1

2 MS. ELENI LETO PETROU (Orcid ID : 0000-0001-7811-9288)

3

4

5 Article type : Resource Article

6

7

8 **Intraspecific DNA contamination distorts subtle population structure in a marine fish:**
9 **decontamination of herring samples before restriction-site associated (RAD) sequencing**
10 **and its effects on population genetic statistics**

11

12 Eleni L. Petrou[1], Daniel P. Drinan[1], Robert Kopperl[2], Dana Lepofsky[3], Dongya Yang[3], Madonna
13 L. Moss[4], Lorenz Hauser[1]

14 [1] School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street,
15 Seattle WA, 98105, USA

16 [2] Willamette Cultural Resources Associates Ltd., 650 S. Orcas St., Ste. 201 Seattle, Washington
17 98108, USA

18 [3] Department of Archaeology, Simon Fraser University, Education Building 9635, 8888
19 University Dr. Burnaby, BC V5A 1S6, Canada

20 [4] Department of Anthropology, University of Oregon, Eugene, OR 97403, USA

21

22 **Corresponding author:**

23 Eleni L. Petrou

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12978

This article is protected by copyright. All rights reserved

24 School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Seattle
25 WA, 98105, USA
26 elpetrou@uw.edu
Author Manuscript
27
28
29
30
31
32
33

34 Abstract

35 Wild specimens are often collected in challenging field conditions, where samples may
36 be contaminated with the DNA of conspecific individuals. This contamination can result in false
37 genotype calls, which are difficult to detect, but may also cause inaccurate estimates of
38 heterozygosity, allele frequencies, and genetic differentiation. Marine broadcast spawners are
39 especially problematic, because population genetic differentiation is low and samples are often
40 collected in bulk and sometimes from active spawning aggregations. Here, we used contaminated
41 and clean Pacific herring (*Clupea pallasi*) samples to test (i) the efficacy of bleach
42 decontamination, (ii) the effect of decontamination on RAD genotypes, and (iii) the
43 consequences of contaminated samples on population genetic analyses. We collected fin tissue
44 samples from actively spawning (and thus contaminated) wild herring and non-spawning
45 (uncontaminated) herring. Samples were soaked for 10 minutes in bleach or left untreated, and
46 extracted DNA was used to prepare DNA libraries using a restriction-site associated DNA
47 (RAD) approach. Our results demonstrate that intraspecific DNA contamination affects patterns
48 of individual and population variability, causes an excess of heterozygotes, and biases estimates
49 of population structure. Bleach decontamination was effective at removing intraspecific DNA
50 contamination and compatible with RAD sequencing, producing high-quality sequences,

This article is protected by copyright. All rights reserved

51 reproducible genotypes, and low levels of missing data. Although sperm contamination may be
52 specific to broadcast spawners, intraspecific contamination of samples may be common and
53 difficult to detect from high-throughput sequencing data, and can impact downstream analyses.

54 **Keywords:**

55 DNA contamination, RAD sequencing, population genetics, Pacific herring, heterozygosity

56 Introduction

57 High-throughput DNA sequencing has advanced the field of molecular ecology by
58 enabling comprehensive investigations of genetics and genomics in non-model species
59 (Allendorf, Hohenlohe, & Luikart, 2010; Andrews, Good, Miller, Luikart, & Hohenlohe, 2016;
60 Ekblom & Galindo, 2011). However, high-throughput sequencing is sensitive to the
61 contamination of samples with exogenous (non-target) DNA. Errors introduced by interspecific
62 DNA contamination have been identified in whole genome assemblies (Koutsovoulos et al.,
63 2016; Longo, O'Neill, & O'Neill, 2011), ancient DNA (Campana, Robles García, Rühli, &
64 Tuross, 2014), and metagenomic datasets (Schmieder & Edwards, 2011). To address the problem
65 of interspecific contamination, bioinformatic tools have been developed to remove exogenous
66 DNA from sequence data (Schmieder & Edwards, 2011) before contaminated sequences are
67 incorporated into downstream analyses. These methods typically identify non-target sequences
68 by aligning them to databases of common contaminating species; as a result, they cannot be used
69 to detect intraspecific contamination caused by the unintentional mixing of DNA between
70 individual samples of the same species.

71 Intraspecific contamination may profoundly affect downstream analysis, even though it
72 can be hard to detect in raw data. False heterozygotes inflate measures of observed
73 heterozygosity (Jun et al., 2012) and genetic diversity, and can lead to biased estimates of allele
74 frequencies and genetic differentiation. In species with weak population structure, contamination
75 may either obscure true differentiation or, alternatively, suggest significant genetic
76 differentiation where none exists.

77 Some bioinformatic tools have been developed to screen sequences for intraspecific DNA
78 contamination (Flickinger, Jun, Abecasis, Boehnke, & Kang, 2015; Jun et al., 2012), but these
79 tools were primarily developed for human re-sequencing studies; as such, they require pre-

Author Manuscript

This article is protected by copyright. All rights reserved

80 existing baseline data on population allele frequencies or high-coverage individual genotypes to
81 identify contaminated individuals. These types of genomic resources are oftentimes unavailable
82 for non-model species and consequently little attention has been given to the potential problem
83 of intraspecific DNA contamination in most molecular ecology studies.

84 Intraspecific contamination can be particularly problematic in studies of wild populations
85 of non-model organisms. First of all, samples are often collected in challenging or remote field
86 conditions, where access to resources such as sterile water and clean tools is limited. In addition,
87 field sampling can involve the bulk collection of multiple individuals. For example, animals such
88 as fish or insects may be caught in nets where numerous individuals are in close contact with
89 each other's tissues or bodily fluids, increasing the risk of intraspecific contamination
90 (Greenstone, Weber, Coudron, Payton, & Hu, 2012; D. Mitchell, McAllister, Stick, & Hauser,
91 2008). More generally, laboratory errors during sample handling or DNA library preparation can
92 also result in intraspecific DNA contamination (Sehn et al., 2015), and the common use of
93 Illumina adapters during high-throughput sequencing (such as restriction site-associated DNA
94 (RAD) sequencing (Baird et al., 2008), means that any exogenous DNA present in a sample
95 could be amplified during PCR.

96 One of the standard methods to decontaminate samples is treatment with bleach; this
97 approach has been used to clean bone samples before sequencing of ancient DNA (Kemp &
98 Smith, 2005; Yang & Watt, 2005), as well as fresh tissue samples for microsatellite (D. Mitchell
99 et al., 2008) and mitochondrial analysis (Greenstone et al., 2012). However, traditional
100 microsatellite and mitochondrial sequencing, as well as high-throughput sequencing of ancient
101 DNA, can utilize short DNA fragments as template. In contrast, RAD sequencing requires very
102 high-quality DNA with intact restriction sites, otherwise there is a dramatic reduction in the
103 number of raw sequences produced (Graham et al., 2015). Given that bleach decontaminates
104 samples by degrading surface DNA (Kemp & Smith, 2005), the effect of bleach on the quality
105 and quantity of endogenous sequence reads produced by RAD sequencing is currently unknown.
106 Therefore, bleach treatment may affect downstream analyses, even if decontamination were
107 successful.

108 Here, we used contaminated and clean Pacific herring (*Clupea pallasi*) samples to test (i)
109 the efficacy of bleach decontamination, (ii) the effect of decontamination on RAD genotypes,

Author Manuscript

This article is protected by copyright. All rights reserved

110 and (iii) the consequences of contaminated samples on population genetic analyses. By
111 combining these results, we identified the impacts of contamination on population genetic
112 analyses and empirically validated an approach aimed at minimizing contamination that is
113 compatible with RAD sequencing.

114 Materials and Methods

115 Sample collection

116 Sexually mature Pacific herring were collected immediately prior to or during active
117 spawning events using seine nets or hook and line fishing gear (Table 1). Adult herring were
118 sampled from genetically differentiated populations with different spawn timing (Beacham,
119 Schweigert, MacConnachie, Le, & Flostrand, 2008; D. M. Mitchell, 2006; Small et al., 2005);
120 our study included samples from the "primary-spawning" populations of Quilcene Bay (WA)
121 and Spiller Channel (BC), and the "late-spawning" population from Cherry Point (WA). The
122 sexual maturity of each individual was visually determined following the guidelines described in
123 Bucholtz et al. (2008). During sampling, herring sperm was clearly visible in the water column
124 and fish readily released gametes when slight pressure was applied to their abdomen. The density
125 of sperm in the water column during a herring spawn may be as high as 80-210 sperm/mL
126 (Hourston & Rosenthal, 1976), resulting in considerable intraspecific DNA contamination (D.
127 Mitchell et al., 2008). Thus, our samples were likely contaminated with the DNA of multiple
128 herring. Fin or muscle tissue samples were taken from each individual and immediately stored in
129 100% ethanol in individual vials.

130 Captive juvenile herring that were sexually immature were used as an uncontaminated
131 control group. Juvenile herring were reared at the United States Geological Survey (USGS)
132 Marrowstone Marine Field Station, WA, from fertilized eggs collected at Cherry Point, WA
133 (Table 1). Herring were individually caught from aquaria and euthanized using tricaine
134 methanesulfonate (MS-222). Fin tissue from each individual fish was sampled immediately and
135 samples were preserved in 100% ethanol. To minimize the risk of cross-contamination during
136 sampling, a new scalpel was used for each fish, and other sampling equipment (e.g., tweezers,
137 cutting mats) was cleaned with 10% bleach solution followed by three rinses of distilled water
138 and flame sterilization.

This article is protected by copyright. All rights reserved

Author Manuscript

139 ## Experimental assessment of bleach treatment

140 Tissue samples taken from wild adults (N = 17) and captive juveniles (N = 20) were split
141 into two pieces (approximately 2 mm^2) and exposed to the following experimental treatments:

142 1. Null treatment: samples were stored in 100% ethanol until DNA extraction.

