Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 5 2012 *Article* 12

Large-scale Parentage Inference with SNPs: an Efficient Algorithm for Statistical Confidence of Parent Pair Allocations

Eric C. Anderson, *Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA*

Recommended Citation:
Anderson, Eric C. (2012) "Large-scale Parentage Inference with SNPs: an Efficient Algorithm for Statistical Confidence of Parent Pair Allocations," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 5, Article 12.
DOI: 10.1515/1544-6115.1833

Large-scale Parentage Inference with SNPs: an Efficient Algorithm for Statistical Confidence of Parent Pair Allocations

Eric C. Anderson

Abstract

Advances in genotyping that allow tens of thousands of individuals to be genotyped at a moderate number of single nucleotide polymorphisms (SNPs) permit parentage inference to be pursued on a very large scale. The intergenerational tagging this capacity allows is revolutionizing the management of cultured organisms (cows, salmon, etc.) and is poised to do the same for scientific studies of natural populations. Currently, however, there are no likelihood-based methods of parentage inference which are implemented in a manner that allows them to quickly handle a very large number of potential parents or parent pairs. Here we introduce an efficient likelihood-based method applicable to the specialized case of cultured organisms in which both parents can be reliably sampled. We develop a Markov chain representation for the cumulative number of Mendelian incompatibilities between an offspring and its putative parents and we exploit it to develop a fast algorithm for simulation-based estimates of statistical confidence in SNP-based assignments of offspring to pairs of parents. The method is implemented in the freely available software SNPPIT. We describe the method in detail, then assess its performance in a large simulation study using known allele frequencies at 96 SNPs from ten hatchery salmon populations. The simulations verify that the method is fast and accurate and that 96 well-chosen SNPs can provide sufficient power to identify the correct pair of parents from amongst millions of candidate pairs.

KEYWORDS: pedigree reconstruction, Baum algorithm, relationship inference

Author Notes: Veronica Mayorga programmed a version of the forward-backward algorithm in Java. Results from this program were useful for testing and debugging the version in SNPPIT. I also acknowledge helpful discussions with Matthew Stephens, Elizabeth Thompson, Noah Rosenberg, Robin Waples, Marc Mangel, and Carlos Garza. Two anonymous referees provided insightful comments that strengthened the paper. I am grateful to the members of the Southwest Fisheries Science Center, Molecular Ecology and Genetic Analysis Team for providing me with the allele frequencies used to parameterize the simulations. Financial support was provided by the US Section of the Chinook Technical Committee of the International Pacific Salmon Commission.

1 Introduction

Likelihood-based pedigree reconstruction methods are increasingly used in studies of natural populations (Pemberton, 2008). The scale of these studies is growing as molecular ecologists adopt new and more efficient genotyping technologies from the fields of human and medical genetics. In particular, the recent development of single nucleotide polymorphism (SNP) markers in non-model organisms has enabled the rapid genotyping of many thousands of individuals in commercially important species such as Pacific salmon (Elfstrom, Smith, and Seeb, 2006), Atlantic salmon (Hayes, Nilsen, Berg, Grindflek, and Lien, 2007), beef cattle (Heaton, Harhay, Bennett, Stone, Grosse, Casas, Keele, Smith, Chitko-McKown, and Laegreid, 2002), and pigs (Fahrenkrug, Freking, Smith, Rohrer, and Keele, 2002). This capacity makes it possible to reconstruct pedigree relationships with genetic data from amongst tens of thousands of candidate parents (Anderson and Garza, 2006) and is revolutionizing the management of livestock operations and hatchery-propagated fish populations. It is only a matter of time before similar genotyping capacity is realized in many other species; however, the likelihood-based methods in use today were not designed for parentage inference on a very large scale, and many are not computationally efficient enough to handle large quantities of data. This paper introduces a novel computational approach that allows the rapid and accurate calculation of statistical confidence for likelihood-based, individual parentage assignments, even when the number of candidate parents is very large. It is specifically tailored to situations, as are often encountered with cultured organisms, where a large fraction of parent pairs (as opposed to just single parents) can be sampled, even if the total sampling rate is not high.

Thompson (1976) introduced likelihood methods for pedigree reconstruction in human populations. These methods were first adapted and applied to non-human populations by Meagher and Thompson (1987) who identified likely parents by means of the LOD score: the log of the probability of the offspring and putative-parent genotypes under the hypothesis of parentage divided by the probability under the hypothesis that the offspring is unrelated to the putative parents. Marshall, Slate, Kruuk, and Pemberton (1998) extended the likelihood-based approaches of Meagher and Thompson (1987) by allowing for genotyping error and by developing a Monte Carlo scheme to estimate statistical confidence in parentage assignments. Their method and its revisions (Kalinowski, Taper, and Marshall, 2007), implemented in the software program CERVUS, are the best-known of a class of parentage inference methods called "categorical allocation" methods (Jones and Ardren, 2003) and have been used in hundreds (if not thousands) of natural population studies.

Operationally, the approach implemented in CERVUS first allocates an offspring to the candidate parent (or parent pair) with the highest LOD, say LOD*. Then, the confidence of that assignment is assessed by simulating genotypes, conditional on the sample allele frequencies, and recording the fraction X of simulated, non-parentally-related pairs (or trios) that have a LOD score that exceeds LOD*. The confidence in the assignment is then expressed as a posterior predictive value which is computed assuming that the fraction of offspring whose parents are expected to be included in the pool of candidates is known (see Marshall et al. 1998 for details).

Here we develop a method for assigning offspring to parent pairs using SNP markers. Our target application is the identification of parent pairs of hatchery-born Pacific salmon. Our approach is similar in spirit to CERVUS in that it is a categorical allocation method that assesses confidence using Monte Carlo techniques; however it is efficient enough to handle parentage problems with a large number of candidate pairs of parents. For example, in the management of populations of Pacific salmon, it is conceivable that as many as 100 million candidate parent-pairs must be investigated for every offspring. Current implementations of parentage inference software such as CERVUS, COLONY (Wang, 2003, Wang and Santure, 2009), and FRANZ (Riester, Stadler, and Klemm, 2009) either run out of RAM or require a prohibitive amount of time to analyze such data sets.

We achieve computational efficiency by ensuring that both our search for likely parent pairs *and* our Monte Carlo simulations to assess confidence are performed conditional on the putative parental trios possessing a small number of Mendelian incompatibilities. We introduce a hidden Markov representation of the number of Mendelian incompatibilities in a trio that permits this to be done quickly. In addition, we implement a False Discovery Rate approach that allows confidence in parent-pair allocations to be assessed without assuming that the fraction of sampled parents is known, and we allow for parents to be derived from multiple populations with different allele frequencies. We assess the method, implemented in our software SNPPIT (SNP Program for Intergenerational Tagging) with a large scale simulation demonstrating that it is possible to accurately identify the parents of individual salmon caught in the ocean with only 96 SNPs.

2 Methods

2.1 Data, notational conventions, and preliminaries

We assume that individuals in the study are diploids with genetic data at L independently segregating SNP loci. At each locus ℓ there are two alleles: one labeled 0

and having frequency q_ℓ in the population under study and the other labeled 1 and having the frequency $p_\ell = 1 - q_\ell$. At each locus the genotype g of an individual is the number of 1 alleles it carries (*i.e.*, $g = 0$, 1, or 2), or, if the individual was not successfully genotyped at the locus, $g = \bullet$. We have a list $\mathscr{O}$ of offspring individuals whose parents we wish to infer from amongst a collection of possible fathers $\mathscr{S}$ (for "sires") and mothers $\mathscr{D}$ (for "dams"). There may be individuals in $\mathscr{O}$ whose parents are not in $\mathscr{S}$ or $\mathscr{D}$. Our goal is to infer, for every individual i in $\mathscr{O}$ the *pair* of i's parents, if present in both $\mathscr{S}$ and $\mathscr{D}$. In our application, we are not interested, for example, in inferring only the father when the mother is not in $\mathscr{D}$ because in fish hatcheries one can ensure that the mates of every female included in $\mathscr{D}$ are in $\mathscr{S}$, and vice-versa. There might additionally be information about the possible matings $\mathscr{C}$ (for "crosses") between the members of $\mathscr{S}$ and $\mathscr{D}$ and there may be additional data, collectively $\mathscr{H}$, such as age information, that can be used to exclude certain individuals from parentage.

A basic unit of interest is a trio of putative youth, father, and mother. At the ℓ^{th} locus the genotype of such a trio is the triplet $a_\ell = (g_\ell^{\text{kid}}, g_\ell^{\text{pa}}, g_\ell^{\text{ma}})$, the values of which we will write without commas (*e.g.*, $a_\ell = 000$ or $a_\ell = 10\bullet$). Note that the superscript kid, pa, and ma refer to the *putative* youth, *putative* father and *putative* mother, respectively. When all individuals are successfully genotyped the 27 possible states of a_ℓ are the set $\mathscr{A} = \{000, 001, 002, 010, \ldots, 221, 222\}$. When as many as three individuals in the trio can have missing data at the locus the 64 possible states are the set $\mathscr{A}^\bullet = \{000, 001, 002, 00\bullet, 010, \ldots, \bullet\bullet 2, \bullet\bullet\bullet\}$. For $a_\ell \in \mathscr{A}$ the probability of a_ℓ depends on the allele frequency p_ℓ, the genotyping error rate at the locus μ_ℓ, and the true relationship r of the members of the trio. We denote this probability $P(a_\ell|r)$ taking the dependence on p_ℓ and μ_ℓ as implicit. These probabilities are easily computed for any possible single-locus model of genotyping error and any r by simply summing over the genotypes of any relevant but unobserved individuals in the pedigree describing r and over the unobserved true genotypic states underlying the observed, possibly erroneous genotypes. Details can be found in the appendix of Anderson and Garza (2006). We will also make use of $P(a_\ell|r, g_\ell^{\text{kid}})$, the conditional probability of a_ℓ given r and the genotype of the kid in the trio. This probability is proportional to $P(a_\ell|r)$ for all states a_ℓ consistent with g_ℓ^{kid} and 0 otherwise, so is also computed easily. Over L independently-segregating loci which are not in linkage disequilibrium in the population, the probability of $\boldsymbol{a} = (a_1, \ldots, a_L)$ given r is simply a product, $P(\boldsymbol{a}|r) = \prod_{\ell=1}^{L} P(a_\ell|r)$. Dependence on $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)$ and $\boldsymbol{p} = (p_1, \ldots, p_L)$ is implicit in that notation.

