

# Parentage and sibship inference from markers in polyploids

JINLIANG WANG

*Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom*

KIM T. SCRIBNER

*Department of Fisheries and Wildlife and Department of Zoology, Michigan State University,  
East Lansing, MI, USA*

*Left running head:* J Wang & KT Scribner

*Right running head:* Parentage and sibship inference in polyploids

*Key words:* Polyploids, parentage, sibship, genetic markers, simulations, maximum likelihood

*Corresponding author:*

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: [jinliang.wang@ioz.ac.uk](mailto:jinliang.wang@ioz.ac.uk)

## **Abstract**

Many plants and some animal species are polyploids. Non-disomically inherited markers (e.g. microsatellites) in such species cannot be analysed directly by standard population genetics methods developed for diploid species. One solution is to transform the polyploid codominant genotypes to pseudo diploid dominant genotypes, which can then be analysed by standard methods for various purposes such as spatial genetic structure, individual relatedness and relationship. Although this data transformation approach has been used repeatedly in the literature, no systematic study has been conducted to investigate how efficient it is, how much marker information is lost and thus how much analysis accuracy is reduced. More specifically, it is unknown whether or not the transformed data can be used to infer parentage and sibship jointly, and how different sampling schemes (number and polymorphism of markers, number of individuals) and ploidy level affect the inference accuracy. This study analyzes both simulated and empirical data to examine the effects of polyploid levels, actual pedigree structures, and marker number and polymorphism on the accuracy of joint parentage and sibship assignments in polyploid species. We show that sibship, parentage and selfing rates in polyploids can be inferred accurately from a typical set of microsatellite loci. We also show that inferences can be substantially improved by allowing for a small genotyping error rate to accommodate the distortion in assumed Mendelian inheritance of the converted markers when large sibship groups are involved. The results are discussed in the context of polyploid data analysis in molecular ecology.

## **Introduction**

Many population genetics methods have been widely applied in ecology, evolutionary and conservation biology to infer population structure (e.g. Pritchard *et al.* 2000; Vekemans & Hardy 2004; Waples & Gaggiotti 2006; Guillot *et al.* 2009), individual relatedness (Queller & Goodnight 1989; Lynch & Ritland 1999; Wang 2002), and relationship (Marshall *et al.* 1998; Goodnight & Queller 1999; Wang 2004) from a sample of marker genotypes. Almost invariably these methods are developed for diploid species, by calculating a likelihood or moment estimator derived explicitly from Mendelian principles for diploids. Without modifications, these established diploid model based methods do not apply to polyploid species in which markers show non-disomically inheritance.

There are several difficulties in extending these methods to apply to polyploid species. One primary difficulty comes from the identification of individual marker genotypes. For a  $k$ -allele codominant marker such as microsatellites, all possible  $k(k+1)/2$  genotypes, including  $k$  homozygotes and  $k(k-1)/2$  heterozygotes, are distinguishable in diploid species in the absence of null alleles. Allele frequencies can thus be calculated directly from allele counting in genotype data, and expected genotype frequencies can be calculated from allele frequencies with or without deviation from Hardy-Weinberg proportions. In polyploid species, in contrast, a phenotypic heterozygote may have several possible underlying genotypes that vary in the number of copies of one or more alleles. In tetraploids, for example, the 3-band phenotype ABC at a microsatellite locus can have 3 alternative genotypes ABBC, AABC and ABCC which cannot be easily and reliably distinguished experimentally. In octaploids, the same phenotype ABC can have 21 possible genotypes which are indistinguishable using current experimental approaches.

The uncertainty and polyploidy of the genotype data make most population genetics analyses developed for diploid species inapplicable to polyploids. This is unfortunate because polyploids are estimated to be 30–80% among plant species (Meyers & Levin 2006; Rieseberg & Willis 2007), and less frequent (Mable 2004) but an evolutionarily significant (Ohno 1999) factor associated with diversification among animal species. Many important crops, such as cotton, wheat, sweet potato, kiwifruit and certain strawberries, are polyploids. Polyploidy is widespread and has evolved repeatedly during the development and diversification of fishes (Leggatt & Iwama 2003), especially in primitive taxa such as sturgeons (family *Acipenseriformes*, Ludwig *et al.* 2001).

How to analyze polyploid genotype data in applications widely used in diploids is an important question. One solution is to modify each population genetics method developed for diploids to accommodate the polyploid inheritance model and the uncertainty of genotype data. While not impossible, this is a formidable task as there are many population genetics methods developed and enjoyed by researchers working with diploid species and each needs to be modified individually. Furthermore, for some sophisticated methods involving a high computational load such as the Bayesian population clustering analysis (Pritchard *et al.* 2000) and the joint likelihood sibship and parentage analysis (Wang 2004; Wang & Santure 2009), accounting for polyploid inheritance and genotype uncertainty could incur a dramatic increase in computational cost (see discussion below).

Another solution is to convert the polyploid genotypes to pseudo diploid genotypes such that many methods developed for diploids are applicable without modification. In theory, this option is inferior to the alternative one because any transformation of data would mean a loss of information and thus a reduction in analysis accuracy. However, this option is appealing because it is simple and universally applicable to many methods. Rodzen *et al.* (2004) proposed such a scheme to convert microsatellite genotypes in the polyploid white sturgeon (*Acipenser transmontanus*) to pseudo diploid dominant genotypes. Essentially the same data transformation scheme was independently applied in several earlier studies (e.g. Mengoni *et al.* 2000; Buteler *et al.* 2002). In this scheme, effectively each allele (or band) at a codominant marker locus is treated as an independent dominant “locus” with 2 alleles (dominant and recessive), 3 genotypes, and 2 phenotypes (band present and absent). The converted data can then be analysed by methods developed for diploid dominant markers for various purposes as reviewed by Bonin *et al.* (2007), such as population structure (Zhivotovsky 1999; Falush *et al.* 2007; Milot *et al.* 2008), relatedness and relationship (Lynch & Milligan 1994; Hardy 2003; Gerber *et al.* 2003; Wang 2004; Nybom 2004; Kosman & Leonard 2005), and hybridization (Anderson 2008; Sun & Lo 2011). Indeed several studies (Mengoni *et al.* 2000; Buteler *et al.* 2002; Rodzen *et al.* 2004; Hanson *et al.* 2008) have demonstrated using empirical data that accurate parentage assignments and relatedness estimates for polyploids could be obtained from the converted data using standard methods developed for diploid dominant markers.

The data transformation by the scheme of Rodzen *et al.* (2004) leads to an inherent loss of information on heterozygosity (Hanson *et al.* 2008), and possibly some reduced inferential power and accuracy when applied, for example, to parentage and population structure analyses. No systematic study has been conducted to investigate how efficient this data conversion is, and how much marker information is lost and thus how much analysis accuracy is reduced by this approach. More specifically, it is unknown whether or not the transformed data can be used to infer full and half sibship directly rather than indirectly by clustering individuals based on pairwise relatedness (as per Rodzen *et al.* 2004), and to infer parentage jointly with sibship. This study aims to fill the gaps. By applying Rodzen’s *et al.* (2004) scheme to both simulated and empirical data in polyploids, we investigate the effects of polyploid levels and marker number and polymorphism on the accuracy of parentage and sibship assignments and of selfing rate estimates. We show that sibship and parentage assignments can be substantially improved by allowing for a small genotyping error rate to

accommodate the distortion in the assumed Mendelian inheritance of the converted markers when large sibship is involved. The results are discussed in the context of polyploid data analyses used in molecular ecology.

## Methods

### *Data conversion*

Following Rodzen *et al.* (2004), we convert a polyploid individual phenotype at a  $k$ -allele codominant locus to diploid phenotypes at  $k$  dominant “loci”. Each converted locus has 2 “alleles”, one dominant (indexed by 1) and the other recessive (indexed by 0). The locus has thus 2 possible phenotypes determined by three genotypes. The dominant phenotype (presence of a band, denoted by 1) has two unordered genotypes {1,1} and {0,1}, while the recessive phenotype (absence of a band, denoted by 0) has a single genotype {0,0}. Indexing the  $k$  bands as independent loci by 1, 2, ...,  $k$ , we obtain an individual’s diploid phenotype at locus  $i$  ( $=1, 2, \dots, k$ ) as 1 and 0 when band  $i$  is present in and absent from the individual’s original polyploid phenotype, respectively. For example, an individual phenotype showing bands 2, 4 and 7 at an 8-band microsatellite locus will thus have phenotypes {0, 1, 0, 1, 0, 0, 1, 0} at 8 pseudo diploid dominant loci.

### *Simulations*

Simulated data were generated and analyzed to investigate the accuracy of sibship, parentage, and selfing rate inferred from the pseudo-diploid dominant marker phenotypes converted from polyploid phenotypes at codominant marker loci. We considered the impact of species polyploidy, dioecy and monoecy with selfing, marker information content (numbers of alleles and loci), and the actual family structure in the data. To understand how much information is lost and how much inference accuracy is thus reduced by the data conversion, we also considered diploid species and comparatively analysed the original codominant phenotypes and the converted dominant phenotypes. Such a comparative analysis of the converted and non-converted data would indicate the loss of information and accuracy due purely to data transformation rather than polyploidy. The level varies between diploids ( $2N$ ) to decaploids ( $10N$ ) for ploidy, dioecy vs monoecy, 5 to 25 for the number of alleles per locus, and 5 to 25 for the number of loci, small (family size  $\leq 8$  siblings) and large (family size = 125 siblings) full sibships and half sibships for the actual relationship structure. For monoecy, the level of selfing rate varies from low (0.05), medium (0.33) to high (0.50). These levels of marker

information cover the typical range commonly used for microsatellite loci in the literature (Blouin 2003). The dioecy and monoecy with variable selfing rates, the polyploid models and the family structures evaluated in simulations also represent a comprehensive suite of empirical possibilities in nature (Brown 1989; Otto & Whitton 2000; Leggett & Iwama 2003).