143 2. Bleach treatment: Following a modified protocol of Mitchell et al. (2008), samples were
144 placed in individual tubes and immersed in 180 μL of 0.12% sodium hypochlorite (bleach)
145 (Sigma-Aldrich, St. Louis, MO, USA) for ten minutes. During bleach incubation, samples were
146 vortexed at medium-high speed. Subsequently, we removed bleach from the tubes and added 200
147 μL of Milli-Q purified water (Millipore, Bedford, MA, USA). Samples were vortexed for one
148 minute at medium-high speed, after which Milli-Q water was removed and fresh Milli-Q water
149 was added to the tube. This water rinse was repeated five times, and samples were stored in
150 100% ethanol until DNA extraction.

151 To estimate genotyping error rates within and between treatment groups, five juvenile
152 herring were subsampled in replicate, and both subsamples were subjected to both experimental
153 treatments. In addition, we also created four "dirty cocktails" as reference positive controls for
154 DNA contamination. Each dirty cocktail contained 25 ng/ μL of DNA from four different
155 juvenile herring in equal proportions.

156 We tested the reproducibility of the bleach treatment by implementing it on a large
157 number of spawning adult herring (N = 194). These fish were sampled from the same geographic
158 location as the herring that were used in the null and bleached treatments (Table 1).

159 ## DNA library preparation and sequencing

160 Genomic DNA was extracted from each subsample using the Qiagen DNeasy Blood and
161 Tissue Kit (Qiagen, Valencia, CA, USA). DNA was visualized with agarose gel electrophoresis
162 to assess DNA quality and quantified with the PicoGreen dsDNA Assay Kit (Invitrogen,
163 Waltham, MA, USA). We standardized the DNA concentration of each sample to 25 ng/ μL.

164 As an initial check for contamination, six microsatellite loci (*Cpa-8, Cpa-104, Cpa-113*
165 (Miller, Laberee, Schulze, & Kaukinen, 2001) and *Cpa-106, Cpa-107a, Cpa-111* (Olsen, Lewis,
166 Kretschmer, Wilson, & Seeb, 2002)) were used by the Washington Department of Fish and
167 Wildlife Molecular Genetics Laboratory to screen every sample that was present in both the

This article is protected by copyright. All rights reserved

Author Manuscript

168 bleach and null treatment groups (N = 37), following the protocol of Olsen et al. (2002). Alleles
169 were scored on Peak Scanner 2 (Life Technologies, Carlsbad, CA, USA). In the microsatellite
170 data, we defined contaminated samples as those containing more than two alleles at any locus.

171 We followed the protocol of Etter at al. (2011) to prepare DNA libraries for restriction-
172 site associated (RAD) sequencing. Depending on availability, 200 to 500 ng (depending on
173 availability) of genomic DNA per individual was digested with the restriction enzyme *SbfI* (New
174 England Biolabs, Ipswich, MA). Samples were individually labeled using a custom set of 96
175 barcodes (Integrated DNA Technologies, San Diego, CA) and groups of 12 samples were pooled
176 into libraries that were sheared to a length of approximately 500 base pairs (bp) using a
177 Bioruptor sonicator (Diagenode, Denville, NJ). We modified the Etter et al. (2011) protocol by
178 using AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) to size-select DNA
179 fragments (300-500 bp) and purify DNA products. However, all other steps (blunt-end repair, 3'-
180 dA overhang addition, P2 adapter ligation, PCR) were conducted as described in Etter at al.
181 (2011). After PCR, the DNA concentration of each library was quantified using the PicoGreen
182 dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA). We standardized the concentration
183 of each library to 10 *nM* and pooled libraries such that 48 individuals were sequenced per lane of
184 an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA) at the University of Oregon Genomics
185 Core Facility. The resulting sequences were single-end and 100 bp in length.

186 ## Bioinformatics analyses

187 We used the *process_radtags* script in *Stacks* version 1.39 (Catchen, Hohenlohe,
188 Bassham, Amores, & Cresko, 2013) to demultiplex individual samples, remove sequences with
189 low quality scores (Phred score < 10), and trim sequences to a length of 90 base pairs. The
190 quality of sequencing data was assessed using FastQC
191 (http://www.bioinformatics.babraham.ac.uk /projects/fastqc/).

192 Following the protocol of Brieuc et al. (2014), we created a reference database of herring
193 RAD loci to facilitate sequence assembly and locus identification. The reference database was
194 built using juvenile samples (null treatment) that had at least 1.5 million sequences (N = 19).
195 First, we assembled sequences and identified loci in these samples using the *de novo* locus
196 discovery pipeline in *Stacks*. Loci within each sample were allowed to have up to three
197 nucleotide mismatches (*ustacks*, M = 3) and each allele had to be sequenced at a minimum depth

This article is protected by copyright. All rights reserved

Author Manuscript

198 of 5X to be retained in the analysis (*ustacks*, m = 5). Subsequently, we removed loci with tandem
199 repeat units using *Blast* version 2.2.25 (Altschul, Gish, Miller, Myers, & Lipman, 1990) and
200 *bowtie* version 0.12.7 (Langmead, Trapnell, Pop, & Salzberg, 2009) as described in Brieuc et al.
201 (2014).

202 All sequenced samples (N = 280) were aligned to the reference database of RAD loci
203 using *bowtie*, allowing up to three nucleotide mismatches between the reference and query
204 sequences. Sequences that aligned to the database were subsequently processed with the *pstacks*
205 script in *Stacks* to identify loci in each sample (minimum depth of coverage to report a stack =
206 10; SNP model, alpha = 0.05). We filtered out low-quality samples by only retaining those that
207 contained at least 20,000 RAD loci after *pstacks*. To maximize the number of loci retained, a
208 catalog of loci was constructed in *cstacks* using a subset of the ten most deeply sequenced
209 individuals (bleach treatment) from each sampling location. All samples were genotyped using
210 *sstacks* and we only retained loci that were present in 80% of samples from each treatment
211 group.

212 We removed possible sequencing errors by filtering the SNPs discovered by *Stacks*. A
213 custom python script published in Brieuc et al. (2014) were used to retain only loci with two
214 haplotypes and to re-score genotypes. This method designates a heterozygote genotype if each
215 allele is sequenced at least twice and the locus is sequenced to a depth of at least ten reads.
216 Subsequently, we filtered out loci and individuals that had more than 20% missing data. Loci
217 characterized by very low minor allele frequencies were filtered from the final dataset; a minor
218 allele had to be present in at least one of the treatment groups at a frequency of 0.05 for that
219 locus to be retained in downstream analyses. Finally, we tested for deviations from Hardy-
220 Weinberg equilibrium (HWE) using the exact test based on 1,000 Monte Carlo permutations of
221 alleles, as implemented in the R package *pegas* (Paradis, 2010). Loci that were out of HWE in
222 every one of the population genetic samples (Cherry Point, Quilcene Bay, and Spiller Channel)
223 were removed from the analysis. As a final assessment of locus assembly, we followed the
224 recommendations of Paris et al. (2017) and aligned the filtered set of loci to the Atlantic herring
225 genome using *bowtie2* version 2.2.6 (Langmead & Salzberg, 2012). We also estimated per-locus
226 F_{IS} at each sampling location using *Genepop* version 4 (Rousset, 2008).

Author Manuscript

This article is protected by copyright. All rights reserved

227 Individual multilocus heterozygosity (H_I), the number of heterozygous loci divided by
228 the total number of loci genotyped, was calculated for each sample. Our expectation was that
229 contaminated samples would be characterized by higher values of H_I than the uncontaminated
230 control group (juvenile herring) because they would contain alleles from multiple individuals.
231 Variation in multilocus heterozygosity among uncontaminated individuals and populations was
232 expected to be small, as Pacific herring are characterized by large population sizes, low
233 inbreeding and low genetic population differentiation (Small et al. 2005, Mitchell 2006,
234 Beacham et al. 2008).

235 In addition, we tested whether bleach degraded target DNA and introduced error to the
236 data by comparing the genotypes of identical juvenile herring in the null and bleach treatment
237 groups (N = 20). This error was quantified as the number of genotype mismatches observed
238 between replicate extractions from the same individual (N = 5). A Wilcoxon signed-rank test was
239 used to assess whether the mean genotype mismatch rate differed between replicate samples and
240 treatment groups ($\alpha = 0.05$).

241 Population structure

242 We investigated the effect of intraspecific DNA contamination on patterns of population
243 structure by analyzing samples in the null and bleached treatment groups in combination with the
244 larger number of bleached samples. First, we conducted a principal components analysis (PCA)
245 using the R-package *adegenet* (Jombart, 2008). We also conducted an analysis with *Structure*
246 version 2.3.4 (Pritchard, Stephens, & Donnelly, 2000) using two different subsets of the data: the
247 first set included all samples, while the second included only bleached samples whose H_I was
248 within the range observed in uncontaminated juvenile samples. We implemented the admixture
249 model and allowed allele frequencies to be correlated among populations. Sampling location was
250 used as prior information (LOCPRIOR model), which can help detect clusters when population
251 structure is weak (Hubisz, Falush, Stephens, & Pritchard Jonathan, 2009). Three repetitions of
252 the model were run for each value of K (number of clusters) ranging from one to six. All runs
253 consisted of 20,000 burn-in steps followed by 50,000 Markov chain Monte Carlo steps. We
254 subsequently used *structure harvester* (Earl & vonHoldt, 2012) to visualize likelihood values for
255 different values of K and calculate the ad-hoc statistic ΔK to identify the highest hierarchical
256 level of clustering in our data set (Evanno, Regnaut, & Goudet, 2005).

