Purging putative siblings from population genetic datasets:

A skeptical view

Robin S. Waples1 and Eric C. Anderson2

1NOAA Fisheries, Northwest Fisheries Science Center

2725 Montlake Blvd. East, Seattle, WA 98112 USA

2NOAA Fisheries, Southwest Fisheries Science Center

110 Shaffer Road, Santa Cruz, CA 95060 USA

Running head: Removing siblings

Key words:

corresponding author:

Robin S. Waples

2725 Montlake Blvd. East

Seattle, WA 98112 USA

FAX +1 206 860-3335

[robin.waples@noaa.gov](mailto:robin.waples@noaa.gov)

September 2016

*Molecular Ecology or ME Resources*

**Introduction**

Random sampling is a convenient theoretical and statistical construct. It is easy to implement in computer simulations but difficult or impossible to achieve in the real world. A truly random sample has two important features: *equal opportunity*—every individual in the focal population must have the same probability of appearing in the sample; and *independence*—the probability that an individual will be sampled does not depend on whether any other individuals are in the sample. A violation of the second criterion occurs when relatives are likely to be sampled together. Recent attention related to non-random sampling has focused on three key points:

* Non-random samples, with some families being over-represented, are common in natural populations, particularly for species with relatively high fecundity that are sampled at early life stages (Hansen et al. 1997; Goldberg and Waits 2010);
* Presence of family structure in samples can affect population genetic analyses (Anderson and Dunham 2008; Rodríguez-Ramilo and Wang 2012);
* Recent advances in DNA technology and analytical methodology have improved the ability to identify relatives in natural populations (Ashley et al. 2009; Jones and Wang 2010; Almudevar and Anderson 2012).

Based on these three observations, many researchers now routinely remove all but one member of a putative sibling group before using population genetic data in downstream analyses (e.g. Hess et al. 2015)—an approach that is now regarded by some as “best practices” (Peterman et al. 2016). The premises are sound and the conclusion seems logical—what could go wrong?

Actually, a lot can go wrong. At least three major problems can arise when one attempts to purge putative siblings from samples. (1) Siblings occur naturally in all finite populations at frequencies that are inversely related to effective population size (*Ne*)—and this fact forms the basis of the sibship method for estimating *Ne* (Wang 2009). Indiscriminant removal of all putative siblings thus risks erasing part of the evolutionary signal of small populations, making them look more like large or infinitely large ones. (2) Even if sibling removal is effective in reducing the appearance of non-random sampling, it also reduces the sample size, which sets up an inevitable tradeoff that should be formally considered. (3) Methods for sibling inference are not infallible, particularly for identification of half-siblings or other more distant relatives. Results of sibship reconstruction often differ depending on the method used (Ringler et al. 2015) and, for a given method, can differ depending on the type of markers used (Linløkken et al. 2016). Therefore, researchers interested in pursuing removal of putative siblings should consider the consequences of imperfect ability to identify relatives.

As far as we can determine, the only analyses for which sibling removal has been convincingly demonstrated to improve performance are Bayesian clustering methods such as STRUCTURE (Pritchard et al. 2000), for which family groups can be mistaken for separate “populations” (Anderson and Dunham 2008; Rodríguez-Ramilo and Wang 2012). Rodríguez-Ramilo et al. (2014) found that some other methods that do not make assumptions about Hardy-Weinberg equilibrium and linkage equilibrium were much less sensitive to family groups than STRUCTURE. This suggests that the effects of siblings are case-specific and depend on the particular analyses involved. Because related individuals in a sample create a mismatch between the variance of estimated allele (or genotype or gamete) frequencies and the variance predicted under the assumption of random, independent sampling, it is clear that the presence of siblings can bias the results of statistical tests that compare the observed deviations in statistics (like allele frequencies) with their expectations under random sampling theory. For example, over 35 years ago Allendorf and Phelps (1981) noted that the occurrence of siblings amongst samples of juvenile salmonids could inflate Type I error rates for tests of population differentiation. Accordingly, researchers must remain vigilant about the effect of siblings on statistical hypothesis tests. However this does not mean that eliminating siblings from samples is necessarily and universally a “best practice.” In particular, although sibling elimination has been advoacated for improving point estimation of allele frequencies (CITATION) it may actually be detrimental to accuracy. This in turn suggests that there can be a downside to sibling elimination and that, depending on circumstances, it may not be warranted to routinely remove putative siblings from genetic datasets without a clear understanding of the consequences.

Here, we use computer simulations to evaluate the practical consequences of removing some or all siblings from both random and non-random samples, generated under several mating models. We focus on two single-locus metrics (allele frequency, one of the most fundamental parameters in population genetics; and *FST*, a widely-used measure of genetic differentiation) and one two-locus metric (linkage disequilibrium between pairs of loci, which is increasingly important in analyses of genomics-scale datasets). Results indicate that even if sibling relationships are known without error, the elimination of sibships can degrade precision in the estimation of allele frequencies and Fst, and bias estimates of Ne. The results argue for caution by researchers who are tempted to try purging siblings from samples from real populations.

To understand these simulation results, mathematically, we next undertake an analytical treatment of sibling elmination’s effect on allele frequency estimation. We consider sibling elimination as a special (extreme) case of unequal weighting of the information from each individual in a sample. Use of such weighting schemes to obtain the best linear unbiased estimator (BLUE) for a parameter are well known in statistics, and we review McPeek et al. (2004)’s application of the BLUE approach for estimating allele frequency in the presence of related individuals. Using the same mathematical machinery, we obtain an algorithm to find the optimal sibling elimination scheme. And finally, through gene-dropping simulations we assess the utility of the BLUE approach and sibling elimination for estimating allele frequencies from multiple collections of highly related juvenile coho salmon. Our results indicate that, if the true relationships are known, the BLUE approach has great merit (considerably more than sibling elimination MUST VERIFY THAT IN THE SIMS), but that misidentification of relationships can reduce its utility.