For dioecious species, the simulated data contain either full sib families only (FS model) when both sexes are monogamous, or both half and full sib families (HS model) when one or both sexes are polygamous. The FS model considered two family structures. In structure 1 (denoted by FS1), a sample contained 4 sets of families, with set  $i$  ( $i=1\sim 4$ ) consisting of  $2^{6-i}$  full sib families and each family having  $2^{i-1}$  full siblings. The offspring sample has thus 128 individuals, distributed in 32, 16, 8 and 4 families of size 1, 2, 4 and 8 siblings, respectively. In structure 2 (denoted by FS2), a sample contains  $5^{i-1}$  full sib families and each family has  $5^{4-i}$  full siblings, where  $i=1\sim 4$ . The offspring sample has thus 600 individuals, distributed in 125, 25, 5 and 1 families of size 1, 5, 25 and 125 siblings, respectively. Both samples in FS1 and FS2 contain mixed full sibships of various sizes, including numerous singletons and a very large full sib family (FS2 only). Such samples can be realistic and common in practice, and are particularly challenging for marker-based sibship reconstruction because both false sibship assignments (i.e. true non-siblings being assigned to a sibship) and false sibship exclusions (i.e. true siblings being not assigned to a sibship) are potentially common (Wang 2013). FS2 has a much larger sample size than FS1, and contains a very large sibship that might be spuriously split into 2 or more reconstructed sibships when marker information is insufficient (Wang 2013).

Many species have a polygamous mating system, and both full and half siblings can exist in a sample. Half siblings are intermediate in relatedness between non-siblings and full siblings, and are thus more challenging to identify (Blouin 2003). We considered a HS model in which a sample contains 8 half sib families. In each family, each of 3 males mates with each of 3 females. The full sib family from the mating between male  $i$  ( $=1, 2, 3$ ) and female  $j$  ( $=1, 2, 3$ ) has 1 and 2 offspring when  $i \neq j$  and  $i = j$ , respectively. Thus, each half sib family has 12 offspring, and a sample has a total number of 96 offspring.

In both FS1 and HS models, male parents were not sampled. Female parents were sampled in half of the full sib families in each set for the FS1 model, and for the first 4 of the 8 half sib families for the HS model, respectively. Sampled female parents were genotyped and included in the candidate mother sample. Additionally, 100 unrelated individuals of each

sex were included in the candidate samples. Thus the candidate father sample has 100 individuals, who are unrelated to any individual in the offspring and candidate samples. The candidate mother sample has 130 and 112 individuals, including 30 and 12 true mothers, for the FS1 and HS models respectively. No parents and no candidates were included for FS2.

For monoecious species with a mixed outcrossing and selfing mating system (Clegg 1980), we considered an offspring sample containing 6 half sib families. In each family, 3 outbred parents ( $i=1, 2, 3$ ) contributed both male and female gametes which combined to form the sampled offspring. The number of offspring from the mating between individual  $i$  and  $j$  was always 1 when  $i \neq j$  (outcrossing), and variable when  $i = j$  (selfing). For the case of a low selfing rate, each of 2 of the 18 parents in the 6 families contributed 1 selfing offspring. For the cases of medium and high selfing rates, each of the 18 parents in the 6 families contributed 1 and 2 selfing offspring, respectively. In the low, medium, and high selfing scenarios, therefore, a sample has 38, 54 and 72 offspring including 2, 18, and 36 selfing offspring, respectively, resulting in an actual selfing rate of 0.05, 0.33, and 0.50, respectively. No true parents and no candidates were included, which made the inference of selfing rate from pedigree reconstruction even more challenging.

The frequencies of the  $k$  alleles at a locus were assumed to be in a triangular distribution, with  $p_i = i/(k(k+1)/2)$  for allele  $i (=1, 2, \dots, k)$ . Given allele frequencies, the genotype of a parent at a locus is generated by sampling randomly and independently  $2N$  gene copies. Genotypes at multiple loci were generated independently. A gamete of  $N$  homologues was obtained by random sorting of the parental  $2N$  homologues, and an offspring of  $2N$  homologues was formed by combining a male gamete and a female gamete. The possible values of  $2N$  considered in the simulations were 2, 4, 6, 8, and 10.

The simulated polyploid parent and offspring phenotypes at any locus were perfect, having no genotyping errors, mutations, and missing data. These phenotypes at a  $k$ -allele codominant locus were converted to phenotypes at  $k$  diploid dominant “loci” as described above before conducting relationship analysis. The data were analysed for relationship assuming the absence of genotyping errors, except for the case of FS2 that involves very large full sib families. For FS2, although no genotyping errors were introduced in simulating the data, a variable mistyping rate (0-0.16) was used for each converted locus when analyzing the data. The allowance of (false) genotyping errors was used to accommodate expected Mendelian segregation distortions created by the data conversion (more details below).

There are so many possible parameter combinations involving family structures, ploidy levels, selfing rates, number of loci and number of alleles, that it is impossible to explore all of them even in a simulation study. We chose to investigate the impact of each factor individually, fixing all the other factors. Except when explicitly stated, the standard levels of the factors were octaploids ( $8N$ ), dioecious species, 10 microsatellite loci, 10 alleles per locus, and the FS1 family structure.

For each parameter combination, 100 datasets were simulated and analysed by the full likelihood (FL) and pairwise likelihood (PL) methods implemented in the computer program Colony version 2.0.4.4 (Jones & Wang 2002). The FL method assigns all sampled individuals to candidate relationships (parent-offspring, full siblings, half siblings, unrelated) jointly. It uses a simulated annealing algorithm to construct relationship configurations and search for the one that has the maximum likelihood (Wang 2004; Wang & Santure 2009). The PL method assigns sibship or parentage to a pair or trio of individuals in isolation (e.g. McPeck & Son 2000; Marshall *et al.* 1998), with or without controlling for false assignments. Most of the default parameter settings (e.g. diploid, no inbreeding, a single run of medium length, medium likelihood precision, no sibship prior, no update of allele frequencies) in Colony were accepted in analysing the data. Specifically, the probabilities of a male and female parent included in the candidates were set as 0.1 and 0.5 respectively, both sexes were set as monogamous and polygamous for the FS and HS models respectively, and dioecious and monoecious models were adopted for such simulated data. Except when FL and PL are compared in accuracy, analysis results are presented for the FL method only.

#### *Accuracy measurements*

The accuracy of sibship and parentage assignments was measured by the statistic,  $P(a/b)$ , the frequency of dyads assigned relationship  $a$  when their actual relationship is  $b$  (Thomas & Hill 2000; Wang 2004). Thus,  $P(a/b)$  gives the frequency of correct and incorrect inference of dyadic relationship  $b$  when  $b=a$  and  $b\neq a$ , respectively. For sibship inference among offspring, accuracy is measured by  $P(\text{FS}|\text{FS})$ ,  $P(\text{HS}|\text{HS})$ , and  $P(\text{UR}|\text{UR})$ , where FS, HS, and UR are full sibs, half sibs, and unrelated offspring, respectively. For parentage inference, accuracy is measured by the frequencies that parentage is correctly assigned,  $P(\text{PO}|\text{PO})$ , or correctly excluded (unassigned),  $P(\text{XO}|\text{XO})$ , when the actual parent is included in and excluded from the candidate pool, respectively. The values of  $P(a/b)$  were calculated for each replicate, and averaged across replicates in the report.

The accuracy of selfing rate estimates was also assessed for monoecious species. The FL method infers the parents of each offspring, who is thus inferred to come from selfing and outcrossing reproduction when its two parents are identical and different, respectively. The proportion of the inferred selfing offspring in a sample acts as an estimate of selfing rate,  $\hat{s}$ , of the population from which the sample is drawn representatively or at random. We calculated the mean and root mean square error (RMSE) of  $\hat{s}$  across  $n=100$  replicates as

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n \hat{s}_i,$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{s}_i - s)^2},$$

where  $s$  is the true simulated selfing rate. A comparison between  $\bar{s}$  and  $s$  shows whether  $\hat{s}$  is biased or not, and the direction and extent of bias. RMSE measures the overall accuracy of  $\hat{s}$ , including both sampling errors and bias.

### *Empirical data analyses*

Simulations are valuable in evaluating the accuracy, robustness, computation efficiency and statistical behaviour of an estimator, because many different population and sampling scenarios can be considered easily and the true (simulated) parameter values are known. However, the assumptions of a simulation model may be unrealistic and thus the simulated data may deviate from the reality, which is usually much more complicated. How well an estimator fares with a real dataset is of more interest to empiricists, and is the ultimate question to be addressed before the estimator is recommended for application in practice.

