bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Grandparent inference from genetic data: The potential for parentage-based tagging programs to
2 identify offspring of hatchery strays

3

4 Thomas A. Delomas[1*] and Matthew Campbell[2]

5

6 [1]Pacific States Marine Fisheries Commission / Idaho Department of Fish and Game, Eagle Fish
7 Genetics Laboratory, 1800 Trout Rd, Eagle, ID 83616

8

9 [2]Idaho Department of Fish and Game, Eagle Fish Genetics Laboratory, 1800 Trout Rd, Eagle,
10 ID, 83616

11

12 *Corresponding author: thomas.delomas@usda.gov; current address: USDA Agricultural
13 Research Service, National Cold Water Marine Aquaculture Center, 483 CBLS, 120 Flagg Road,
14 Kingston, RI, 02881

15

16 Running head: Grandparent-grandchild trio inference

17

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18 Abstract

19 Fisheries managers routinely use hatcheries to increase angling opportunity. Many
20 hatcheries operate as segregated programs where hatchery-origin fish are not intended to spawn
21 with natural-origin conspecifics in order to prevent potential negative effects on the natural-
22 origin population. Currently available techniques to monitor the frequency with which hatchery-
23 origin strays successfully spawn in the wild rely on either genetic differentiation between the
24 hatchery- and natural-origin fish or extensive sampling of fish on the spawning grounds. We
25 present a method to infer grandparent-grandchild trios using only genotypes from two putative
26 grandparents and one putative grandchild. We developed estimators of false positive and false
27 negative error rates and showed that genetic panels containing 500 - 700 single nucleotide
28 polymorphisms or 200 - 300 microhaplotypes are expected to allow application of this technique
29 for monitoring segregated hatchery programs. We discuss the ease with which this technique can
30 be implemented by pre-existing parentage-based tagging programs and provide an R package
31 that applies the method.

32

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

33 [A]Introduction

34 Fisheries managers have long used hatcheries to increase angling opportunity and to
35 compensate for anthropogenic impacts that have decreased population sizes (Waples et al. 2007).
36 In some situations, it has been observed that hatchery-origin fish have lower fitness in the wild
37 relative to natural-origin conspecifics, potentially due to selection in the hatchery environment or
38 the use of hatchery strains that are not locally adapted (Ford 2002; Miller et al. 2004; Araki et al.
39 2007a, 2007b, 2008; Christie et al. 2014). In order to prevent negative effects of hatchery- and
40 natural-origin fish interbreeding, it has been recommended that hatcheries operate as either
41 integrated or segregated programs (Hatchery Scientific Review Group 2009). Integrated
42 programs aim to balance the proportion of natural-origin fish in the hatchery broodstock and the
43 proportion of hatchery-origin fish spawning naturally to minimize the effect of domestication
44 (Goodman 2005). Segregated programs intend to minimize gene flow between hatchery- and
45 natural-origin populations (Mobrand et al. 2005; Hatchery Scientific Review Group 2009). These
46 strategies, developed in the context of Pacific salmon hatcheries, represent alternative
47 approaches to both protect natural-origin populations and provide harvest opportunities for
48 anglers.

49 In order to evaluate the efficacy of segregated hatchery programs, a method of
50 monitoring gene flow from hatchery-origin fish to nearby natural-origin populations is needed.
51 This has been previously estimated in several ways. Some studies estimate the proportion of fish
52 on the spawning grounds that are of hatchery origin (pHOS) through observation of marks and/or
53 tags present on hatchery fish (Dauer et al. 2009; Tattam and Ruzycki 2020). This metric is
54 important as it demonstrates the potential for competition between hatchery- and natural-origin
55 fish on the spawning grounds and suggests that hatchery introgression may be occurring.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

56 Alternatively, if the hatchery- and natural-origin populations show sufficient genetic
57 differentiation, migration rates can be estimated from temporal samples (van Doornik et al.
58 2013), or genetic structure of hatchery- and natural-origin populations can be evaluated to
59 determine whether the pattern indicates hatchery introgression (Matala et al. 2012; Ozerov et al.
60 2016; Lehnert et al. 2020). All of these approaches have drawbacks. Observing the proportion of
61 hatchery-origin fish on the spawning grounds, while suggestive, does not directly assess gene
62 flow as the reproductive success of hatchery-origin fish is unknown. Methods utilizing genetic
63 differentiation are not applicable to cases without sufficient differentiation between the hatchery-
64 and natural-origin populations, which is common where hatchery stocks were derived from
65 nearby natural-origin populations.

66 An alternative technique uses genetic samples to infer relationships between hatchery
67 broodstock and individuals sampled in the wild. Hatchery broodstock can be genetically sampled
68 at the time of spawning, and their genotypes later used to infer whether a given fish is a
69 descendent of hatchery broodstock. Parentage-based tagging (PBT) uses this approach to identify
70 offspring of the hatchery broodstock for monitoring and management of hatchery stocks
71 (Anderson and Garza 2005, 2006). Parentage-based tagging has been implemented and validated
72 on large and small scales for a variety of species (DeHaan et al. 2008; Denson et al. 2012; Evans
73 et al. 2018; Bingham et al. 2018; Campbell et al. 2019; Vandeputte et al. 2021), most notably
74 Pacific salmonids (Steele et al. 2013, 2019; Beacham et al. 2019).

75 With appropriate methods for statistical inference, the general approach of PBT can be
76 extended to identify grandchildren of hatchery broodstock. Genetic samples can be taken from
77 hatchery broodstock, and samples from natural-origin fish can later be assessed to determine if
78 they are grandchildren of those broodstock (and therefore had an unsampled, hatchery-origin

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

79 parent). The relationship being inferred is a grandparent-grandchild trio consisting of one
80 grandchild and two grandparents on the same side (i.e., either both maternal or both paternal
81 grandparents). The other two grandparents and the parents are unsampled and therefore have
82 unknown genotypes (Figure 1). While comprehensive sampling in the hatchery is
83 straightforward, similar sampling of adults spawning naturally is often logistically prohibitive.
84 The ability to infer recent hatchery ancestry without sampling naturally spawning parents would
85 overcome this issue.

86 Methods have previously been developed for inferring grandparent-grandchild
87 relationships, but these methods are not optimal for inferring grandparent-grandchild trios.
88 Letcher and King (2001) describe a method for inferring the relationship between all four
89 grandparents and a grandchild. To identify offspring of hatchery strays, this method requires
90 sampling natural-origin grandparents as well as hatchery broodstock which is not feasible in
91 many situations. Christie et al. (2011) describe a method to infer grandparent-grandchild trios by
92 identifying trios with no observed Mendelian incompatibilities (or less than a specified number).
93 While this method has been successfully applied to situations where one parent is known
94 (Christie et al. 2011; Sard et al. 2016), the lower power of exclusionary methods compared to
95 likelihood-based methods (Anderson and Garza 2006) make exclusionary methods less feasible
96 when no parents are sampled.

97 We here report a likelihood-based method to identify grandparent-grandchild trios using
98 genotypes. Additionally, we develop techniques to estimate assignment error rates for this
99 method and provide an R package implementing the described techniques
100 (https://github.com/delomast/gRandma).
101

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

102 [A]Methods

103 [C]*Inferring grandparent-grandchild relationships*.—

104 Relationships have previously been inferred from genetic data using likelihood ratios to
105 compare the putative relationship (e.g. parent-offspring) to an alternative relationship (typically
106 that the individuals are unrelated) (Marshall et al. 1998; Anderson and Garza 2006; Kalinowski
107 et al. 2007; Anderson 2012). Methods for calculating likelihoods of relationships (simple
108 pedigrees) have previously been detailed (Thompson 1976, 2000; SanCristobal and Chevalet
109 1997; Anderson and Garza 2006), and we extended this approach specifically to grandparent-
110 grandchild trios. Given allele frequencies for a locus under Hardy-Weinberg equilibrium (HWE),
111 the likelihood of the three individuals being unrelated was calculated as the product of the
112 probability of sampling each genotype from the population. The likelihood of the individuals
113 being a grandparent-grandchild trio was calculated by utilizing allele frequencies and the laws of
114 Mendelian inheritance. Genotyping error was accounted for by marginalizing the true genotypes
115 utilizing estimates of genotyping error (Anderson and Garza 2006). This allows genotyping error
116 to be flexibly modeled in any way that yields, for a given true genotype, the probability of
117 observing each genotype. With the R package we provide, users can specify these probabilities or
118 utilize a default error model. The default error model, used in the analyses described in this
119 study, is described in supplementary file 1. Loci were considered to be independent, as in
120 previous methods of relationship inference and pedigree reconstruction (Marshall et al. 1998;
121 Riester et al. 2009; Jones and Wang 2010; Anderson 2012; Huisman 2017).

