


RESOURCE ARTICLE

Intraspecific DNA contamination distorts subtle population structure in a marine fish: Decontamination of herring samples before restriction-site associated sequencing and its effects on population genetic statistics

Eleni L. Petrou¹  | Daniel P. Drinan¹ | Robert Kopperl² | Dana Lepofsky³ | Dongya Yang³ | Madonna L. Moss⁴ | Lorenz Hauser¹

¹School of Aquatic and Fishery Sciences, University of Washington, Seattle, Washington

²Willamette Cultural Resources Associates Ltd., Seattle, Washington

³Department of Archaeology, Simon Fraser University, Burnaby, British Columbia, Canada

⁴Department of Anthropology, University of Oregon, Eugene, Oregon

Correspondence

Eleni L. Petrou, School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA.

Email: elpetrou@uw.edu

Funding information

Washington Sea Grant, University of Washington; National Oceanic and Atmospheric Administration, Grant/Award Number: NA14OAR4170078.; Natural Sciences and Engineering Research Council of Canada; National Science Foundation, Grant/Award Number: 1203868 and 1068839

Abstract

Wild specimens are often collected in challenging field conditions, where samples may be contaminated with the DNA of conspecific individuals. This contamination can result in false genotype calls, which are difficult to detect, but may also cause inaccurate estimates of heterozygosity, allele frequencies and genetic differentiation. Marine broadcast spawners are especially problematic, because population genetic differentiation is low and samples are often collected in bulk and sometimes from active spawning aggregations. Here, we used contaminated and clean Pacific herring (*Clupea pallasii*) samples to test (a) the efficacy of bleach decontamination, (b) the effect of decontamination on RAD genotypes and (c) the consequences of contaminated samples on population genetic analyses. We collected fin tissue samples from actively spawning (and thus contaminated) wild herring and nonspawning (uncontaminated) herring. Samples were soaked for 10 min in bleach or left untreated, and extracted DNA was used to prepare DNA libraries using a restriction site-associated DNA (RAD) approach. Our results demonstrate that intraspecific DNA contamination affects patterns of individual and population variability, causes an excess of heterozygotes and biases estimates of population structure. Bleach decontamination was effective at removing intraspecific DNA contamination and compatible with RAD sequencing, producing high-quality sequences, reproducible genotypes and low levels of missing data. Although sperm contamination may be specific to broadcast spawners, intraspecific contamination of samples may be common and difficult to detect from high-throughput sequencing data and can impact downstream analyses.

KEYWORDS

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

1 | INTRODUCTION

High-throughput DNA sequencing has advanced the field of molecular ecology by enabling comprehensive investigations of genetics

and genomics in nonmodel species (Allendorf, Hohenlohe, & Luikart, 2010; Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Ekblom & Galindo, 2011). However, high-throughput sequencing is sensitive to the contamination of samples with exogenous (nontarget) DNA. Errors

introduced by interspecific DNA contamination have been identified in whole genome assemblies (Koutsovoulos et al., 2016; Longo, O'Neill, & O'Neill, 2011), ancient DNA (Campana, Robles García, Rühli, & Tuross, 2014) and metagenomic data sets (Schmieder & Edwards, 2011). To address the problem of interspecific contamination, bioinformatic tools have been developed to remove exogenous DNA from sequence data (Schmieder & Edwards, 2011) before contaminated sequences are incorporated into downstream analyses. These methods typically identify nontarget sequences by aligning them to databases of common contaminating species; as a result, they cannot be used to detect intraspecific contamination caused by the unintentional mixing of DNA between individual samples of the same species.

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

Some bioinformatic tools have been developed to screen sequences for intraspecific DNA contamination (Flickinger, Jun, Abecasis, Boehnke, & Kang, 2015; Jun et al., 2012), but these tools were primarily developed for human resequencing studies; as such, they require pre-existing baseline data on population allele frequencies or high-coverage individual genotypes to identify contaminated individuals. These types of genomic resources are oftentimes unavailable for nonmodel species, and consequently, little attention has been given to the potential problem of intraspecific DNA contamination in most molecular ecology studies.