**Materials and methods**

*Computer simulations*

*Reproduction.* All simulated populations had discrete generations, a constant number (*N*) of mature individuals, and separate sexes with an equal number (*N*/2) of males and females. Selfing was not allowed. Genotypes at 100 diallelic (SNP) loci were tracked in each individual. For each parameter set, a population of *N* individuals was initialized with a distribution of minor allele frequencies in the range 0.2-0.5. Then, 100 replicate daughter populations were created, each of which was allowed to reproduce in isolation for 10 generations before generating samples, as described below.

Three mating systems were considered: 1) Monogamy (the same male always mated with the same female, but which pair produced each offspring was a random draw, with replacement); 2) Random Mating (the male and female parents of each offspring were chosen randomly and independently, with replacement); and 3) Mixed. In the Mixed model, the male and female parents were drawn randomly and independently, as in the Random Mating model; however, with probability α the chosen pair was allowed to produce more than one offspring. If so, the number of offspring produced by that pair was chosen by another random draw from a uniform distribution of integers [2 … *X*], where *X* was the maximum family size. Parameters used for the Mixed mating model were α = 0.5 and *X* = 9, except in one extreme scenario which used α = 0.1 and *X* = 30. As necessary, family size of the last pair was truncated to keep population size constant.

*Sampling*. In each of the three models, offspring that represented a random sample of progeny were generated as described above, with the exception that the number of individuals (*S*) in the sample from the final generation could be smaller than, larger than, or equal to *N*. In addition, non-random, family-correlated samples were also generated for the Monogamy and Random Mating models. Under Monogamy, after each pair of parents was selected, the number of offspring that pair would produce was chosen randomly from integer values 1..*X*. This mimicked a situation where individuals are more likely to be sampled if one or more full siblings also are sampled. Under Random Mating, to generate family-correlated samples the female parent for each offspring was chosen randomly, and a random number chosen from integer values 1..*X* determined how many offspring that female would produce. The male parent of each of these offspring was then chosen randomly and with replacement from all the males. This produced mostly maternal half-sib families, with occasional production of full siblings. After the female produced the specified number of offspring, another female was randomly selected (with replacement) to produce the next set of offspring. Family sizes considered were *X* = 3 or 9, which produced mean family sizes for each reproductive event of 2 or 5, respectively. Note that total family sizes for the sample as a whole could exceed these values if the same individuals was chosen to be parents more than once. The Mixed mating model already generates considerable family structure from the mating scheme, so only random samples of offspring were generated.

In addition to considering the full samples, data analysis (see below) considered subsets of the samples from which some or all siblings were removed. Removal was based on the recorded pedigree and did not consider potential errors in sibling identification. In fractional removals, individuals from the full sample were evaluated one at a time to determine whether they should be included in the reduced sample. If the individual was a relative of any other individual already in the reduced sample, it was excluded with probability *β*. Values evaluated were *β* = 0.25, 0.5, 0.75, 0.9, 0.95, and 1. With *β* = 1, all but one member of each family was excluded. In the Random and Mixed mating models, sibship exclusion was considered two ways: excluding all siblings, or only full siblings.

To recap, in the Monogamy and Random models, reproduction was by Wright-Fisher ideal populations with and without monogamy, respectively (so *Ne* = *N*). In the final generation, samples of *S* progeny were produced in two ways: 1) using the same two ideal mating systems (random samples), or 2) by allowing some pairs of parents to produce large families (to mimic non-random, family correlated samples). In the latter case, the effective number of parents that produced the sample (*Nb*) was less than *Ne*. In the Mixed model of reproduction, some pairs of parents were allowed to produce > 1 offspring per mating episode, so variance in reproductive success was overdispersed, leading to *Ne* < *N* each generation.

All simulations and data analyses were conducted in R (www.rproject.org) using code that is available on request. The different scenarios that were simulated are summarized in Table 1.

*Sibling production*

The Random Mating and Mixed models produced three classes of offspring based on their one-generation pedigree: full siblings (FS), half siblings (HS), and unrelated (U). Only FS and U offspring were produced in the Monogamy model. The proportions of siblings produced are expected to be simple functions of *Ne* and the mating model. This relationship, based on a simplification of Equation 10 in Wang (2009) that ignores the (generally small) correction for departures from Hardy-Weinberg equilibrium, is:

, (1)

where *QHS* is the fraction of pairs that are half siblings (maternal and paternal half siblings combined) and *QFS* is the fraction that are full siblings. It can be shown (Ackerman et al., in press) that Equation 1 yields the same estimate of *Ne* as the parentage-analysis-without parents (PWOP) approach of Waples and Waples (2011), which calculates inbreeding *Ne* based on the vector of numbers of offspring produced by each parent. Therefore, in cases where 100% of siblings are correctly identified, the estimate of *Ne* from the sibship method (Wang 2009) will be identical to that from PWOP calculated on the true pedigree. Simple rearrangement of Equation 1 produces the expected frequencies of sibships:

. (2)

Observed fractions of siblings produced in the simulations were tracked and compared with the expected fractions, based on Equation 2 and realized *Ne* calculated from the pedigree using PWOP.

*Data analysis*

Analyses of simulated data focused on three metrics.

*Allele frequency*. For each locus in each replicate, true allele frequency (*TrueP)* was calculated as the mean frequency in the *N* final-generation parents. Estimated allele frequency () for each locus was calculated as the mean across all individuals in each sample of offspring. For each replicate, the root-mean-squared-error (RMSE) of allele frequency was calculated as

,

with the summation taken across all 100 loci. An overall RMSE was then calculated as the mean RMSE across 100 replicate simulations. The consequences of sibling removal for allele frequency estimation were quantified as the relative RMSE (*ϕ*), which is the ratio of RMSE for the subsample and the full sample. Values of *ϕ* < 1 indicate that sibling removal increased precision of , while values > 1 indicate that sibling removal degraded performance.

With random mating and random sampling of *S* progeny across one generation of genetic drift, the expected MSE of  is just the binomial sampling variance:

.