122 Calculating likelihoods for large numbers of possible trios can be computationally
123 intensive. We therefore implemented a preliminary screen based on the observed number of
124 Mendelian incompatibilities (MI) in a grandparent-grandchild trio. Trios with more than m MIs

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

125 were excluded from consideration. The value of m was chosen so that the probability of
126 excluding a true grandparent-grandchild trio was less than 0.0001 for a trio with no missing
127 genotypes. With a constant value of m, the probability of rejecting a trio with one or more
128 missing genotypes is smaller, as a locus with a missing genotype cannot be considered a
129 Mendelian incompatibility. The probability of rejecting a true trio given a value of m was
130 calculated by representing the number of MIs as a Markov Chain, following the method detailed
131 by Anderson (2012) but extended to the relationship of grandparent-grandchild trio.
132 Assigning grandparent-grandchild relationships can be treated as a hypothesis test by
133 comparing the calculated log-likelihood ratio (LLR) to a critical value, c (SanCristobal and
134 Chevalet 1997; Anderson and Garza 2006). If the LLR was greater than or equal to c, the trio
135 was considered related; otherwise, the trio was considered unrelated. The value of c can be
136 chosen to achieve a desired balance of false negative and false positive error rates.
137
138 [C]*False positive error rates*.—
139 Per-comparison false positive error rates (probability of assigning a relationship to a trio
140 that is not a grandparent-grandchild trio) are dependent upon the true relationship for a trio. We
141 have implemented methods to assess false positive error rates for unrelated trios (typically the
142 most important) as well as 13 other types of relationships. In these relationships, the two putative
143 grandparents were considered unrelated to each other, but the putative grandparents had different
144 relationships to the putative grandchild. The relationships considered represented combinations
145 of true grandparents, individuals unrelated to the putative grandchild, great-aunts, half-great-
146 aunts, and first cousins of the putative grandchild's true grandparent.
147 Estimating the per-comparison false positive error rate for a given value of c has been

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

148 demonstrated for parent-offspring and sibling relationships using importance sampling
149 (Anderson and Garza 2006; Baetscher et al. 2018), a Monte Carlo variance reduction technique.
150 Variance reduction is needed as a naïve Monte Carlo approach would be inefficient at estimating
151 very small false positive rates (Anderson and Garza 2006). Small error rates can be meaningful
152 as the experiment-wide false positive error rate is estimated by the product of the per-comparison
153 false positive error rate and the number of trios (with the corresponding true relationship)
154 considered. We extended this approach to the current application of assessing grandparent-
155 grandchild trios. We also implemented an alternative method utilizing stratified sampling,
156 another Monte Carlo variance reduction technique, because specific implementations of
157 importance sampling can produce unreliable results (Owen 2013).
158
159 [C]*Importance sampling*.—
160 A standard Monte Carlo estimator of false positive rates would be obtained by simulating
161 genotypes for trios of a given relationship (such as unrelated) and recording how many fit the
162 criteria to be considered related. Importance sampling can be thought of as focusing the
163 simulation on producing mostly genotypes that do assign and then correcting for this
164 modification. To implement importance sampling, we simulated genotypes from the distribution
165 of genotypes in true grandparent-grandchild trios. Missing genotypes were accounted for
166 utilizing the forward-backward algorithm described below with the state of the Markov chain
167 representing the number of missing genotypes in one individual. If the simulated trio had fewer
168 than m MIs and the calculated LLR was greater than or equal to c, the observation was recorded
169 as a false positive with the appropriate importance sampling weight (Owen 2013).
170

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

171 [C]*Stratified sampling*.—

172 Similar to importance sampling, the goal of stratified sampling was to focus simulation
173 effort on categories (strata) that produce false positives. We stratified the distribution of trio
174 genotypes by the number of observed MIs. This was a natural choice because the algorithm we
175 use explicitly filters possibilities based on the number of observed MIs. Therefore, we can
176 eliminate simulating genotypes for most strata. Sampling effort can then be focused on trios with
177 m or fewer MIs. This method requires calculating the probability that a trio of given relationship
178 has a given number of observed MIs and the ability to simulate genotypes for a trio given a
179 relationship and number of MIs. Genotypes for trios were simulated in each stratum and the false
180 positive rates were recorded. Utilizing the probabilities that a trio has each number of MIs (i.e.,
181 the size of the strata), the overall false positive rate was then calculated.

182 To calculate the probability a trio has a given number of observed MIs, we represented
183 the observation of MIs and missing genotypes as a Markov chain and utilized the forward step of
184 the forward-backward algorithm. We extended the approach described by Anderson (2012) to
185 account for the common practice of only analyzing samples given a maximum number of
186 missing genotypes, d, and to fit the target relationships. The value of d in the current analyses
187 was 10% of loci. Let s_i be the state, describing the number of observed MIs and missing
188 genotypes, after observing locus i. Prior to observing any loci, $s_0 = (0,0,0,0)$, representing the
189 number of observed incompatibilities and number of missing genotypes for the three individuals.
190 Let a_i be a vector indicating whether an incompatibility or any missing genotypes are observed
191 at locus i, in the same order as s_i. Assuming HWE, known allele frequencies, known locus-
192 specific probabilities of a genotype being missing, and observation of missing genotypes is
193 independent across loci and individuals, the probabilities of each possibility for a_i can be

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

194 calculated according to standard probability arguments for a given true relationship. The
195 probability of being in state x after a given locus can then be calculated as

$$\mathrm{P}(\mathrm{s}_{\mathrm{i+1}} = x) = \mathrm{P}(\mathrm{s}_{\mathrm{i}}) \sum_{\mathrm{a}} \mathrm{P}(\mathrm{a}_{\mathrm{i+1}}) \mathrm{I}\{\mathrm{s}_{\mathrm{i}} + \mathrm{a}_{\mathrm{i+1}} = x\}. \quad (1)$$

196 This can be evaluated recursively to obtain the probabilities of each final state. The probability a
197 trio has a given number of observed MIs conditional on d can then be calculated.

198 However, memory constraints make saving all the probabilities at each step impractical
199 even with moderate numbers of loci. We can save only the probabilities of states with m or fewer
200 MIs and with all three individuals having d or fewer missing genotypes to reduce memory usage.
201 These values are divided by the probability that all members have d or fewer missing genotypes
202 to obtain probabilities of being in each state conditional on d. The probability of observing an
203 individual with more than d missing genotypes can be obtained through the same algorithm
204 (forwards step), but with s and a now only representing missing genotypes in one individual.
205 Because we assumed that missing genotypes are independent between individuals (given locus-
206 specific rates of missing genotypes), the probability of all three samples having d or fewer
207 missing genotypes is straightforward to calculate using the obtained probability that one
208 individual has more than d missing genotypes.