We assume that whether or not data are missing at a locus is independent of the unobserved genotype at that locus. If the data are missing amongst some members of the trio at a locus then we compute the probability of the observed data at that locus by summing the values of $P(a_\ell|r)$ for $a_\ell \in \mathscr{A}$ over the unobserved

members. For example, in the situation where pa is missing data at locus ℓ, we define

$$P(a_\ell = 1\bullet 2|r) = \sum_{k=0}^{2} P(g^{\text{kid}} = 1, g^{\text{pa}} = k, g^{\text{ma}} = 2|r)$$

and the extensions to data missing at other members (or at more members) of the trio or to conditioning on g_{kid} are obvious.

Let $\boldsymbol{v}(a_\ell)$ be a binary 3-vector whose values indicate the manner in which the observed genotypes of a trio are, or are not, compatible with Mendelian inheritance between a mother, father, and offspring. $\boldsymbol{v}(a_\ell)$ always depends on a_ℓ, and so the a_ℓ may sometimes be dropped from the notation. The first element of $\boldsymbol{v}$ is 1 if pa and kid are Mendelian-incompatible and 0 otherwise; the second element of $\boldsymbol{v}$ is 1 if ma and kid are incompatible and 0 otherwise; the third element of $\boldsymbol{v}$ is 1 if either pa or ma are incompatible, considered alone, with kid, or, when taken together, pa and ma are not compatible as a pair of parents for kid. Values of $\boldsymbol{v}$ are written with commas like $\boldsymbol{v} = (1,0,1)$. An individual with missing data at a locus is deemed to provide no evidence that can be used to declare Mendelian incompatibility. There are 5 possible values of $\boldsymbol{v}$, each one corresponding to a subset of $\mathscr{A}$ and $\mathscr{A}^\bullet$ as summarized in Table 1.

The probability that $\boldsymbol{v}$ at locus ℓ takes a value $\boldsymbol{v}^*$ is computed by a sum over genotype states a_ℓ:

$$P(\boldsymbol{v}(a_\ell) = \boldsymbol{v}^*|r) = \sum_{a':\boldsymbol{v}(a'_\ell)=\boldsymbol{v}^*} P(a'_\ell|r). \tag{1}$$

The same holds when conditioning on g_ℓ^{kid}:

$$P(\boldsymbol{v}(a_\ell) = \boldsymbol{v}^*|r, g_\ell^{\text{kid}}) = \sum_{a':\boldsymbol{v}(a'_\ell)=\boldsymbol{v}^*} P(a'_\ell|r, g_\ell^{\text{kid}}).$$

The quantity $\boldsymbol{v}^{(\ell)}(\boldsymbol{a}) = \sum_{k=1}^{\ell} \boldsymbol{v}(a_k)$ is the cumulative number of Mendelian incompatibilities observed at the first ℓ SNP loci in a trio having genotypes $\boldsymbol{a}$. It may be written simply as $\boldsymbol{v}^{(\ell)}$. The components of this vector are written as $v_1^{(\ell)}$, $v_2^{(\ell)}$, and $v_3^{(\ell)}$, and if we write $\boldsymbol{v}^{(\ell)} \le \boldsymbol{v}^{(\ell)*}$ it means that $v_1^{(\ell)} \le v_1^{(\ell)*}$, $v_2^{(\ell)} \le v_2^{(\ell)*}$, and $v_3^{(\ell)} \le v_3^{(\ell)*}$ for some value $\boldsymbol{v}^{(\ell)*}$. The analogous convention holds if we write $\boldsymbol{v}^{(\ell)} \ge \boldsymbol{v}^{(\ell)*}$.

In many parentage applications, the log likelihood ratio or LOD, $\Lambda(\boldsymbol{a}) = \log[P(\boldsymbol{a}|\mathrm{C}_{\mathrm{Se}}^{\mathrm{Se}})/P(\boldsymbol{a}|\mathrm{C}_{\mathrm{U}}^{\mathrm{U}})]$, is employed to compare candidate parents of an individual. Following Anderson and Garza (2006), $\mathrm{C}_{\mathrm{Se}}^{\mathrm{Se}}$ denotes the hypothesis that pa and ma are the true parents of kid and $\mathrm{C}_{\mathrm{U}}^{\mathrm{U}}$ denotes the hypothesis that pa and ma are completely unrelated to kid. This statistic is most appropriate when all trios in the

Table 1: Patterns of Mendelian incompatibility, $\boldsymbol{v}$, with corresponding trio genotype states a_ℓ.

$\boldsymbol{v}$	$a_\ell \in \mathscr{A}$			Additional $a_\ell \in \mathscr{A}^\bullet$					
$(0,0,0)$	000	001	010	00•	01•	0•0	0•1	0••	10•
	011	101	102	11•	12•	1•0	1•1	1•2	1••
	110	111	112	21•	22•	2•1	2•2	2••	•00
	120	121	211	•01	•02	•0•	•10	•11	•12
	212	221	222	•1•	•20	•21	•22	•2•	••0
				••1	••2	•••			
$(0,0,1)$	100	122							
$(0,1,1)$	002	012	210	0•2	2•0				
	220								
$(1,0,1)$	020	021	201	02•	20•				
	202								
$(1,1,1)$	022	200							

sampleare either $\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}$ or $\mathrm{C}^{\mathrm{U}}_{\mathrm{U}}$, which will seldom be the case because many individuals in a finite population will be related to some degree. In such cases a preferable test statistic will be the posterior probability of parentage for a trio, as suggested by Thompson and Meagher (1987). Denoting by $\mathscr{R}$ the set of relationships amongst a trio that will be considered, and assuming that a prior probability π_r is available for all $r \in \mathscr{R}$, and $\sum_{r\in\mathscr{R}} \pi_r = 1$, the posterior probability of parentage is:

$$P(\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}|\boldsymbol{a},\boldsymbol{\pi}) = \frac{\pi_{\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}} P(\boldsymbol{a}|\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}})}{\sum_{r\in\mathscr{R}} \pi_r P(\boldsymbol{a}|r)}. \tag{2}$$

Using (2) as a test statistic is functionally equivalent to using a LOD or a likelihood ratio criterion for an alternative hypothesis of $\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}$ versus a null hypothesis of "non-parental" relationship (*i.e.*, $\{\mathscr{R}\backslash\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}\}$) because (2) is monotonically increasing with the likelihood ratio

$$\frac{P(\boldsymbol{a}|\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}})}{(1-\pi_{\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}})^{-1}\sum_{r\in\{\mathscr{R}ash\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}\}} \pi_r P(\boldsymbol{a}|r)}.$$

In fish hatchery applications, there is typically enough information on sizes of spawning populations in the past that a reasonable estimate of $\boldsymbol{\pi} = (\pi_r)_{r\in\mathscr{R}}$ can be made. In the populations that we study, we determined by simulation that there are 18 trio relationship categories which, given their expected chances of occurrence and their probability of being mistaken for a parental trio, should be included in $\mathscr{R}$ (See Appendix A). Note that a summary of notation appears in Table 2.

2.2 Overview of the method

Here we give an overview of our method, providing further detail on certain aspects and calculations in subsequent sections. The main steps are as follows:

1. Data are read in and values of parameters are initialized:
   - The genotypes of the individuals in $\mathscr{S}$ and $\mathscr{D}$ are used together to make an estimate of p_ℓ for each locus by the posterior mean given a Beta$(\frac{1}{2},\frac{1}{2})$ prior and the data in $\mathscr{S}$ and $\mathscr{D}$. This estimate is taken to be the value p_ℓ used in all probability calculations in the preceding and following sections.
   - Values of μ_ℓ, $\ell = 1,\ldots,L$, are assumed known from other sources of data, experiments, or prior beliefs.
   - Values of $\boldsymbol{\pi}$ are estimated from demographic data and from assumptions or estimates of variance in reproductive success. These estimates of $\boldsymbol{\pi}$ are used in the method as if known without error.
2. A value of $\boldsymbol{v}^{(L)}$, denoted $\boldsymbol{v}^{(L)\max}$, is chosen such that given $\boldsymbol{\mu}$ and $\boldsymbol{p}$ there is only a small probability, β^{MI}, that a truly parental trio will have a $\boldsymbol{v}^{(L)} > \boldsymbol{v}^{(L)\max}$. That is:

$$1-\beta^{\mathrm{MI}} = \sum_{a_\ell\in\mathscr{A},\ \ell=1,\ldots,L} \delta[\boldsymbol{v}^{(L)}(\boldsymbol{a}) \le \boldsymbol{v}^{(L)\max}] P(\boldsymbol{a}|\mathrm{C}^{\mathrm{Se}}_{\mathrm{Se}}) \tag{3}$$

   where $\delta[x]$ is the indicator function returning 1 if x is true and 0 otherwise. β^{MI} is the rate at which truly parental trios will not be identified because they have too many Mendelian incompatibilities. In practice, we use values of β^{MI} on the order of 0.001. The sum in (3) is calculated efficiently via a recursion which is the forward step of the forward-backward algorithm described in Section 2.3.
3. Each individual i in $\mathscr{O}$ is compared against every male in $\mathscr{S}$ that is a potential father of i according to $\mathscr{H}$, and a list $\mathrm{Pas}^{(i)}$ is maintained of potential fathers having no more than $v_1^{(L)\max}$ Mendelian incompatibilities with i. Likewise,

each i is compared to every female in $\mathscr{D}$ that is a potential mother of i according to $\mathscr{H}$, and a list $\text{Mas}^{(i)}$ is made of potential mothers with no more than $v_2^{(L)\max}$ Mendelian incompatibilities with i.