Author Manuscript

This article is protected by copyright. All rights reserved

257 To further investigate the effects of contamination and bleach treatment on measures of
258 population structure, populations were divided into 39 subsamples of approximately six
259 individuals (range = 4 to 7 individuals), the sample size of the smallest collection of
260 contaminated individuals from a single location. A recent study (Willing, Dreyer, & van
261 Oosterhout, 2012) showed that a small number of individuals (N= 4-6) can be used to obtain
262 unbiased estimates of F_{ST} when large numbers of loci (N > 1,000) are genotyped. Pairwise F_{ST}
263 (B. Weir & C. Cockerham, 1984) between subsamples was calculated in Genepop version 4 and
264 used for non-metric multidimensional scaling (nMDS) in Primer 6 (Clarke & Gorley, 2006).
265 Observed and expected heterozygosity were calculated in GenAlEx version 6.5 (Peakall &
266 Smouse, 2012), and F_{IS} (B. S. Weir & C. C. Cockerham, 1984) was estimated in Genepop
267 version 4 (Rousset, 2008). To compare differentiation with and without contaminated
268 individuals, hierarchical AMOVAs were calculated in Arlequin version 3.52 (Excoffier &
269 Lischer, 2010), using two alternative groupings. In the first comparison, groups were defined by
270 population (Cherry Point; Quilcene Bay; Spiller Channel) and subgroups consisted of the two
271 different treatments (bleach, null). In the second comparison, groups were defined by population
272 and subgroups consisted of subsamples of individuals (N= 4-7); different iterations of this
273 AMOVA were conducted excluding untreated individuals and H_I outliers.

274 Results

275 Sequencing and genotyping

276 We successfully genotyped 92% of individuals at three or more microsatellite loci. Six
277 out of 17 adult herring in the null treatment group displayed more than two alleles per
278 microsatellite locus, indicating that they were contaminated with the DNA of multiple herring.
279 Treatment with bleach appeared to remove contamination from all but one of the samples. None
280 of the 20 juvenile herring had more than two microsatellite alleles after either treatment,
281 demonstrating lack of contamination and confirming our hypothesis that sample contamination
282 was caused by the presence of sperm in the water column in wild spawning aggregations.

283 A reference database of RAD loci was built using sequences from 19 juvenile herring in
284 the null treatment group; one individual was excluded from the database because it contained
285 fewer than 1.5 million raw sequences. A total of 29,551 putative loci were initially identified,

This article is protected by copyright. All rights reserved

Author Manuscript

286 and 28,997 loci were retained in the reference database after filtering out loci with tandem
287 repeats and highly repetitive sequences.

288 After removing loci that were out of HWE in every population, we identified 3,502
289 biallelic RAD loci that were sequenced at a minimum read depth of 10 sequences in more than
290 80% of individuals and had a minor allele frequency that exceeded 0.05 in at least one of the
291 populations. We found that 93% of these loci aligned exactly once to the closely-related Atlantic
292 herring genome. Locus-specific estimates of F_{IS} were distributed around zero (Supplemental
293 Figure 1), which is concordant with expectations under HWE. A total of 240 herring had less
294 than 20% missing genotypes and were retained in the final data set.

295 Sequencing quality was robust and genotyping error was low for juvenile samples in the
296 null and bleached treatment groups. Juvenile samples treated with bleach were characterized by
297 slightly more sequences containing the restriction site (RADtags), loci per sample, and average
298 read depth (Figure 1). However, the genotype mismatch rate between treatments in the replicated
299 juvenile individuals was very low (1.8 ± 1.4%, mean ± SD), and similar to repeated bleach
300 treatments (1.4 ± 1.3%). The distribution of genotype mismatches did not differ statistically
301 between replicate individuals in the same (bleached) or across (null vs. bleached) treatment
302 groups (Wilcoxon sign rank test, $p = 0.55$), indicating that treatment with bleach does not alter
303 the endogenous ("true") genotype of a sample.

304 Impacts of contamination on individual level variation

305 As expected, multilocus individual heterozygosity (H_I) was higher in the untreated adult
306 samples than in any samples that were cleaned with bleach (Figure 2). Samples in the dirty
307 cocktail group (N= 4) exhibited high H_I (median = 0.45) but low variation in H_I among
308 individuals (25th and 75th quantiles = 0.44 - 0.46). In comparison, adult herring samples in the
309 null treatment group (N =11) had slightly lower but more variable H_I (median = 0.41, 25th and
310 75th quantiles = 0.31-0.42), but the maximum H_I observed in this group was as high as 0.60.
311 Adult herring samples treated with bleach (N = 174), were characterized by much lower H_I
312 (median = 0.18, 25th and 75th quantiles = 0.17 - 0.20). These values were similar to that observed
313 for non-spawning juvenile herring (N = 20), in the null (median H_I = 0.18, 25th and 75th quantiles
314 = 0.17 – 0.19) and bleach (median H_I = 0.18, 25th and 75th quantiles = 0.18 – 0.20) treatments.
315 However, there was some evidence for residual contamination in cleaned adult samples, as 8%

Author Manuscript

This article is protected by copyright. All rights reserved

316 (14/174) of those samples had H_I that was above the range observed in juvenile samples (Figure
317 2).

318 Intraspecific contamination affected patterns of individual differentiation, as shown by
319 PCA (Figure 3A, B, and C). When all samples were included in the same analysis, most of the
320 variation was driven by contaminated adult samples (Figure 3A). When these contaminated
321 samples were removed from the analyses, less variation was explained by the first axis but
322 outlier samples were still evident (Figure 3B). These samples consisted of 14 adult herring that
323 were treated with bleach but whose H_I was relatively high (between 0.25 and 0.34) and exceeded
324 the maximum value observed in juvenile samples (0.23); we hereinafter refer to these samples as
325 H_I outliers. Once these H_I outliers were removed from the analysis, Cherry Point adults and
326 juveniles clustered separately from Quilcene Bay and Spiller Channel samples (Figure 3C).
327 Furthermore, cleaned adult samples collected from two different years at Cherry Point clustered
328 together with juvenile samples originating from the Cherry Point population.

329 Multiple runs of *Structure* identified $K = 2$ as the most likely number of groups when
330 only cleaned data were included in the analysis. This result was supported by estimates of the
331 posterior probability of the data given K clusters ($LnP(D)$) and ΔK (Figure 4A). Fish collected at
332 Cherry Point (adults and juveniles) formed a distinct cluster, while fish collected at Quilcene Bay
333 and Spiller Channel strongly assigned to a second cluster. In contrast, when all samples
334 (including contaminated adults) were included in the same *Structure* analysis, $LnP(D)$ and ΔK
335 did not converge on the same answer (Figure 4B and C). The posterior probability of the data
336 given K clusters was highest at $K = 4$, while the distribution of ΔK showed peaks at both $K = 2$
337 and $K = 4$ (Supplemental Figure 2). At $K = 2$, the estimated ancestry coefficient of bleached
338 samples was symmetric across all sampling locations ($Q = 0.82 \pm 0.02$, mean ± SD), while it was
339 quite different for contaminated samples (Figure 4B). At $K = 4$ the same pattern was observed,
340 although population differentiation was more apparent in both clean and contaminated samples
341 (Figure 4C). In all cases, however, all individuals appeared to be highly admixed, most likely
342 because of low population differentiation.

343 ## Impacts of contamination on estimates of population structure

344 Similar and considerable effects of contamination were apparent for population
345 parameters (H_e, F_{IS}, F_{ST}) estimated from subsamples of individuals drawn from each herring

This article is protected by copyright. All rights reserved

Author Manuscript

346 population (Cherry Point, Quilcene Bay, and Spiller Channel). All contaminated subsamples and
347 the 'dirty cocktail' had a more negative F_{IS} (indicating an excess of heterozygotes) and higher
348 expected heterozygosity values than bleached adult subsamples lacking H_I outliers
349 (Supplemental Figure 3). In addition, subsamples of juvenile herring had similar values of
350 heterozygosity and F_{IS} before and after bleaching. Most adult subsamples had similar
351 heterozygosity and an F_{IS} close to zero after bleaching, especially when H_I outliers were
352 removed.

353 Contamination also had a clear effect on genetic differentiation between subsamples of
354 individuals selected from the same population. Subsamples containing highly contaminated
355 individuals were outliers in the nMDS analysis (Figure 3D, E, and F). Both the 'dirty cocktail'
356 and the unbleached adult subsamples exhibited high differentiation from bleached subsamples
357 taken from the same population (Fig 3D, F_{ST}= 0.015 - 0.070, Supplemental Table 1). After
358 bleaching, adult herring subsamples taken from the same population were less differentiated
359 from each other (Fig 3F, F_{ST}= -0.009 - 0.019, Supplemental Table 1), although subsamples
360 containing H_I outliers exhibited higher differentiation (Fig 3E, F_{ST}= 0.016-0.028 Supplemental
361 Table 1). The lowest F_{ST} values were observed between the bleached and unbleached replicate
362 subsamples of the same juvenile individuals (Supplemental Table 1).

363 Hierarchical AMOVAs demonstrated that contamination can inflate underlying genetic
364 population differentiation (Table 2). When contaminated individuals were included in
365 comparisons of population and treatment (Table 2, AMOVA 1), the differentiation between
366 treatment groups from the same population (F_{SC}) was greater than the differentiation observed
367 between distinct populations (F_{CT}). When contaminated individuals were included in an
368 AMOVA using subsamples of individuals (Table 2, AMOVA 2), contamination inflated the
369 overall F_{ST}. Contamination also increased the differentiation between population groups (F_{CT}) as
370 well as the differentiation among subsamples within a population (F_{SC}). Adding individual level
371 analyses into the AMOVA did not change these trends, although the presence of contaminated
372 samples was clearly indicated by more negative F_{IS} values.