209 To utilize stratified sampling, we need to simulate genotypes for trios with a specified
210 number of MIs. The backwards step of the forward-backward algorithm fills this need. Given L
211 loci, a value of s_L is chosen given the number of MIs by sampling a categorical distribution with
212 probabilities proportional to the probability of each s_L that has the specified number of MIs (and
213 allowable number of missing genotypes). Next, for each locus and iterating backwards, a value
214 for a_i is chosen by sampling a categorical distribution with

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$P(a_i|s_{i+1}) \propto P(a_i)P(s_i = s_{i+1} - a_i). \quad (2)$$

215 Once a_i is chosen, genotypes are sampled using the genotype frequencies for the true
216 relationship calculated from the allele frequencies, HWE, laws of Mendelian inheritance, and
217 genotyping error rates. If all three genotypes are observed (i.e., no missing genotypes in the
218 chosen a_i), then the genotypes are either sampled conditional upon an MI being present or not.
219
220 [C]*False negative error rate*.—
221 To estimate the per-comparison false negative error rate (probability of failing to assign a
222 relationship to a true grandparent-grandchild trio) for a given value of *c*, Monte Carlo methods
223 have been previously used for other relationships (Anderson and Garza 2006; Baetscher et al.
224 2018) and we adopted this strategy here. Genotypes of grandparent-grandchild trios are
225 simulated and loci with missing genotypes for each individual are chosen by representing the
226 observation of missing genotypes as a Markov chain and utilizing the forward-backward
227 algorithm, as described above with the state representing the number of missing genotypes in one
228 individual. Log-likelihood ratios were calculated for the simulated genotypes, and the proportion
229 of trios with the number of MIs greater than *m* or LLR less than *c* was the estimate of the false
230 negative rate.
231
232 [C]*Panel size simulations and error rate estimator evaluation*.—
233 A key question for designing experiments implementing this method is, how many loci
234 need to be genotyped to obtain reliable assignments? To help answer this question, we estimated
235 error rates for panels of different size. We simulated panels containing 100, 300, 500, 700, and
236 900 biallelic single nucleotide polymorphisms (SNPs) and panels containing 100, 200, 300, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

237 400 triallelic microhaplotypes. The SNPs and microhaplotypes were simulated with expected
238 heterozygosity of 0.22 (allele frequencies of 0.125 and 0.872) and 0.42 (allele frequencies of
239 0.13, 0.13, and 0.74), respectively. This choice reflects the mean expected heterozygosities for
240 SNPs and microhaplotypes in a comparison of both marker types for relationship inference in
241 rockfish (Baetscher et al. 2018). The power of a given marker for kinship inference is directly
242 related to its variability (Anderson and Garza, 2006). While we investigated marker panels of
243 differing size but fixed variability, similar power to the simulated panels could be achieved with
244 fewer, more variable markers. The probability of a missing genotype at a locus was set at 3% in
245 these analyses. It is common practice to remove any samples with more than a threshold number
246 of missing genotypes, and so we restricted all simulated genotypes to have 90% or more
247 genotypes present. Given a true relationship of unrelated, false negative and false positive error
248 rates were estimated for a range of c values and the relationship between error rates was
249 compared between panels. Integer values of c were chosen starting at 0 and increasing until the
250 estimated false negative rate was above 0.05. Estimates of false negative rates and estimates
251 from the importance sampling method were derived from 10,000 and 1,000,000 Monte Carlo
252 iterations, respectively. Estimates from the stratified sampling routine were derived from
253 1,000,000 iterations for each stratum (number of observed Mendelian incompatibilities) less than
254 or equal to m.

255 To compare performance of the two methods for estimating false positive error rates, we
256 compared the estimated error rates for unrelated trios between the methods with all four
257 simulated microhaplotype panels. Estimates for other true relationships were compared using the
258 300 locus microhaplotype panel. Finally, to examine the importance of modelling the presence of
259 missing genotypes, we compared importance sampling estimates using the 300 locus

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

260 microhaplotype panel with the probability of a genotype being missing equal to 3% and 0%.

261 The scripts used to perform these simulations and their outputs are available at

262 https://github.com/delomast/gpError2021.

263

264 [C]*Example analyses*.—

265 To fully evaluate false positive per-comparison error rates, one needs to have a general

266 idea of the size of the analysis (number of comparisons) being attempted. We consider two

267 examples modeled around steelhead *Oncorhynchus mykiss* hatchery programs in the Snake River

268 basin using data collected during 2018. The Upper Salmon B-run (USB) represents a smaller

269 hatchery program and spawned 66 steelhead in 2018. The Dworshak National Fish Hatchery

270 (DNFH) represents a larger hatchery program and spawned 1,778 steelhead in 2018. Both of

271 these hatchery programs take genetic samples from all broodstock and record the day of

272 spawning, phenotypic sex, and crosses made. Data collected at the hatchery (or a genetic sex

273 marker) can be used to constrain the number of possible pairs of grandparents considered in an

274 analysis. For each hatchery program, we consider the effect on the desired per-comparison false

275 positive rate of using no data, phenotypic sex, spawn day, phenotypic sex and spawn day, or

276 cross records.

277 In the example analysis, we assume that natural-origin juveniles are sampled and that

278 exact age is unknown but is constrained to one, two, or three years old. The effect of this

279 assumption is that three potential years of parents must be considered for each juvenile.

280 Hatchery-origin steelhead in the Snake River basin return almost exclusively as three and four

281 year old fish (Warren et al. 2017), so this translates to four years of potential grandparents that

282 must be considered.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

283 The total number of comparisons in these analyses is the product of the number of
284 possible pairs of grandparents per year, the number of years of potential grandparents being
285 considered (four), and the number of potential grandchildren evaluated (assumed here to be 200).
286 We then calculate the desired per-comparison false positive (true relationship of unrelated) error
287 rate to achieve an expected number of false positive assignments of 0.1 (0.1 / number of
288 comparisons) assuming all trios are unrelated. This ignores false positives arising from trios of
289 other relationships. In some situations, error rates for alternative relationships are important to
290 consider, but in analyses of segregated hatchery programs (where no breeding of hatchery-origin
291 and natural-origin fish is desired), false positives arising from trios of other relationships are not
292 necessarily harmful as the purpose is to identify fish with recent hatchery-origin ancestry.
293
294 [A]Results
295 Estimated error rates declined with increasing panel size, and the microhaplotype panels
296 showed lower error rates than SNP panels of similar size (Figure 2). For example, false positive
297 error rates (true relationship of unrelated) below 10^{-10} were achieved at false negative error rates
298 below 0.05 with SNP panels containing 700 or more loci and microhaplotype panels containing
299 300 or more loci. False positive rates for related trios decreased with decreasing relatedness
300 between individuals in the trio (Supplemental Figure 1).
301 False positive rates estimated by importance sampling and stratified sampling were
302 practically identical when estimates were above approximately 10^{-8} (Figure 3). As the false
303 positive rate decreased below this, the importance sampling method estimated false positive rates
304 higher than the stratified sampling method. In these cases, the stratified sampling method
305 estimated false positive rates of 0 within one or more strata (i.e., no false positives were sampled

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

306 out of the 1,000,000 iterations). Similar results were obtained for false positive rates estimated
307 for trios with relationships other than unrelated (Figure 4) except for two trio types (true
308 grandparent and unrelated; true grandparent and cousin of grandparent) that had noticeably
309 different estimates between the two methods and did not have a low false positive rate.

310 Incorporating a 3% missing genotype rate in the error rate estimation had a moderate
311 effect on the results (Figure 5). For example, at a false negative rate of 0.05, the false positive
312 (true relationship of unrelated) error rates were approximately (derived from linear interpolation
313 with neighboring points) $1.3 \cdot 10^{-11}$ and $2.1 \cdot 10^{-12}$ when missing genotypes had rates of 3% and
314 0%, respectively.

315 The example hatchery programs considered show that analysis of smaller programs (e.g.,
316 USB) require per-comparison false positive rates on the order of 10^{-6} - 10^{-8} depending on what
317 data, if any, are available to reduce the number of comparisons (Table 1). Larger programs,
318 similar to DNFH, require rates on the order of 10^{-7} - 10^{-10} if the number of comparisons can be
319 reduced by one of the data sources considered. Error rates estimated for the simulated panels
320 (Figure 2) suggest that these rates can be achieved with panels containing 500 - 700 SNPs or 200
321 - 300 microhaplotypes. If no data are available to reduce the number of pairs of grandparents that
322 must be considered, then hatchery programs similar in size to DNFH (1,700 fish spawned / year)
323 will require error rates on the order of 10^{-11}, which is achievable with panels containing 700 -
324 900 SNPs or 300 - 400 microhaplotypes.