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

One of the standard methods to decontaminate samples is treatment with bleach; this approach has been used to clean bone samples before sequencing of ancient DNA (Kemp & Smith, 2005; Yang & Watt, 2005), as well as fresh tissue samples for microsatellite (Mitchell et al., 2008) and mitochondrial analysis (Greenstone et al., 2012). However, traditional microsatellite and mitochondrial sequencing, as

well as high-throughput sequencing of ancient DNA, can utilize short DNA fragments as template. In contrast, RAD sequencing requires very high-quality DNA with intact restriction sites; otherwise, there is a dramatic reduction in the number of raw sequences produced (Graham et al., 2015). Given that bleach decontaminates samples by degrading surface DNA (Kemp & Smith, 2005), the effect of bleach on the quality and quantity of endogenous sequence reads produced by RAD sequencing is currently unknown. Therefore, bleach treatment may affect downstream analyses, even if decontamination were successful.

Here, we used contaminated and clean Pacific herring (*Clupea pallasii*) samples to test (a) the efficacy of bleach decontamination, (b) the effect of decontamination on RAD genotypes and (c) the consequences of contaminated samples on population genetic analyses. By combining these results, we identified the impacts of contamination on population genetic analyses and empirically validated an approach aimed at minimizing contamination that is compatible with RAD sequencing.

2 | MATERIALS AND METHODS

2.1 | Sample collection

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

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

TABLE 1 Sampling locations and associated collection information for samples used in this study. 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

2.2 | Experimental assessment of bleach treatment

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

1. Null treatment: samples were stored in 100% ethanol until DNA extraction.
2. Bleach treatment: following a modified protocol of Mitchell et al. (2008), samples were placed in individual tubes and immersed in $180 \mu\text{l}$ of 0.12% sodium hypochlorite (bleach) (Sigma-Aldrich, St. Louis, MO, USA) for ten minutes. During bleach incubation, samples were vortexed at medium–high speed. Subsequently, we removed bleach from the tubes and added $200 \mu\text{l}$ of Milli-Q purified water (Millipore, Bedford, MA, USA). Samples were vortexed for one minute at medium–high speed, after which Milli-Q water was removed and fresh Milli-Q water was added to the tube. This water rinse was repeated five times, and samples were stored in 100% ethanol until DNA extraction.

To estimate genotyping error rates within and between treatment groups, five juvenile herring were subsampled in replicate, and both subsamples were subjected to both experimental treatments. In addition, we also created four “dirty cocktails” as reference positive controls for DNA contamination. Each dirty cocktail contained $25 \text{ ng}/\mu\text{l}$ of DNA from four different juvenile herring in equal proportions.

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

2.3 | DNA library preparation and sequencing

Genomic DNA was extracted from each subsample using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA). DNA was visualized with agarose gel electrophoresis to assess DNA quality

and quantified with the PicoGreen dsDNA Assay Kit (Invitrogen, Waltham, MA, USA). We standardized the DNA concentration of each sample to $25 \text{ ng}/\mu\text{l}$.

As an initial check for contamination, six microsatellite loci (*Cpa-8*, *Cpa-104*, *Cpa-113* (Miller, Laberee, Schulze, & Kaukinen, 2001) and *Cpa-106*, *Cpa-107a*, *Cpa-111* (Olsen, Lewis, Kretschmer, Wilson, & Seeb, 2002)) were used by the Washington Department of Fish and Wildlife Molecular Genetics Laboratory to screen every sample that was present in both the bleach and null treatment groups ($N = 37$), following the protocol of Olsen et al. (2002). Alleles were scored on Peak Scanner 2 (Life Technologies, Carlsbad, CA, USA). In the microsatellite data, we defined contaminated samples as those containing more than two alleles at any locus.