4. The genotype of each i in $\mathscr{O}$ is compared to every pair (j,k) of $j \in \text{Pas}^{(i)}$ and $k \in \text{Mas}^{(i)}$ such that j and k are a possible mated pair according to $\mathscr{C}$, and a list $\text{Pairs}^{(i)}$ is maintained of all parent pairs such that the trio they form with i has $\boldsymbol{v}^{(L)} \leq \boldsymbol{v}^{(L)\max}$. Let $N^{(i)}$ denote the number of elements in $\text{Pairs}^{(i)}$, and take the elements of $\text{Pairs}^{(i)}$ to be sorted by the largest to smallest value of $P(\text{C}_{\text{Se}}^{\text{Se}}|\boldsymbol{a},\boldsymbol{\pi})$. Thus, $\text{Pairs}_1^{(i)}$ is the pair of potential parents with highest posterior probability of parentage to offspring i.
5. The pair $\text{Pairs}_1^{(i)}$ is assigned parentage to i, and the statistical confidence in that assignment is assessed by comparison to a "null" distribution approximated via Monte Carlo by simulating for M replicates $N^{(i)}$ pairs of non-parental genotypes drawn conditional on $\boldsymbol{\pi}$ and the fact that they must have no more than $\boldsymbol{v}^{(L)\max}$ incompatibilities with i. For each of the M replicates the highest value of $P(\text{C}_{\text{Se}}^{\text{Se}}|\boldsymbol{a},\boldsymbol{\pi})$ amongst the $N^{(i)}$ simulated values is recorded and the fraction (out of M) of these which exceed the value of $P(\text{C}_{\text{Se}}^{\text{Se}}|\boldsymbol{a},\boldsymbol{\pi})$ achieved by $\text{Pairs}_1^{(i)}$ is interpreted as a p-value for the confidence in the parentage assignment. This simulation, described fully in Section 2.4, makes extensive use of the forward-backward algorithm of Section 2.3.
6. The p-values of Step 5 are used control the false discovery rate (Benjamini and Hochberg, 1995). Even when an estimate of the fraction of sampled parents is not available, use of the adaptive procedure of Benjamini and Hochberg (2000) provides a reasonable estimate of the fraction of Pairs_i observed that are true parent pairs, and this allows for more powerful control of the rate of incorrect parentage assignments. This is described in Section 2.5.

As in most methods for inferring parent pairs, we first identify individual males and females with a good chance of being parents, and then we restrict our attention to the pairs formed from that small group of males and females. However, instead of using both Mendelian incompatibility *and* the parent-offspring LOD to initially screen individual males and females (as done in Meagher and Thompson 1987), we screen candidate males and females solely on the basis of the number of loci with Mendelian incompatibilities. At first this may seem disadvantageous compared to using LODs, however it allows the assessment of statistical significance of individual parentage assignments by performing simulations while conditioning on the fact that only $N^{(i)}$ pairs had sufficiently few Mendelian incompatibilities to be included in $\text{Pairs}^{(i)}$. By contrast, it is not clear how one could efficiently simulate genotypes while conditioning on the LOD exceeding a certain amount.

Typically $N^{(i)}$ is substantially smaller than the number of candidate males or females in the study, so each Monte Carlo replicate from the null distribution requires simulating the genotypes of only $N^{(i)}$ pairs. In large studies this becomes quite important. For example, if there are 10^4 candidate males and 10^4 candidate females, but $N^{(i)}$ is only 10, then each Monte Carlo replicate requires only 10 realizations of genotype pairs. Contrast this with the standard simulation routine of CERVUS: each Monte Carlo replicate requires simulating 10^4 male and 10^4 female genotypes, each of those genotypes must be compared to a single offspring genotype, then all 10^4 males and females must be sorted, and finally some very large number of simulated male-female pairs are compared to an individual offspring genotype.

2.3 Forward-backward algorithm

In this section we show how to evaluate (3) and simulate a pair of genotypes conditional on r and $\boldsymbol{v}^{(L)}(\boldsymbol{a}) \leq \boldsymbol{v}^{(L)\max}$. For convenience we define $\boldsymbol{v}^{(0)} = (0,0,0)$, which asserts merely that, "Before looking at the genetic data at any loci, there are zero Mendelian incompatibilities of any type." It is apparent that, conditional on r, $\boldsymbol{\mu}$, and $\boldsymbol{p}$, the variables $\boldsymbol{v}^{(\ell)}$, $\ell = 1,\ldots,L$, form a Markov chain. That is,

$$P(\boldsymbol{v}^{(\ell)}|\boldsymbol{v}^{(0)},\ldots,\boldsymbol{v}^{(\ell-1)}) = P(\boldsymbol{v}^{(\ell)}|\boldsymbol{v}^{(\ell-1)}) \text{ for } \ell = 1,\ldots,L.$$

The joint distribution of $\boldsymbol{a}$ and all the $\boldsymbol{v}^{(\ell)}$'s respects the directed graph shown in Figure 1(a). The arrows from each μ_ℓ and p_ℓ into each $\boldsymbol{v}^{(\ell)}$ run in the reverse direction of a typical hidden Markov chain, but the moralized undirected graph (Figure 1(b)) is easily recognized as having the same undirected graphical structure as a simple hidden Markov chain, especially when collapsing each $\boldsymbol{v}^{(\ell)}$ and a_ℓ, and each μ_ℓ and p_ℓ into single (composite) variables, as in Figure 1(c). Therefore, we can employ the familiar forward-backward family of algorithms (Baum, Petrie, Soules, and Weiss, 1970) to efficiently compute the marginal probability of $\boldsymbol{v}^{(L)}$ and to simulate values of $\boldsymbol{a}$ conditional on $\boldsymbol{v}^{(L)}(\boldsymbol{a}) \leq \boldsymbol{v}^{(L)\max}$.

Let $\mathscr{V}$, with no superscript, refers to the five possible values of $\boldsymbol{v}$ (see Table 1). When adorned with a locus superscript, $\mathscr{V}^{(\ell)}$ refers to sets of vectors representing the cumulative number of Mendelian incompatibilities up to and including the locus. $\mathscr{V}^{(\ell)\downarrow}$ denotes the set of all vectors $\boldsymbol{v}^{(\ell)} \leq \boldsymbol{v}^{(L)\max}$ that have non-zero probability. Likewise, $\mathscr{V}^{(\ell)\uparrow}$ denotes the set of all vectors $\boldsymbol{v}^{(\ell)}$ such that $\boldsymbol{v}^{(\ell)} > \boldsymbol{v}^{(L)\max}$. For the current discussion, we will assume that data are not missing at any loci at any of the trio members (*i.e.*, $a_\ell \in \mathscr{A}$). We discuss treatment of missing data in

Table 2: Notation used in the paper

g_ℓ	genotype at locus ℓ: the number of "1" alleles, or $\bullet$ if missing.
p_ℓ	the relative frequency of the "1" allele at locus ℓ in the population.
q_ℓ	relative frequency of the "0" allele at locus ℓ. $q_\ell = 1 - p_\ell$.
$\mathscr{O}$	list of offspring whose parents are to be inferred.
$\mathscr{S}$	list of possible fathers (sires).
$\mathscr{D}$	list of possible mothers (dams).
$\mathscr{C}$	information, if available, detailing which members of $\mathscr{S}$ and $\mathscr{D}$ could have mated (crosses).
$\mathscr{H}$	other information, like age data, which, if available, could be used to exclude some parents from parentage with particular offspring.
kid, pa, pa	an offspring and a *putative* father and mother respectively.
a_ℓ	genotypes at locus ℓ in a kid, pa, and ma: $(g_\ell^{\text{kid}}, g_\ell^{\text{pa}}, g_\ell^{\text{ma}})$.
$\mathscr{A}$	the 27 possible states a_ℓ can take with no missing data.
$\mathscr{A}^\bullet$	the 64 possible states of a_ℓ when missing data are allowed.
μ_ℓ	the rate of genotyping error at locus ℓ.
L	the total number of SNPs in the data set.
$\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{\mu}$	$(a_1, \ldots, a_L)$, $(p_1, \ldots, p_L)$, and $(\mu_1, \ldots, \mu_L)$, respectively.
r	generically, a relationship between a trio of kid, pa, and ma.
$\mathscr{R}$	the set of relationships r given positive prior probability.
π_r	the prior probability that a kid, pa, and ma drawn at random from the population have relationship r.
$\boldsymbol{v}(a_\ell)$ or $\boldsymbol{v}$	vector of three binary indicators describing patterns of Mendelian incompatibility in a kid-pa-ma trio at locus ℓ.
$\boldsymbol{v}^{(\ell)}(\boldsymbol{a})$ or $\boldsymbol{v}^{(\ell)}$	cumulative number of Mendelian incompatibilities (of certain types) at loci 1 through ℓ. $\boldsymbol{v}^{(\ell)}(\boldsymbol{a}) = \boldsymbol{v}^{(\ell)} = \sum_{k=1}^{\ell} \boldsymbol{v}(a_k)$.
$\boldsymbol{v}^{(k)} \leq \boldsymbol{v}^{(k)*}$	shorthand for componentwise equality/inequality, meaning: $v_1^{(k)} \leq v_1^{(k)*}$, $v_2^{(k)} \leq v_2^{(k)*}$, and $v_3^{(k)} \leq v_3^{(k)*}$.
$\boldsymbol{v}^{(L)\max}$	max allowed number of Mendelian incompatibilities in a trio.
$\mathscr{V}$	Possible states of $\boldsymbol{v}(a_\ell)$. See Table 1
$\mathscr{V}^{(\ell)\downarrow}$	All possible values of $\boldsymbol{v}^{(\ell)} \leq \boldsymbol{v}^{(L)\max}$
$\mathscr{V}^{(\ell)\uparrow}$	All possible values of $\boldsymbol{v}^{(\ell)} > \boldsymbol{v}^{(L)\max}$