325

326 [A]Discussion

327 The methods developed here facilitate direct monitoring of introgression between
328 hatchery- and natural-origin populations without relying on genetic differentiation. This fills an

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

329 unmet need for monitoring the segregated hatchery programs upon which many fisheries rely.
330 The method identifies fish with an unsampled, hatchery-origin parent using genetic samples from
331 the hatchery broodstock and estimates error rates given allele frequencies for a population.
332 Another benefit of this method is the ease with which it can be implemented through
333 existing PBT programs because the hatchery and laboratory processes required to sample and
334 genotype broodstock are already in place. Additional requirements for grandparent inference
335 would be the development of a genetic panel with sufficient power (given the size of the relevant
336 hatcheries) and collection of samples from natural-origin fish that are potential grandchildren of
337 the hatchery broodstock. In the past decade, there have been multiple, large-scale demonstrations
338 of the efficacy of PBT for monitoring fisheries (Steele et al. 2013, 2019; Beacham et al. 2019).
339 Numerous PBT programs have been recently reported (Evans et al. 2018; Bingham et al. 2018;
340 Campbell et al. 2019; Vandeputte et al. 2021) and current tagging technologies can be replaced
341 by PBT to provide additional benefits at similar or reduced costs (Beacham 2021). This implies
342 that PBT will continue to grow in usage, and the method described here will become feasible for
343 a larger number of hatchery programs.
344 The simulated microhaplotype panels demonstrated that error rates low enough for
345 relatively large analyses of segregated hatchery programs are achievable with panels containing
346 300 - 400 microhaplotypes. The size/power required from the genetic marker panel will depend
347 on the data available to limit the number of comparisons. As has been previously demonstrated,
348 the availability of accurate hatchery cross-records greatly reduces the power required for
349 grandparent inference (Letcher and King 2001), but we note that simply having sex information
350 associated with a broodstock individual's genotype can have a sizeable effect (Table 1). Panels
351 genotyping hundreds of loci have been created using cost-effective, amplicon sequencing

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

352 techniques (Campbell et al. 2015; Janowitz-Koch et al. 2019), and so we conclude that
353 grandparent inference with this method is feasible using current genotyping techniques.
354 Application of this method at larger scales will require considering potential grandparents
355 from multiple hatcheries. In these situations, having data available to reduce the number of
356 comparisons can be critical for making an analysis feasible. For example, if natural-origin
357 steelhead are sampled at Lower Granite Dam on the Snake River (Hargrove et al. 2021), then all
358 steelhead hatcheries adjacent to and upstream of the dam must be considered (approximately
359 5,000 total steelhead spawned annually). With no data other than the year and hatchery at which
360 fish were spawned, the number of potential pairs of grandparents would be approximately $5.5 \cdot$
361 10^6 each year, while considering phenotypic (or genetic) sex would reduce this to $1.4 \cdot 10^6$
362 (IDFG, unpublished).
363 We expect this method to be most applicable for assessing segregated hatchery programs.
364 For some closely related trios, false positive error rates were relatively high with the 300
365 microhaplotype panel (Supplemental Figure 1). For all analyses, the impact of these error rates is
366 mediated by the infrequency of comparisons involving closely related individuals. When
367 analyzing segregated programs, false positives from closely related trios may not negatively
368 impact conclusions as they indicate recent hatchery ancestry. However, when analyzing
369 integrated hatchery programs, distinguishing trios with different relationships can be important.
370 Application of this method could then either be infeasible or require a more powerful genetic
371 panel. Additionally, the number of relationships for which we provide estimators covers trios
372 with a range of relatedness but is not exhaustive. In some specific cases, other relationships may
373 be present at impactful frequencies. For example, if generations overlap substantially then the
374 impact of trios containing aunts, half-aunts, and first-cousins may need to be considered.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

375 The method developed here addressed shortcomings of previously developed methods for
376 grandparent inference with moderately sized genetic panels. Previous methods required either
377 that all four grandparents were sampled (Letcher and King 2001) or used an exclusionary method
378 that inherently had lower power than likelihood-based methods (Christie et al. 2011).
379 Additionally, the exclusionary method did not provide a formal treatment of genotyping error.
380 Applications of the exclusionary method (Christie et al. 2011; Sard et al. 2016) have utilized
381 panels of tens of microsatellites. When using SNPs or microhaplotypes, hundreds or thousands
382 of loci are typically genotyped and minimal error rates (e.g., 1%) still result in errors being
383 present in a majority of individuals. An additional effect of ignoring genotyping error in an
384 exclusionary method is that false negatives are implicitly assumed not to occur. It is worth noting
385 that use of a Markov chain to model the number of observed Mendelian incompatibilities and
386 missing genotypes in a trio, as we developed here, could be incorporated into the exclusionary
387 method. This would allow both the formal incorporation of genotyping error and estimation of
388 false positive and false negative error rates.
389 Importance sampling and stratified sampling have different weaknesses, making each
390 better suited to different situations. For example, importance sampling can perform suboptimally
391 when the estimate is dominated by a small fraction of samples (Owen 2013). Stratified sampling
392 does not have this same drawback, but it can fail to give a meaningful estimate when the false
393 positive rate within a particular stratum is low enough that a reasonable number of samples
394 cannot estimate it accurately. Comparison of the importance sampling and stratified sampling
395 methods showed that when the true relationship was unrelated, they estimated essentially the
396 same false positive rate when that rate was above approximately 10^{-8}. This suggests the
397 importance sampling routine performed well for unrelated trios. At lower false positive rates, the

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

398 stratified sampling estimates were lower (and in some cases 0) than those from importance
399 sampling. In these cases, one or more strata had estimates of 0, indicating that the 1,000,000
400 samples taken were not sufficient to observe one or more false positives. This demonstrates one
401 of the strengths of importance sampling compared to stratified sampling - some situations will
402 have false positive rates small enough they cannot be efficiently estimated by this stratified
403 sampling routine. This is further emphasized by the greater computational effort devoted to
404 stratified sampling in this study (1,000,000 iterations in each stratum vs. 1,000,000 iterations
405 total).

406 For false positive error rates under relationships other than unrelated, the importance
407 sampling and stratified sampling estimates were again largely the same. In a few cases where one
408 potential grandparent was the true grandparent and the other was not, the estimates were
409 noticeably different and the stratified sampling method achieved at least 150 false positive
410 observations in each strata. This suggests that for these relationships the importance sampling
411 method may have performed suboptimally. Similar observations were made for importance
412 sampling estimates of false positive error rates in parentage inference, where performance of a
413 given importance distribution varied depending on the true relationship for which an error rate
414 was estimated (Anderson and Garza 2006). One strategy would be to design an alternative
415 importance distribution, but for closely related trios, error rates for most panels will likely be
416 high enough that they are amenable to stratified sampling.

417 The current method assumes loci are in linkage equilibrium. If loci are physically linked
418 (but still in equilibrium) the methods described here can be applied with some additional
419 consideration. Physical linkage of two loci will result in grandchildren being more likely to
420 inherit alleles at both loci from one of the two grandparents in a trio. The estimated false positive

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

421 (when true relationship is unrelated) error rates are not affected because unrelated individuals are
422 not impacted by physical linkage. For the other estimated error rates (false positives for trios
423 with other relationships and false negatives), the simulated LLRs will have lower variance than
424 the true distribution, causing error rates to be underestimated.

425 One drawback of the method as implemented is that computational efficiency decreases
426 with increasing numbers of alleles per locus. This is partly due to the flexibility of the
427 genotyping error model and the need to marginalize over all possible true genotypes. In the
428 current implementation, computation is sped up by precomputing likelihood values for observed
429 trio genotypes at each locus. This works well when the number of alleles per locus is small, but
430 as the number of alleles increases, this can become impractically slow. As such, the current
431 implementation will be most suitable to panels containing biallelic SNPs and microhaplotypes,
432 which in our experience typically have three to five alleles. For panels of highly variable loci, a
433 different implementation of this method, particularly with a more streamlined genotyping error
434 model, would be necessary.