We followed the protocol of Etter, Bassham, Hohenlohe, Johnson, and Cresko (2012) to prepare DNA libraries for restriction site-associated DNA (RAD) sequencing. Depending on availability, 200 to 500 ng (depending on availability) of genomic DNA per individual was digested with the restriction enzyme *SbfI* (New England Biolabs, Ipswich, MA). Samples were individually labelled using a custom set of 96 barcodes (Integrated DNA Technologies, San Diego, CA), and groups of 12 samples were pooled into libraries that were sheared to a length of approximately 500 base pairs (bp) using a Bioruptor Sonicator (Diagenode, Denville, NJ). We modified the Etter et al. (2012) protocol by using AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA) to size-select DNA fragments (300–500 bp) and purify DNA products. However, all other steps (blunt-end repair, 3'-dA overhang addition, P2 adapter ligation and PCR) were conducted as described in Etter et al. (2012). After PCR, the DNA concentration of each library was quantified using the PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA). We standardized the concentration of each library to 10 nM and pooled libraries such that 48 individuals were sequenced per lane of an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA) at the University of Oregon Genomics Core Facility. The resulting sequences were single-end and 100 bp in length.

2.4 | Bioinformatics analyses

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

Following the protocol of Briec, Waters, Seeb, and Naish (2014), we created a reference database of herring RAD loci to facilitate sequence assembly and locus identification. The reference database was built using juvenile samples (null treatment) that had at least 1.5 million sequences ($N = 19$). First, we assembled sequences and identified loci in these samples using the de novo locus discovery pipeline in *Stacks*. Loci within each sample were allowed to have up to three nucleotide mismatches (*ustacks*, $M = 3$), and each allele had to be sequenced at a minimum depth of $5\times$ to be retained in the analysis (*ustacks*, $m = 5$). Subsequently, we removed loci with tandem repeat units using *Blast* version 2.2.25 (Altschul, Gish, Miller, Myers, & Lipman, 1990) and *bowtie* version 0.12.7 (Langmead, Trapnell, Pop, & Salzberg, 2009) as described in Briec et al. (2014).

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

We removed possible sequencing errors by filtering the SNPs discovered by *Stacks*. A custom python script published in Briec et al. (2014) was used to retain only loci with two haplotypes and to rescore genotypes. This method designates a heterozygote genotype if each allele is sequenced at least twice and the locus is sequenced to a depth of at least ten reads. Subsequently, we filtered out loci and individuals that had more than 20% missing data. Loci characterized by very low minor allele frequencies were filtered from the final data set; a minor allele had to be present in at least one of the treatment groups at a frequency of 0.05 for that locus to be retained in downstream analyses. Finally, we tested for deviations from Hardy–Weinberg equilibrium (HWE) using the exact test based on 1,000 Monte Carlo permutations of alleles, as implemented in the R package *pegas* (Paradis, 2010). Loci that were out of HWE in every one of the population genetic samples (Cherry Point, Quilcene Bay and Spiller Channel) were removed from the analysis. As a final assessment of locus assembly, we followed the recommendations of Paris, Stevens, & Catchen (2017) and aligned the filtered set of loci to the Atlantic herring genome using *bowtie2* version 2.2.6 (Langmead

& Salzberg, 2012). We also estimated per-locus F_{IS} at each sampling location using *Genepop* version 4 (Rousset, 2008).

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

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

2.5 | Population structure

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

To further investigate the effects of contamination and bleach treatment on measures of population structure, populations were divided into 39 subsamples of approximately six individuals (range = 4–7 individuals), the sample size of the smallest collection of contaminated individuals from a single location. A recent study (Willing, Dreyer, & van Oosterhout, 2012) showed that a small number of individuals ($N = 4$ –6) can be used to obtain unbiased estimates of F_{ST} when large numbers of loci ($N > 1,000$) are genotyped. Pairwise F_{ST} (Weir & Cockerham 1984) between subsamples