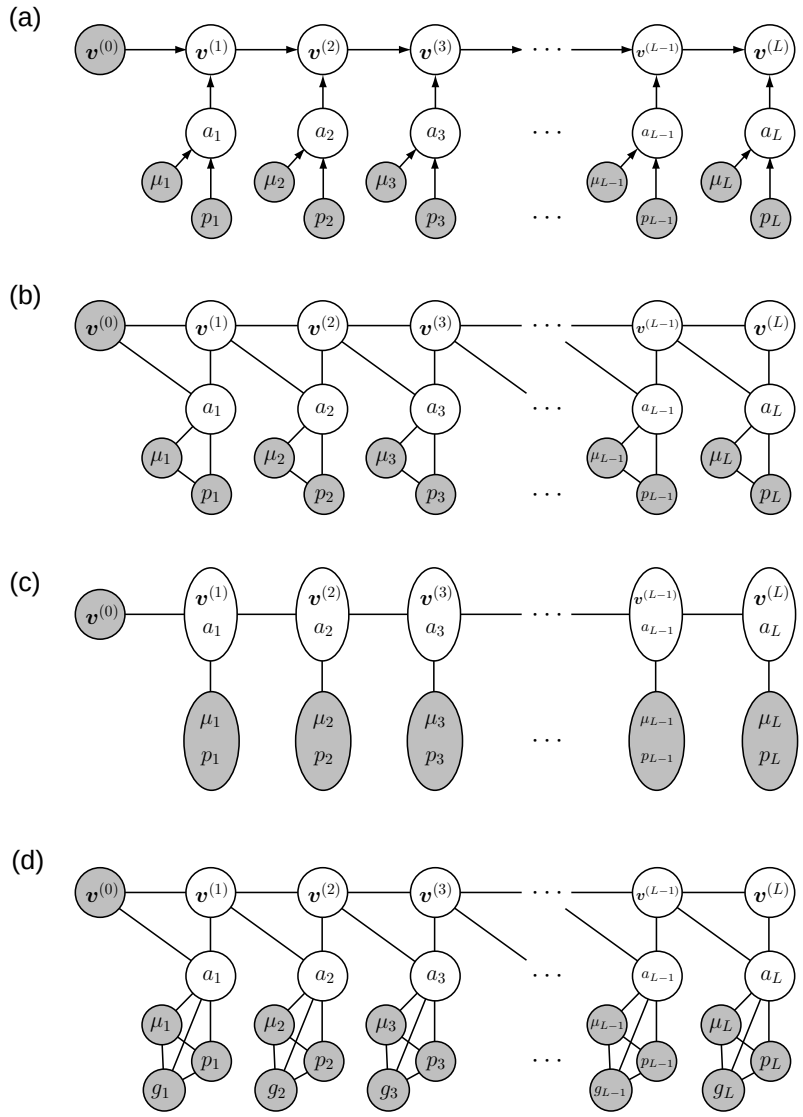

Figure 1: Graphical depictions of the dependence between trio genotypic states, $(a_1, \ldots, a_L)$, and the vectors, $\boldsymbol{v}^{(\ell)}$. Shaded nodes (μ_ℓ = mutation rate, p_ℓ = allele frequency) represent known or fixed quantities to be conditioned upon; unshaded nodes represent variables that we wish to do inference for or that we shall sum over. The dependence on some trio relationship category r is implicit. (*a*) The directed graph. (*b*) Moralized undirected graph. (*c*) With variables merged into nodes representing several variables together, this more obviously has a hidden Markov chain structure. (*d*) Conditioning on the offspring genotype is straightforward with the addition of nodes g_ℓ for the offspring genotype.

Section 2.6. The probability that $\boldsymbol{\nu}^{(L)}$ takes a certain value in $\mathscr{V}^{(L)\downarrow}$ can be computed by the forward step recursion:

$$P(\boldsymbol{\nu}^{(\ell)} = \boldsymbol{\nu}^{(\ell)*}|r) = \sum_{\boldsymbol{\nu}^{(\ell-1)} \in \mathscr{V}^{(\ell-1)\downarrow}} \sum_{\boldsymbol{\nu}' \in \mathscr{V}} P(\boldsymbol{\nu}^{(\ell-1)}|r) P(\boldsymbol{\nu}(a_\ell) = \boldsymbol{\nu}'|r) \delta[\boldsymbol{\nu}^{(\ell-1)} + \boldsymbol{\nu}' = \boldsymbol{\nu}^{(\ell)*}] \quad (4)$$

for any value $\boldsymbol{\nu}^{(\ell)*} \in \mathscr{V}^{(\ell)\downarrow}$, for $\ell = 1, \ldots, L$. In practice this sum can be calculated for all values of $\boldsymbol{\nu}^{(\ell)*} \in \mathscr{V}^{(\ell)\downarrow}$ by iterating over all the terms in the sum only once. Additionally, it is important to note that since $\boldsymbol{\nu}^{(\ell)} \geq \boldsymbol{\nu}^{(\ell-1)}$ for all ℓ, there is zero probability of reaching any state in $\mathscr{V}^{(t)\downarrow}$ from any state in $\mathscr{V}^{(\ell)\uparrow}$ for any t, ℓ. Hence, so long as the elements of $\boldsymbol{\nu}^{(L)\max}$ are not large, the sums in (4) can be evaluated quite rapidly.

This recursion is evaluated from $\ell = 1$ to L, and the values of $P(\boldsymbol{\nu}^{(\ell)} = \boldsymbol{\nu}^{(\ell)*}|r)$ are stored for later use in the backward step. At the end of the forward step, one has obtained $P(\boldsymbol{\nu}^{(L)} = \boldsymbol{\nu}^{(L)*}|r)$ for $\boldsymbol{\nu}^{(L)*} \in \mathscr{V}^{(L)\downarrow}$. Summing these values yields the probability that a trio of relationship r, given allele frequencies $\boldsymbol{p}$ and genotyping error rates $\boldsymbol{\mu}$, will have no more than $\boldsymbol{\nu}^{(L)\max}$ Mendelian incompatibilities:

$$P(\boldsymbol{\nu}^{(L)} \leq \boldsymbol{\nu}^{(L)\max}|r) = \sum_{\boldsymbol{\nu}^{(L)*} \in \mathscr{V}^{(L)\downarrow}} P(\boldsymbol{\nu}^{(L)} = \boldsymbol{\nu}^{(L)*}|r). \quad (5)$$

The conditional probability of each $\boldsymbol{\nu}^{(L)}$, given that it is in $\mathscr{V}^{(L)\downarrow}$ is

$$P(\boldsymbol{\nu}^{(L)} = \boldsymbol{\nu}^{(L)*}|r, \boldsymbol{\nu}^{(L)} \in \mathscr{V}^{(L)\downarrow}) = \frac{P(\boldsymbol{\nu}^{(L)} = \boldsymbol{\nu}^{(L)*}|r)}{P(\boldsymbol{\nu}^{(L)} \leq \boldsymbol{\nu}^{(L)\max}|r)}. \quad (6)$$

With (6) specified, we now proceed to the backward step.

The goal of the backward step is to simulate a realization of $\boldsymbol{a}$ from its distribution given r, $\boldsymbol{p}$, $\boldsymbol{\mu}$, and conditional on $\boldsymbol{\nu}^{(L)} \leq \boldsymbol{\nu}^{(L)\max}$. The backward step commences by simulating a value of $\boldsymbol{\nu}^{(L)*}$ from $P(\boldsymbol{\nu}^{(L)} = \boldsymbol{\nu}^{(L)*}|r, \boldsymbol{\nu}^{(L)} \in \mathscr{V}^{(L)\downarrow})$. It proceeds iteratively with two steps for each ℓ from L to 1:

1. Simulate a value $\boldsymbol{\nu}^*(a_\ell)$ for the pattern of Mendelian incompatibility in the trio at locus ℓ, given the realized cumulative value $\boldsymbol{\nu}^{(\ell)*}$ and the probabilities of $P(\boldsymbol{\nu}^{(\ell-1)} = \boldsymbol{\nu}^{(\ell-1)*}|r)$ computed and stored in the forward step. $\boldsymbol{\nu}^*(a_\ell)$ is drawn from:

   $$P(\boldsymbol{\nu}(a_\ell) = \boldsymbol{\nu}^*(a_\ell)|r, \boldsymbol{\nu}^{(\ell)*}) \propto P(\boldsymbol{\nu}^{(\ell-1)} = \boldsymbol{\nu}^{(\ell)*} - \boldsymbol{\nu}^*(a_\ell)|r) P(\boldsymbol{\nu}(a_\ell) = \boldsymbol{\nu}^*(a_\ell)|r),$$

   where the first term on the right hand side is computed and stored during the forward step, and the second term is computed with (1). The value of $\boldsymbol{\nu}^*(a_\ell)$, once realized, also determines the value of $\boldsymbol{\nu}^{(\ell-1)*}$ for the next iteration.