We applied the data transformation scheme to a lake sturgeon dataset before conducting a sibship analysis. Lake sturgeon (*Acipenser fulvescens*) was reported to have an octoploid (Blackledge & Bidwell 1993) or tetraploid (Fontana *et al.* 2004) genome, but microsatellites in the species can be either disomically or non-disomically inherited (e.g. Pyatskowitz *et al.* 2001; Welsh & May 2006). Larvae from 7 full-sib families of a lake sturgeon population were sampled and genotyped at 10 non-disomically inherited microsatellite loci (details in online document). Each of 5 families has 8 offspring, and each of 2 families has 7 offspring included in the sample, resulting in a total sample size of 54. The numbers of alleles at the 10 loci observed in the 54 individuals were 3, 6, 8, 9, 14, 14, 15, 15, 20, and 21. Transforming the phenotypes to pseudo diploid dominant marker data, we obtain individual phenotypes at 125 loci, which were analysed in two ways.

First, the phenotypes at each locus were permuted among individuals to obtain 54 new multilocus phenotypes. These 54 artificial individuals are obviously unrelated (UR), but were analysed by program Colony to check whether false sibships were inferred. A variable number of the original 125 loci were used in the permutation simulations to investigate how false sibship classification rates varied with the number of markers. For a given number of loci  $L$ , a permuted dataset was generated by first selecting at random  $L$  out of the 125 loci, and then permute the original phenotypes of the 54 individuals at each selected locus. The permuted 54 multilocus phenotypes at the selected  $L$  loci were then analysed for sibship. For each  $L$ , 100 permuted datasets were generated and analysed to obtain an average value of  $P(\text{UR}|\text{UR})$ .

Second, the original transformed data were analysed by using a variable number of loci,  $L$ , to see how much marker information is needed to fully recover the actual sibship structure. For each  $L$  ( $\leq 125$ ), 100 bootstrapping samples were formed and analysed. Each sample was generated by drawing at random  $L$  loci, and the individual phenotypes of these selected  $L$  loci were used in sibship reconstruction. Accuracy was measured by average  $P(\text{FS}|\text{FS})$  and  $P(\text{UR}|\text{UR})$  values across 100 replicate samples.

## **Results**

### *Effects of number of loci*

The accuracy for both full sibship and parentage assignments increased rapidly with an increasing number of loci (Figure 1). Five microsatellites, each having 10 alleles, was insufficient for accurate relationship inference, but at least 90% of all of the possible dyadic relationships including PO, XO, FS, UR were correctly identified with 10 loci.

It is revealing to compare the parentage assignment and exclusion accuracies for offspring with different numbers of siblings included in a sample, and between the full (FL) and pairwise (PL) likelihood methods (Figure 2). As expected, FL method correctly assigns and excludes, when true parents are included in and excluded from candidates respectively, parentage at an increasing frequency for offspring in sibships with an increasing size. This is because more siblings, when considered jointly, provide more information about their common parentage and thus allow for more accurate parentage assignments or exclusions. In contrast, the PL method yields essentially the same parentage inference accuracy regardless of sibship sizes, because it considers each offspring in isolation no matter how many siblings

are present in the sample. For singletons, FL and PL have similar parentage assignment accuracies,  $P(\text{PO}|\text{PO})$ . For offspring with one or more siblings, FL yields much more accurate parentage assignments using any number of loci. FL gives an increasing frequency of correct parentage exclusion,  $P(\text{XO}|\text{XO})$ , with an increasing number of loci, regardless of sibship size. PL shows an interesting property that  $P(\text{XO}|\text{XO})$  is high when the number of loci is either low ( $L=5$ ) or high ( $L=25$ ), and is low when the number of loci is intermediate ( $L=10\sim 15$ ). This is because PL, designed to assign parentage with confidence, tends to be conservative for parentage assignments and liberal for parentage exclusion, especially when marker information is scarce. Even when  $L=25$ , however, the PL method still has a  $P(\text{XO}|\text{XO})$  value of around 0.9, much lower than the FL method. This means that PL tends to yield a substantial proportion of false parentage assignments even when a large number of loci are used in polyploid species.

#### *Effects of number of alleles*

The inference accuracy for all types of relationship increases rapidly with an increasing number ( $k$ ) of alleles per locus (Figure 3). When marker information is scarce ( $k=5$ ), most parents are unassigned although they are included in the candidates, resulting in a low value of  $P(\text{PO}|\text{PO}) = 0.22$ . This value jumps to 0.91 when each locus has 10 alleles. Similar to the results shown in Figure 2, the pairwise likelihood approach to parentage assignments proposed by Marshall *et al.* (1998), implemented in Colony (Wang 2004) and FAMOZ (Gerber *et al.* 2003) programs for both dominant and codominant markers, yields lower  $P(\text{PO}|\text{PO})$  values than the full likelihood approach that considers joint sibship and parentage assignments among all sampled individuals, and gives a value of 0.01 and 0.73 for  $k=5$  and 10, respectively.

#### *Effects of polyploid levels*

For diploids, transforming codominant marker phenotypes to dominant ones incurs a sizable decrease in parentage assignment and exclusion accuracies (Figure 4), but has a much smaller effect on sibship assignments for the FS1 model. Polyploid levels also have a much larger impact on parentage than sibship analyses. True parents included in the candidates become unassigned at an increasing frequency and sibship inference deteriorates with an increasing polyploid level. This is unsurprising given that the original polyploid phenotypes are increasingly uncertain in their underlying genotypes, and the transformed phenotypes display more and more distorted Mendelian proportions (see below), with increasing levels of ploidy.

To achieve the same sibship and parentage inference accuracy, more marker information is needed for species with a higher level of ploidy.

### *Effects of large full sib families*

The transformation of polyploid codominant phenotypes to diploid dominant phenotypes could cause an apparent distortion of Mendelian segregation, and thus may lead to the spurious split of large full sib families in the likelihood sibship reconstruction. The higher the polyploid levels and the larger the true sibship size, the more severe will be the distortion and the more likely will be the sibship splitting. Using hexaploid species as an example, parents of the same genotype AAABBB will generate offspring free of allele A at a probability of 1/400 in the absence of double reductions. The frequencies of converted phenotypes 1 (band presence) and 0 (band absence) of the offspring at the transformed locus corresponding to allele A are therefore 399/400 and 1/400, respectively. In contrast, heterozygous parents at a true diploid dominant locus will produce 1 and 0 phenotype offspring with a probability of 3/4 and 1/4, respectively. This means, when the true sibship with heterozygous polyploid parents is large, the frequency of the common phenotype can be too high and the frequency of the rare phenotype can be too low compared with the expectations of Mendelian segregation laws for diploids. As a result, this phenotype configuration will have a very low probability calculated according to Mendelian inheritance on diploids. Splitting the sibship into two full sib families, each containing individuals with identical phenotypes at this A locus, is highly likely to result in a higher likelihood. As a result, a large sibship could be split into 2 or more full sib families in reconstruction. However, when a small mistyping rate is permitted at each locus, then the rare phenotype will be considered to be due to genotyping errors, and the large sibship will not be incorrectly split in reconstruction.

Simulations verify our reasoning above. When the transformed data were regarded as perfect diploid data without mistypings (i.e.  $e = 0$ ), large sibships were split by the full likelihood method, resulting in a low  $P(\text{FS}|\text{FS})$  value of 0.55 (Figure 5). Sibship reconstruction improves rapidly by allowing for a small mistyping rate, and the actual sibship structure was almost completely recovered when  $e = 0.04$ . Further increase in  $e$  incurs a slight decrease in  $P(\text{UR}|\text{UR})$ , meaning some unrelated individuals were mis-assigned as siblings.

For comparison, Figure 5 also shows the accuracy of the pairwise likelihood (PL) method, which considers the likelihood of full-sib and unrelated relationships for each pair of individuals in isolation of other individuals. With an increasing value of  $e$ ,  $P(\text{FS}|\text{FS})$

increases but  $P(\text{UR}|\text{UR})$  decreases rapidly. There seems to be no benefit by introducing a small mistyping rate in the PL sibship analysis. This is in sharp contrast with the FL method, and is understandable because the apparent Mendelian segregation distortion due to data transformation becomes severe only when sibship size is large. The distortion has trivial effects on the PL method because it considers just 2 individuals each time. Allowing for a small error rate in analysis may well reduce the overall accuracy of the PL method because most often a sample contains more non-sib than sib dyads.

#### *Effects of half siblings in polygamous mating systems*

Half sibship can also be inferred from the transformed data for polyploids (Figure 6). Comparing analyses of diploids with and without data transformation, however, it becomes clear that the transformation incurs a substantially reduced accuracy in both full and half sibship inferences, while it has little effect on parentage inference accuracy. This is in contrast to the results in Figure 4, where parentage assignments suffer much more than full sibship assignments. This discrepancy comes from the difference in the simulated true sibship structures. In Figure 4, a simulated sample contains numerous singleton offspring who have no siblings, and who have or have no parents included in the candidates. Parentage assignment or exclusion for a singleton is difficult, because the information comes from just this single offspring and a candidate parent. In such a case, the full likelihood method for sibship and parentage assignments is similar to the pairwise likelihood approach and has no extra information to use. In Figure 6, each offspring has 3 paternal siblings and 3 maternal siblings, and therefore its parentage assignment or exclusion, when the true parent is included in or excluded from the candidates respectively, is made much easier as the siblings also provide parentage information. Therefore, even if marker information is scarce or reduced by the transformation, parentage can still be reliably inferred.