Author Manuscript

This article is protected by copyright. All rights reserved

373 # Discussion

374 ## Effects of contamination

375 Our results demonstrate that intraspecific DNA contamination affects patterns of
376 individual and population variability, causes an excess of heterozygotes, and biases estimates of
377 population structure. However, contamination could be easily removed, and treatment of tissues
378 with bleach did not affect the quality of resulting sequencing results. Our results therefore
379 highlight the importance of identifying and removing contamination in tissues intended for RAD
380 sequencing.

381 Signals of intraspecific DNA contamination are more subtle in SNPs compared to
382 microsatellite loci. In highly variable markers such as microsatellites, heavily contaminated
383 individuals are easily identified by the presence of more than two alleles (in a diploid species) at
384 a single locus (D. Mitchell et al., 2008). In contrast, contaminated samples genotyped at biallelic
385 SNPs simply exhibited higher individual heterozygosity (H_I) relative to uncontaminated sample.
386 Nevertheless, SNP data appeared more sensitive to contamination than microsatellites: while
387 only 35% of unbleached adult herring had three or more microsatellite alleles per locus, 82% of
388 those same samples exhibited elevated H_I relative to juvenile herring.

389 These findings underscore the utility of using clean samples to estimate empirical
390 distributions of H_I. A modest number of clean reference samples can be used to construct a
391 baseline for comparison with potentially contaminated samples using the simple metric of H_I.
392 Furthermore, H_I is a standard metric that is commonly reported in population genetic studies
393 (Hoffman et al., 2014; Kjeldsen et al., 2016; Tarpey et al., 2017). To our knowledge, this is the
394 one of the first studies of wild populations to examine patterns of H_I as a quality-control
395 measure, even though a related metric (ratio of heterozygous/non-reference homozygous sites) is
396 commonly used in the quality control of human genomic data (Wang, Raskin, Samuels, Shyr, &
397 Guo, 2015). We recommend that researchers examine the distribution of H_I in their data across
398 individuals and populations, and carefully consider whether outlier samples could be caused by
399 intraspecific DNA contamination.

400 However, we recognize that interpreting H_I outliers in species with very small effective
401 population sizes or inbreeding could be more complicated. Individual heterozygosity and

This article is protected by copyright. All rights reserved

Author Manuscript

402 inbreeding are strongly correlated with each other when population sizes are very small and
403 mating systems are highly skewed (e.g. polygyny, selfing) (Balloux, Amos, & Coulson, 2004;
404 Hoffman et al., 2014). Therefore, if individual heterozygosities were highly variable between
405 individuals and/or populations, higher values of H_I in outbred individuals, immigrant
406 individuals, or highly diverse populations could be mistaken for a signal of contamination. If
407 those individuals were removed from a dataset because they were mistaken for contamination, it
408 would lead to be a reduction in the average heterozygosity of that population and bias sampling.
409 For species with large populations and potentially high gene flow, such as herring (Beacham et
410 al., 2008; Lamichhaney et al., 2017; Limborg et al., 2012) and many other marine fishes
411 (Knutsen et al., 2011; Reiss, Hoarau, Dickey-Collas, & Wolff, 2009), variability in individual
412 heterozygosity should be low. Our results suggest that F_{IS} estimated even in relatively small
413 subsamples of individuals (*N*= 4-7) is a sensitive indicator of contamination, that may be useful
414 when H_I is variable.

415 Marine species are characterized by weak population differentiation that is sensitive to
416 sampling errors (Waples, 1998). A possible consequence of contamination would be that "noise"
417 introduced into a dataset through contaminating alleles would overwhelm faint signals of genetic
418 differentiation between populations. Indeed, this hypothesis was confirmed by our results;
419 contaminated samples appeared as outliers in every analysis, and led to inflated estimates of
420 population differentiation (F_{ST}) and differentiation among subsamples within a population (F_{SC})
421 in an AMOVA framework. Clustering approaches were also strongly affected by contamination:
422 heavily contaminated individuals and population subsamples were outliers in PCA and nMDS
423 analyses, and may thus impact the interpretation from such approaches. *Structure* results were
424 also dominated by contaminated samples, and *LnP(D)* and ΔK did not converge on the same
425 value of K when these contaminated samples were included in the data. Without contaminated
426 samples, *Structure* detected subtle but clear population structure. Contamination can therefore
427 distort true population structure, which is especially problematic in the context of conservation
428 genetics and resource management, as genetic data are often used to help delineate conservation
429 or management units (Funk, McKay, Hohenlohe, & Allendorf, 2012; Palsbøll, Bérubé, &
430 Allendorf, 2007; Scribner et al., 2016). Thus, it is possible that contaminated genotypes could
431 lead to the erroneous designation of management units and the accidental overexploitation of
432 harvested populations.

This article is protected by copyright. All rights reserved

Author Manuscript

433 Efficacy of bleach treatment

434 Our research also confirms the efficacy of bleach treatment as a method to decontaminate
435 tissue samples collected for RAD sequencing in challenging field conditions. Bleach removed
436 the majority of contaminant DNA on samples collected from spawning adult herring; using this
437 method, we were able to salvage 92% of adult samples collected during active spawn events and
438 discover 3,502 polymorphic RAD loci in Pacific herring. After decontamination with bleach,
439 only one sample was identified by microsatellites as being contaminated. However, a modest
440 number of bleached adult samples (8%) were characterized by elevated values of H_I, which
441 could be indicative of small amounts of residual contamination. It is possible that the
442 concentration and/or duration of the bleach treatment was insufficient to remove all traces of
443 contamination, and that low levels of residual contamination were still detectable in RAD
444 sequences generated from these samples.

445 Once contaminated individuals were removed from the data set, subsamples of
446 individuals taken from the same location produced very concordant estimates of F_{ST}, even
447 though subsample sizes were tiny (*N*= 4- 7). However, it has been shown that reliable F_{ST}
448 estimates can be obtained from very few individuals if loci can be sampled without bias (Willing
449 et al., 2012). Furthermore, while the separation between Cherry Point herring and samples from
450 Quilcene Bay and Spiller Channel in both individual and subsample clustering approaches
451 confirms previous studies of Pacific herring (Beacham et al., 2008; D. M. Mitchell, 2006; Small
452 et al., 2005), which found that Cherry Point herring were reproductively isolated from other
453 populations due to differences in their spawn timing, an nMDS and AMOVA based on
454 subsamples detected subtle but significant differentiation between Quilcene Bay and Spiller
455 Channel, which were previously both considered part of the same population (D. M. Mitchell,
456 2006; Small et al., 2005). This result indicates that analyses based on small subsamples of
457 individuals may be more powerful than those based on full samples, as suggested by (Nielsen et
458 al., 2012).

459 Previous research has shown that RAD sequencing requires very high-quality DNA as
460 input, otherwise there is a significant reduction in the number of raw sequences produced
461 (Graham et al., 2015). Treating tissue samples in a dilute solution of bleach did not hinder the
462 construction of RAD sequencing libraries, reduce the number of loci discovered in each sample,

Author Manuscript

This article is protected by copyright. All rights reserved

463 or affect the quality of sequence reads. Instead, juvenile samples treated with bleach yielded
464 slightly more loci and were characterized by greater read depth per locus when compared to the
465 same sample in the null treatment. This is most likely due to batch effects caused by slight
466 differences in the amplification success of pooled DNA libraries, which exclusively contained
467 either samples from the null or bleached treatment group. Importantly, we found that bleach did
468 not degrade the endogenous DNA of tissue samples; on average, 98% of loci had matching
469 genotypes when we compared replicate extractions from the same juvenile herring (across and
470 within treatment groups). This genotyping error rate is similar to rates observed in conventional
471 RAD sequencing studies (Fountain, Pauli, Reid, Palsbøll, & Peery, 2016; Mastretta-Yanes et al.,
472 2015). In addition, the fact that juvenile samples (from either treatment) and cleaned adult
473 samples (from both sampling years) from Cherry Point clustered together lends further support
474 that bleach treatment did not degrade endogenous DNA and cause false patterns of genetic
475 differentiation.

476 Although the problem of sperm contamination may be specific to broadcast spawners,
477 intraspecific DNA contamination remains a possible source of error for wild-caught specimens
478 of most species. Therefore, researchers will have to evaluate the risk of contamination on a case-
479 by-case basis. While treatment with bleach is a relatively simple and cost-effective way to clean
480 adult tissue samples, it might only be appropriate for studies where robust pieces of tissue are
481 available. For example, when we applied this method to delicate one-day-old herring larvae,
482 almost no DNA could be recovered (data not shown). Thus the concentration and/or duration of
483 bleach treatment might have to be adjusted for studies targeting very delicate samples. In
484 addition, special consideration should be given to sampling conditions, such as the bulk
485 collection (Greenstone, Weber, Coudron, & Payton, 2011; King et al., 2011) or storage of
486 specimens that could result in the accidental mixing of bodily fluids or cells. For example, in
487 forensic science, considerable attention has been given to the potential of intraspecific
488 contamination during sample collection (Cale, Earll, Latham, & Bush, 2016) and sample
489 processing in the laboratory (Vandewoestyne et al., 2011), though such practices are less
490 common in molecular ecology.

491 In conclusion, we show that intraspecific DNA contamination can affect subtle patterns
492 of population structure that are characteristic of many marine fish. We verified that treatment

Author Manuscript

This article is protected by copyright. All rights reserved

with bleach is an appropriate method for removing surface contamination from tissue samples without degrading endogenous DNA, resulting in reproducible genotypes from RAD sequencing. Our approach is likely to be applicable to tissue samples from other species.