As the size of a random sample becomes arbitrarily large,  for samples of progeny will converge on *TrueP*. Let *K* = *TrueP*(*1- TrueP*). Then,

. (3)

Now consider a reduced sample size *S*\* = *xS*, where 0 < *x* < 1. For this smaller sample,

. (4)

Therefore, assuming random mating and random sampling, the expected relative RMSE of  for a reduced sample of size *xS* is given by

.

We can also calculate an effective sample size index (*x*\*), which is given by

= Observed(*ϕ*), so . Whereas *x* is the actual proportional reduction in sample size caused by sibling removal, *x*\* is the proportional reduction that is consistent with the observed *ϕ.* This leads naturally to the definition of an effective sample size (ESS) = *Sx*\* = the size of an ideal, random sample that would be expected to produce the observed value of *ϕ*.

[add info about empirical coho example and BLUE method]

*Population differentiation*. For each parameter set, the 100 replicate daughter populations were divided into 50 pairs, and Nei’s (1973) *FST* was calculated for each locus as

,

where *HS* is the average expected heterozygosity within the two populations and *HT* is the total expected heterozygosity based on mean allele frequencies across the populations. For each pair of populations, an overall mean *FST* () was calculated across all 100 loci using data for all *N* parents, and this was considered the parametric (true) *FST*. Unbiased estimates of *FST* () from samples of progeny were adjusted for sampling error by subtracting the quantity 1/(2*S*) (Wright 1978; Chakraborty and Leimar 1987). Loci monomorphic in both samples were excluded. Across the 50 pairs of populations, and for each sibling-reduced sample size, RMSE of *FST* was calculated as

.

*Effective population size*. True effective size under the Monogamy and Random Mating models was *Ne* = *N* + 0.5 + 1/(2*N*) (Balloux 2004); these values are shown in Table 1, but for simplicity in the text only the whole number is used. In the Mixed Mating model, the possibility that a parental pair would be allowed to produce multiple offspring at each draw led to overdispersed variance in reproductive success and *Ne* < *N*, with the magnitude of reduction in *Ne*/*N* determined by choice of *N*, α, and *X*. Realized *Ne* in the Mixed model was calculated each generation using PWOP. PWOP effective size was calculated separately for each sex (*NeM*, *NeF*) and the overall *Ne* was calculated as *Ne* = 4*NeMNeF*/(*NeM*+*NeF*) (Wright 1938). PWOP was also used to track the realized effective number of parents that produced the family-correlated (non-random) samples generated under the Monogamy and Random Mating models.

For each sample, *Ne* was estimated using the linkage disequilibrium (LD) method (Waples and Do 2008), and an overall harmonic mean  was calculated across all replicates. Alleles with frequencies < 0.05 were excluded. In addition, RMSE of 1/ was calculated as described above for , and results were used to calculate a relative RMSE of estimates of effective size.

**Results**

Results are summarized in Figures 1-6 and S1-S5; see Table S1 (under construction) for more details.

*Sibling frequency and family size*

Observed frequency of siblings in the samples agreed closely with those expected from Equation 2 (Figure S1). Although expected overall frequencies of siblings depend only on realized *Ne*, the mix of full and half siblings depends on the mating system and the type of samples. Furthermore, the distribution of family sizes in the sample depends strongly on the sample size. Both random and family-correlated samples showed a strong, positive relationship between maximum family size and the ratio *S/Ne* (Figure 1). For the same *S/Ne*, random samples have consistently smaller maximum family sizes. However, a random sample for a scenario in which sample size is large compared to effective size can have larger family sizes than are found in some non-random samples. Frequencies of full and half siblings in the complete samples for the three mating models are shown in Figure S2.

*Reduced sample sizes*

100% sibling exclusion reduced the final sample size by up to 90% or more when *S* was larger than *Ne* and all siblings were excluded (e.g., Scenarios J and H; Figure 2). At the other extreme, in some scenarios with very small *S/Ne* and/or exclusion only of full siblings (e.g., Scenario S; Figure 2), the final sample sizes were reduced by less than 10%. Partial removal of siblings had predictably intermediate consequences for sample size.

*Allele frequency*

Removing siblings from random samples of progeny generally reduced precision, such that RMSE of  was higher after sibling removal than it was for the full dataset (Figure 3, top). This effect was non-linear, with RMSE rising faster for high levels of sibling removal. With 100% sibling removal, RMSE of  could be over twice as large as for the unpurged dataset. In some scenarios, RMSE of  increased only very slightly with removal of siblings. This occurred when few siblings were produced in the first place (Scenario G) or most siblings were half-siblings but only full siblings were removed (Scenario F), but also in some mixed-mating scenarios (e.g., Scenario I) where sampling was random but many siblings were produced. In one extreme scenario using the Mixed Mating model (Scenario Y), where probability that a chosen pair of parents would be allowed to produce multiple offspring was low (α = 0.1) but if they were lucky family size could be as high as 30, RMSE of  dropped by 30% with removal of all siblings (Figure 3, top).

Surprisingly, the same patterns were found in most scenarios with non-random samples: removing siblings increased RMSE (by up to ~50%) and reduced precision compared to the full samples (Figure 4, top). In one case (Scenario Q, random mating with family-correlated sampling that primarily produced half siblings), removing 100% of the full siblings actually reduced RMSE of  by 3%.

In every scenario considered, regardless whether sampling was random or not, the ratio ESS/S\* was > 1, indicating that effective sample size for sibling-reduced samples was larger than would be expected for random reductions of sample size of the same magnitude (Figure 5, top). However, with the exception of Scenario Y, the ratio ESS/S was < 1, indicating that effective sample size after sibling removal was less than the original sample size (Figure 5, bottom).