435

436 [A]Acknowledgements

437 Funding for this project was provided by Bonneville Power Administration Project 2010-031-00.
438 We would like to thank John Powell, John Hargrove, Matt Corsi, and Eric Anderson for
439 providing comments that improved this manuscript. There is no conflict of interest declared in
440 this article.

441

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

442 [A]References

443 Anderson, E. C. 2012. Large-scale parentage inference with SNPs: An efficient algorithm for
444 statistical confidence of parent pair allocations. Statistical Applications in Genetics and
445 Molecular Biology 11(5):12.
446 Anderson, E. C., and J. C. Garza. 2005. A description of full parentage genotyping. Vancouver,
447 British Columbia.
448 Anderson, E. C., and J. C. Garza. 2006. The power of single-nucleotide polymorphisms for
449 large-scale parentage inference. Genetics 172(4):2567–2582.
450 Araki, H., W. R. Ardren, E. Olsen, B. Cooper, and M. S. Blouin. 2007a. Reproductive success of
451 captive-bred steelhead trout in the wild: Evaluation of three hatchery programs in the Hood
452 River. Conservation Biology 21(1):181–190.
453 Araki, H., B. A. Berejikian, M. J. Ford, and M. S. Blouin. 2008. Fitness of hatchery-reared
454 salmonids in the wild. Evolutionary Applications 1(2):342–355.
455 Araki, H., B. Cooper, and M. S. Blouin. 2007b. Genetic effects of captive breeding cause a rapid,
456 cumulative fitness decline in the wild. Science 318(5847):100–103.
457 Baetscher, D. S., A. J. Clemento, T. C. Ng, E. C. Anderson, and J. C. Garza. 2018.
458 Microhaplotypes provide increased power from short-read DNA sequences for relationship
459 inference. Molecular Ecology Resources 18(2):296–305.
460 Beacham, T. D. 2021. Parentage-based tagging combined with genetic stock identification is a
461 cost-effective and viable replacement for coded-wire tagging in large-scale assessments of
462 Canadian salmon fisheries. Fisheries Research 239:105920.
463 Beacham, T. D., C. Wallace, K. Jonsen, B. McIntosh, J. R. Candy, D. Willis, C. Lynch, J.-S. S.
464 Moore, L. Bernatchez, and R. E. Withler. 2019. Comparison of coded-wire tagging with

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

465 parentage-based tagging and genetic stock identification in a large-scale coho salmon
466 fisheries application in British Columbia, Canada. Evolutionary Applications 12(2):230–
467 254.
468 Bingham, D. M., P. C. Gerrity, and S. Painter. 2018. Genetic tagging is an effective way to
469 monitor survival of released hatchery saugers: Conservation efforts in the Wind River,
470 Wyoming. Environmental Practice 20(4):92–103.
471 Campbell, M. R., N. V. Vu, A. P. LaGrange, R. S. Hardy, T. J. Ross, and S. R. Narum. 2019.
472 Development and application of single-nucleotide polymorphism (SNP) genetic markers for
473 conservation monitoring of burbot populations. Transactions of the American Fisheries
474 Society 148(3):661–670.
475 Campbell, N. R., S. A. Harmon, and S. R. Narum. 2015. Genotyping-in-Thousands by
476 sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon
477 sequencing. Molecular Ecology Resources 15(4):855–867.
478 Christie, M. R., M. J. Ford, and M. S. Blouin. 2014. On the reproductive success of early-
479 generation hatchery fish in the wild. Evolutionary Applications 7(8):883–896.
480 Christie, M. R., M. L. Marine, and M. S. Blouin. 2011. Who are the missing parents?
481 Grandparentage analysis identifies multiple sources of gene flow into a wild population.
482 Molecular Ecology 20(6):1263–1276.
483 Dauer, M. B., T. R. Seamons, L. Hauser, T. P. Quinn, and K. A. Naish. 2009. Estimating the ratio
484 of hatchery-produced to wild adult steelhead on the spawning grounds using scale pattern
485 analyses. Transactions of the American Fisheries Society 138(1):15–22.
486 DeHaan, P. W., G. R. Jordan, and W. R. Ardren. 2008. Use of genetic tags to identify captive-
487 bred pallid sturgeon (*Scaphirhynchus albus*) in the wild: Improving abundance estimates for

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

488 an endangered species. Conservation Genetics 9(3):691–697.
489 Denson, M. R., K. Brenkert, W. E. Jenkins, and T. L. Darden. 2012. Assessing red drum juvenile
490 stocking in a South Carolina estuary using genetic identification. North American Journal of
491 Fisheries Management 32(1):32–43.
492 van Doornik, D. M., D. L. Eddy, R. S. Waples, S. J. Boe, T. L. Hoffnagle, E. A. Berntson, and P.
493 Moran. 2013. Genetic monitoring of threatened chinook salmon populations: estimating
494 introgression of nonnative hatchery stocks and temporal genetic changes. North American
495 Journal of Fisheries Management 33(4):693–706.
496 Evans, H. K., K. B. Carlson, R. Wisser, M. E. Raley, K. M. Potoka, and K. J. Dockendorf. 2018.
497 Genetics and hatchery management: A parentage-based tagging approach to blueback
498 herring conservation. Journal of Fish and Wildlife Management 9(1):1–13.
499 Ford, M. J. 2002. Selection in captivity during supportive breeding may reduce fitness in the
500 wild. Conservation Biology 16(3):815–825.
501 Goodman, D. 2005. Selection equilibrium for hatchery and wild spawning fitness in integrated
502 breeding programs. Canadian Journal of Fisheries and Aquatic Sciences 62(2):374–389.
503 Hargrove, J. S., C. A. Camacho, W. C. Schrader, J. H. Powell, T. A. Delomas, J. E. Hess, S. R.
504 Narum, and M. R. Campbell. 2021. Parentage-based tagging improves escapement
505 estimates for ESA-listed adult Chinook salmon and steelhead in the Snake River basin.
506 Canadian Journal of Fisheries and Aquatic Sciences 78(4):349–360.
507 Hatchery Scientific Review Group. 2009. Columbia River hatchery reform system-wide report.
508 Huisman, J. 2017. Pedigree reconstruction from SNP data: parentage assignment, sibship
509 clustering and beyond. Molecular Ecology Resources 17(5):1009–1024.
510 Janowitz-Koch, I., C. Rabe, R. Kinzer, D. Nelson, M. A. Hess, and S. R. Narum. 2019. Long-

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

511 term evaluation of fitness and demographic effects of a chinook salmon supplementation
512 program. Evolutionary Applications 12(3):456–469.
513 Jones, O. R., and J. Wang. 2010. COLONY: A program for parentage and sibship inference from
514 multilocus genotype data. Molecular Ecology Resources 10(3):551–555.
515 Kalinowski, S. T., M. L. Taper, and T. C. Marshall. 2007. Revising how the computer program
516 CERVUS accommodates genotyping error increases success in paternity assignment.
517 Molecular Ecology 16(5):1099–1106.
518 Lehnert, S. J., S. M. Baillie, J. MacMillan, I. G. Paterson, C. F. Buhariwalla, I. R. Bradbury, and
519 P. Bentzen. 2020. Multiple decades of stocking has resulted in limited hatchery
520 introgression in wild brook trout (*Salvelinus fontinalis*) populations of Nova Scotia.
521 Evolutionary Applications 13(5):1069–1089.
522 Letcher, B. H., and T. L. King. 2001. Parentage and grandparentage assignment with known and
523 unknown matings: Application to Connecticut River Atlantic salmon restoration. Canadian
524 Journal of Fisheries and Aquatic Sciences 58(9):1812–1821.
525 Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton. 1998. Statistical confidence for
526 likelihood-based paternity inference in natural populations. Molecular Ecology 7(5):639–
527 655.
528 Matala, A. P., S. R. Narum, W. Young, and J. L. Vogel. 2012. Influences of hatchery
529 supplementation, spawner distribution, and habitat on genetic structure of Chinook salmon
530 in the South Fork Salmon River, Idaho. North American Journal of Fisheries Management
531 32(2):346–359.
532 Miller, L. M., T. Close, and A. R. Kapuscinski. 2004. Lower fitness of hatchery and hybrid
533 rainbow trout compared to naturalized populations in Lake Superior tributaries. Molecular