Acknowledgements

We thank colleagues from the Washington Department of Fish and Wildlife (Dayv Lowry, Adam Lindquist), Department of Fishes and Oceans Canada (Terry Beacham), United States Geological Survey (Paul Hershberger), and the Heiltsuk Integrated Resource Management Department (Mike Reid) for collecting samples and sharing them with us. Isadora Jimenez-Hidalgo and Mary Fisher provided indispensable laboratory support, while Marine Brieuc and Charles D. Waters contributed advice on bioinformatic analyses. We thank Todd Seamons of the Molecular Genetics Lab of the Washington Department of Fish and Wildlife for running the microsatellite analyses. This work was funded in part by a grant from Washington Sea Grant, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA14OAR4170078, project No. R/HCE-3. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies. Additional support was provided by the Natural Sciences and Engineering Research Council of Canada Strategic Partnership Grant (*Understanding the Ecosystem Role of Pacific Herring in Coupled Social-ecological Systems: Advancing Forage Fish Science*) and a US National Science Foundation (NSF) award # 1203868. ELP received additional support from the University of Washington Program on Ocean Change Integrative Graduate Education and Research Traineeship (IGERT), funded by the NSF award # 1068839.

Author Manuscript

This article is protected by copyright. All rights reserved

520 # References

521 Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation
522 genetics. *Nature Reviews Genetics, 11*(10), 697-709. doi:10.1038/nrg2844
523 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search
524 tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:http://dx.doi.org/10.1016/S0022-
525 2836(05)80360-2
526 Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of
527 RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics, 17*(2), 81-92.
528 doi:10.1038/nrg.2015.28
529 http://www.nature.com/nrg/journal/v17/n2/abs/nrg.2015.28.html#supplementary-information
530 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., . . . Johnson, E. A. (2008).
531 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE, 3*(10),
532 e3376. doi:10.1371/journal.pone.0003376
533 Balloux, F., Amos, W., & Coulson, T. (2004). Does heterozygosity estimate inbreeding in real
534 populations? *Molecular Ecology, 13*(10), 3021-3031. doi:10.1111/j.1365-294X.2004.02318.x
535 Beacham, T. D., Schweigert, J. F., MacConnachie, C., Le, K. D., & Flostrand, L. (2008). Use of
536 microsatellites to determine population structure and migration of Pacific herring in British
537 Columbia and adjacent regions. *Transactions of the American Fisheries Society, 137*(6), 1795-
538 1811. doi:10.1577/T08-033.1
539 Brieuc, M. S. O., Waters, C. D., Seeb, J. E., & Naish, K. A. (2014). A dense linkage map for Chinook salmon
540 (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole
541 genome duplication event. *G3: Genes|Genomes|Genetics, 4*(3), 447-460.
542 doi:10.1534/g3.113.009316
543 Bucholtz, R. H., Tomkiewicz, J., & Dalskov, J. (2008). *Manual to determine gonadal maturity of herring
544 (Clupea harengus L.)*. Charlottenlund, Denmark: DTU Aqua, National Institute of Aquatic
545 Resources
546 Cale, C. M., Earll, M. E., Latham, K. E., & Bush, G. L. (2016). Could secondary DNA transfer falsely place
547 someone at the scene of a crime? *Journal of Forensic Sciences, 61*(1), 196-203.
548 doi:10.1111/1556-4029.12894

Author Manuscript

This article is protected by copyright. All rights reserved

549 Campana, M. G., Robles García, N., Rühli, F. J., & Tuross, N. (2014). False positives complicate ancient
550 pathogen identifications using high-throughput shotgun sequencing. *BMC Research Notes, 7*(1),
551 111. doi:10.1186/1756-0500-7-111
552 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool
553 set for population genomics. *Molecular Ecology, 22*(11), 3124-3140.
554 Clarke, K. R., & Gorley, R. N. (2006). *PRIMER v6: User Manual/Tutorial*. Plymouth: PRIMER-E.
555 Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing
556 STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources,*
557 *4*(2), 359-361. doi:10.1007/s12686-011-9548-7
558 Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of
559 non-model organisms. *Heredity, 107*(1), 1-15.
560 Etter, P., Bassham, S., Hohenlohe, P., Johnson, E., & Cresko, W. (2011). SNP discovery and genotyping for
561 evolutionary genetics using RAD sequencing. In *Molecular Methods for Evolutionary Genetics*
562 (Vol. 772, pp. 157-178): Humana Press.
563 Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the
564 software STRUCTURE: a simulation study. *Molecular Ecology, 14*(8), 2611-2620.
565 Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform
566 population genetics analyses under Linux and Windows. *Molecular Ecology Resources, 10*(3),
567 564-567. doi:10.1111/j.1755-0998.2010.02847.x
568 Flickinger, M., Jun, G., Abecasis, Gonçalo R., Boehnke, M., & Kang, Hyun M. (2015). Correcting for
569 sample contamination in genotype calling of DNA sequence data. *The American Journal of*
570 *Human Genetics, 97*(2), 284-290. doi:http://dx.doi.org/10.1016/j.ajhg.2015.07.002
571 Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). Finding the right coverage:
572 the impact of coverage and sequence quality on single nucleotide polymorphism genotyping
573 error rates. *Molecular Ecology Resources, 16*(4), 966-978. doi:10.1111/1755-0998.12519
574 Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for
575 delineating conservation units. *Trends in Ecology & Evolution, 27*(9), 489-496.
576 doi:10.1016/j.tree.2012.05.012
577 Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., . . . Somers, C. M.
578 (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq).
579 *Molecular Ecology Resources, 15*(6), 1304-1315. doi:10.1111/1755-0998.12404

Author Manuscript

This article is protected by copyright. All rights reserved

580 Greenstone, M. H., Weber, D. C., Coudron, T. A., Payton, M. E., & Hu, J. S. (2012). Removing external
581 DNA contamination from arthropod predators destined for molecular gut-content analysis.
582 *Molecular Ecology Resources, 12*(3), 464-469. doi:10.1111/j.1755-0998.2012.03112.x
583 Greenstone, M. H., Weber, D. C., Coudron, T. C., & Payton, M. E. (2011). Unnecessary roughness?
584 Testing the hypothesis that predators destined for molecular gut-content analysis must be hand-
585 collected to avoid cross-contamination. *Molecular Ecology Resources, 11*(2), 286-293.
586 doi:10.1111/j.1755-0998.2010.02922.x
587 Hoffman, J. I., Simpson, F., David, P., Rijks, J. M., Kuiken, T., Thorne, M. A. S., . . . Dasmahapatra, K. K.
588 (2014). High-throughput sequencing reveals inbreeding depression in a natural population.
589 *Proceedings of the National Academy of Sciences, 111*(10), 3775-3780.
590 Hourston, A. S., & Rosenthal, H. (1976). Sperm density during active spawning of Pacific herring (*Clupea*
591 *harengus pallasi*). *Journal of the Fisheries Research Board of Canada, 33*(8), 1788-1790.
592 doi:10.1139/f76-226
593 Hubisz, M., J., Falush, D., Stephens, M., & Pritchard Jonathan, K. (2009). Inferring weak population
594 structure with the assistance of sample group information. *Molecular Ecology Resources, 9*(5),
595 1322-1332. doi:10.1111/j.1755-0998.2009.02591.x
596 Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers.
597 *Bioinformatics, 24*(11), 1403-1405.
598 Jun, G., Flickinger, M., Hetrick, Kurt N., Romm, Jane M., Doheny, Kimberly F., Abecasis, Gonçalo R., . . .
599 Kang, Hyun M. (2012). Detecting and estimating contamination of human DNA samples in
600 sequencing and array-based genotype data. *The American Journal of Human Genetics, 91*(5),
601 839-848. doi:10.1016/j.ajhg.2012.09.004
602 Kemp, B. M., & Smith, D. G. (2005). Use of bleach to eliminate contaminating DNA from the surface of
603 bones and teeth. *Forensic Science International, 154*(1), 53-61.
604 doi:https://doi.org/10.1016/j.forsciint.2004.11.017
605 King, R. A., Davey, J. S., Bell, J. R., Read, D. S., Bohan, D. A., & Symondson, W. O. C. (2011). Suction
606 sampling as a significant source of error in molecular analysis of predator diets. *Bulletin of*
607 *Entomological Research, 102*(3), 261-266. doi:10.1017/S0007485311000575
608 Kjeldsen, S. R., Zenger, K. R., Leigh, K., Ellis, W., Tobey, J., Phalen, D., . . . Raadsma, H. W. (2016).
609 Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population
610 variability across its range. *Conservation Genetics, 17*(2), 337-353. doi:10.1007/s10592-015-
611 0784-3

This article is protected by copyright. All rights reserved

Author Manuscript

612 Knutsen, H., Olsen, E. M., Jorde, P. E., Espeland, S. H., André , C., & Stenseth, N. C. (2011). Are low but
613 statistically significant levels of genetic differentiation in marine fishes 'biologically meaningful'?
614 A case study of coastal Atlantic cod. *Molecular Ecology, 20*(4), 768-783. doi:10.1111/j.1365-
615 294X.2010.04979.x
616 Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., . . . Blaxter, M. (2016). No
617 evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius
618 dujardini. *Proceedings of the National Academy of Sciences, 113*(18), 5053-5058.
619 doi:10.1073/pnas.1600338113
620 Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., . . .
621 Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations
622 on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences,*
623 *114*(17), E3452-E3461.
624 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Meth, 9*(4), 357-
625 359. doi:10.1038/nmeth.1923
626 http://www.nature.com/nmeth/journal/v9/n4/abs/nmeth.1923.html#supplementary-information
627 Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of
628 short DNA sequences to the human genome. *Genome Biology, 10*(3), R25. doi:10.1186/gb-2009-
629 10-3-r25
630 Limborg, M. T., Helyar, S. J., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R. O. B., . . . Bekkevold, D.
631 (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish,
632 the Atlantic herring (*Clupea harengus*). *Molecular Ecology, 21*(15), 3686-3703.
633 doi:10.1111/j.1365-294X.2012.05639.x
634 Longo, M. S., O'Neill, M. J., & O'Neill, R. J. (2011). Abundant human DNA contamination identified in
635 non-primate genome databases. *PLoS ONE, 6*(2), e16410. doi:10.1371/journal.pone.0016410
636 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015).
637 Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly
638 optimization for population genetic inference. *Molecular Ecology Resources, 15*(1), 28-41.
639 doi:10.1111/1755-0998.12291
640 Miller, K. M., Laberee, K., Schulze, A. D., & Kaukinen, K. H. (2001). Development of microsatellite loci in
641 Pacific herring (*Clupea pallasi*). *Molecular Ecology Notes, 1*(3), 131-132. doi:10.1046/j.1471-
642 8278.2001.00048.x