was calculated in Genepop version 4 and used for nonmetric multidimensional scaling (nMDS) in Primer 6 (Clarke & Gorley, 2006). Observed heterozygosity and expected heterozygosity were calculated in GenAEx version 6.5 (Peakall & Smouse, 2012), and F_{IS} (Weir & Cockerham 1984) was estimated in Genepop version 4 (Rousset, 2008). To compare differentiation with and without contaminated individuals, hierarchical AMOVAs were calculated in Arlequin version 3.52 (Excoffier & Lischer, 2010), using two alternative groupings. In the first comparison, groups were defined by population (Cherry Point; Quilcene Bay; Spiller Channel) and subgroups consisted of the two different treatments (bleach, null). In the second comparison, groups were defined by population and subgroups consisted of subsamples of individuals ($N = 4-7$); different iterations of this AMOVA were conducted excluding untreated individuals and H_i outliers.

3 | RESULTS

3.1 | Sequencing and genotyping

We successfully genotyped 92% of individuals at three or more microsatellite loci. Six out of 17 adult herring in the null treatment group displayed more than two alleles per microsatellite locus, indicating that they were contaminated with the DNA of multiple herring. Treatment with bleach appeared to remove contamination from all but one of the samples. None of the 20 juvenile herring had more than two microsatellite alleles after either treatment, demonstrating lack of contamination and confirming our hypothesis that sample

contamination was caused by the presence of sperm in the water column in wild spawning aggregations.

A reference database of RAD loci was built using sequences from 19 juvenile herring in the null treatment group; one individual was excluded from the database because it contained fewer than 1.5 million raw sequences. A total of 29,551 putative loci were initially identified, and 28,997 loci were retained in the reference database after filtering out loci with tandem repeats and highly repetitive sequences.

After removing loci that were out of HWE in every population, we identified 3,502 biallelic RAD loci that were sequenced at a minimum read depth of 10 sequences in more than 80% of individuals and had a minor allele frequency that exceeded 0.05 in at least one of the populations. We found that 93% of these loci aligned exactly once to the closely related Atlantic herring genome. Locus-specific estimates of F_{IS} were distributed around zero (Figure S1), which is concordant with expectations under HWE. A total of 240 herring had less than 20% missing genotypes and were retained in the final data set.

Sequencing quality was robust, and genotyping error was low for juvenile samples in the null and bleached treatment groups. Juvenile samples treated with bleach were characterized by slightly more sequences containing the restriction site (RADtags), loci per sample and average read depth (Figure 1). However, the genotype mismatch rate between treatments in the replicated juvenile individuals was very low ($1.8\% \pm 1.4\%$, mean \pm SD) and similar to repeated bleach treatments ($1.4\% \pm 1.3\%$). The distribution of genotype mismatches did not differ statistically between