[summarize results of BLUE and coho example here]

*Population differentiation*

Results for estimation of *FST* largely paralleled those for estimation of allele frequency: in most scenarios with both random samples and non-random samples, removing siblings increased RMSE of  (Figures 3 and 4, bottom). Two differences in results for  and  are worth noting, however. First, the effects were magnified for *FST*: purging all siblings led to almost a 10-fold increase in RMSE for some scenarios with both random and non-random sampling. Second, in some scenarios using the mixed or monogamy mating models, RMSE of  was reduced when intermediate levels of siblings were removed. The best result occurred for extreme Scenario Y noted above, where RMSE of  was reduced by 55-60% when 50-75% of siblings were removed. However, in this scenario precision of was reduced compared to that of the full sample if all siblings were removed.

*Effective population size*

Harmonic mean  calculated using PWOP closely tracked the theoretical true *Ne* in scenarios with random sampling and random mating or monogamy (Figure S3). In the scenarios with non-random sampling and in the Mixed mating model (where sampling was random but mating was not), PWOP estimated realized *Ne* based on the pedigree for the samples, and  from LDNe closely tracked PWOP (Figure S4). When sibling removal was less than about 80%,  from LDNe was generally a few percent higher than PWOP. This reflects a slight upward bias in the LDNe version of the LD method, which has been documented elsewhere (e.g., Waples and Do 2010). With 100% sibling removal, the point estimate by PWOP is infinity (see Equation 2). Because sibling removal was probabilistic, on occasion all siblings were removed from samples by chance when the probability of removal was high but less than 1. Infinite estimates (recorded here as 99999 for each sex) increased harmonic mean PWOP, such that  from PWOP generally exceeded that from LDNe under very aggressive purging of siblings (Figure S4).

Removing siblings had the expected effect of increasing  because truncating largest family sizes reduced disparities in reproductive success among parents. With random sampling,  from LDNe was essentially unbiased using the full sample but became increasingly upwardly biased as a larger fraction of siblings was removed (Figure 6, top). For example, Scenario G (with a large *Ne* = 500 and a small *S* = 40) did not produce many siblings, but removing all of the few siblings that did occur by chance produced an estimate that was over 6 times the true *Ne*.

When sampling was non-random, estimates of *Ne* based on full samples were all downwardly biased, and  increased with more aggressive purging of siblings (Figure 6, bottom). When initial samples are non-random, there is a transition from under-estimation to overestimation of *Ne* somewhere along the continuum of fractional sibling removal. This means that (in theory at least) an unbiased estimate could be obtained from non-random samples by removing exactly the right fraction of siblings. This “sweet spot,” however, varied widely among scenarios. In Scenario B, where the sample was only moderately non-random, removal of half of the siblings produced an unbiased  (Figure 6, bottom). But for Scenarios J, A, M, and K,  did not become unbiased until about 80%, 90%, 95%, and >95% of siblings were removed, respectively.

Removing siblings from random samples always sharply increased relative RMSE of

, except in scenarios where there were few siblings to remove (Figure S5, top). With non-random samples, removing siblings generally lowered RMSE until sibling removal hit the sweet spot where the estimates of *Ne* became largely unbiased, after which RMSE rose again, in some cases sharply (Figure S5, bottom).

**Mathematical analysis of allele frequency estimation under sibling elimination**

The maximum likelihood estimate and the method of moments estimate of the frequency of an allele from a sample of independent and randomly sampled diploids are both found as

where is a random variable that records the number of copies of the allele carried by the -th member of the sample, divided by 2. The sibling-eliminated version of the allele frequency estimator can be written as

where each is a 1 or 0, indicating that individual has either been included in or eliminated from the sample, respectively. Note that if every then this estimator clearly reduces to the simple estimator, . An obvious generalization of such an estimator would be

in which the weights are non-negative numbers not constrained to be only 0 or 1. We can refer to the weights, , by the values which have been normalized so that they sum to unity.

It is straightforward to show that any choice of (so long as at least one element is nonzero) yields an unbiased estimate of ; however different choices of will produce estimators with different variances. The variance of can be found with the standard formula for the variance of a sum:

which can be more compactly expressed in matrix notation as

where is a column vector of the normalized weights and is the variance-covariance matrix of the 's. From this it is clear that the variance of depends on the normalized weights (the 's), the individual variances of each and also on all the pairwise covariances of the 's. The variances and covariances are, in turn, determined by the true allele frequency and the inbreeding coefficients of each individual and by their pairwise coefficients of kinship. In fact, the covariance matrix is

where is the matrix whose off-diagonal elements are twice the pairwise coefficients of kinship between the members of the sample, and whose diagonal elements in each row, , are equal to where is the inbreeding coefficient of individual (CITATION). The matrix does not depend on the allele frequency itself and it can be easily found by recursive calculation on any pedigree that connects the members of the sample (Thompson 2000), and so can be computed in the case of known (or suspected) siblings in a sample. Hence we can determine the variance and the effective sample size of a sibling-eliminated sample or of one in which the individuals are assigned non-integer weights. This variance is

If all the members of the sample were independent, the variance of would be from which it follows that the effective size (i.e., the size of an independent sample that would provide the same variance) of a weighted or sibling-eliminated sample is where each element of is , if considering a weighting scheme, or if we are considering, specifically, a scenario in which siblings are eliminated.

McPeek et al. (2004) exploited these relationships to derive a Best Linear Unbiased Estimator (BLUE) for allele frequencies when the sample consists of individuals of known relationship. They define as above and use the weights known from the theory of best unbiased linear estimation to minimize the variance of , namely:

In other words, the optimal (unnormalized) weights are the sums of the columns of the inverse of the matrix . The estimator in general can be expressed in matrix notation as

where is the column vector . Plugging into (1), the variance of the BLUE estimator can be derived as:

and the effective sample size of the BLUE estimator is .

Since the weights in are chosen to minimize the variance, it should be clear to the reader that if the matrix reflects the *true* relationship of the individuals in the sample, the variance of will always be smaller than the variance of ---the estimate obtained by any sibling elimination scheme. Nonetheless, it will be instructive in our analysis to consider the optimal sibling elimination scheme: the set of weights that minimizes . Given a sample of individuals with relationship and inbreeding implying a matrix we can perform the analysis above for any set of weights. Hence, for any , it is possible to find the optimal sibling elimination scheme (if any). As far as we can tell, finding the optimal is a combinatorial optimization problem akin to quadratic programming (CITATION) but with integer constraints. We are not aware of a straightforward approach to this problem, but propose a simple greedy solution to approximate the optimal .