Author Manuscript

This article is protected by copyright. All rights reserved

643 Mitchell, D., McAllister, P., Stick, K., & Hauser, L. (2008). Sperm contamination in archived and
644 contemporary herring samples. *Molecular Ecology Resources, 8*(1), 50-55. doi:10.1111/j.1471-
645 8286.2007.01840.x
646 Mitchell, D. M. (2006). *Biocomplexity and metapopulation dynamics of Pacific herring (Clupea pallasii) in*
647 *Puget Sound, Washington.* (Master of Science), University of Washington, Seattle, WA.
648 Nielsen, E. E., Cariani, A., Aoidh, E. M., Maes, G. E., Milano, I., Ogden, R., . . . Carvalho, G. R. (2012).
649 Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nat*
650 *Commun, 3*, 851.
651 doi:http://www.nature.com/ncomms/journal/v3/n5/suppinfo/ncomms1845_S1.html
652 Olsen, J. B., Lewis, C. J., Kretschmer, E. J., Wilson, S. L., & Seeb, J. E. (2002). Characterization of 14
653 tetranucleotide microsatellite loci derived from Pacific herring. *Molecular Ecology Notes, 2*(2),
654 101-103.
655 Palsbøll, P. J., Bérubé, M., & Allendorf, F. W. (2007). Identification of management units using
656 population genetic data. *Trends in Ecology & Evolution, 22*(1), 11-16.
657 doi:10.1016/j.tree.2006.09.003
658 Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach.
659 *Bioinformatics, 26*(3), 419-420. doi:10.1093/bioinformatics/btp696
660 Paris, J., R., Stevens, J., R., & Catchen, J., M. (2017). Lost in parameter space: a road map for stacks.
661 *Methods in Ecology and Evolution, 8*(10), 1360-1373. doi:10.1111/2041-210X.12775
662 Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software
663 for teaching and research—an update. *Bioinformatics, 28*(19), 2537-2539.
664 doi:10.1093/bioinformatics/bts460
665 Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus
666 Genotype Data. *Genetics, 155*(2), 945-959.
667 Pritchard, J. K., Wen, X., & Falush, D. (2010). Documentation for *structure* software: Version 2.3. In.
668 Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine
669 fish: mismatch between biological and fisheries management units. *Fish and Fisheries, 10*(4),
670 361-395. doi:10.1111/j.1467-2979.2008.00324.x
671 Rousset, F. (2008). Genepop'007: a complete reimplementation of the Genepop software for Windows
672 and Linux. *Molecular Ecology Resources, 8*, 103-106. doi:10.1111/j.1471-8286.2007.01931.x

Author Manuscript

This article is protected by copyright. All rights reserved

673 Schmieder, R., & Edwards, R. (2011). Fast Identification and Removal of Sequence Contamination from
674 Genomic and Metagenomic Datasets. *PLoS ONE, 6*(3), e17288.
675 doi:10.1371/journal.pone.0017288
676 Scribner, K. T., Lowe, W. H., Landguth, E., Luikart, G., Infante, D. M., Whelan, G. E., & Muhlfeld, C. C.
677 (2016). Applications of genetic data to improve management and conservation of river fishes
678 and their habitats. *Fisheries, 41*(4), 174-188. doi:10.1080/03632415.2016.1150838
679 Sehn, J. K., Spencer, D. H., Pfeifer, J. D., Bredemeyer, A. J., Cottrell, C. E., Abel, H. J., & Duncavage, E. J.
680 (2015). Occult specimen contamination in routine clinical next-generation sequencing testing.
681 *American Journal of Clinical Pathology, 144*(4), 667-674. doi:10.1309/ajcpr88wdjjldmbn
682 Small, M. P., Loxterman, J. L., Frye, A. E., Von Bargen, J. F., Bowman, C., & Young, S. F. (2005). Temporal
683 and Spatial Genetic Structure among Some Pacific Herring Populations in Puget Sound and the
684 Southern Strait of Georgia. *Transactions of the American Fisheries Society, 134*(5), 1329-1341.
685 doi:10.1577/T05-050.1
686 Tarpey, C. M., Seeb, J. E., McKinney, G. J., Templin, W. D., Bugaev, A. V., Sato, S., & Seeb, L. W. (2017).
687 SNP data describe contemporary population structure and diversity in allochronic lineages of
688 pink salmon (Oncorhynchus gorbuscha). *Canadian Journal of Fisheries and Aquatic Sciences*.
689 doi:10.1139/cjfas-2017-0023
690 Vandewoestyne, M., Van Hoofstat, D., De Groote, S., Van Thuyne, N., Haerinck, S., Haerinck S, . . .
691 Deforce, D. (2011). Sources of DNA contamination and decontamination procedures in the
692 forensic laboratory. *Journal of Forensic Research, S2*(001). doi:doi:10.4172/2157-7145.S2-001
693 Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control
694 are dependent on gene function and ancestry. *Bioinformatics, 31*(3), 318-323.
695 doi:10.1093/bioinformatics/btu668
696 Waples, R. S. (1998). Separating the wheat from the chaff: Patterns of genetic differentiation in high
697 gene flow species. *Journal of Heredity, 89*(5), 438-450.
698 Weir, B., & Cockerham, C. (1984). Estimating F-Statistics for the analysis of population structure.
699 *Evolution, 38*(6), 1358-1370.
700 Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure.
701 *Evolution, 38*(6), 1358-1370.
702 Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by
703 FST do not necessarily require large sample sizes when using many SNP markers. *PLOS ONE,
704 7*(8), e42649. doi:10.1371/journal.pone.0042649

Author Manuscript

This article is protected by copyright. All rights reserved

705 Yang, D. Y., & Watt, K. (2005). Contamination controls when preparing archaeological remains for
706 ancient DNA analysis. *Journal of Archaeological Science, 32*(3), 331-336.
707 doi:http://dx.doi.org/10.1016/j.jas.2004.09.008

708

709 ## Data accessibility

710 Sequence data (individual .fastq files) are available in the NCBI Sequence Read Archive under
711 accession PRJNA508972. Sample metadata, RAD genotypes, and the custom python genotyping
712 script are available in DRYAD under doi:10.5061/dryad.g28rh86.

713 ## Author contributions

714 ELP, LH, RK, DL, MM, and DY designed research. ELP and DD performed research and
715 analyzed the data. ELP, LH, and DD wrote the paper.

716

717

718

719

720

721

722

723

724

725

726

727

728

Author Manuscript

This article is protected by copyright. All rights reserved

729

730 ## Tables and Figures

731 *Tables*

732 **Table 1.** Sampling locations and associated collection information for samples used in this study.
733 Approximate GPS coordinates are provided for herring collected from Spiller Channel in 2001.

Sampling location	Latitude	Longitude	Sampling dates	Sexual maturity	Treatment groups	Sample size
Spiller Channel, BC	52.372	-128.188	3/14/2001, 4/4/2014	Spawning adult	Null, Bleach	11
Quilcene Bay, WA	47.808	-122.860	3/8/2012	Spawning adult	Null, Bleach	6
Cherry Point, WA	48.932	-122.798	9/21/2015	Juvenile	Null, Bleach	20
Spiller Channel, BC	52.372	-128.188	4/3/2015	Spawning adult	Bleach	48
Quilcene Bay, WA	47.808	-122.860	4/7/2014	Spawning adult	Bleach	48
Cherry Point, WA	48.932	-122.798	5/12/2014, 5/9/2016	Spawning adult	Bleach	98

734

735

736 **Table 2**. AMOVA results using two different hierarchical groupings. In AMOVA 1, groups are
737 defined by population (Cherry Point; Quilcene Bay; Spiller Channel) and subgroups consist of

Author Manuscript

This article is protected by copyright. All rights reserved

738 the two different treatments (bleach, null). In AMOVA 2, groups are defined by population and
739 subgroups consist of subsamples of individuals (N = 4-6); different iterations of this AMOVA
740 were conducted excluding untreated individuals and H_I outliers. F_{SC} is the differentiation among
741 subsamples within a group, while F_{CT} represents the differentiation among groups (i.e. among
742 the three populations). **Bold formatting**: $P< 0.001$, no formatting: $P> 0.05$

743

	Without individual level			*With individual level*		
AMOVA 1	F_{ST}	F_{SC}	F_{CT}	F_{IS}	F_{SC}	F_{CT}
All individuals	**0.0270**	**0.0414**	-0.0150	**-0.1034**	**0.0414**	-0.0139
AMOVA 2						
All individuals	**0.0255**	**0.0046**	**0.0209**	**-0.1100**	**0.0145**	**0.0210**
Bleached individuals	**0.0204**	**0.0010**	**0.0194**	**-0.0604**	**0.0065**	**0.0194**
Bleached individuals, no H_I outliers	**0.0206**	**0.0007**	**0.0199**	**-0.0356**	**0.0041**	**0.0199**

744

745 *Figure legends*

746 **Figure 1.** Sequencing quality data for juvenile herring in the null (black) and bleach (grey)
747 treatment groups. Each dot represents an individual herring sample. A) Number of raw
748 sequences per sample containing a restriction site, B) number of RAD loci identified in each
749 sample by *pstacks*, and C) average read depth per locus for each sample.