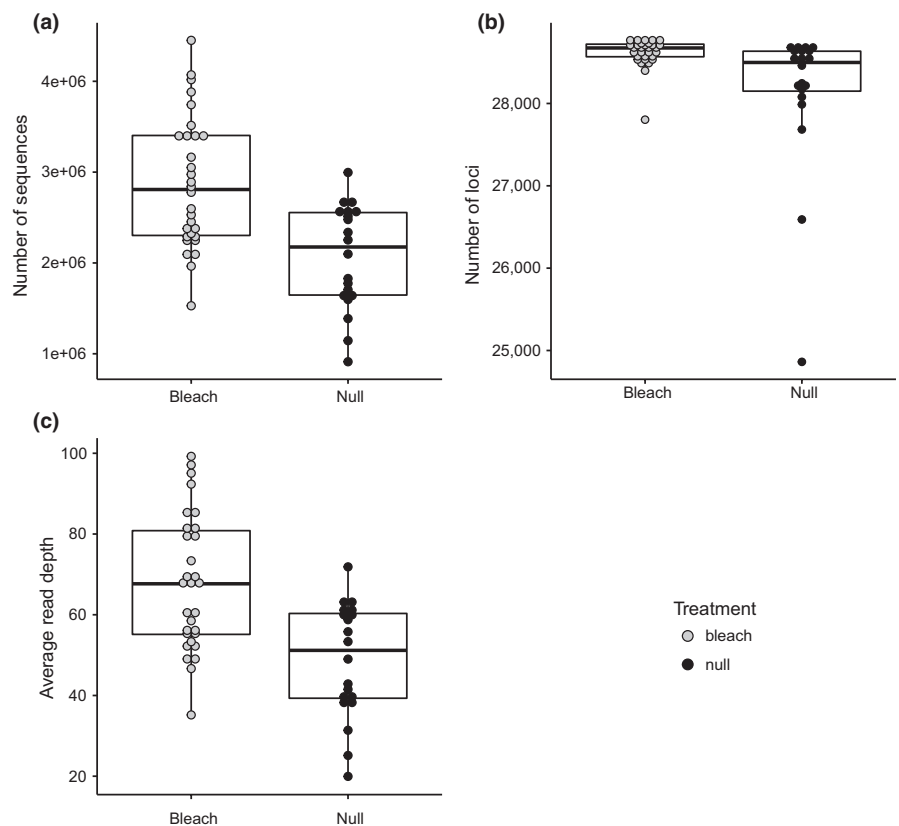


FIGURE 1 Sequencing quality data for juvenile herring in the null (black) and bleach (grey) treatment groups. Each dot represents an individual herring sample. (a) Number of raw sequences per sample containing a restriction site, (b) number of RAD loci identified in each sample by *pstacks* and (c) average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

replicate individuals in the same (bleached) or across (null vs. bleached) treatment groups (Wilcoxon sign rank test, $p = 0.55$), indicating that treatment with bleach does not alter the endogenous ("true") genotype of a sample.

3.2 | Impacts of contamination on individual-level variation

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

evidence for residual contamination in cleaned adult samples, as 8% (14/174) of those samples had H_I that was above the range observed in juvenile samples (Figure 2).

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

Multiple runs of *Structure* identified $K = 2$ as the most likely number of groups when only cleaned data were included in the analysis. This result was supported by estimates of the posterior probability of the data given K clusters ($\text{LnP}(D)$) and ΔK (Figure 4a).