Similar to the FS1 model shown in Figure 4, an increasing ploidy level leads to a decreasing accuracy in full and half sibship inference. Parentage inferences are little affected by polyploidy, because as explained above little marker information is needed for the full likelihood method to infer parentage when siblings are present. For accurate half and full sibship inference, more markers are needed for species with a higher ploidy level.

#### *Accuracy of selfing rate estimates in monoecious species*

Selfing rates are estimated with a decreasing bias and an increasing accuracy for both diploids and octaploids with an increasing number of loci ( $L$ ), and almost perfect (no bias and very low RMSE) estimates are obtained when  $L \geq 15$  (Figure 7). The estimates obtained with  $L=10$  are also satisfactorily accurate, no matter the actual selfing rate is high, medium or low. When marker information is scarce ( $L \sim 5$ ), however, selfing rate is underestimated and overestimated when it is high and low respectively, and RMSE is always high. Overall, ploidy levels have little effect on the quality of selfing rate estimates.

#### *Empirical lake sturgeon data analyses*

Almost perfect sibship reconstruction was achieved using just 80 of the 125 transformed loci (Figure 8), which means about 4 microsatellites are needed to recover the sibships completely. This high assignment power is not surprising given the uniformly large full sib families (7-8 siblings per family) of the sample, and the fact that the full likelihood method is particularly powerful for such a case. Permuting genotypes among individuals at each locus generates a dataset containing unrelated individuals (singletons). Analyses of these permuted datasets show that the rate of false sibship discovery is low, only about 1% when all 125 transformed loci are used (Figure 9). Combining the results in Figures 8-9 suggests that 10 or more microsatellites are required to accurately infer full sibships if the sample is large and contains many singletons, but only ~4 microsatellites are needed if full sibships in a sample are uniformly large. When full sibships are very large (>100 siblings per family), more than 4 microsatellites would be necessary to prevent the splitting of large sibship in reconstruction.

#### **Discussion**

In this study, simulated and empirical data were analysed to investigate the accuracy of sibship, parentage and selfing rate inference from the pseudo diploid dominant marker data that were transformed from disomically or non-disomically inherited codominant marker data in diploids and polyploids. In reality, most polyploid species may fall somewhere on a continuum between the two extreme inheritance modes. Autopolyploids result from genome duplication events and should in principle follow the non-disomical inheritance. However, accrual of mutational differences in microsatellite motif size and at priming sites (Estoup & Cornuet 1999) can lead to loss of amplification products in different gene copies, and to an apparently disomical inheritance of markers. Allopolyploids are due to hybridization events and thus should in principle follow the disomical inheritance. However, the hybridizations are typically between closely related species with similar genomes, and thus allopolyploids may

still display non-disomical inheritance to some extent. In many vertebrates there is evidence for partial diploidiation (Wolfe 2001). Our results show that in both inheritance modes, the marker data transformation proposed by Rodzen *et al.* (2004) works well for both sibship and parentage assignments. With a typical suit of 10-20 microsatellites, sibship and parentage can be accurately recovered for polyploids using the full likelihood methods developed for diploids. Therefore, polyploid marker data can always be transformed by the scheme of Rodzen *et al.* (2004) before conducting a relationship analysis, when the marker inheritance mode is non-disomical, unknown, uncertain or partially disomical. Polyploid marker data in disomical inheritance can be analysed similarly, but are better analyzed without applying the transformation because of the loss of information.

The accurate sibship and parentage inferences from the transformed data shown in the present simulation study do not mean the transformation scheme (Rodzen *et al.* 2004) is perfect and works for other analyses. This data transformation is not biologically accurate in several aspects, and at least it leads to some information loss. Different bands at a codominant marker locus are obviously dependent in frequencies which sum to 1. Converting these bands to different and independent “loci” not only violates the dependence, but also changes “allele” and “genotype” frequencies dramatically. These frequencies are inevitably increased on average by the conversion. Consequently, some analyses such as estimating inbreeding ( $F_{IS}$ ) from genotype frequencies are severely affected and become invalid (see below). Another consequence is that the transformed pseudo diploid dominant genotypes do not follow the exact Mendelian inheritance of either diploids or polyploids. Fortunately, some analyses, such as relatedness and relationship as demonstrated by the simulations, are little affected by this data conversion, because they rely mainly on the genotypic similarity between individuals which largely survives the data conversion. The data conversion leads to some distortion in Mendelian segregation proportions, which becomes severe and more frequent in large full sib families. However, we demonstrate that this apparent distortion can be affectively accounted for by introducing a small mistyping rate in data analysis.

Similar to the case of diploid species (Wang & Santure 2009), the power of a sibship/parentage analysis in polyploid species depends on many interacting factors, as shown in this simulation study. First, marker informativeness, determined by the number, polymorphism as well as the missing and mistyping rates of the markers, is the primary factor that affects the inference accuracy of all types of relationships. Usually less than 5 microsatellites or 50 SNPs are insufficient for relationship inference, except for a sample

consisting of uniformly large full sibships. Second, among different close relatives, parentage is the easiest to infer, followed by the inferences of full sibship and half sibship relationships. This is because the probabilities of sharing 2, 1, 0 alleles that are identical by descent (IBD) at a diploid locus are 0, 1 and 0 for a parent-offspring (PO) dyad, 0.25, 0.5, and 0.25 for a full sib (FS) dyad, and 0, 0.5, 0.5 for a half sib (HS) dyad. PO dyads have a high coancestry coefficient (0.5) and no variation in IBD sharing patterns, which make them much easier to infer. HS dyads have a lower coancestry coefficient (0.25) and higher variation in IBD sharing patterns, which make them much more difficult to distinguish from other dyads such as unrelated and FS dyads. Our simulations (Figures 4 and 6) showed that allowing for polygamy to infer half sibship reduces the inferential power substantially, and more markers would be needed to reach the same power as that for monogamy. Third, family size is also critically important in determining the power, as shown in Figure 2. More offspring sharing the same parent (half siblings) or parents (full siblings) provide more information about their common parentage (parental genotypes), and allow better inference of both sibship and parentage. Family size has, however, little effect on the accuracy of pairwise approaches, which infer the relationship of a pair of individuals in isolation. Unfortunately, in practice the family types (full or half sibship), structures, and sizes in a dataset are often unknown and are indeed the particular subjects to infer from marker data.

As a result of accurate sibship and parentage assignments, selfing rate in monoecious polyploid species is also accurately estimated from the transformed marker data. There are several marker based methods available for estimating selfing rates in diploid species, and the pedigree reconstruction (PR) method performs the best when marker information is sufficient (Wang *et al.* 2012). The simulations in the present study further confirm that this method works well also for polyploid species using transformed data. The inbreeding ( $F_{IS}$ ) method estimates  $s$  based on the excess of homozygosity caused by selfing. When a large diploid population reproduces at a constant selfing rate  $s$  for a sufficient number of generations, it will reach an inbreeding equilibrium at which the average inbreeding is  $F_{IS} = s/(2 - s)$  (Hedrick & Cockerham 1986). Therefore, one can infer  $s$  by estimating  $F_{IS}$  from marker data and solving for  $s$  (e.g. Clegg 1980; Penteado *et al.* 1996). However, this inbreeding method does not apply to dominant marker data, because allele frequencies at a dominant marker locus have to be estimated assuming Hardy-Weinberg equilibrium (i.e. the absence of inbreeding). With such estimated allele frequencies,  $F_{IS}$  is expected to be zero, irrespective of the mating system or actual selfing rate. The identity disequilibrium (ID) method (David *et al.*

2007) also relies on diploid codominant markers to work, and does not apply to dominant marker data. The Bayesian population assignment (PA) method that allows for inbreeding could potentially use dominant marker data to infer selfing rates. However, the current implementation of this method in the program InStruct (Gao *et al.* 2007) does not have this capability. Currently, there are no alternative methods for estimating  $s$  from dominant markers with which our pedigree reconstruction method can be compared.

Our simulated data of monoecious species with a mixed mating system were generated without assuming inbreeding equilibrium (IE). Unlike some other methods (e.g.  $F_{IS}$ , ID) which require IE, our PR method and the PA method of Gao *et al.* (2007) do not rely on IE and apply to populations with or without IE (Wang *et al.* 2012). The simulations considered a single population practicing mixed selfing and outcrossing reproduction without biparental inbreeding. However, previous simulations showed that both PR and PA methods were robust to the presence of biparental inbreeding (Wang *et al.* 2012), because an offspring from a single parent (i.e. selfing) is distinguishable from an offspring from two parents (i.e. outcrossing) even if they are highly related (e.g. full siblings). In contrast, both  $F_{IS}$  and ID methods are sensitive to the presence of biparental inbreeding, which could lead to an overestimate of selfing. Similarly, the presence of multiple populations should have little effect on the PR method, except when the populations are highly differentiated. The PR method also makes no assumption about the relatedness among pollen received by the same individual maternal parent (Schoen & Clegg 1984). The pollen may come from different related or unrelated individuals (as in wind-pollinated plants), or from the same individual (as in certain insect-pollinated plants). The higher accuracy and robustness and fewer assumptions of PR method in comparison with other methods makes it a good choice in studying plant and animal mating systems (Wang *et al.* 2012). However, this method is computationally much more demanding, especially for a large sample of individuals. It also requires more marker information to obtain unbiased and accurate inferences.