750

751 **Figure 2.** Distribution of H_I in each treatment group. Colors represent different treatments and
752 the dashed line shows the upper limit of H_I observed in the juvenile samples. Bleached adult
753 samples to the right of the dashed line are "H_I outliers" that likely contain residual
754 contamination.

755

Author Manuscript

This article is protected by copyright. All rights reserved

756 **Figure 3.** PCA (panels A, B, C) and nMDS (panels D, E, F) plots of herring genotyped at 3,502
757 RAD loci. In the PCA, each point represents an individual herring, while in the nMDS each point
758 represents a subsample of multiple herring (N= 4-7). Different colors depict the population from
759 which the samples were collected, while shapes (circle or triangle) are indicative of treatment
760 group. Note that juvenile herring samples (in both null and bleach treatments) cluster together
761 with adult samples collected from the same population (Cherry Point). A) PCA of all samples, B)
762 PCA of bleached samples; H_I outliers are circled in red, C) PCA of bleached samples when H_I
763 outliers are removed, D) nMDS of all samples, E) nMDS of bleached samples; H_I outliers are
764 circled in red, F) nMDS of bleached samples when H_I outliers are removed.

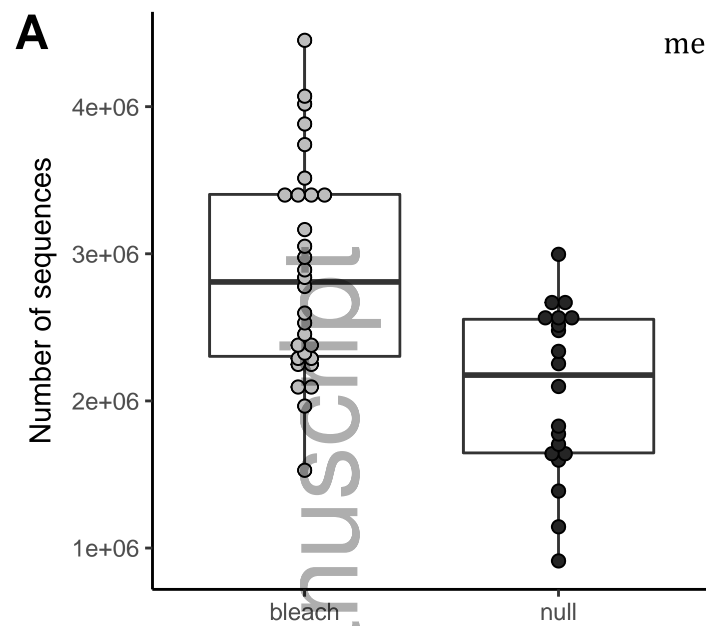
765

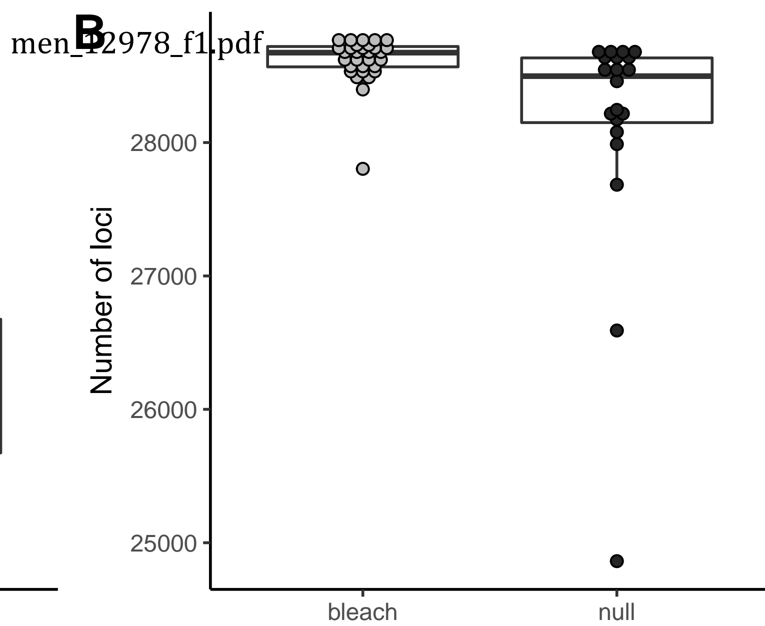
766 **Figure 4.** Population structure estimated using *Structure*. Each sample is portrayed by a vertical
767 line which consists of colored segments, representing the estimated fraction of an individual's
768 ancestry (*Q*) belonging to *k* clusters. Individuals represented by transparent bars are
769 contaminated adult samples. A) *Structure* analysis using only bleached samples and no H_I
770 outliers; *LnP(D)* and *ΔK* unambiguously identify $K = 2$ as the most likely number of clusters.
771 These clusters correspond to the major known spawning phenotypes of Pacific herring ("late-
772 spawners" and "primary-spawners"). B) *Structure* analysis using all samples and $K = 2$. The
773 presence of contaminated samples alters the values of *LnP(D)* and *ΔK,* compared to the clean
774 data set. C) *Structure* analysis using $K = 4$ and all samples.