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

534 Ecology 13(11):3379–3388.
535 Mobrand, L. E., J. Barr, L. Blankenship, D. E. Campton, T. T. P. Evelyn, T. A. Flagg, C. V. W.
536 Mahnken, L. W. Seeb, P. R. Seidel, and W. W. Smoker. 2005. Hatchery Reform in
537 Washington State. Fisheries 30(6):11–23.
538 Owen, A. B. 2013. Monte Carlo theory, methods and examples.
539 Ozerov, M. Y., R. Gross, M. Bruneaux, J. P. Vähä, O. Burimski, L. Pukk, and A. Vasemägi. 2016.
540 Genomewide introgressive hybridization patterns in wild Atlantic salmon influenced by
541 inadvertent gene flow from hatchery releases. Molecular Ecology 25(6):1275–1293.
542 Riester, M., P. F. Stadler, and K. Klemm. 2009. FRANz: Reconstruction of wild multi-generation
543 pedigrees. Bioinformatics 25(16):2134–2139.
544 SanCristobal, M., and C. Chevalet. 1997. Error tolerant parent identification from a finite set of
545 individuals. Genetical Research 70(1):53–62.
546 Sard, N. M., D. P. Jacobson, and M. A. Banks. 2016. Grandparentage assignments identify
547 unexpected adfluvial life history tactic contributing offspring to a reintroduced population.
548 Ecology and Evolution 6(19):6773–6783.
549 Steele, C. A., E. C. Anderson, M. W. Ackerman, M. A. Hess, N. R. Campbell, S. R. Narum, and
550 M. R. Campbell. 2013. A validation of parentage-based tagging using hatchery steelhead in
551 the Snake River basin. Canadian Journal of Fisheries and Aquatic Sciences 70(7):1046–
552 1054.
553 Steele, C. A., M. Hess, S. Narum, and M. Campbell. 2019. Parentage-based tagging: Reviewing
554 the implementation of a new tool for an old problem. Fisheries 44(9):412–422.
555 Tattam, I. A., and J. R. Ruzycki. 2020. Smolt Transportation Influences Straying of Wild and
556 Hatchery Snake River Steelhead into the John Day River. Transactions of the American

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

557 Fisheries Society 149(3):284–297.

558 Thompson, E. 1976. Inference of genealogical structure. Social Science Information 15(2–
559 3):477–526.

560 Thompson, E. 2000. Statistical inference from genetic data on pedigrees. Institue of
561 Mathematical Statistics, Beachwood, OH.

562 Vandeputte, M., A. Bestin, L. Fauchet, J. M. Allamellou, S. Bosc, O. Menchi, P. Haffray, and C.
563 Saavedra. 2021. Can we identify wild-born salmon from parentage assignment data? A case
564 study in the Garonne-Dordogne rivers salmon restoration programme in France. Aquatic
565 Living Resources 34:7.

566 Waples, R. S., M. J. Ford, and D. Schmitt. 2007. Empirical results of salmon supplementation in
567 the northeast pacific: a preliminary assessment. Pages 383–403 *in* T. M. Bert, editor.
568 Ecological and Genetic Implications of Aquaculture Activities.

569 Warren, C., S. Rosenberger, and F. Bohlen. 2017. 2015 calendar year hatchery steelhead report:
570 IPC and LSRCP monitoring and evaluation programs for the state of Idaho. IDFG Report
571 Number 17-05

572

Hatchery	Data used	Number of possible crosses per year	Years of potential grandparents	Number of potential grandchildren	Number of comparisons	Desired per-comparison false positive error rate
USB	None	4,290	4	200	3,432,000	$2.91 \cdot 10^{-8}$
	Sex	1,088	4	200	870,400	$1.15 \cdot 10^{-7}$
	Spawn day	584	4	200	467,200	$2.14 \cdot 10^{-7}$
	Spawn day and sex	162	4	200	129,600	$7.72 \cdot 10^{-7}$
	Cross records	34	4	200	27,200	$3.68 \cdot 10^{-6}$
DNFH	None	3,134,670	4	200	2,507,736,000	$3.99 \cdot 10^{-11}$
	Sex	769,348	4	200	615,478,400	$1.62 \cdot 10^{-10}$
	Spawn day	284,270	4	200	227,416,000	$4.40 \cdot 10^{-10}$
	Spawn day and sex	69,709	4	200	55,767,200	$1.79 \cdot 10^{-9}$
	Cross records	1,003	4	200	802,400	$1.25 \cdot 10^{-7}$

1 Table 1. Desired per-comparison false positive (true relationship of unrelated) error rates for analyses using different data sources to

2 reduce the number of comparisons.

3

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Figure Legends

2 Figure 1. The pedigree the proposed method could be used to infer. H: hatchery-origin, W:
3 natural-origin

4 Figure 2. Error rates estimated by importance sampling for simulated (A) SNP and (B)
5 microhaplotype panels with varying numbers of loci.

6 Figure 3. Comparison of false positive (true relationship of unrelated) error rates for simulated
7 microhaplotype panels estimated by importance sampling and stratified sampling. The black line
8 represents y = x. Points shown on the y-axis had an estimated error rate of 0 from stratified
9 sampling.

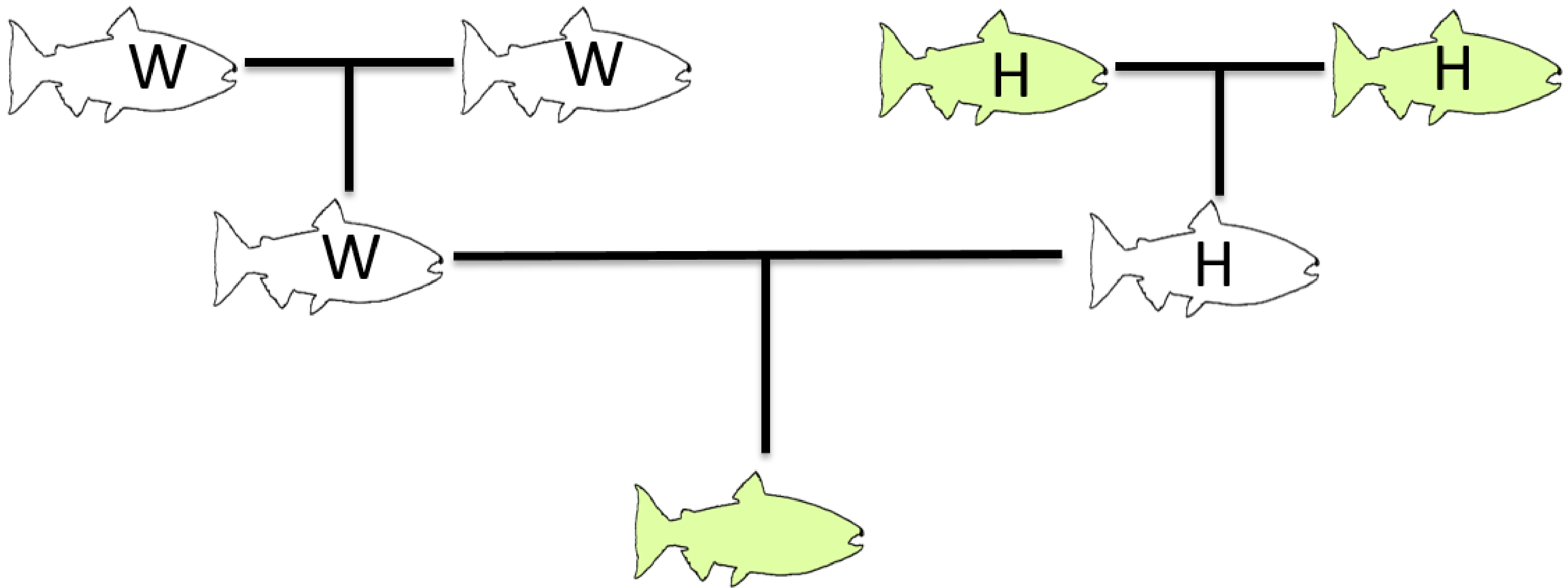
10 Figure 4. Comparison of false positive error rates estimated by importance sampling and
11 stratified sampling for related trios with the simulated 300 locus microhaplotype panel. The
12 black line represents y = x. The two putative grandparents were unrelated to each other. The
13 relationship labels indicate the relationship of the two putative grandparents to the putative
14 grandchild. Where the label is only one relationship, both putative grandparents have the same
15 relationship to the putative grandchild. True is true grandparent, GAunt is great-aunt, HGAunt is
16 half great-aunt, GpCous is first-cousin to the putative grandchild's true grandparent, Unrel is
17 unrelated. The inset gives a magnified view of the region containing error rates close to 1.