The comparison between analyses of untransformed and transformed data in diploids verifies some loss of information and thus some reduction in sibship and parentage inference accuracy (Figures 4 and 6). This is understandable because diploid individuals have unambiguous genotypes at codominant loci that would allow both parentage and sibship exclusions. This exclusion power is lost for sibship and much reduced for parentage inferences once the data are transformed to dominant loci. For polyploid species, the transformation is likely to incur less information loss because the original untransformed

genotypes may be uncertain anyway. A proper evaluation of the extent of information loss and thus accuracy reduction by the data transformation for polyploids as shown for diploids in Figures 4 and 6 requires the development and implementation of a polyploid likelihood model for relationship inference from codominant markers (see below).

The data transformation causes an apparent distortion of Mendelian segregation, where the dominant and recessive phenotypes of the offspring of heterozygous parents do not appear in the expected proportions of  $\frac{3}{4}$  and  $\frac{1}{4}$  for diploids at a dominant locus. The higher the ploidy level is, the lower will be the probability of the recessive phenotype. This problem becomes severe when sibship size is large. For the pairwise likelihood method, this problem is irrelevant, because no matter how large the sibship is, the method deals with each pair of individuals in isolation. For the full likelihood method that considers the entire sample of individuals jointly for sibship assignments, however, this problem may lead to the spurious splitting of a large sibship. We show in simulations (Figure 5) that this problem can be satisfactorily resolved by allowing for a small mistyping rate in data analysis. In a large full sib family with a very low but non-zero proportion of the recessive phenotype, the rare recessive phenotype would be treated effectively as due to mistyping when it is permitted in analysis, and the sibship will not be spuriously split in reconstruction.

In polyploidy species, allele dosage in microsatellite profiles could be useful to obtain genotypes (rather than phenotypes, each of which may contain several alternative genotypes). The peak height for each allele could be measured and used to infer the number of allele copies at a locus in each individual (after calibration). The uncertainty of the allele copy number can be accounted for by using a suitable genotyping error model similar to the stepwise mutation model. To analyse such polyploid genotype data, a likelihood method specifically designed for joint sibship and parentage assignments needs to be developed and implemented. In principle, such a method should be very similar to that developed for diploids (Wang 2004; Wang & Santure 2009), only the likelihood function needs to be modified. The likelihood function for a pure full sibship with unknown parental genotypes at a single  $2N$ -ploidy locus of  $k$  alleles is

$$L(\text{FS}|g, m) = \Pr(g, m|\text{FS}) = \sum_{x1=1}^k p_{x1} \sum_{x2=1}^k p_{x2} \cdots \sum_{xa=1}^k p_{xa} \sum_{y1=1}^k p_{y1} \sum_{y2=1}^k p_{y2} \cdots \sum_{ya=1}^k p_{ya} \prod_{l=1}^d \Pr(g_l|x1, x2, \dots, xa; y1, y2, \dots, ya)^{m_l}$$

where  $a=2N$ , data  $g = (g_1, g_2, \dots, g_d)$  and  $m = (m_1, m_2, \dots, m_d)$  are the  $d$  distinctive polyploid phenotypes and their counts respectively observed among the inferred siblings,  $x_i$  and  $y_j$  index the alleles in the two parents (where  $i, j=1 \sim a$ ). The probability of an offspring phenotype  $g_l$  given parental alleles  $x_i$  and  $y_j$  ( $i, j=1 \sim a$ ),  $\Pr(g_l | x_1, x_2, \dots, x_a; y_1, y_2, \dots, y_a)$ , can be derived assuming independent sorting of the  $a$  homologous genes in gamete formation, with the possibility of accommodating genotyping errors (Wang 2004). For the more complicated cases of half sibship and assigned parents with genotype data, similar equations to those for diploids can also be derived. However, the computation of these likelihood functions is highly computationally costly for polyploids, and quickly becomes prohibitive with increasing ploidy (value of  $a$ ). The number of parental genotype combinations that must be considered in calculating the likelihood function is in the order of  $k^{2a}$ . For a locus with  $k=10$  alleles, the numbers of parental genotype combinations are  $10^4$  and  $10^{16}$  for diploid and octoploid species, respectively. Although the number of parental genotype combinations can be reduced substantially by merging unobserved alleles in a sibship when mistypings are absent or comply with certain simple models (Wang 2004), it still increases very rapidly with  $a$  and quickly becomes too large to deal with in likelihood computation. Furthermore, marker inheritance in many polyploid species is unknown, but may fall into the continuum between disomical and non-disomical (Hanson *et al.* 2008). It is therefore of dubious value to develop such a highly computationally demanding method for parentage and sibship inference in polyploids given the uncertain and variable inheritances of markers.

In conclusion, our simulation and empirical data analyses show that accurate sibship and parentage assignments as well as good selfing rate estimates can be made based on a typical set of microsatellites (e.g. 10 markers, each with 10 alleles) using the full likelihood method and the data transformation scheme of Rodzen *et al.* (2004) for polyploid species. The transformation does cause some loss of information, and distortion of Mendelian segregation proportions for diploids. The latter may cause the splitting of large sibship in reconstruction, but can be effectively accommodated by allowing for a small mistyping rate in data analysis.

In this study we have focused on sibship and parentage assignments and evaluated the effect of various factors on the assignment accuracy. It is worth pointing out that in practice sibship and parentage assignments are not the final goal of many practical applications, but are used for further analyses such as estimating the effective number of breeders and the current effective population size (Wang 2009; Wang *et al.* 2010), male and female mating

system and reproductive skews (e.g. Gottelli *et al.* 2007), genetic parameters of quantitative traits (e.g. Thomas 2005), and migration rate (e.g. Saenz-Agudelo *et al.* 2009). In all these downstream analyses, accurate sibship and parentage assignments and exclusions are critically important in determining the biasness, precision and power. In effective population size ( $N_e$ ) inference, for example, false sibship assignments and false sibship exclusions will cause an overly high and overly low frequency of siblings and thus an underestimation and overestimation of  $N_e$ , respectively (Wang 2009).

## Acknowledgements

Funding for this project was provided by the Great Lakes Fishery Trust, Michigan Department of Natural Resources through the Partnership for Ecosystem Research and Management (PERM) program and B.C. Hydo. Jeannette Kanefsky conducted the genotyping for the empirical lake sturgeon dataset.

## Literature

- Anderson EC. 2008. Bayesian inference of species hybrids using multilocus dominant genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**: 2841-2850.
- Blackledge KH, Bidwell CA. 1993. Three ploidy levels indicated by genome quantification in *Acipenseriformes* of North America. *J. Hered.* **84**: 427–430.
- Blouin MS. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution* **18**: 503-511.
- Bonin A, Ehrlich, D, Manel S. 2007. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology* **16**: 3737-3758.
- Brown AHD. 1989. Genetic characterization of plant mating systems. Pp 145-162. In: *Plant population genetics, breeding and genetic resources*. Brown AHD, Clegg MT, Kahler AL, Weir BS (eds). Sinauer Associates, Sunderland, MA.
- Buteler MI, Bonte DRL, Jarret RL, Macchiavelli RE. 2002. Microsatellite-based paternity analysis in polyploidy sweet potato. *J Am Soc Hort Sci* **127**: 392–396.
- Clegg MT. 1980. Measuring plant mating systems. *BioScience* **30**: 814-818.

- David P, Pujol B, Viard F, Castella V, Goudet J. 2007. Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology* **16**: 2474-2487.
- Estoup A, Cornuet J-M. 1999. Microsatellite evolution: inferences from population data. Pp 49-65. In Goldstein DB and Schlotterer (eds). *Microsatellites: Evolution and Applications*. Oxford University Press, Oxford, UK
- Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* **7**: 574-578.
- Fontana F, Lanfredi M, Chicca M, Beltrami N, Congiu L. 2004. Karyotype characterization of the lake sturgeon, *Acipenser fulvescens* (Rafinesque 1817) by chromosome banding and fluorescent in situ hybridization. *Genome* **47**, 742-746.
- Gao H, Williamson S, Bustamante CD. 2007. An MCMC approach for joint inference of population structure and inbreeding rates from multi-locus genotype data. *Genetics* **176**: 1635-1651.
- Gerber S, Chabrier P, Kremer A. 2003. FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes* **3**: 479-481.
- Goodnight KF, Queller DC. 1999. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology* **8**: 1231-1234.
- Gottelli D, Wang J, Bashir S, Durant SM. 2007. Genetic analysis reveals promiscuity among female cheetahs. *Proceedings of the Royal Society B: Biological Sciences* **274**: 1993-2001.
- Guillot G, Leblois RL, Coulon AL, Frantz AC. 2009. Statistical methods in spatial genetics. *Molecular Ecology* **18**: 4734-4756
- Hanson TR, Brunsfeld SJ, Finegan B, Waits LP. 2008. Pollen dispersal and genetic structure of the tropical tree *Dipteryx panamensis* in a fragmented Costa Rican landscape. *Molecular Ecology* **17**: 2060-2073.
- Hardy OJ. 2003. Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Molecular Ecology* **12**: 1577-1588.
- Hedrick PW, Cockerham CC. 1986. Partial inbreeding: equilibrium heterozygosity and the heterozygosity paradox. *Evolution* **40**: 856-861.
- Jones, O.R., Wang, J. 2002. COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources* **10**: 551-555.