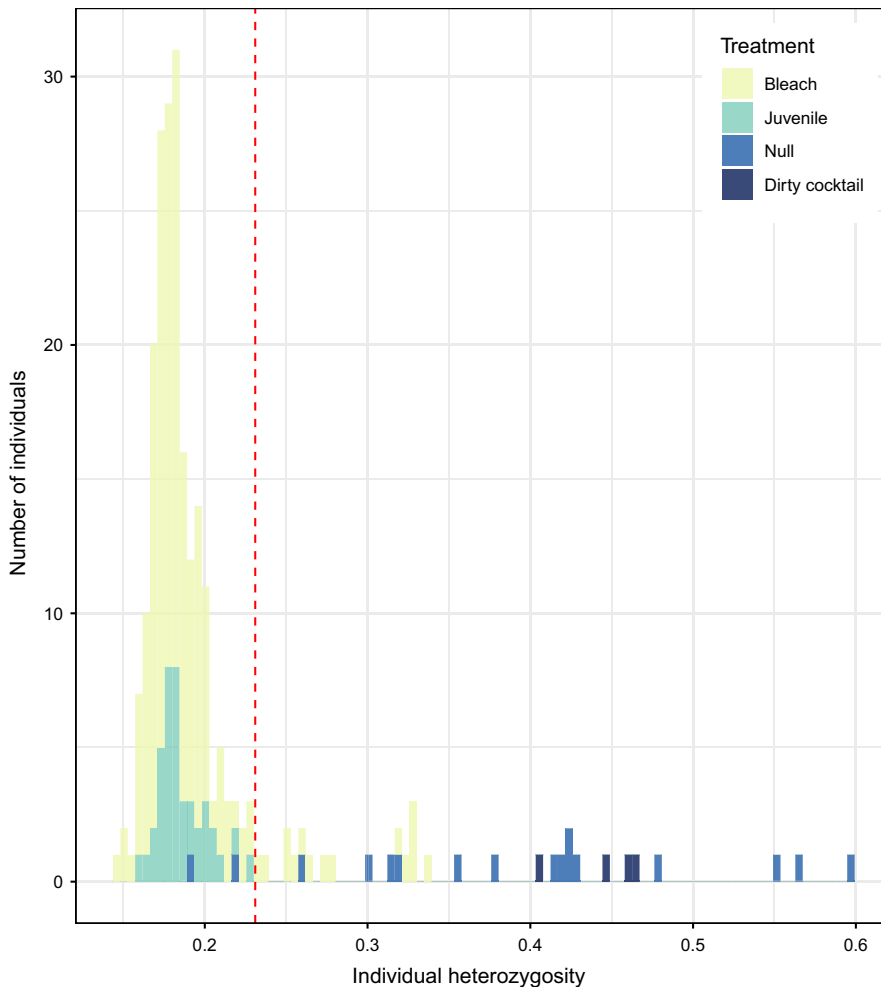


FIGURE 2 Distribution of H_I in each treatment group. Colours represent different treatments and the dashed line shows the upper limit of H_I observed in the juvenile samples. Bleached adult samples to the right of the dashed line are " H_I outliers" that likely contain residual contamination [Colour figure can be viewed at wileyonlinelibrary.com]