1. Initialize: Set , set a vector of ones, compute using equation 1, and let .
2. Set , set and .
3. Identify individual with the ``most relatedness and inbreeding'' in the sample as the individual whose row in has the largest sum.
4. Modify by setting .
5. Using and equation XX compute the variance .
6. If then remove row and column from and go back to step 2. If not, then return

**Investigation of Sibling Elimination and Weighting Schemes in an Empirical Dataset**

The foregoing provides us a way to rapidly (without any simulation) compute the effective sample size of a weighted or sibling-eliminated sample given any known true pattern of relationship between the sample members. We now investigate the outcome of weighting and sibling elimination in a large empirical data set of juvenile coho salmon. The practice of sibling elimination has been used extensively in the analysis of such data because collecting juvenile salmon from small streams can yield samples containing a great many siblings (Garza et al. 2014; Gilbert-Horvath et al. 2016). Here, we compute the effective sample size under different weighting and sibling elimination schemes for 77 collections of coho salmon taken from 47 different streams in California and Southern Oregon (described in Gilbert-Horvath et al. 2016). Each collection, which varied in size from 2 to 150 individuals (mean = 55) was genotyped at 95 SNPs and analyzed with Colony (Version 2, Wang 2004, Wang and Santure 2009) to infer full- and half-siblings. For our “Sample Related” scenario, we assumed that the pedigree structure inferred by Colony within each collection was correct and hence was inferred without error. We then used the inferred pedigree to compute the effective sample size under 5 different weighting schemes: 1) **naïve**: no weighting, 2) **BLUE**: weighting according to the best unbiased linear estimator, 3) **optimal-z**: the minimum variance sibling-elimination scheme found using our greedy algorithm, 4) **sib-yank-1**: individuals were eliminated randomly from sibships of size greater than or equal to 3 until only 1 member of the sibship was remaining. 5) **sib-yank-2**: individuals were eliminated randomly from sibships of size greater than or equal to 3 until only 2 members of the sibship were remaining. The sib-yank-1 scheme was used in Garza et al (2014) and Gilbert-Horvath et al. (2016). The sib-yank-2 scheme is a variant that eliminates fewer individuals from the sample.

The “Sample-Related” scenario represents a best-case scenario for the efficacy of sibling elimination–one in which the true relationships between half and full siblings are inferred without error. However, if relationships are inferred incorrectly, sibling elimination and application of the BLUE have the potential to degrade the estimation of allele frequencies. We investigated this possibility with our “Sample-Unrelated” scenario. In this scenario, the gene copies carried by each member of each collection were permuted within loci (holding missing data positions constant) yielding for each collection a permuted sample in which there is no longer any relationship between the sample members. These permuted samples were analyzed with Colony 2 and the 5 different weighting schemes were determined using the Colony-inferred pedigree. The effective sample size, however, was computed using (1) assuming the sample consisted of unrelated individuals, consistent with the fact that the data had been permuted. In this case, when Colony spuriously identifies related individuals in the permuted samples, it can potentially reduce the effective sample size of the weighted estimators.

The results of our investigations are shown in Figure XX. Under the “Sample-Related” scenario, not surprisingly, application of the BLUE maximized the effective sample size in every case. Also, as expected, the optimal-z sibling elimination strategy never decreases the effective sample size, but, in some collections, yields substantial increases in effective sample size. The sib-yank-2 strategy almost universally yielded a higher effective sample size than the naïve estimator. The sib-yank-1 strategy occasionally yielded a smaller effective sample size than the naïve estimator, but, more often than not, increased the effective sample size, sometimes significantly so. This suggests that if the degree of relatedness is high, as it is in these collections of juvenile coho salmon, sibling elimination can actually be a beneficial procedure for improving estimates of allele frequencies. The results of the “Sample-Unrelated” scenario show that using the BLUE or the optimal-z strategy can reduce the effective sample size when Colony errs by identifying spurious relationships in a sample. Colony is much more likely to spuriously identify half-sibling relationships or small (<3 individuals) full-sibling groups than large full sibling groups from data where none exist. As a consequence the sib-yank-1 and sib-yank-2 strategies often yield results identical to the naïve (and, in the “Sample-Unrelated” case, correct) estimator, since sibships of size less than three are not affected in these strategies. Thus, the sib-yank-1 and sib-yank-2 strategies , yielded the maximal effective sample size in almost every collection under the “Sample Unrelated” scenario.

These results show clearly that, although our initial simulations suggested that sibling elimination would seldom be a useful strategy, in real world samples containing many related individuals, like the coho salmon data that we investigated, sibling elimination can be beneficial for decreasing the variance of estimates of allele frequency. While it might be tempting to apply the BLUE or try to find the optimal elimination scheme given the inferred pedigree of sample members, our investigation shows that such approaches are not robust to errors in the identification of siblings using molecular data. Rather, the sib-yank-1 approach taken previously in Garza et al. (2014) and Gilbert-Horvath et al. (2016) was found to, in most cases, decrease variance of allele frequency estimates while being robust to inaccuracies in pedigree inference. However, the sib-yank-2 procedure provides an improvement on that, having higher effective sample size than sib-yank-1 in almost every case.

**Figure Caption**: Effective sample sizes (ESSs) of various sibling elimination and weighting strategies applied to empirical data from coho salmon. In each case, collections of juvenile coho have been arranged along the *x*-axis in increasing order of the effective sample size under the naïve scenario. In (a), the “Sample-Related” scenario, estimators perform worse than the naïve estimator for collections at which their lines drop below the naïve estimator line. This occurs occasionally for sib-yank-1 and rarely for sib-yank-2, and never for any of the other estimators. In (b), the “Sample-Unrelated” scenario, the naïve estimator has the highest ESS, but its line is covered almost entirely by sib-yank-1 and sib-yank-2 which, for every collection except one, achieve this maximal ESS. The BLUE and optimal-z estimators, by contrast, have ESSs that are reduced by inaccuracies in pedigree inference. See text for full explanation of estimators and scenarios.