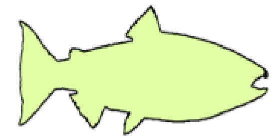
18 Figure 5. Error rates estimated by importance sampling for the simulated 300 locus
19 microhaplotype panel with two rates of missing genotypes: 3% (Missing) and 0% (No missing).
20 False positive error rate is the rate for unrelated trios.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

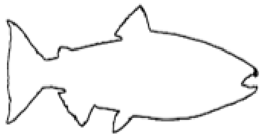
21 Supplemental Figure 1. Error rates estimated by stratified sampling for the simulated 300 locus
22 microhaplotype panel for select related trios. The two putative grandparents were unrelated to
23 each other. The relationship labels indicate the relationship of the two putative grandparents to
24 the putative grandchild. True is true grandparent, GAunt is great-aunt, HGAunt is half great-aunt,
25 Unrel is unrelated.

26





Sampled



Unsampled

A

False positive

1e−02
1e−06
1e−10
1e−14

0.00 0.01 0.02 0.03 0.04 0.05

False negative

Number of loci

100
300
500
700
900

B

False positive

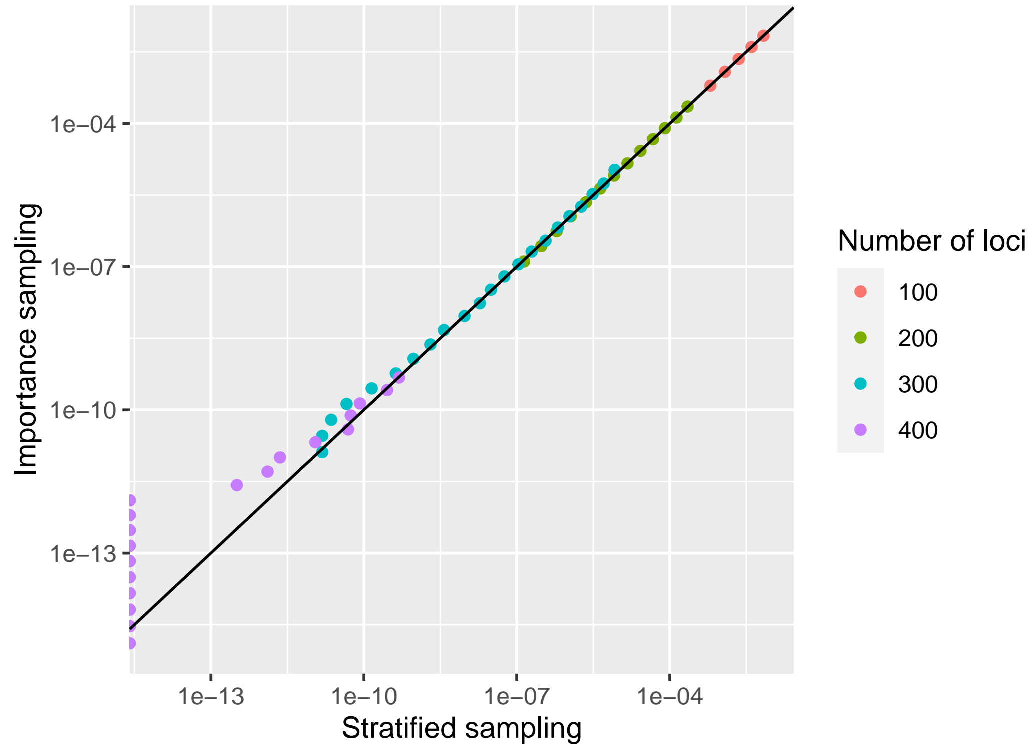
1e−02
1e−06
1e−10
1e−14

0.00 0.01 0.02 0.03 0.04 0.05

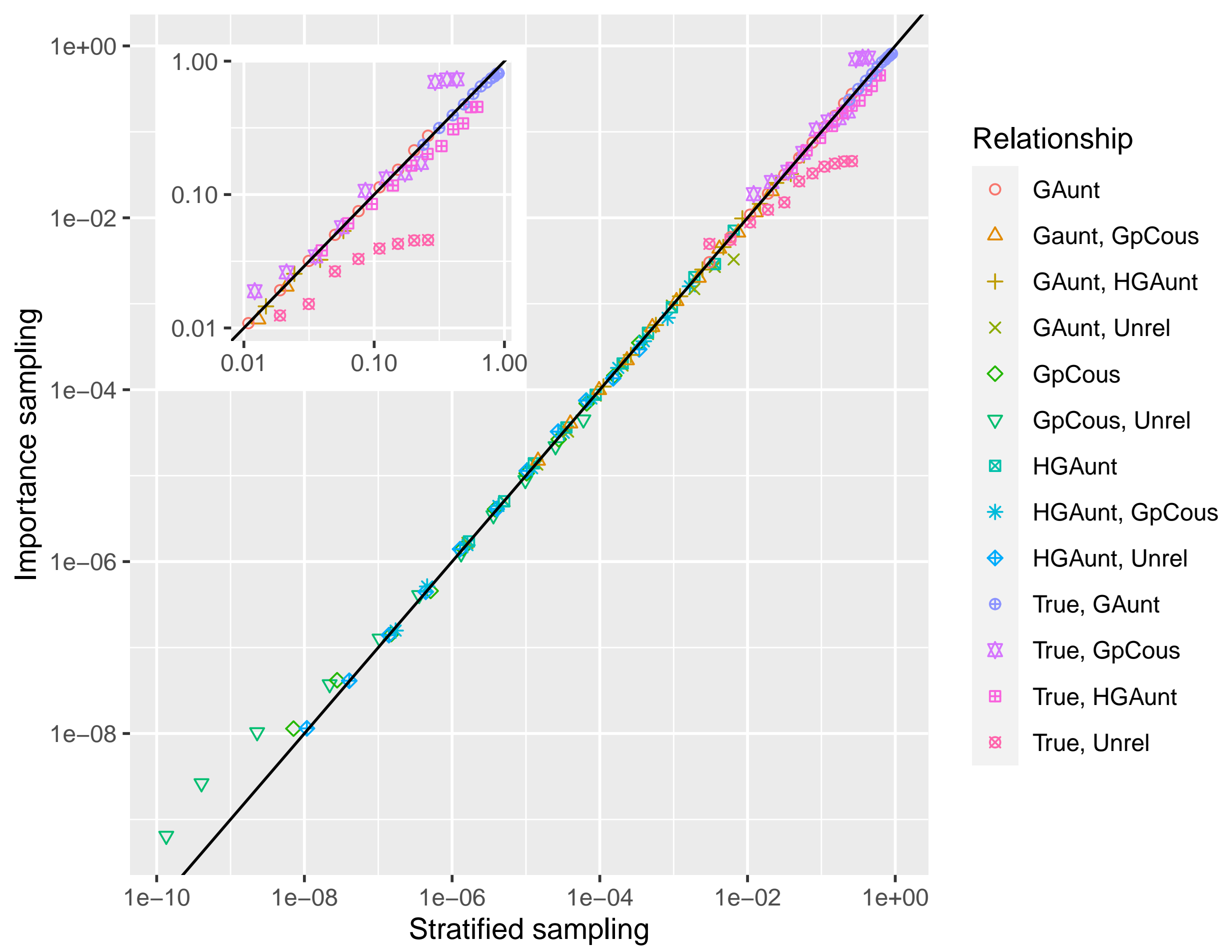
False negative

Number of loci

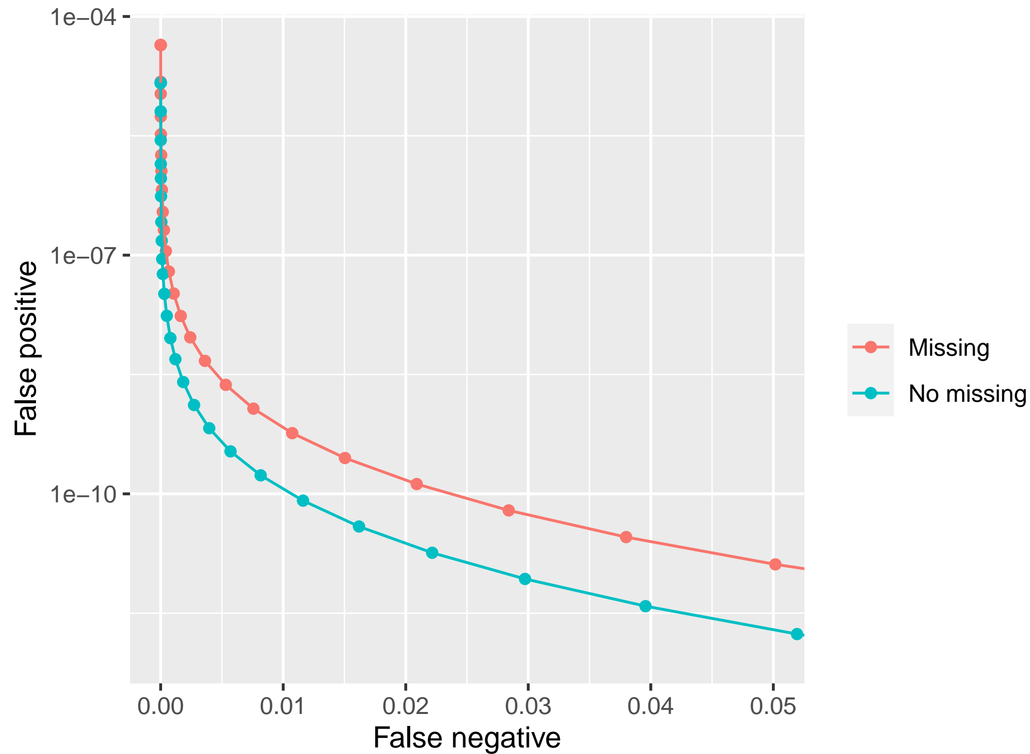
100
200
300
400

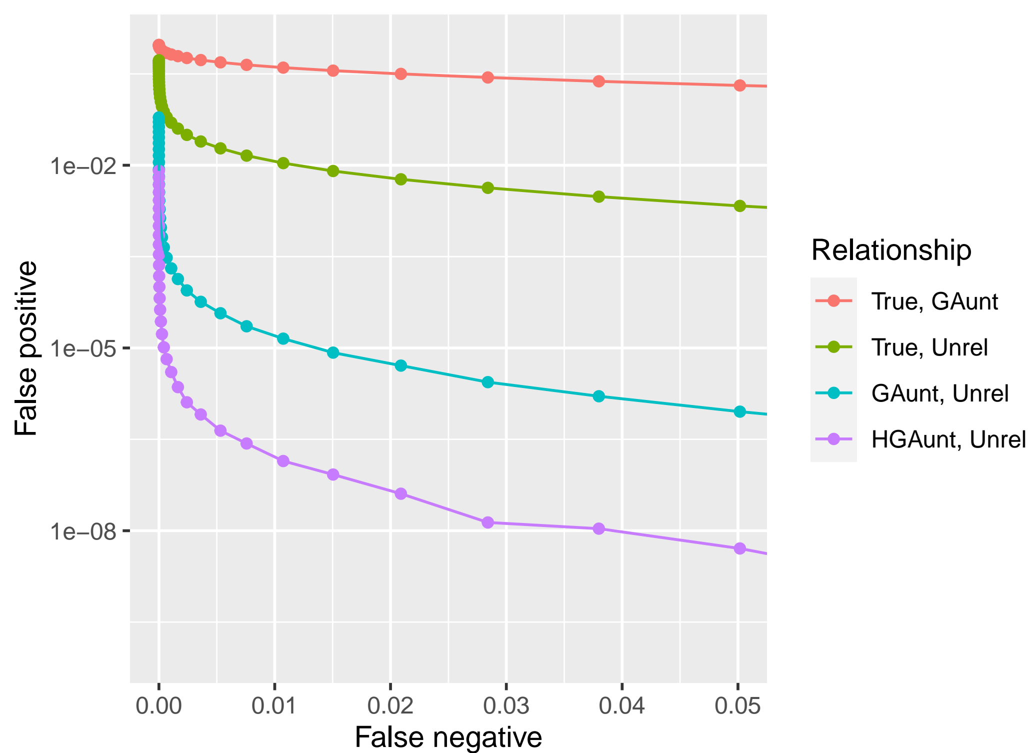


Importance sampling
Stratified sampling
1e−04
1e−07
1e−10
1e−13
Number of loci
100
200
300
400



Importance sampling
Stratified sampling
1e+00
1e−02
1e−04
1e−06
1e−08
1e−10
1.00
0.10
0.01
Relationship
GAunt
Gaunt, GpCous
GAunt, HGAunt
GAunt, Unrel
GpCous
GpCous, Unrel
HGAunt
HGAunt, GpCous
HGAunt, Unrel
True, GAunt
True, GpCous
True, HGAunt
True, Unrel


1e−04
1e−07
1e−10
False positive
0.00
0.01
0.02
0.03
0.04
0.05
False negative
Missing
No missing



False positive
1e−02
1e−05
1e−08
0.00
0.01
0.02
0.03
0.04
0.05
False negative
Relationship
True, GAunt
True, Unrel
GAunt, Unrel
HGAunt, Unrel

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Supplementary file 1

2

1 Default error model for gRandma