- Kosman E, Leonard KJ. 2005. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology* **14**: 415-424.
- Leggatt RA, Iwama GK. 2003. Occurrence of polyploidy in the fishes. *Reviews in Fish Biology and Fisheries* **13**: 237-246.
- Ludwig A, Belfiore NM, Pitra C, Svirsky V, Jenneckens I. 2001. Genome duplication events and functional reduction of ploidy levels in sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*). *Genetics* **158**: 1203-1215.
- Lynch M, Milligan BG. 1994. Analysis of population genetic structure with RAPD markers. *Molecular Ecology* **3**: 91-99.
- Lynch M, Ritland K. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753-1766.
- Mable BK. 2004. Why polyploidy is rarer in animals than in plants: myths and mechanisms. *Biol. J. Linn. Soc.* **82**: 453-466.
- Marshall TC, Slate J, Kruuk LEB, Pemberton JM. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* **7**: 639-655.
- McPeck MS, Sun L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* **66**: 1076-1094.
- Mengoni A, Gori A, Bazzicalupo M. 2000. Use of RAPD and microsatellite (SSR) variation to assess genetic relationships among populations of tetraploid alfalfa, *Medicago sativa*. *Plant Breeding* **119**: 311-317.
- Meyers LA, Levin DA. 2006. On the abundance of polyploids in flowering plants. *Evolution* **60**: 1198-206.
- Milot E, Weimerskirch H, Bernatchez L. 2008. The seabird paradox: dispersal, genetic structure and population dynamics in a highly mobile, but philopatric albatross species. *Molecular Ecology* **17**: 1658-1673.
- Nybom H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology* **13**: 1143-1155.
- Ohno S. 1999. Gene duplication of vertebrate genomes circa 1970-1999. *Seminars in Cell and Developmental Biology* **10**: 517-522.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401-437.
- Penteado MI, Garcia P, Pérez de la Vega M. 1996. Genetic variability and mating system in three species of the genus *Centrosema*. *Journal of Heredity* **87**: 124-130.

- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- Pyatskowitz JD, Krueger CC, Kincaid HL, May B. 2001. Inheritance of microsatellite loci in the polyploid lake sturgeon (*Acipenser fulvescens*). *Genome* **44**: 185-191.
- Queller DC, Goodnight KF. 1989. Estimating relatedness using molecular markers. *Evolution* **43**: 258-275.
- Rieseberg LH, Willis JH. 2007. Plant speciation. *Science* **317**: 910–914.
- Rodzen JA, Famula TR, May B. 2004. Estimation of parentage and relatedness in the polyploid white sturgeon (*Acipenser transmontanus*) using a dominant marker approach for duplicated microsatellite loci. *Aquaculture* **232**: 165-182.
- Saenz-Agudelo P, Jones GP, Thorrold SR, Planes S. 2009. Estimating connectivity in marine populations: an empirical evaluation of assignment tests and parentage analysis under different gene flow scenarios. *Molecular Ecology* **18**: 1765–1776.
- Schoen DJ, Clegg MT. 1998. Estimation of mating system parameters when outcrossing events are correlated. *Proc. Natl. Acad. Sci.* **81**: 5258-5262.
- Sun M, Lo EYY. 2011. Genomic markers reveal introgressive hybridization in the Indo-West pacific mangroves: A Case Study. *PLoS ONE* **6**: e19671.
- Thomas SC. 2005. The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**: 1457-1467.
- Thomas SC, Hill WG. 2000. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**: 1961-1972.
- Vekemans X, Hardy OJ. 2004. New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**: 921–935.
- Wang J. 2002. An estimator for pairwise relatedness using molecular markers. *Genetics* **160**: 1203-1215.
- Wang J. 2004. Sibship reconstruction from genetic data with typing errors. *Genetics* **166**: 1963-1979.
- Wang J. 2009. A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology* **18**: 2148-2164.
- Wang J, Brekke P, Huchard E, Knapp LA, Cowlishaw G. 2010. Estimation of parameters of inbreeding and genetic drift in populations with overlapping generations. *Evolution* **64**: 1704-1718.
- Wang J, Santure AW. 2009. Parentage and sibship inference from multilocus genotype data

- under polygamy. *Genetics* **181**: 1579-1594.
- Wang J. 2013. An improvement on the maximum likelihood reconstruction of pedigrees from marker data. *Heredity* (in press).
- Wang J, El-Kassaby YA, Ritland K. 2012 Estimating selfing rates from reconstructed pedigrees using multilocus genotype data. *Molecular Ecology* **21**: 100-116.
- Waples R, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* 15: 1419–1439.
- Welsh A, May B. 2006. Development and standardization of disomic microsatellite markers for lake sturgeon genetic studies. *Journal of Applied Ichthyology* **22**, 337–344.
- Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2: 333-341.
- Zhivotovsky LA. 1999. Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* 8: 907-913.
- 

Both authors conceived the project. K.T.S designed and conducted the sturgeon experiment and collected the microsatellite data. J.W. conducted the simulation study and analysed the simulated and sturgeon datasets. Both authors wrote the paper.

---

### **Data Accessibility**

A file of multilocus genotypes of 54 sturgeon individuals from 7 full-sib families at 10 microsatellite loci has been deposited at Dryad: [doi:10.5061/dryad.9006](https://doi.org/10.5061/dryad.9006).

### **Supporting Information**

Additional Supporting Information may be found in the online version of this article: Supplemental materials on empirical data collection and laboratory genotyping.



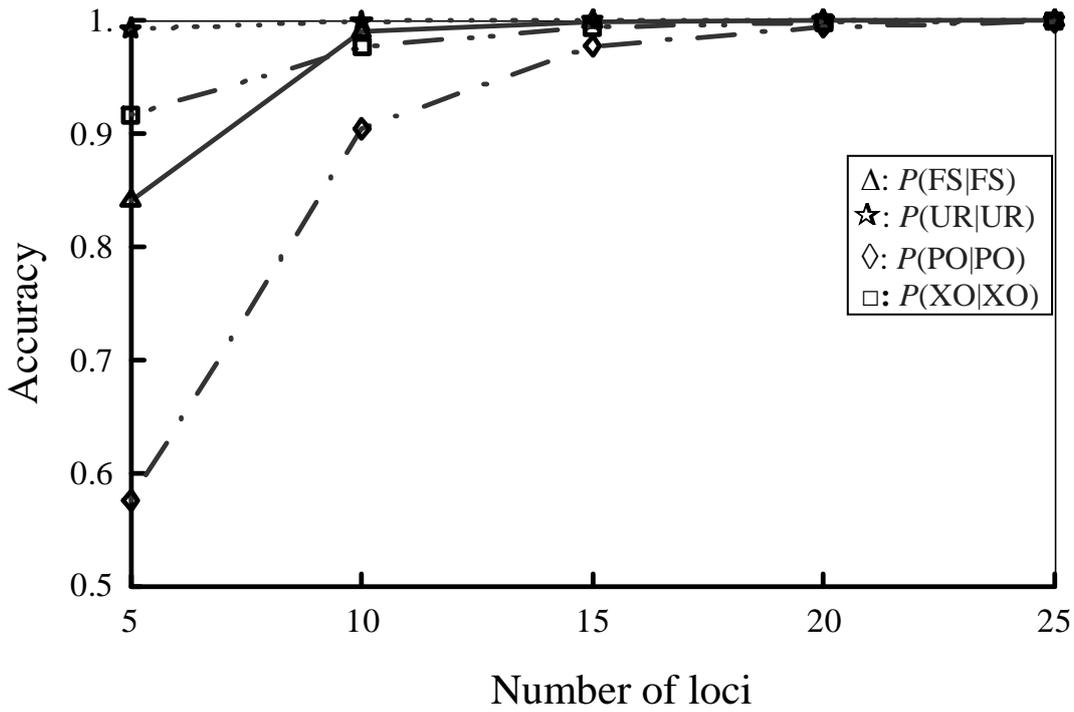


Figure 1: Accuracy of full-sib and parentage assignments as a function of the number of loci. Each locus has 10 codominant alleles in a triangular frequency distribution. Individual genotypes at a variable number of microsatellite loci ( $x$  axis) were simulated using the FS1 model for octaploids, and were converted to diploid dominant phenotypes before conducting relationship analysis.

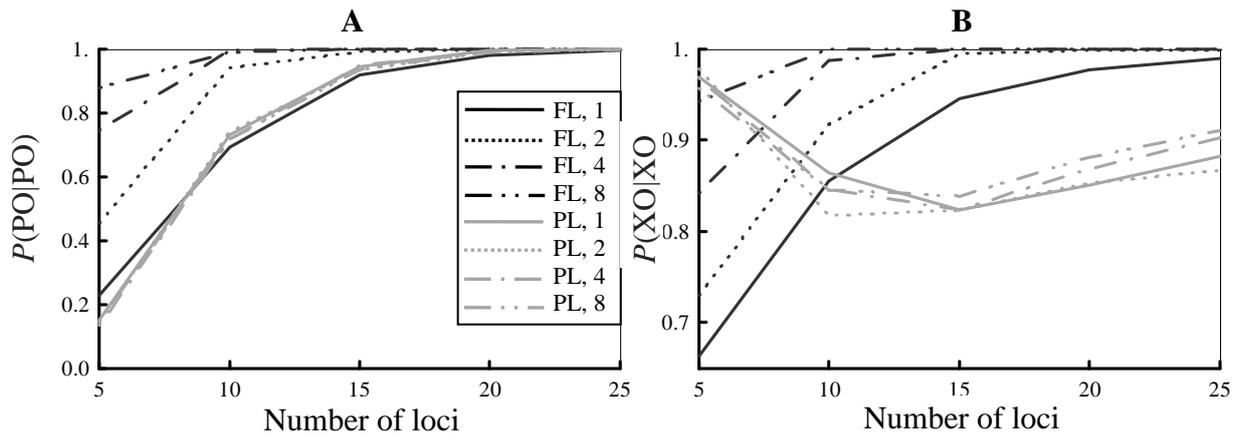


Figure 2: Comparison of parentage inference accuracies between methods for sibships of different sizes. Correct parentage assignment (A:  $P(PO|PO)$ ) and exclusion (B:  $P(XO|XO)$ ) frequencies from the full (FL) and pairwise (PL) likelihood methods for offspring in sibships of various sizes (1, 2, 4, 8) are plotted as a function of the number of loci used in the inferences. Each locus has 10 codominant alleles in a triangular frequency distribution. Individual genotypes at a variable number of microsatellite loci ( $x$  axis) were simulated using the FS1 model for octaploids, and were converted to diploid dominant phenotypes before conducting relationship analysis.