**Discussion**

*Allele frequency*

Presence of family structure in a sample does not lead to systematic bias in estimation of population allele frequency, in the sense that large numbers of siblings do not consistently lead to under- or over-estimation of *TrueP*. However, for any given sample, the presence of siblings will tend to skew the estimate  toward frequencies of the parents responsible for the large families, and that increases RMSE of, and reduces precision of, . The rationale for removing siblings to estimate allele frequency is twofold: 1) reducing large families to one or a small number of representatives produces a more balanced picture of allele frequencies in all the parents, and 2) siblings provide partially redundant information about parental allele frequencies, especially when family size is large, so removing them does not sacrifice much useful information.

Our results provide three important insights on this issue. First, in every scenario evaluated, and for every level of sibling purging, the effective sample size of the sibling-reduced sample was larger than the number of individuals remaining in the sample (all ESS/S\* > 1; see Figure 5, top). This confirms the partial redundancy of siblings with respect to allele-frequency estimation of the parents, as ESS did not decline as fast with removal of siblings as would be expected if each of the individuals removed provided completely independent information. The second key result is that in most cases this partial redundancy was not strong enough to fully offset the consequences of reducing overall sample size. When sibling removal had any appreciable effect on performance, it always reduced precision of . The only exception to this pattern was for the rather extreme Scenario Y, which led to greatly overdispersed variance in reproductive success among the parents. This means that purging putative siblings is not a sound general strategy to improve estimates of *TrueP*; it seems much more likely to make things worse than better. Finally, the empirical example using the BLUE approach, which calculates a weighted  by down-weighting individuals that are closely related to others in the sample, suggests that this method … better general option.

*Population differentiation*

Our results for *FST* estimation largely paralleled those for allele frequency, which suggests that they might be generally applicable to other types of analyses that strongly depend on population allele frequencies, such as …X,Y,Z—but that is only a conjecture that requires empirical evaluation. We did find that a wider range of scenarios showed improved performance of  under partial removal of siblings (see Figures 3 and 4). We suspect this might reflect an interaction between ESS and the adjustment to  to make it unbiased (i.e., by subtracting the quantity 1/(2*S*) to account for sampling). If a sample of *S* individuals contains relatives, its ESS will be < *S*, which means that subtracting the quantity 1/(2*S*) will not be sufficient to produce an unbiased estimator of *FST*. Purging some siblings with partially redundant information reduces the number of individuals in the sample faster than it reduces ESS and hence brings the effective and actual sample sizes into closer agreement.

*Effective population size*

Presence of siblings affects estimation of *Ne* differently (in two ways) than it does estimation of allele frequency. First, whereas family structure only affects precision of , it can severely bias estimates of *Ne*. Second, whether samples are collected randomly or not has a much larger effect on  than . For the LD (and sibship) methods, family structure that occurs naturally in all finite populations is part of the signal that allows one to estimate effective size. Non-random sampling that tends to collect groups of siblings together creates additional family structure that is mistaken for drift, which downwardly biases . In theory, an unbiased sample could be reconstructed by removing siblings collected in excess of the random expectation (Figure 6). In practice, however, this will be difficult or impossible to accomplish. It is clear even from our limited evaluations that no generic, one-size-fits-all strategy of fractional sibling removal will be effective in every scenario. The location of the sweet spot that produces a completely unbiased estimate varies widely, depending on (at least) the mating system, type of sampling, values of *S* and *Ne*, and whether all siblings or only full siblings are removed. Furthermore, the Goldilocks zone (an area around the sweet spot that represents “just the right amount” of bias adjustment to produce a reasonable estimate) can be very narrow, such that the consequences of small errors in identifying the optimal fraction of siblings to remove can be harsh. For example, under Scenario M, an unbiased estimate can be achieved by removing 90-95% of all siblings, but removing only 75% leaves  downwardly biased by almost 50%, while removing all siblings produces an estimate that is over 5 times as large as true *Ne* (Figure 6). Researchers interested in pursuing this option are faced with a chicken-and-egg conundrum: only if one knows the true effective population size can one determine the precise amount of sibling reduction that will produce an unbiased estimate of *Ne*.

We used the LD method to estimate *Ne*, but similar results can be expected for the other two widely-used single sample estimators of effective size. OneSamp (Tallmon et al. 2008) uses an approximate-bayesian-computation approach with several summary statistics, but the most important signal is from LD. When all sibling assignments are 100% accurate, the estimates from Wang’s (2009) sibship method are the same as calculated using PWOP, except for the random-mating adjustment employed by Wang (2009) (see Ackerman et al. in press). However, the temporal method for estimating Ne depends only on allele frequency, so using the BLUE procedure to obtain a pedigree-weighted estimate of TrueP could be an effective way of dealing with family structure in samples (but this should be verified empirically).

*Random and non-random sampling*

We see no realistic way to distinguish random and non-random samples based on patterns of relatedness, even if family structure can be reconstructed with 100% accuracy. For any given array of sibling relationships that can be generated with non-random sampling, it is possible to imagine a random sample that could produce the same result. For example, consider the extreme Scenario Y, in which most of the *N* potential parents produced no offspring that appeared in the samples, while one or a few lucky pairs of parents produced large numbers. This is exactly the kind of result that could arise if offspring of a species with Type III survival are sampled at an early age before they have properly mixed (as could easily occur with many fish, marine invertebrates, amphibians, insects, etc.). On the other hand, this pattern of family structure (realized *Ne* was < 2% of *N* for Scenario Y; Table 1) also conforms to predictions of Hedgecock’s (1994) hypothesis of sweepstakes reproductive success (reviewed by Hedgecock and Pudovkin 2011), which has been postulated to be responsible for a number of tiny estimates of the *Ne*/*N* ratio in marine species (Hedrick 2005; Hauser and Carvalho 2008; Waples in press). A sample with this type of family structure could therefore represent an extreme case of non-random sampling, or an extreme case of overdispersed variance in reproductive success that characterizes a novel and important evolutionary phenomenon. Without independent information about the nature of the samples, it generally will not be possible to distinguish these scenarios.