2. Simulate a_ℓ^* from the conditional probability of the observed trio genotypes given the pattern of Mendelian incompatibility at the single locus ℓ:

$$P(a_\ell = a_\ell^* | r, \boldsymbol{v}(a_\ell) = \boldsymbol{v}^*(a_\ell)) = \frac{P(a_\ell = a_\ell^* | r)}{\sum_{a':\boldsymbol{v}(a')=\boldsymbol{v}^*(a_\ell)} P(a_\ell = a' | r)}. \quad (7)$$

At the end of this, $\boldsymbol{a}^* = (a_1^*, \ldots, a_L^*)$ is a realization from $P(\boldsymbol{a} | r, \boldsymbol{v}^{(L)}(\boldsymbol{a}) \leq \boldsymbol{v}^{(L)\max})$—the distribution of $\boldsymbol{a}$ conditional on having no more than $\boldsymbol{v}^{(L)\max}$ Mendelian incompatibilities.

As is evident in the graphical structure of Figure 1*d*, the forward-backward algorithm above extends immediately to the case of conditioning both on r and g_{kid}, rather than simply conditioning on r alone. Thus, the meaning of expressions like $P(\boldsymbol{v}^{(L)} \leq \boldsymbol{v}^{(L)\max} | r, g_{\text{kid}})$ and $P(\boldsymbol{a} | r, g_{\text{kid}}, \boldsymbol{v}^{(L)}(\boldsymbol{a}) \leq \boldsymbol{v}^{(L)\max})$ should be clear.

2.4 Simulation assessment of error rates for a single kid i

For a given kid, i, the ma and pa in $\text{Pairs}_1^{(i)}$ are designated as the best candidates to be the true parents. Let $P(\text{C}_{\text{Se}}^{\text{Se}} | \boldsymbol{a}, \boldsymbol{\pi})$ (equation 2) of this pair have the value $P^{(1)}$. If we declare ma and pa in $\text{Pairs}_1^{(i)}$ the true parents, we risk making the (Type I) error of incorrectly rejecting the null hypothesis of non-parentage, when, in fact $\text{Pairs}_1^{(i)}$ are not the true parents of i. To assess this possibility, we compute a Type I error rate or "p-value" associated with assigning parentage of each kid, i, to $\text{Pairs}_1^{(i)}$. This p-value is the probability that *at least* one pair of potential parents that are not both parents of i has a value of $P(\text{C}_{\text{Se}}^{\text{Se}} | \boldsymbol{a}, \boldsymbol{\pi})$ with kid i that exceeds $P^{(1)}$. In typical implementations of likelihood based parentage (e.g. CERVUS) this probability is approximated via Monte Carlo methods by repeatedly simulating X genotypes of pairs of individuals that are not parents of kid i and recording the fraction of simulations in which at least one pair $P(\text{C}_{\text{Se}}^{\text{Se}} | \boldsymbol{a}, \boldsymbol{\pi})$ exceeds $P^{(1)}$. Traditionally, X is the number of all possible parent pairs in the actual data set. This is computationally very demanding for large scale problems. In our approach, because we can simulate genotypes for trios conditional on $\boldsymbol{v}^{(L)}(\boldsymbol{a}) \leq \boldsymbol{v}^{(L)\max}$, we need only focus on simulating a number of parent pairs equal to the number of pairs not excluded by Mendelian incompatibility with kid i. The procedure for doing so is as follows:

1. *Initialize variables*. Set EXCEED to 0. Recall that $N^{(i)}$ is the number of possible parent pairs having fewer than $\boldsymbol{v}^{(L)\max}$ Mendelian incompatibilities with kid i.

2. *Perform the forward step calculations.* For each $r \in \mathscr{R}$ compute $P(\boldsymbol{v}^{(L)} \leq \boldsymbol{v}^{(L)\max} | r, g_{\text{kid } i})$. Note that this probability is conditional on the offspring genotype.
3. *Compute π_r^* the expected fraction of trios with relationship r conditional on them having fewer than $\boldsymbol{v}^{(L)\max}$ Mendelian incompatibilities with* kid *i.* This involves a simple reweighting of $\boldsymbol{\pi}$:

$$\pi_r^* = \frac{\pi_r P(\boldsymbol{v}^{(L)} \leq \boldsymbol{v}^{(L)\max} | r, g_{\text{kid } i})}{\sum_{k \in \mathscr{R}} \pi_k P(\boldsymbol{v}^{(L)} \leq \boldsymbol{v}^{(L)\max} | r = k, g_{\text{kid } i})} \quad , \quad \forall r \in \mathscr{R}$$

4. *Repeat the following steps REPS times*:
   - *Repeat the following $N^{(i)}$ times*:
     - Simulate a relationship r^* from $\boldsymbol{\pi}^*$
     - Using the backward algorithm, simulate the genotypes $\boldsymbol{a}^*$ of a trio from the distribution $P(\boldsymbol{a} | r^*, g_{\text{kid } i}, \boldsymbol{v}^{(L)}(\boldsymbol{a}^*) \leq \boldsymbol{v}^{(L)\max})$
     - Compute $P(\mathrm{C}_{\mathrm{Se}}^{\mathrm{Se}} | \boldsymbol{a}^*, \boldsymbol{\pi})$ (using $\boldsymbol{\pi}$, not $\boldsymbol{\pi}^*$) for this simulated genotype.
   - *If any of the $N^{(i)}$ values of $P(\mathrm{C}_{\mathrm{Se}}^{\mathrm{Se}} | \boldsymbol{a}^*)$ exceeded $P^{(i)}$, add 1 to* EXCEED.
5. *Compute the p-value.* At the end, EXCEED/REPS is a Monte Carlo estimate of the Type I error for assigning kid *i* to the ma and pa of Pairs$^{(i)}$.

2.5 Using *p*-values in the False Discovery Rate procedure

After computing *p* values as described above for every fish *i* in $\mathscr{O}$, we use the False Discovery Rate procedure (FDR) to control our rate of False Discoveries (*i.e.*, the fraction of offspring assigned to parent pairs that are not both parents of *i*). Let *m* be the total number of offspring with $\boldsymbol{v}^{(L)} \leq \boldsymbol{v}^{(L)\max}$ for at least one pair of putative parents, and let $m_0 \leq m$ be the unknown number of those offspring for whom pa and ma in Pairs$^{(1)}$ are not the true parental pair. Then, order these *m* offspring from smallest to largest *p*-value, letting (i) denote the offspring with the i^{th} smallest *p*-value, $p^{(i)}$. Benjamini and Hochberg (1995) showed that, in expectation, a false discovery rate less than α_{fdr} can be achieved by declaring parentage to offspring $(1), \ldots, (k)$, where (k) is the largest value such that

$$p^{(i)} < \frac{i}{m} \alpha_{\text{fdr}}.$$

A more powerful approach is possible if the number m_0 is known or can be estimated. Benjamini and Hochberg (2000) provide an *ad hoc*, but general, graphically-inspired method for estimating m_0 that we use. With an estimate of m_0 the FDR can

be controlled by assigning parentage to offspring $(1),\ldots,(k)$ where (k) is the largest value such that:

$$p^{(i)} < \frac{i}{m_0}\alpha_{\text{fdr}}.$$

The above expressions can be easily inverted to express the FDR as a function of $p^{(i)}$:

$$\alpha_{\text{fdr}} \approx p^{(i)}\frac{m_0}{i}.$$

This quantity is reported in the output of SNPPIT, so users can see the FDR implied by any choice of i.

2.6 Treatment of missing data and extension to multiple populations

There are many ways in which missing data might be handled in the above procedures. We have chosen a way that gives reasonable results without creating too much computational overhead. First, as described above, we can compute the probability of a trio, given r, with missing data by simply marginalizing over the missing genotypes. For the forward step while assesssing p-values, however, we condition only on the missing data in the offspring. This is done via a straightforward side effect of the fact that we condition on the offspring genotype when doing the forward step for assessing p-values. In the backward step, we incorporate the occurrence of missing data in the members of the Pairs_i list by masking the genotypes of the simulated trios by the pattern of missing data found in each member of Pairs_i. To be more explicit, on each of the REPS replicates, each of the trio genotypes $j \in 1,\ldots,N^{(i)}$ is first simulated by the backward algorithm, then holes are created in each simulated genotype j according to where data are missing in the trio formed by g_{kid} and $\text{Pairs}_i^{(j)}$. If there is more than a small amount of missing data in any individual, we find that the genotype calls in that individual are prone to high rates of error, so we have a missing data threshhold that the user may set in our software. If an individual has more missing data than this threshold (set by default to be 10 SNPs) then it is discarded from further consideration.

The extension to multiple populations of parents in the parent data base is also quite simple. If it is unknown *a priori* which population a collection of offspring came from, then each offspring in that collection is compared to every parent from every population in the parent data base. Each individual i is assigned to the population that $\text{Pairs}_i^{(1)}$ belongs to, and then the analysis proceeds as before, assuming that all $N^{(i)}$ pairs in Pairs_i are from the same population as $\text{Pairs}_i^{(1)}$, even

Table 3: Population allele frequencies used in simulations. Columns headed by a give the number of SNPs with minor allele frequency $\geq a$ and $< a + 0.1$.

Name[a]	Size	ID	0.0	0.1	0.2	0.3	0.4
Feather R H Fa	Large	L1	13	15	22	20	27
Kalama R H	Large	L2	15	13	19	22	28
Cowlitz R H	Large	L3	14	17	15	30	21
Feather R H Sp	Medium	M1	10	16	19	25	27
Klamath R Iron Gate H	Medium	M2	22	16	15	21	23
Trinity R H Fa	Medium	M3	23	19	18	17	20
Rogue R Sp	Medium	M4	15	13	24	23	22
Upper Sacramento R LF	Small	S1	13	16	23	26	19
Sacramento R Wi	Small	S2	26	17	13	25	16
Chetco R	Small	S3	10	16	27	26	18

[a] R = river, H = hatchery, Fa = fall run, Sp = spring run, LF = late fall run, Wi = winter run

if they were not. The false discovery rates are then accordingly computed as FDRs for the individuals non-excluded from a particular population.