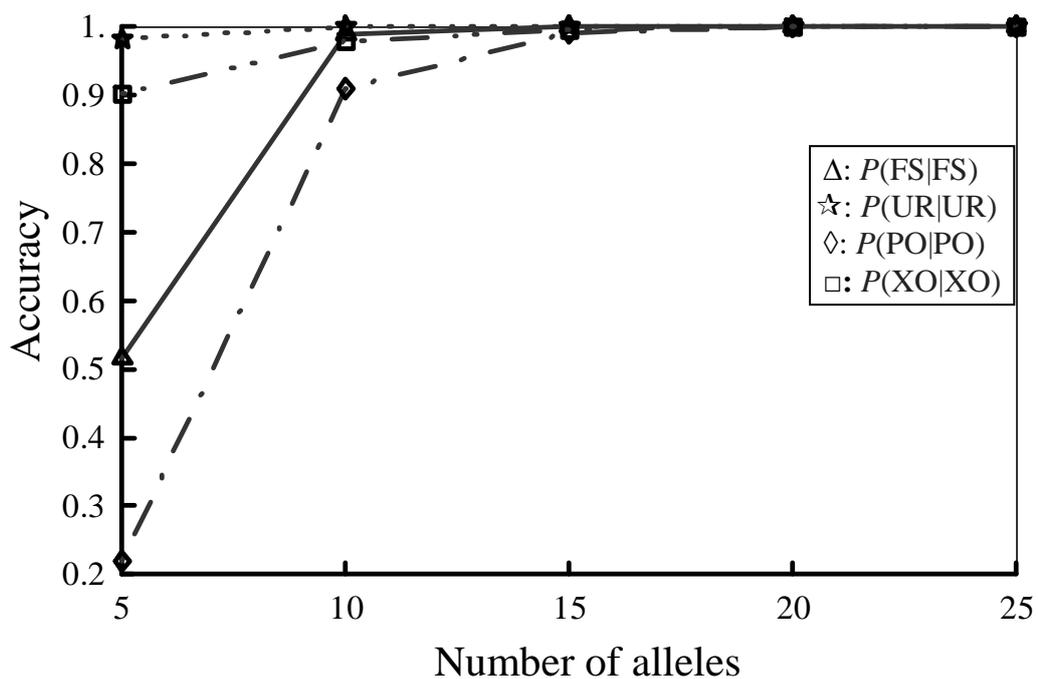


Figure 3: Accuracy of full-sib and parentage assignments as a function of the number of alleles per locus. Data were simulated using the FS1 model for octaploids. Each individual was genotyped at 10 loci, with each locus having a variable number ( $x$  axis) of codominant alleles in a triangular frequency distribution.

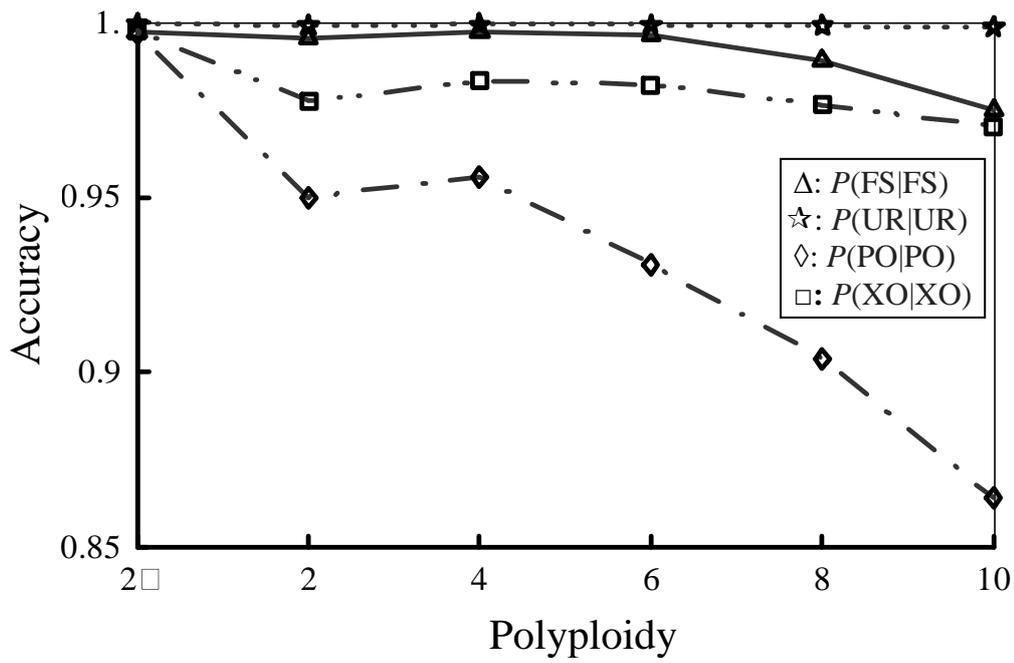


Figure 4: Accuracy of full-sib and parentage assignments as a function of species ploidy levels. For diploid species, individual phenotypes at 10 codominant marker loci were either untransformed or transformed to phenotypes to 100 dominant marker loci, indicated by 2\* and 2 on the  $x$  axis, respectively. Data were simulated using the FS1 model and 10 loci, each having 10 codominant alleles in a triangular frequency distribution.

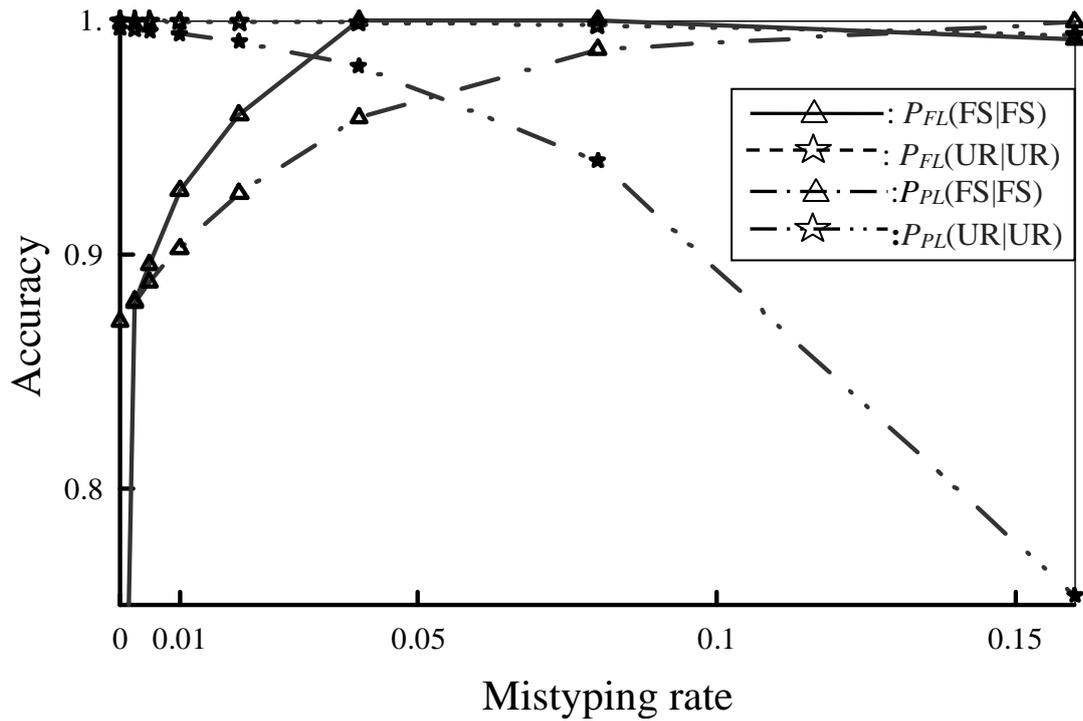


Figure 5: Accuracy of full-sib assignments as a function of mistyping rate at each locus used in data analysis. Genotype data were simulated, under the FS2 model for octaploids, for 10 loci, each having 10 alleles in a triangular frequency distribution. The polyploid codominant genotypes were generated with no mistyping, but were converted to diploid dominant genotypes and were analysed assuming a variable mistyping rate ( $x$  axis) by the full likelihood (FL) and pairwise likelihood (PL) sibship assignment methods.