*Conclusions and future directions*

The paradigm that seems to be emerging in applied conservation genetics is that family structure in genetic datasets is an aberration or disease that has to be cured by removing siblings. We think it is dangerous to adopt this as a general perspective, even though it might be effective in some specific applications, such as Bayesian clustering. But if programs like STRUCTURE are not designed to deal properly with family structure that occurs naturally in all finite populations, perhaps the problem lies not with the data but the programs. As noted by Anderson and Dunham (2008), in theory it would be possible to modify Bayesian clustering methods to better account for family structure and such developments are being pursued in human genetics (Conomos et al. 2016).

Our results show that purging putative siblings is not a good general strategy for either of the following:

* Improving estimates of allele frequency or Fst; it is more likely to adversely affect performance. Using a BLUE approach to weight individuals before estimating allele frequencies appears to hold considerable promise for dealing with family structure in both random and non–random samples, but empirical evaluations are needed to rigorously evaluate the consequences for downsteam analyses that depend on allele frequency.
* Dealing with problems arising from non-random sampling. Even if sibship reconstruction is 100% accurate and an optimal adjustment for allele frequency can be made, the result will still be lower precision than could have been obtained with a random sample. And for analyses such as Ne estimation that can be biased by excess family structure in non-random samples, it will be difficult or impossible to determine the optimal fraction of siblings to remove. Therefore, if one strongly suspects they have non-random, family-correlated samples, by far the best strategy therefore is to go back and obtain a random sample. Of course, that is easier said than done in many natural populations.

Pessimistic as these results might appear, they nevertheless represent an overly optimistic view of the effectiveness of adjustments to account for family structure. In all results presented here, it was assumed that every true full and half sibling could be correctly identified, and that no unrelated pairs were incorrectly identified as siblings. Although that level of precision in sibling identification might be approached under some ideal conditions, in general there will be uncertainties associated with many sibling assignments, so a rigorous assessment of performance must include realistic assumptions about the reliability of sibling reconstruction.

By no means should results presented here be considered comprehensive; other mating systems could be modeled, and only a few selected parameter values and two types of analyses were considered. Furthermore, complications such as age structure and migration were ignored. It should be apparent that a great deal more research needs to be done before precise guidance can be given about how to deal with the complex problem of family structure in population genetic datasets. Can classes of analytical methods be identified that respond in predictable ways? If (as seems likely) different analyses show different sensitivities to family structure, then a likely result would be that different methods are used with different subsets of the original data. How can these differences be accommodated in trying to integrate results to draw population-level inferences about key evolutionary and ecological processes?

These are complex issues, and simple answers are not likely to emerge soon. In the meantime, it would be prudent for researchers tempted to try this approach to take some guidance from the Hippocratic Oath: “first, do no harm.” That is, the first objective should be to ensure that any data manipulations do not make problems associated with family structure worse.

Finally, all of the above analyses related to allele frequency assume that the goal is to accurately estimate the "true" allele frequency, which is defined to be P in the entire population of N potential parents.  But what if instead one wanted to estimate a weighted parental allele frequency, with weights being proportional to each parents' contribution of offspring to the next generation?  This could be thought of as the P that characterizes the effective population that actually produces the next generation.  If one wanted to estimate that weighted parental allele frequency, which in terms of expectation can be thought of as estimating the parametric allele frequency in the offspring generation, the best strategy would be to take a large, random sample of progeny and weight all individuals equally, regardless of family structure.  So, the two perspectives about what quantity we should be trying to estimate lead to different conclusions regarding handling of siblings.  Can we say one perspective is right and the other wrong?

**References (pending)**

Ackerman, MW, BK Hand, RK Waples, G Luikart, RS Waples, C Steele, BA Garner, J McCane, and M Campbell. In press. Effective number of breeders from sibship reconstruction: empirical evaluations using hatchery steelhead. Evolutionary Applications.

Anderson EC, Dunham KK. 2008. The influence of family groups on inferences made with the program Structure. Molecular Ecology Resources 8:1219\_1229

Ballloux 2004.

Chakraborty, R., and Leimar, O. 1987. Genetic variation within a subdivided population. *In* Population genetics and fishery management. *Edited by* N. Ryman and F. Utter. University of Washington Press, Seattle, Wash. pp. 89–102.

Goldberg CS, Waits LP. 2010. Quantification and reduction of bias from sampling larvae to infer population and landscape genetic structure. Molecular Ecology Resources 10:304\_313

Hansen MM, Nielsen EE, Mensberg KLD. 1997. The problem of sampling families rather than populations: relatedness among individuals in samples of juvenile brown trout Salmo trutta L. Molecular Ecology 6:469\_474

Hauser, L, & Carvalho, G.R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish and Fisheries 9, 333–362.

Hedgecock, D. (1994). Does variance in reproductive success limit effective population size of marine organisms? *In:* Beaumont A, editor. Genetics and evolution of aquatic organisms. London: Chapman & Hall. p. 122–134.

Hedgecock, D. & Pudovkin, A.I. (2011). Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bulletin of Marine Science 87, 971-1002.

Hedrick, P. (2005). Large variance in reproductive success and the Ne/N ratio. Evolution 59, 1596–1599.

Hess, J.E., Campbell, N.R., Docker, M.F., Baker, C., Jackson, A., Lampman, R., McIlraith, B., Moser, M.L., Statler, D.P., Young, W.P. and Wildbill, A.J., 2015. Use of genotyping by sequencing data to develop a high‐throughput and multifunctional SNP panel for conservation applications in Pacific lamprey. *Molecular ecology resources*, *15*(1), pp.187-202.