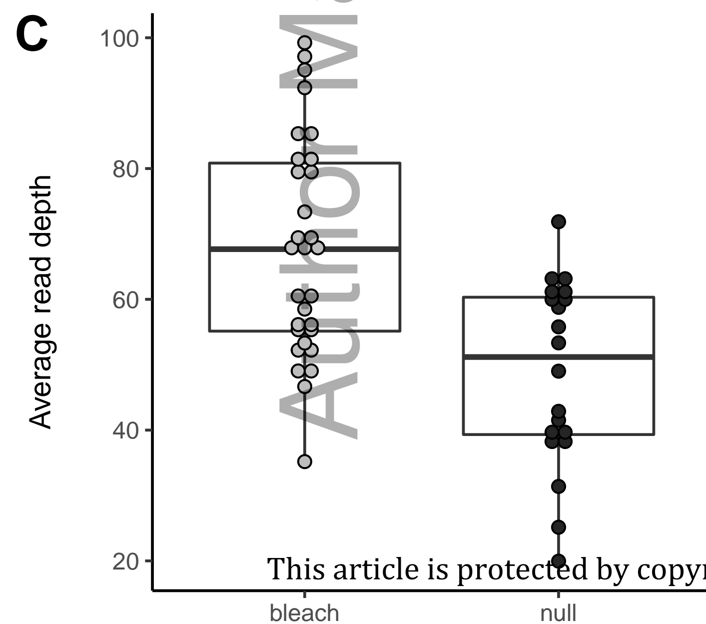
Author Manuscript

This article is protected by copyright. All rights reserved

men_12978_f1.pdf



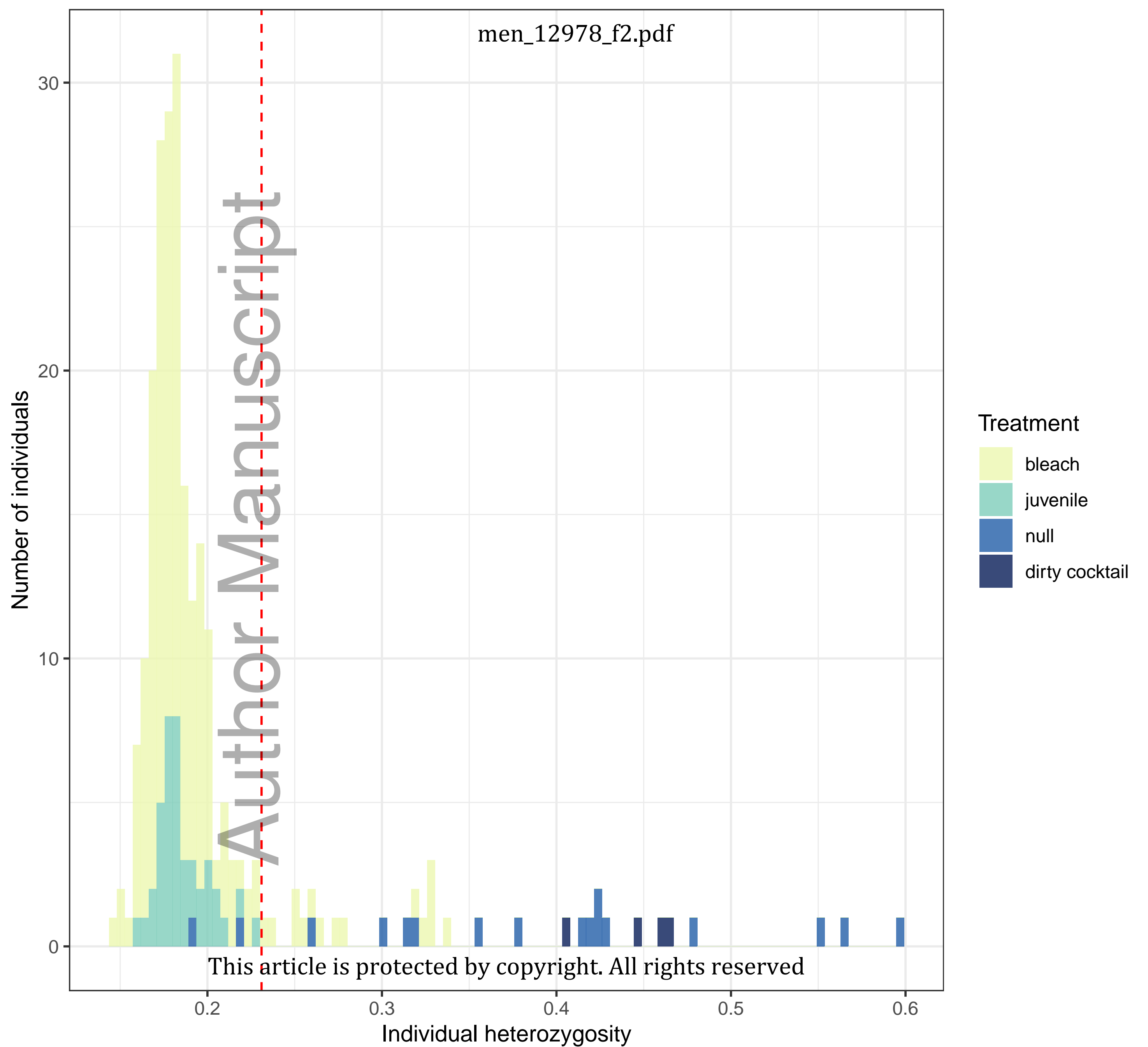




Author Manuscript

This article is protected by copyright. All rights reserved

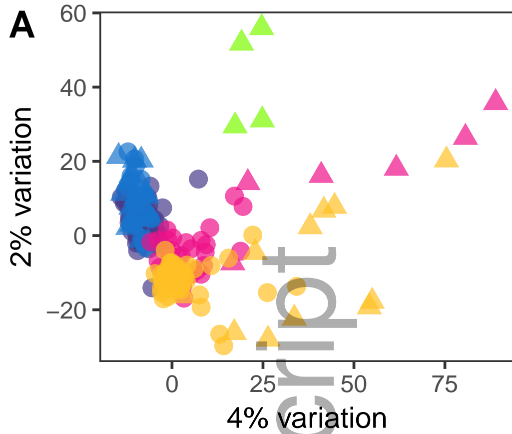
men_12978_f2.pdf

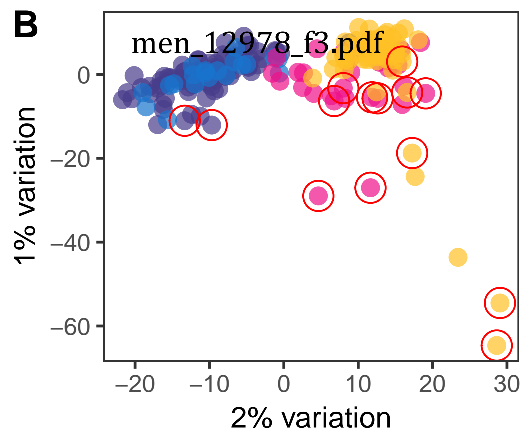


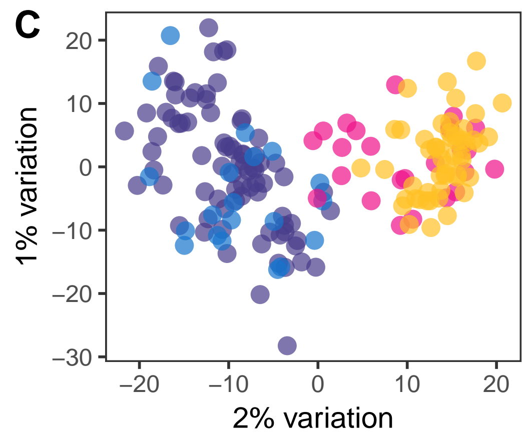
Author Manuscript

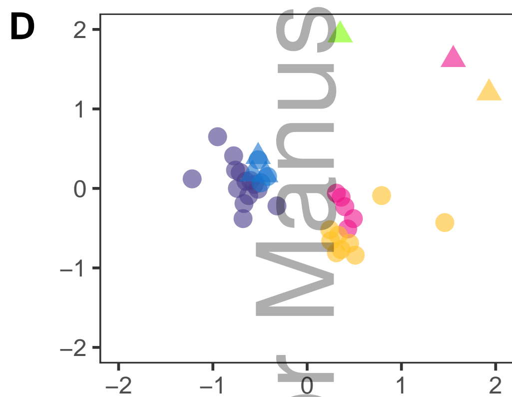
This article is protected by copyright. All rights reserved

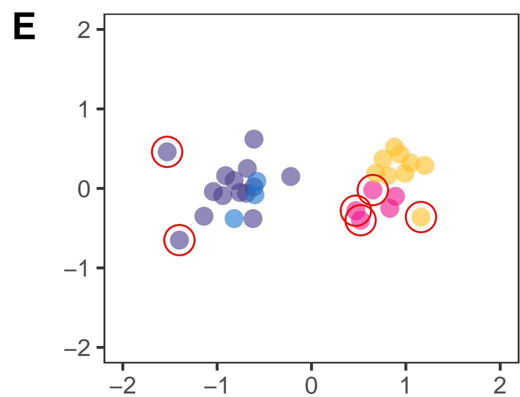
men_12978_f3.pdf

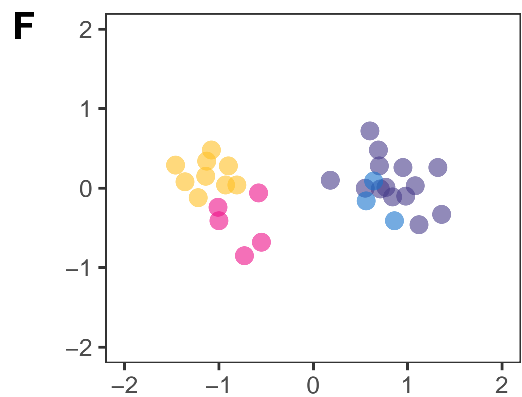












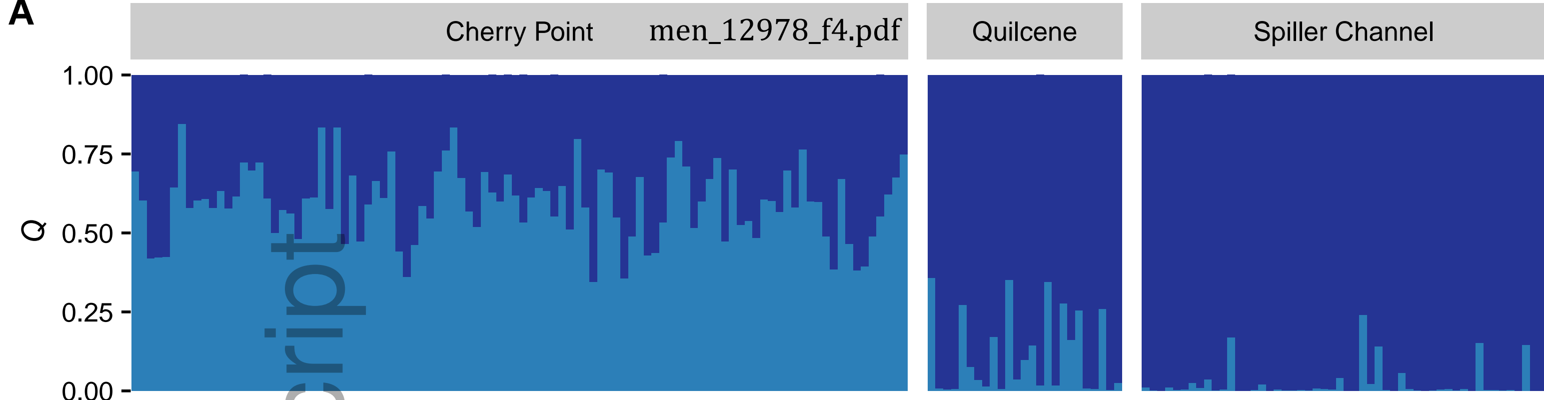
Population
Cherry Point
Cherry Point– juvenile
Dirty cocktail
Quilcene Bay
Spiller Channel

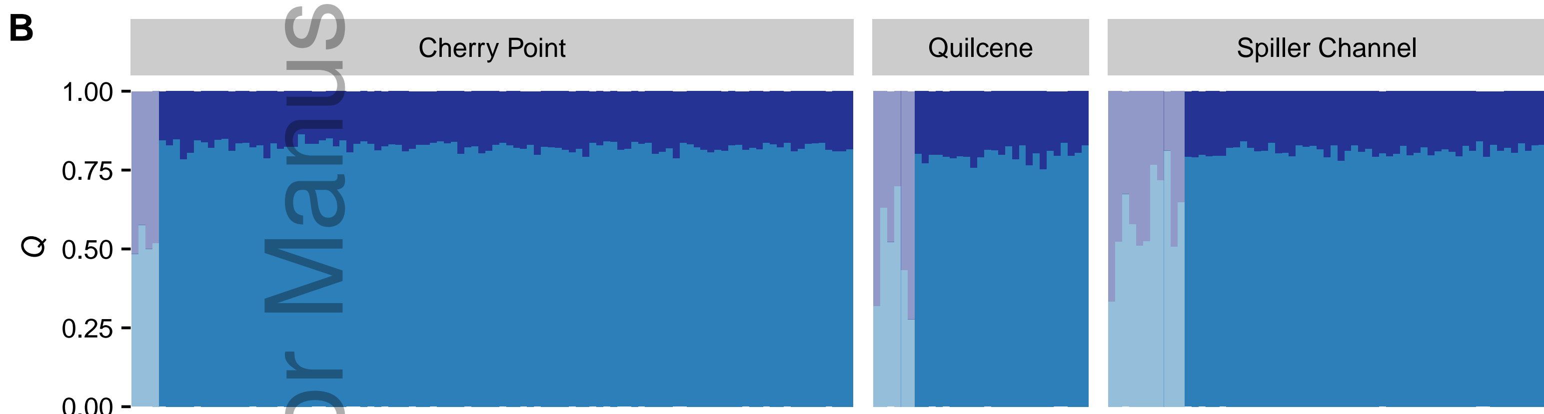
Treatment
bleach
null

Author Manuscript

This article is protected by copyright. All rights reserved

men_12978_f4.pdf





Author Manuscript

This article is protected by copyright. All rights reserved