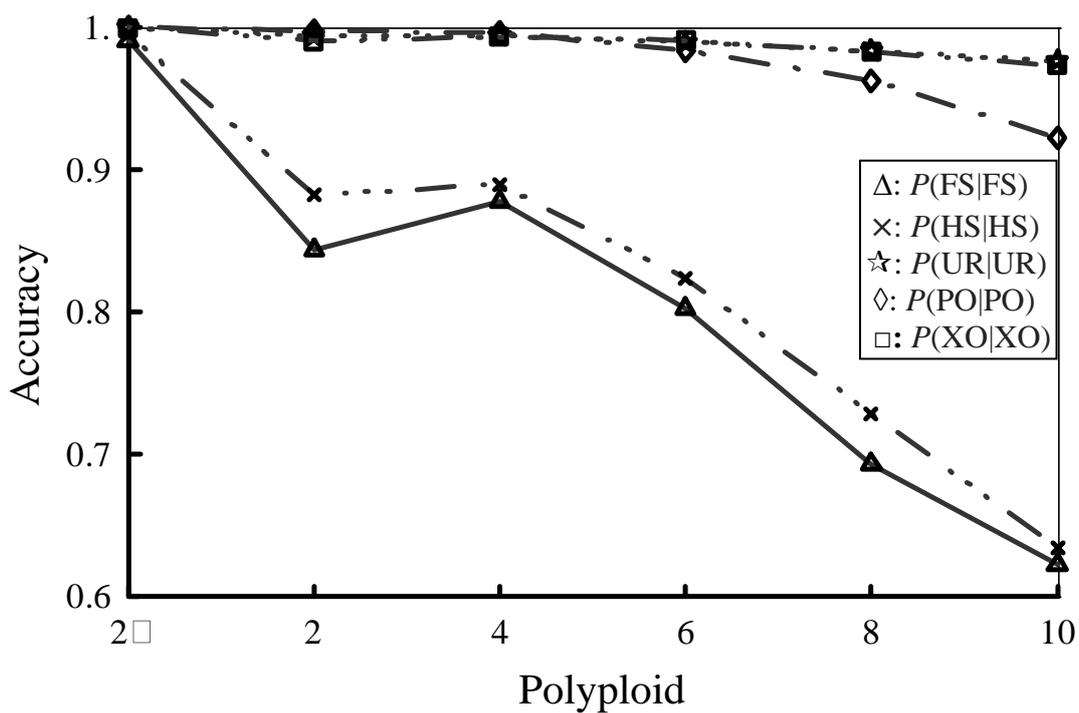


Figure 6: Accuracy of sibship and parentage assignments as a function of species ploidy levels in simulations of the HS model. For diploid species, the phenotype data at 10 codominant marker loci were either untransformed or transformed to phenotypes to 100 dominant marker loci, indicated by 2\* and 2 on the x axis, respectively. Data were simulated using the HS model and 10 loci, each having 10 codominant alleles in a triangular frequency distribution.

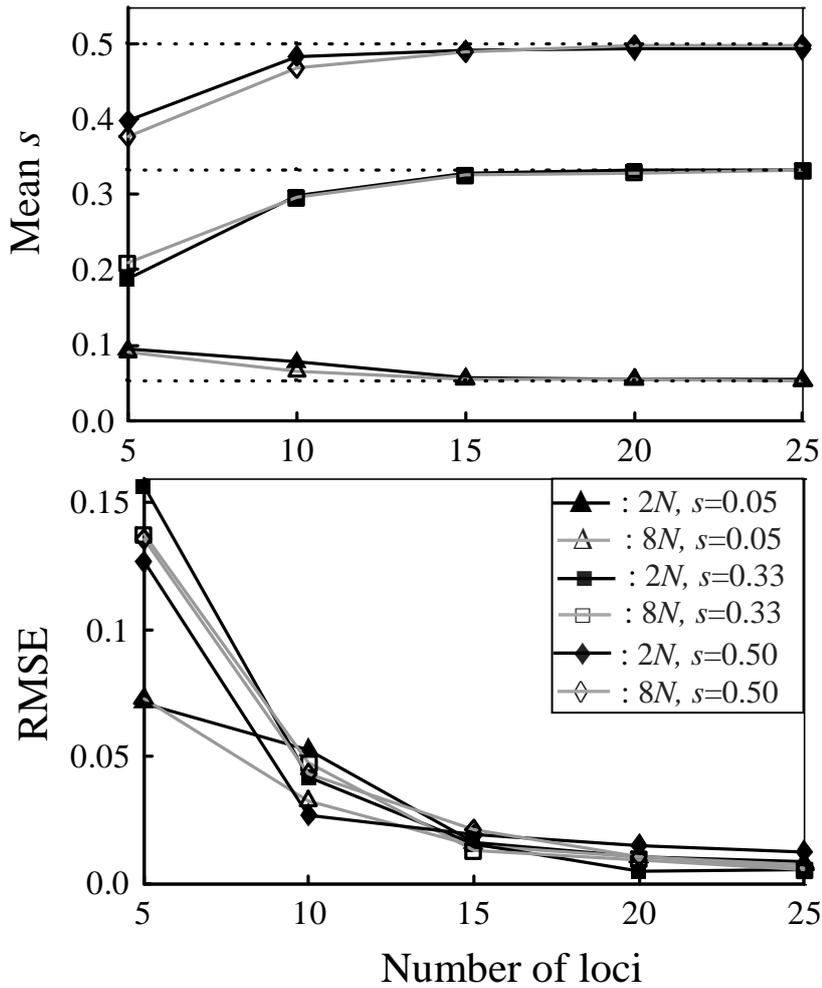


Figure 7: Accuracy of selfing rate estimates as a function of number of loci for diploid ( $2N$ ) and octaploid ( $8N$ ) monoecious species. The actually simulated selfing rates are 0.05 (low), 0.33 (medium) and high (0.50), indicated by horizontal dotted lines. Each locus has 10 codominant alleles in a triangular frequency distribution. The original codominant diploid and octaploid genotype data were converted to pseudo diploid dominant phenotypes before conducting relationship analysis.

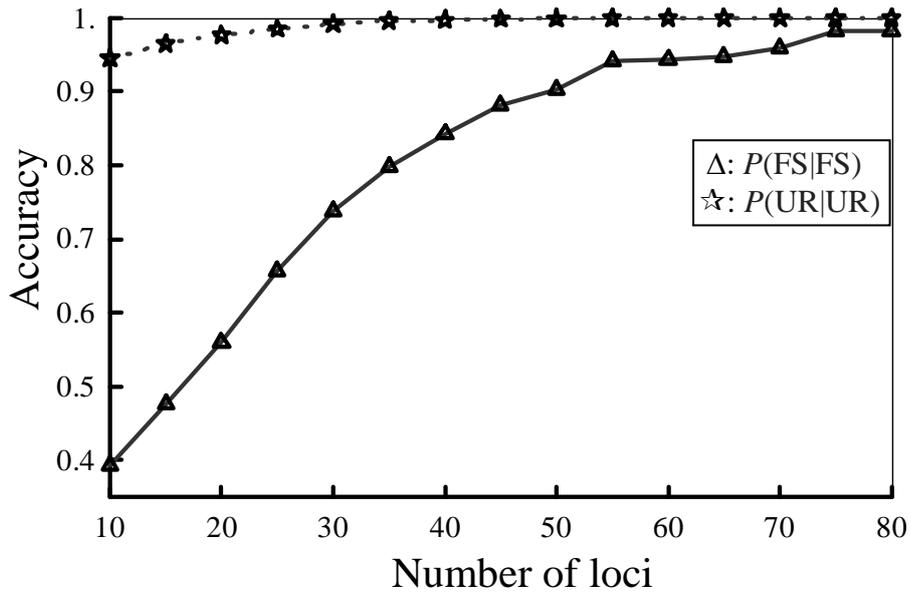


Figure 8: Sibship assignment accuracy in the sturgeon dataset using a variable number of loci (x axis) randomly selected from the original 125 transformed loci.

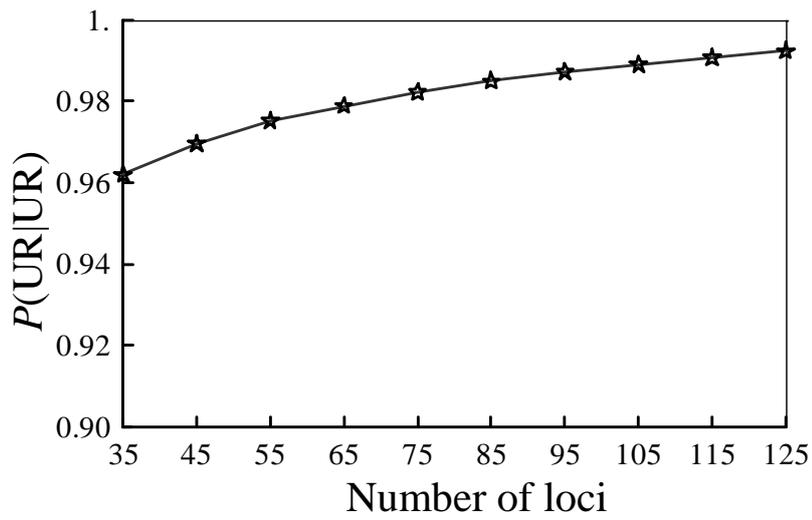


Figure 9: Sibship exclusion accuracy obtained from the sturgeon permuted datasets. Each dataset was generated from genotypes permuted among individuals at each of a variable number of randomly selected loci ( $x$  axis).