3 Simulation Study

A simulation was undertaken to ensure that SNPPIT can handle large inference problems and to assess the expected accuracy of a set of SNP markers available for Chinook salmon. These simulations were parameterized with allele frequencies at 96 SNPs estimated from 10 hatchery salmon populations (Table 3) from California to Washington screened by the Southwest Fisheries Science Center (Clemento et al. in prep). Three hatcheries were designated as large (≈11,000 fish spawned per year), four as medium (≈4,400 fish/yr), and three as small (≈1,100 fish/yr), reflecting the range of actual hatchery sizes (Table 4) in California.

Pedigrees and genetic data were simulated at each hatchery using the program SPIP (Anderson and Dunham, 2005). Parameters chosen reflect a typical Chinook salmon life history in the southern part of their range: maximum age of 5 years; fish spawning at a later age produce more offspring; all fish die after spawn-

Table 4: Average number of spawners in each of the Large, Medium, and Small hatchery scenarios.

	Large		Medium		Small	
Age	Male	Female	Male	Female	Male	Female
2	336	0	135	0	31	0
3	2,594	1,823	1,038	723	253	182
4	1,767	2,314	696	915	174	230
5	999	1,373	401	541	97	136
Total	5,696	5,510	2,270	2,179	555	548

ing; variance in reproductive success between families is such that the effective number of spawners is one fourth of the census number, etc. Approximately 100 males and 100 females were spawned each day in each hatchery. Crosses between parents were either made "1 to 1" between males and females (called SG1), or in a fashion in which 4 males are mated to 4 different females in all 16 possible ways (SG4). A chosen fraction $G \in \{1/4,\ 1/2,\ 1\}$ of the fish spawned at each hatchery were included in the parent data base by sampling genotypes on one fourth, one half, or all of the spawning days. Any male and female spawned on the same day were assumed to be potential mates in $\mathscr{C}$. Thus, for each spawning day with complete sampling ($G = 1$) at a hatchery, an additional 10^4 parent pairs must be considered in the data base (Table 5). See Anderson (2010) for more explicit details of the simulations.

Each simulation was run for 21 years. This provides several generations for the accrual of relatedness between members of the population which should make it more difficult to correctly infer parentage. Every year, fish ≥ 2 years old were subjected to a 10% probability of being captured in a fishery, removed from the population, and sampled. The genotypes of fish sampled this way in spawning year 19 (*i.e.*, those fish that could be 2-year-olds born of parents that spawned at year 17, 3-year-olds from year 16, 4-year-olds from year 15, or 5-year-olds from year 14) were included in the data set. On average there were 18,900 fish in these fishery samples ($\approx$3,850, 1,530, and 390 from each of the Large, Medium, and Small hatcheries, respectively) . We then attempted to infer the parentage of these

Table 5: Representative numbers of fish, N, and possible parent pairs, P (in 1,000s), according to $\mathscr{C}$ in the parent data base from each hatchery of a given size under the approximate sampling fractions G. These are rounded numbers from a single replicate simulation experiment. The total parent data base size is the number of parent genotypes from all possible years of spawners at all the hatcheries.

	$G = 1.0$		$G = 0.5$		$G = 0.25$	
Hatchery Size	N	P	N	P	N	P
Large	44,750	2,210	22,600	1,120	11,500	560
Medium	17,800	875	9,100	450	4,800	240
Small	4,400	207	2,400	119	1,600	79
Total Parent Data Base	219,000	10,760	111,500	5,490	58,600	2,880

fish from amongst the spawners in years 14–17. This represents a difficult case in which the fishery sample is a mixture of unknown proportions of fish from 10 different hatcheries.

For each replicate of the simulation, a data set was compiled in SNPPIT format that included both the parent data base and the fishery sample. With complete sampling ($G = 1.0$), the size of each of these files was about 100 Mb in text format. Genotyping error was simulated by processing the data set once it was in SNPPIT format with a program written in C that changed the type of each SNP allele in the data set, independently, with probability 0.005. This corresponds to a per-locus genotyping error rate of about 1%. Additionally, with probability 2%, each locus was designated as missing data. With complete sampling ($G = 1.0$) SNPPIT required roughly 1.5 hours on a single core of a 2.8 GHz Quad-Core Intel Xeon chip in a Mac Pro computer to analyze each data set. For smaller values of G the size of the parent data base was smaller, and each replicate took less time (roughly 30 minutes for $G = 0.5$ and 10 for $G = 0.25$). We used the `--mi-fnr` option of SNPPIT to set $\beta^{\mathrm{MI}} \leq 0.005$, and we chose a desired FDR of 1 in 200 (0.005). The results were compared with the true simulated pedigrees, and the accuracy of the parentage assignments was compared. Additionally, the number of individuals in the fishery with parents in the parent data base that were not included amongst the set of parentage assignments was recorded.

SNPPIT can take, as an advanced input, the fraction of trios, formed by randomly drawing individuals from the parental generation and from the fishery sample, expected to be of different relationship types $r \in \mathscr{R}$. These fractions were estimated by a simple recursive program (not described here) using the demographic parameters and observed numbers of spawners in the simulated hatchery each year. Thus, the probabilities π_r (for $r \in \mathscr{R}$) were estimated for Large, Medium, and Small hatcheries using data that will typically be available in hatchery programs (approximate number of spawners of different ages each year, average number of male mates per female spawned, approximate N_e/N ratio, *etc.*).

In every simulation, nearly 19,000 fish were compared to the parent data base. Depending on G, 100%, 50% or 25% of these fish had both parents in the parent data base. The fraction of offspring allocated to parents that were allocated to incorrect parents (*i.e.*, the false discovery rate) was less than 1 in 200 in almost all replicates (Figures 2 and 3). As is apparent in the figures, the average realized rate of false discoveries is clearly less than 0.005, as we hope it should be, since the desired false discovery rate was set to 1 in 200. Additionally, in almost all cases the rate at which offspring whose parents were in the parent data base were not assigned parentage (the false negative rate) was less than 0.1. With 100% sampling of the parents, this false negative rate was appreciably lower. There was not a remarkable difference in accuracy between the SG1 and SG4 mating policies.

4 Conclusions

Anderson and Garza (2006) introduced calculations demonstrating that very large parentage inference problems could be solved with relatively few SNPs. Furthermore, they showed that a likelihood-based approach to parentage inference with SNPs would require 30% fewer markers to achieve power comparable to that of a method that relied exclusively on Mendelian incompatibility. However, the scale of such parentage problems is well beyond the capacity of current implementations of likelihood-based parentage inference software. The method presented here represents one possible solution to the large-scale parentage inference problem that combines the speed of a Mendelian incompatibility approach with the accuracy of a likelihood-based approach and which provides a statistical measure of uncertainly in the parent allocations.

The method has been implemented in the software SNPPIT (source and binaries are available for free download from `tinyurl.com/snppit`). We assessed SNPPIT with a simulation study that is, to our knowledge, the largest likelihood-based parentage inference exercise (real or simulated) that has been reported. A comparison of SNPPIT's performance to that of competing softwares (COLONY,