Each locus is treated separately and each genotype is treated as two observations (one for each allele in the presumed diploid). For a given locus, the error model first relies upon a "per allele error rate", ϵ, which represent the probability that, when observing one allele, you observe any allele other than the true one. The default value is 0.005, or 0.5%. This value is used along with a measure of similarity between alleles to calculate the probability of observing each allele given the true allele. For a locus with I alleles, given the true allele i, the probability of correctly observing i is $1 - \epsilon$ and the probability of observing allele j where $j \neq i$ is

$$P(j|i) = \epsilon \cdot \frac{s_{ij}}{\sum_{k=1}^{I} s_{ik}}, \quad (1)$$

where s_{ij} is the similarity between alleles i and j and $s_{ii} = 0$. The measure of similarity between two alleles of a microhaplotype used here was the reciprocal of the number of base pair differences between them.

These probabilities are then used to calculate the probability of observing each genotype given a true homozygous genotype. The probability of observing a genotype of BC given a true genotype of AA is

$$P(BC|AA) = \begin{cases} 2 \cdot P(B|A)P(C|A), & B \neq C \\ P(B|A)^2, & B = C, \end{cases} \quad (2)$$

according to basic probability arguments, where $P(B|A)$ is the probability of observing B when the true allele is A. The probability of observing each possible genotype given a true heterozygous genotype is calculated similarly, but also incorporates a probability of allelic dropout for each allele. Each allele is given a probability of dropping out, d_i for allele i, and in the current study a probability of 0.005 was used for all alleles. It is assumed that allelic dropout events are disjoint, and so the probability of no dropout occurring is $1 - d_i - d_j$. The probability of observing each genotype given no dropout can be calculated using the observation probabilities for each allele calculated in equation 1. The probability of observing a given genotype if a dropout has occurred is taken to be the probability of observing that genotype given a true homozygous genotype for the remaining allele. The total probability of observing a genotype is the sum of

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.11.455946; this version posted October 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24 the probabilities of all three cases (no dropout, allele 1 dropout, allele 2 dropout).