented in a publicly available software package. The reconstruction is highly accurate with only 15-20 polymorphic microsatellite loci (twice as many when age data are not available).

In [Alm03], 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 simulation procedure and the error model described in [KTM07] with the Simulated Annealing algorithm.

The error model was criticized in the literature because of its simplicity. Other programs explicitly model special kinds of errors, for example null alleles [WCK06]. At typical error rates of 1%, however, the number of mismatching loci is low and a detailed modeling seems provide little benefit. More complex error models may be necessary for data with higher error rates, however.

Extensions of the LOD scores for linked loci when the linkage phase is known are proposed in [DRE88]. If the linkage phase and recombination rates are known with high accuracy, the incorporation of this prior information can significantly enhance the perfor-

mance of the parentage assignments [DRE88]. 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 multilocus likelihood calculations [Mea91]. Tightly linked loci in contrast, such as neighboring SNPs, can be combined and treated as one single *pseudolocus*.

Our implementation currently only allows co-dominant markers. In [GMS+00], the original LOD scores for co-dominant markers [MT86] were modified for dominant markers, such as *amplified fragment length polymorphisms* (AFLPs). Statistics for estimating pairwise relationships with dominant markers were proposed e.g. in [Wan04].

The pedigree likelihood function is appealing because of its property being additive over the individuals. This allows very efficient construction algorithms and requires no prior information about the pedigree structure. However, if the genomic signal is low, the likelihood function will fail to construct the correct pedigree, especially when single-parents are considered. This is because the expected number of false positive single-parent arcs becomes large. Age data significantly reduces this effect. The same is true for our full-sib heuristic in particular when large full-sib groups and both of their parents are sampled. Priors about the pedigree structure (the expected inbreeding rates, number of offspring, ...) might further improve the performance. 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 incorporation of full-sib probabilities is a reaction to the concern expressed in [MT86] 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.

Traditional parentage inference methods such as the one described in this paper have been criticized lately [HRB06]. Pedigrees are used to estimate parameters. If the genomic signal is not strong enough, many different pedigrees will have similar likelihood scores. Using only the best pedigree will thus introduce a bias. In [HRB06], it has been proposed to estimate the parameters of interest jointly with the pedigree. This, however, requires that the population's mating behaviour fits the implemented model. `FRANz` can output possible parent combinations, not only the ones of the maximum likelihood pedigree, as a starting point to investigate such a bias [DRE88].

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 [BC97]. A challenging question is then how many unobserved generations we can reconstruct back in time (see [SH06] and [TS07] 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 and Elizabeth Thompson for elaborately answering our questions. This work has been supported by the European Commission NEST Pathfinder initiative on Complexity through project EDEN (Contract 043251).

References

[Alm03] A. Almudevar. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theor Popul Biol*, 63:63–75, Mar 2003.

[BBBE+04] A. Bonin, E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. How to track and assess genotyping errors in population genetics studies. *Mol. Ecol.*, 13:3261–3273, Nov 2004.

[BC97] M. Boehnke and N.J. Cox. Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.*, 61:423–429, Aug 1997.

[Blo03] Michael S. Blouin. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, 18(10):503–511, October 2003.

[BW98] K.W. Broman and J.L. Weber. Estimation of pairwise relationships in the presence of genotyping errors. *Am. J. Hum. Genet.*, 63:1563–1564, Nov 1998.

[BWSD+07] T.Y. Berger-Wolf, S.I. Sheikh, B. DasGupta, M.V. Ashley, I.C. Caballero, W. Chaovalitwongse, and S.L. Putrevu. Reconstructing sibling relationships in wild populations. *Bioinformatics*, 23:49–56, Jul 2007.

[DRE88] B. Devlin, K. Roeder, and N.C. Ellstrand. Fractional paternity assignment: theoretical development and comparison to other methods. *TAG Theoretical and Applied Genetics*, 76(3):369–380, Sep 1988.

[GMS+00] S. Gerber, S. Mariette, R. Streiff, C. Bodenes, and A. Kremer. Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Molecular Ecology*, 9(8):1037–1048, 2000.

[HRB06] J.D. Hadfield, D.S. Richardson, and T. Burke. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol. Ecol.*, 15:3715–3730, Oct 2006.

[JA03] A.G. Jones and W.R. Ardren. Methods of parentage analysis in natural populations. *Mol. Ecol.*, 12:2511–2523, Oct 2003.

[JBCS+00] L. Jin, M.L. Baskett, L.L. Cavalli-Sforza, L.A. Zhivotovsky, M.W. Feldman, and N.A. Rosenberg. Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. *Ann. Hum. Genet.*, 64:117–134, Mar 2000.

[JBMW06] D.R. Jerry, Evansa B.S., Kenwayb M, and K. Wilson. Development of a microsatellite DNA parentage marker suite for black tiger shrimp Penaeus monodon. *Aquaculture*, 255(1-4):542–547, May 2006.

[KGV83] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

[KTM07] S.T. Kalinowski, M.L. Taper, and T.C. Marshall. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.*, 16:1099–1106, Mar 2007.

[LMX06] T.H. Lin, E.W. Myers, and E.P. Xing. Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. *Bioinformatics*, 22:298–306, Jul 2006.

[Mea91] Thomas R. Meagher. Analysis of Paternity within a Natural Population of Chamaelirium luteum. II. Patterns of Male Reproductive Success. *The American Naturalist*, 137(6):738–752, 1991.

[MSKP98] T.C. Marshall, J. Slate, L.E. Kruuk, and J.M. Pemberton. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, 7:639–655, May 1998.

[MT86] Thomas R. Meagher and Elizabeth Thompson. The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, 29(1):87–106, February 1986.

[Off07] Federal Statistical Office. *Statistical Yearbook 2007 For the Federal Republic of Germany*. Number ISBN: 978-3-8246-0803-4. Federal Statistical Office, Wiesbaden, 2007.

[Pem08] J.M. Pemberton. Wild pedigrees: the way forward. *Proc. Biol. Sci.*, 275:613–621, Mar 2008.

[SH06] M. Steel and J. Hein. Reconstructing pedigrees: a combinatorial perspective. *J. Theor. Biol.*, 240:360–367, Jun 2006.

[TH00] S.C. Thomas and W.G. Hill. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, 155:1961–1972, Aug 2000.

[TH02] S.C. Thomas and W.G. Hill. Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genet. Res.*, 79:227–234, Jun 2002.

[Tho76] E.A. Thompson. Inference of genealogical structure. *Social Science Information*, 15(477), 1976.

[TM87] E.A. Thompson and T.R. Meagher. Parental and sib likelihoods in genealogy reconstruction. *Biometrics*, 43:585–600, Sep 1987.

[TS07] B.D. Thatte and M. Steel. Reconstructing pedigrees: A stochastic perspective. *J. Theor. Biol.*, Dec 2007.

[VG06] J.F. Vouillamoz and M.S. Grando. Genealogy of wine grape cultivars: ”Pinot” is related to ”Syrah”. *Heredity*, 97:102–110, Aug 2006.

[Wan04] J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166:1963–1979, Apr 2004.

[WCK06] A.P. Wagner, S. Creel, and S.T. Kalinowski. Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, 97:336–345, Nov 2006.