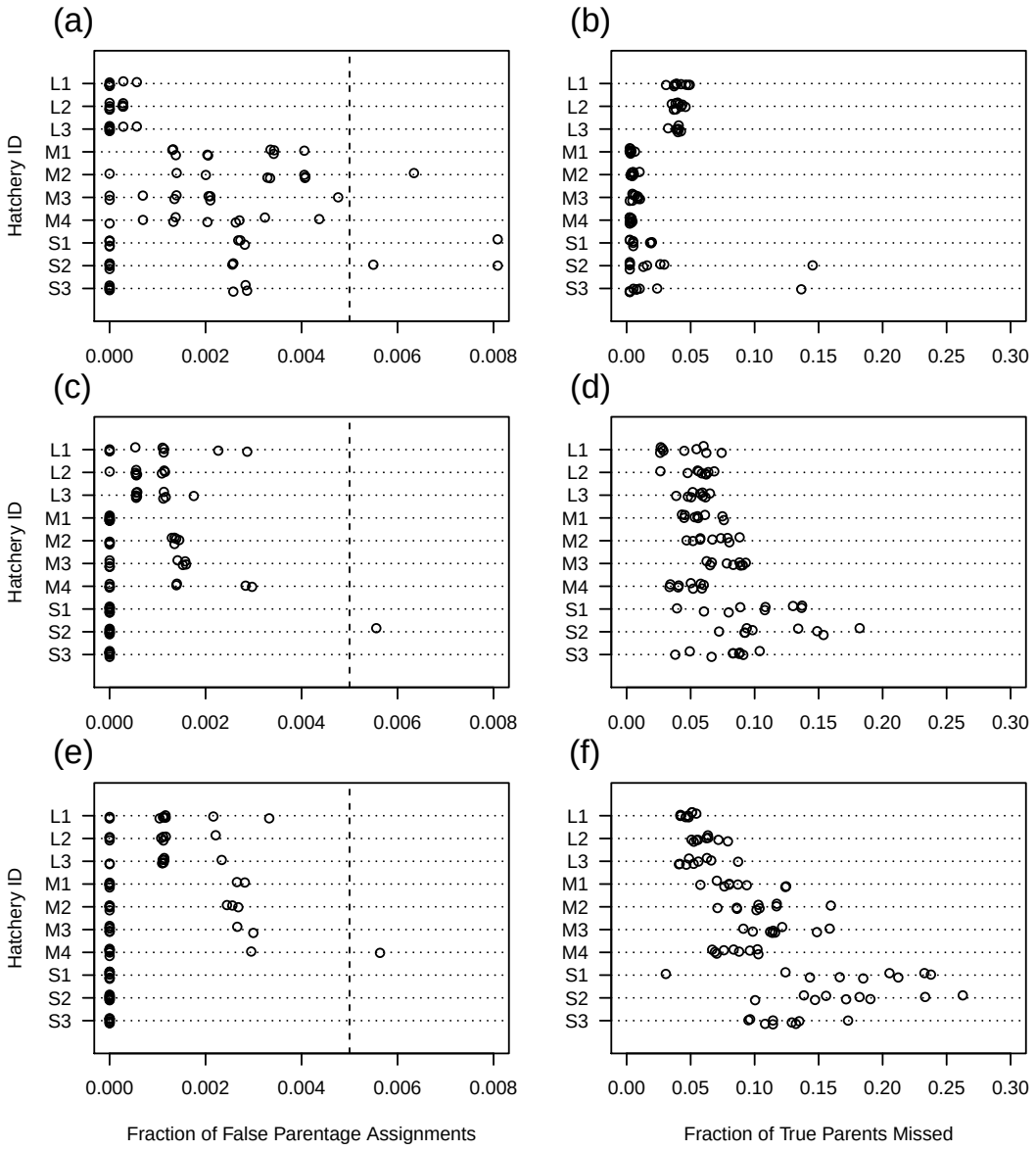

Figure 2: Results from SG1 mating policy. Left column of panels shows the true rate of false discovery (*i.e.*, true fraction of all parentage assignments that were incorrect); right column shows the fraction of offspring with parents in the data base that were not assigned to their parents (the false negative rate). Top row is for $G = 1.0$, middle is $G = 0.5$, and bottom is $G = 0.25$. Each dot in a plot is the result specific to one of the ten hatcheries (indicated by ID on the y-axis, see Table 3) in one of the replicate runs.

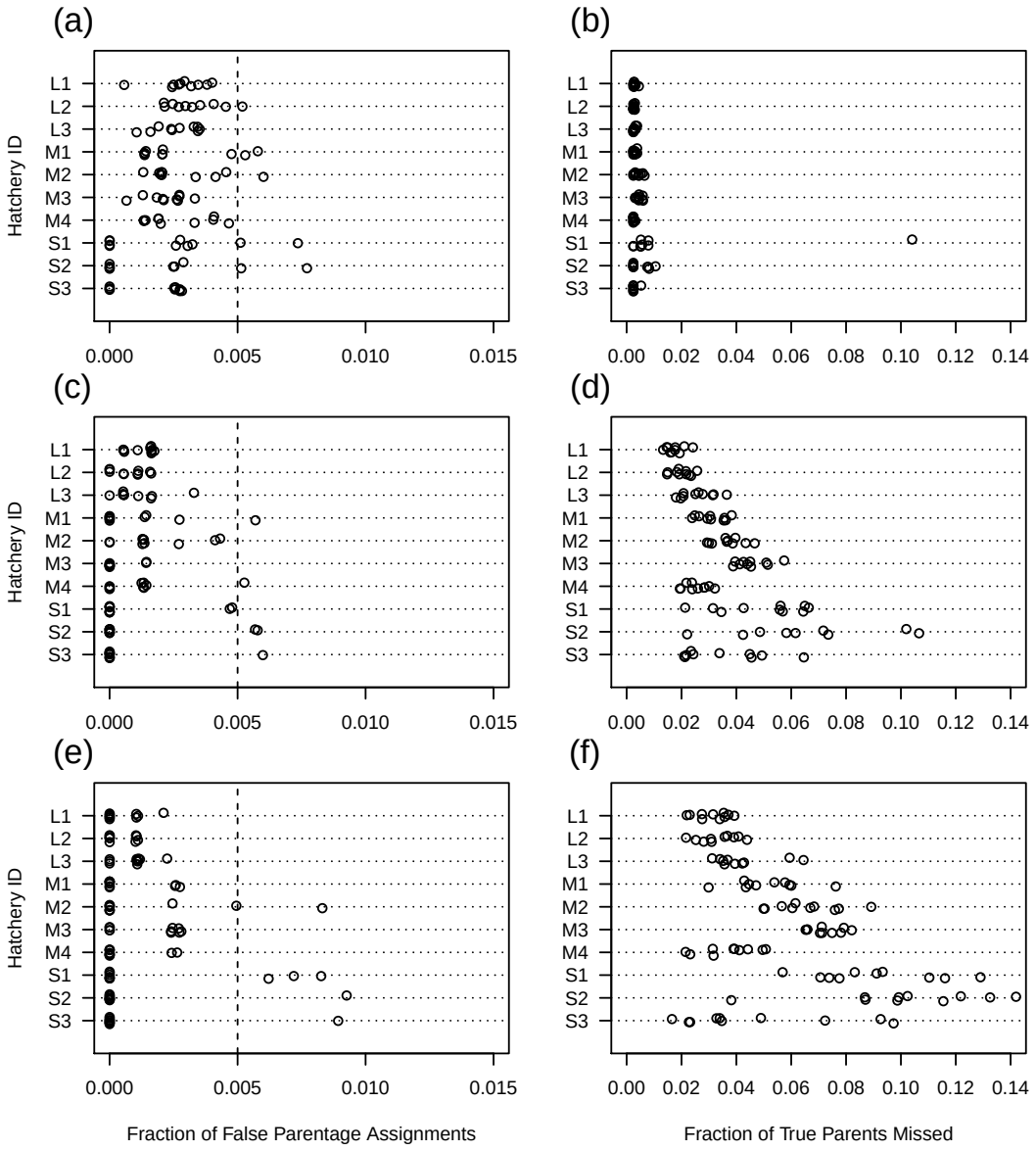

Figure 3: Results from SG4 mating policy. Left column of panels shows the true rate of false discovery (*i.e.*, true fraction of all parentage assignments that were incorrect); right column shows the fraction of offspring with parents in the data base that were not assigned to their parents (the false negative rate). Top row is for $G = 1.0$, middle is $G = 0.5$, and bottom is $G = 0.25$. Each dot in a plot is the result specific to one of the ten hatcheries (indicated by ID on the y-axis, see Table 3) in one of the replicate runs.

FRANZ, CERVUS) on these simulated data sets was not possible because no other software was able to complete the inference. However, the accuracy of SNPPIT for parent pair inference from smaller, real data sets has been found to be comparable to that of other available software (Hauser, Baird, Hilborn, Seeb, and Seeb, 2011). In a small set of simulations on modestly-sized data sets, SNPPIT performed as well or better than FRANz for inference of (and assessment of uncertainty in) parent pairs (unpublished result).

The heart of our method is the forward-backward algorithm for simulating genotypes conditional on the number of Mendelian incompatibilities. The forward step can be seen to be an elaboration of the recursive method of Chakraborty and Schull (1976) for computing the distribution of the number of Mendelian incompatibilities between a pair of individuals. The backward step may prove useful in other contexts where simulating from the distribution of genotypes conditional on the number of Mendelian incompatibilities is desired.

There are broadly three approaches to likelihood-based parentage inference: 1) "categorical allocation" as pursued here and exemplified by Marshall et al. (1998), which essentially treats each comparison of an offspring to possible parents as a hypothesis test of parentage versus non-parentage, 2) "fractional allocation" (Devlin, Roeder, and Ellstrand, 1988, Nielsen, Mattila, Clapham, and Palsboll, 2001, Hadfield, Richardson, and Burke, 2006, Jones, Grossman, Walsh, Porter, Avise, and Fiumera, 2007) in which the genotype of every parent or pair of parents is treated as the parameter of a separate component-specific distribution in a finite mixture model from which the offspring are exchangeable samples, and 3) full pedigree reconstruction (Almudevar, 2003, Riester et al., 2009, Cowell, 2009, Wang and Santure, 2009, Almudevar and LaCombe, 2012), which casts the problem of reconstructing the (possibly multigenerational) pedigree connecting a group of samples as a problem in model selection.

We pursued categorical allocation in the hopes that the simplicity of the approach would lend itself to optimization for computational efficiency; however, categorical allocation is unsatisfying for a number of reasons. For example, categorical allocations of offspring are made independently of one another, which discards some information, especially when parents have widely disparate fertilities (Roeder, Devlin, and Lindsay, 1989). Furthermore, the FDR correction used here, with Benjamini and Hochberg (2000)'s estimate of m_0 is an *ad hoc* approach which could almost certainly be improved upon by a fractional allocation approach which directly estimates the fraction of offspring whose parents are included in the data base (Nielsen et al., 2001). The latter would have the advantage of providing a direct estimate of the false negative rate, which does not seem straightforward in our framework after having conditioned on trios having fewer than β^{MI} Mendelian incompatibilities. A reliable estimate of the false negative rate would be particu-

larly beneficial, as it is clear from Figures 2 and 3 that the false negative rates can be moderately high and also variable depending on the fraction of parents sampled and the size of the hatchery, etc.

It thus seems it would be beneficial to pursue a fractional allocation approach to large scale parentage inference. Doing so would require some approximation to eliminate the very large sums over all possible parent pairs. Furthermore, in the fractional allocation framework, there is not a simple analog to Equation 2; that is, it is not clear how to efficiently deal with the occurrence of individuals who are not parentally related to a putative offspring, but are nonetheless related to it. Such relationships can be dealt with explicitly in the framework of full pedigree reconstruction, but, since such approaches depend on optimization or integration over complex discrete spaces (the space of possible pedigrees) which grow quickly with the sample size, it seems that the extension of full pedigree reconstruction methods to very large scale problems will be problematic, at least in the near future.