Linløkken, A.N., Haugen, T.O., Mathew, P.K., Johansen, W. and Lien, S., 2016. Comparing estimates of number of breeders Nb based on microsatellites and single nucleotide polymorphism of three groups of brown trout (Salmo trutta L.). *Fisheries Management and Ecology*, *23*(2), pp.152-160.

Nei 1973.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

T RODRÃGUEZ-RAMILO, S.I.L.V.I.A., Toro, M.A., Wang, J. and Fernandez, J., 2014. Improving the inference of population genetic structure in the presence of related individuals. *Genetics research*, *96*, e003

Rodríguez-Ramilo, S. T. & Wang, J. (2012). The effect of close relatives on unsupervised Bayesian clustering algorithms in population genetic structure analysis. Molecular Ecology Resources 12, 873–884.

Tallmon DA, Koyuk A, Luikart G, Beaumont MA (2008) OneSamp: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, 8, 299–301.

Wang J (2009) A new method for estimating effective population size from a single sample of multilocus genotypes. *Molecular Ecology*, 18, 2148–2164.

Waples RS, Do C (2008) *LdNe*: A program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8:753-756.

Waples, R.S., and R.K. Waples. 2011. Inbreeding effective population size and parentage analysis without parents. Molecular Ecology Resources 11 (Suppl. 1):162–171.

Wright 1938.

Wright, S. 1978. Evolution and the genetics of populations. Vol. IV. Variability within and among natural populations. University of Chicago Press, Chicago, Ill.

Table 1. Scenarios considered in the simulations. FS = full siblings; HS = half siblings.

[Note: I will re-number the elements in this table before submission)

Sibling

Scenario N Ne S Mating Sampling Removalf

---------------------------------------------------------------------------------------------

D 20 20.5 100 Monogamy Random FS

J 20 20.5 100 Monogamy cF9 FS

L 40 40.5 40 Monogamy Random FS

M 40 40.5 40 Monogamy F9 FS

A 40 40.5 100 Monogamy F9 FS

B 40 40.5 100 Monogamy dF3 FS

C 40 40.5 100 Monogamy Random FS

N 100 100.5 40 Monogamy Random FS

O 250 250.5 50 Monogamy Random FS

G 500 500.5 40 Monogamy Random FS

K 500 500.5 40 Monogamy F9 FS

P 500 500.5 100 Monogamy Random FS

H 40 a21.2 100 bMixedF9 eRandom FS+HS

I 40 a22.0 100 bMixedF9 eRandom FS

W 100 a15.1 40 bMixedF9 eRandom FS+HS

V 100 a15.4 40 bMixedF9 eRandom FS

X 500 a17.8 40 bMixedF9 eRandom FS

Y 500 a8.5 40 gMixedF30 eRandom FS+HS

E 40 40.5 100 Random Random FS+HS

F 40 40.5 100 Random Random FS

R 40 40.5 100 Random F9 FS+HS

Q 40 40.5 100 Random F9 FS

T 100 100.5 40 Random F9 FS+HS

S 100 100.5 40 Random F9 FS

U 500 500.5 40 Random Random FS

a *Ne* < *N* in the mixed mating model because larger families were produced than with random mating

b Parameters: α (probability of each pair producing > 1 offspring) = 0.5; maximum family size = 9 full siblings

c Maximum family size = 9 full siblings

d Maximum family size = 3 full siblings

e Sampling was random but the mixed mating model produced large families

f No half siblings are produced with monogamy

g Parameters: α = 0.1; maximum family size = 30 full siblings



Figure 1. Relationship between maximum family size in the full sample and the ratio of sample size to effective population size (S/Ne). Results are shown for three different mating models and random and non-random samples.



Figure 2. Reduced sample size after sibling removal (*S*\*) as a fraction of the full sample size (*S*). Scenarios are described in Table 1.



Figure 3. Relative RMSE (*θ*) of allele frequency and genetic differentiation as a function of the percentage of siblings removed. Results are for random samples for scenarios with mating models of monogamy (Scenarios C, D, L), random (Scenario E), and mixed (Scenarios I,W,Y). Top panel: *θ* of . Bottom panel: *θ* of *FST* (note the log scale for Y axis).



Figure 4. As in Figure 3, but for non-random samples for mating models of monogamy (Scenarios A, B, J, K) and random (Scenarios Q, R). Note the log scale for Y axis in the bottom panel.



Figure 5. Top: The ratio of effective sample size (ESS) to reduced sample size (*S*\*) after removal of siblings. Values > 1 indicate that increases in RMSE from removing siblings are not as large as they would be expected to be for random samples of the same size. Note the log scale on the Y axis. Bottom: The ratio of ESS to full sample size (S). Values > 1 indicate that removing siblings increased precision of .



Figure 6. Effects of sibling removal on the ratio of  (computed using LDNe) to true *Ne*. Top panel: results based on random samples for scenarios with mating models of monogamy (Scenarios D, G, L), random mating (Scenario F), and mixed (Scenario I). Bottom panel: results based on non-random samples with monogamy mating model. Note the log scale for the Y axis in both panels.

Figures for supporting information



Figure S1. Relationship between observed and expected fractions of siblings in simulated populations. Expected frequencies were calculated from realized Ne using Equation 2. Observed frequencies were calculated as HS + 2\*FS, where HS is the combined frequency of maternal and fraternal half siblings and FS is the frequency of full siblings. Ne for the random samples is shown in Table 1. For the non-random samples, realized Ne was calculated using PWOP.



Figure S2. Proportions of all pairwise relationships that are full siblings and half siblings for three different mating models and random and non-random samples.



Figure S3. Relationship between true Ne and Ne estimated from PWOP. Results are shown for random samples where true Ne = N + 0.5. Note the log scale on both axes.



Figure S4. Relationship between  calculated from LDNe and realized *Ne* calculated using PWOP, for two scenarios using the monogamy mating model. When 100% of siblings are removed, the point estimate from PWOP is infinity (inf).



Figure S5. Relative RMSE of 1/ for random and non-random samples.