Finally, two features of SNPPIT's current implementation may restrict its utility in some sampling contexts. First, it only identifies parent pairs, not single parents whose mates might be absent from the data base. This makes the software most appropriate for controlled breeding situations where all the potential parents can be sampled, rather than sampling in wild populations; although even in controlled situations (like salmon hatcheries), missing data and sample handling problems can reduce the fraction of completely sampled parent pairs in the data base. Second, the forward-backward calculations are linear in the number of markers but quadratic in the cardinality of $\mathscr{V}^{(L)\downarrow}$, and, with constant genotyping error rate, the appropriate cardinality of $\mathscr{V}^{(L)\downarrow}$ increases faster than linearly in L. With higher genotyping error rates (as might be encountered with next-generation genotyping-by-sequencing) the cardinality of $\mathscr{V}^{(L)\downarrow}$ will increase even faster. Thus, as the number of available SNPs increases (especially if they are based on fast but less reliable sequencing technologies), the forward-backward calculations may become cumbersome.

Data sets with more markers may strongly violate the assumption that markers are not physically linked. When loci are not independently segregating, then it is not generally the case that the probability of data at L SNP loci in a trio can be written as a product over loci of the single-locus probabilities. Additionally the state of Mendelian compatibility versus incompatibility is not generally independent across loci. Methods to compute the data at linked markers over a set of related individuals given general pedigrees are available (see, for instance, Thompson 2000), though such methods are considerably more computationally intensive than methods for independent segregation. Fortunately, so long as markers are not in linkage disequilibrium (LD), the genotype probability of a trio of individuals can be written as a product over loci, even in the presence of physical linkage, in two very important

cases: C_{Se}^{Se} and C_{U}^{U} (as can be shown by the fact that for each relationship all possible lines of descent yield the same node-unlabeled founder tree graph (Sobel and Lange, 1996)). Accordingly, the forward calculations related to screening individuals on the basis of Mendelian incompatibility will be correct for C_{Se}^{Se} and C_{U}^{U} trios, even with linked markers (not in LD). Since most trios will be of the C_{U}^{U} category in large problems, this will still provide a principled way of eliminating non-parental trios. Ultimately this will reduce the number of likelihood evaluations necessary while accounting for physical linkage, and represents an interesting direction for future work.

In conclusion, there is still considerable work to be done on the large-scale parentage inference problem, but this paper has introduced one novel approach. The simulations in this paper, as well as ongoing work in Chinook (Clemento et al. *in prep*) and coho (Starks et al. *in prep*) salmon and steelhead trout (Abadía-Cardoso et al. *in prep*) have also demonstrated that parentage at such scales using SNPs is feasible. The intergenerational tagging of individuals that such parentage inference will enable will continue to revolutionize the management of cultured organisms as well as, ultimately, the scientific study of wild populations.

Appendix A: trio relationships

For salmon populations, we included, in $\mathscr{R}$, 18 distinct relationships amongst the three members of a kid-pa-ma trio. In all 18 of these, the individuals are assumed to be noninbred, so we do not consider categories in which, for example, a candidate father is both a sibling and the true father of the putative offspring; however, such trio categories could be accommodated without great difficulty. The first nine trio relationships involve situations in which ma or pa share a unilineal relationship to a noninbred kid through the true parents. Following Anderson and Garza (2006) these are the *C*-type relationships, all of which may be denoted by C_{ma}^{pa} where pa and ma are placeholders for the relationship (Se for self, Si for full sibling, U for unrelated) between pa and a true parent and ma and the other true parent, respectively. For example, C_{Se}^{Se} denotes a trio with two parents and an offspring and C_{U}^{U} denotes a trio of unrelated individuals. The next eight trio relationship categories that we consider are those in which exactly one of ma or pa is related as a full sibling or as a half sibling with kid and the other candidate parent is related, as Se or Si, or is unrelated, U, to a single one of the true parents of kid. We denote these trio relationships by F (for full sibling) or H (for half sibling) adorned with a superscript or subscript Se, Si, or U, if the candidate that does not have the full- or half-sibling relationship with kid is pa or ma, respectively. For example, F_{Si} indicates that pa is the full sibling of kid and ma is the full sibling of the true mother (or the true father), and H^{U} indicates

that ma is a half-sibling of kid and pa is unrelated to either of the true parents. The final trio relationship that we consider is FF—both pa and ma are full siblings of kid. Some of these 18 relationship categories may contain up to two underlying pedigree relationships owing to the fact that, in some cases, the candidate parents may be related to the true parents of like or opposite sex. This distinction becomes important when predicting $\boldsymbol{\pi}$. The 18 categories we used in $\mathscr{R}$ can now be listed as: C_{Se}^{Se}, C_{Si}^{Se}, C_{Se}^{Si}, C_{U}^{Se}, C_{Se}^{U}, C_{Si}^{Si}, C_{U}^{Si}, C_{Si}^{U}, C_{U}^{U}, F^{Se}, F_{Se}, H_{Se}, H^{Se}, F_{Si}, F^{Si}, F_{U}, F^{U}, FF.

References

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

Almudevar, A. and J. LaCombe (2012): "On the choice of prior density for the bayesian analysis of pedigree structure," *Theoretical Population Biology*, 81, 131–143.

Anderson, E. C. (2010): "Computational algorithms and user-friendly software for parentage-based tagging of Pacific salmonids. A final report to the Pacific Salmon Commission's Chinook Technical Committee (US Section). 12 March 2010. `http://tinyurl.com/snppit`," Technical report.

Anderson, E. C. and K. K. Dunham (2005): "SPIP 1.0: a program for simulating pedigrees and genetic data in age-structured populations," *Molecular Ecology Notes*, 5, 459–461.

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

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970): "A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains," *Annals of Mathematical Statistics*, 41, 164–171.

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

Benjamini, Y. and Y. Hochberg (2000): "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83.

Chakraborty, R. and W. Schull (1976): "A note on the distribution of the number of exclusions to be expected in paternity testing," *American Journal of Human Genetics*, 28, 615–618.

Cowell, R. G. (2009): "Efficient maximum likelihood pedigree reconstruction," *Theoretical population biology*, 76, 285–91.

Devlin, B., K. Roeder, and N. C. Ellstrand (1988): "Fractional paternity assignment—theoretical development and comparison to other methods," *Theor Appl Genet*, 76, 369–380.

Elfstrom, C. M., C. T. Smith, and J. E. Seeb (2006): "Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon," *Molecular Ecology Notes*, 6, 1255–1259.

Fahrenkrug, S., B. Freking, T. Smith, G. Rohrer, and J. Keele (2002): "Single nucleotide polymorphism (snp) discovery in porcine expressed genes," *Anim Genet*, 33, 186–195.

Hadfield, J. D., D. S. Richardson, and T. Burke (2006): "Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a bayesian framework," *Molecular Ecology*, 15, 3715–30.

Hauser, L., M. Baird, R. Hilborn, L. W. Seeb, and J. E. Seeb (2011): "An empirical comparison of snps and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*oncorhynchus nerka*) population," *Molecular Ecology Resources*, 11 Suppl 1, 150–61.

Hayes, B. J., K. Nilsen, P. R. Berg, E. Grindflek, and S. Lien (2007): "Snp detection exploiting multiple sources of redundancy in large est collections improves validation rates," *Bioinformatics*, 23, 1692–3.

Heaton, M. P., G. P. Harhay, G. L. Bennett, R. T. Stone, W. M. Grosse, E. Casas, J. W. Keele, T. P. Smith, C. G. Chitko-McKown, and W. W. Laegreid (2002): "Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle," *Mammalian Genome*, 13, 272–81.

Jones, A. G. and W. R. Ardren (2003): "Methods of parentage analysis in natural populations," *Molecular Ecology*, 12, 2511–2523.

Jones, B., G. D. Grossman, D. C. I. Walsh, B. A. Porter, J. C. Avise, and A. C. Fiumera (2007): "Estimating differential reproductive success from nests of related individuals, with application to a study of the mottled sculpin, cottus bairdi," *Genetics*, 176, 2427–39.

Kalinowski, S. T., M. L. Taper, and T. C. Marshall (2007): "Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment," *Mol Ecol*, 16, 1099–1106.

Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton (1998): "Statistical confidence for likelihood-based paternity inference in natural populations," *Molecular Ecology*, 7, 639–655.

Meagher, T. R. and E. A. Thompson (1987): "Analysis of parentage for naturally established seedlings within a population of *Chamaelirium luteum* (Liliaceae)," *Ecology*, 68, 803–812.

Nielsen, R., D. K. Mattila, P. J. Clapham, and P. J. Palsboll (2001): "Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale," *Genetics*, 157, 1673–82.

Pemberton, J. M. (2008): "Wild pedigrees: the way forward," *Proc Roy Soc B*, 275, 613–21.

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

Roeder, K., B. Devlin, and B. G. Lindsay (1989): "Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities," *Biometrics*, 45, 363–380.

Sobel, E. and K. Lange (1996): "Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics," *Am J Hum Genet*, 58, 1323–37.

Thompson, E. A. (1976): "Inference of genealogical structure," *Social Science Information*, 15, 477–526.

Thompson, E. A. (2000): *Statistical Inference from Genetic Data on Pedigrees*, Beachwood, OH: Institute of Mathematical Statistics.

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

Wang, J. (2003): "Maximum-likelihood estimation of admixture proportions from genetic data," *Genetics*, 164, 747–65.

Wang, J. and A. W. Santure (2009): "Parentage and sibship inference from multilocus genotype data under polygamy," *Genetics*, 181, 1579–94.