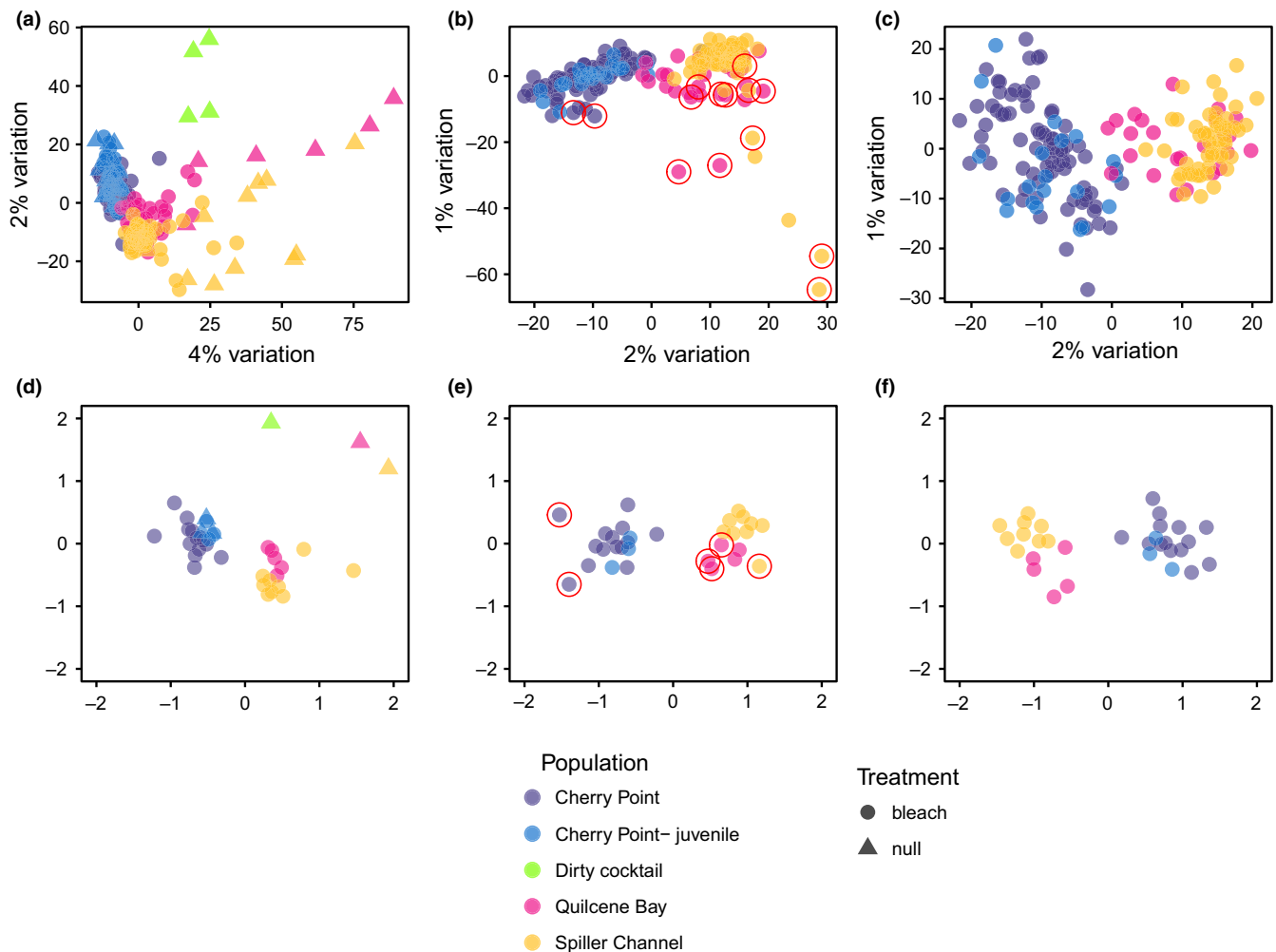


FIGURE 3 Principal component analysis (PCA) (panels a, b, c) and nMDS (panels d, e, f) plots of herring genotyped at 3,502 RAD loci. In the PCA, each point represents an individual herring, while in the nMDS each point represents a subsample of multiple herring ($N = 4-7$). Different colours depict the population from which the samples were collected, while shapes (circle or triangle) are indicative of treatment group. Note that juvenile herring samples (in both null and bleach treatments) cluster together with adult samples collected from the same population (Cherry Point). (a) PCA of all samples; (b) PCA of bleached samples, H_I outliers are circled in red; (c) PCA of bleached samples when H_I outliers are removed; (d) nMDS of all samples; (e) nMDS of bleached samples, H_I outliers are circled in red; (f) nMDS of bleached samples when H_I outliers are removed [Colour figure can be viewed at wileyonlinelibrary.com]

Fish collected at Cherry Point (adults and juveniles) formed a distinct cluster, while fish collected at Quilcene Bay and Spiller Channel strongly assigned to a second cluster. In contrast, when all samples (including contaminated adults) were included in the same *Structure* analysis, $\text{Ln}P(D)$ and ΔK did not converge on the same answer (Figure 4b and c). The posterior probability of the data given K clusters was highest at $K = 4$, while the distribution of ΔK showed peaks at both $K = 2$ and $K = 4$ (Figure S2). At $K = 2$, the estimated ancestry coefficient of bleached samples was symmetric across all sampling locations ($Q = 0.82 \pm 0.02$, mean \pm SD), while it was quite different for contaminated samples (Figure 4b). At $K = 4$, the same pattern was observed, although population differentiation was more apparent in both clean and contaminated samples (Figure 4c). In all cases, however, all individuals appeared to be highly admixed, most likely because of low population differentiation.

3.3 | Impacts of contamination on estimates of population structure

Similar and considerable effects of contamination were apparent for population parameters (H_e , F_{IS} , F_{ST}) estimated from subsamples of individuals drawn from each herring population (Cherry Point, Quilcene Bay and Spiller Channel). All contaminated subsamples and the “dirty cocktail” had a more negative F_{IS} (indicating an excess of heterozygotes) and higher expected heterozygosity values than bleached adult subsamples lacking H_I outliers (Figure S3). In addition, subsamples of juvenile herring had similar values of heterozygosity and F_{IS} before and after bleaching. Most adult subsamples had similar heterozygosity and an F_{IS} close to zero after bleaching, especially when H_I outliers were removed.

Contamination also had a clear effect on genetic differentiation between subsamples of individuals selected from the same

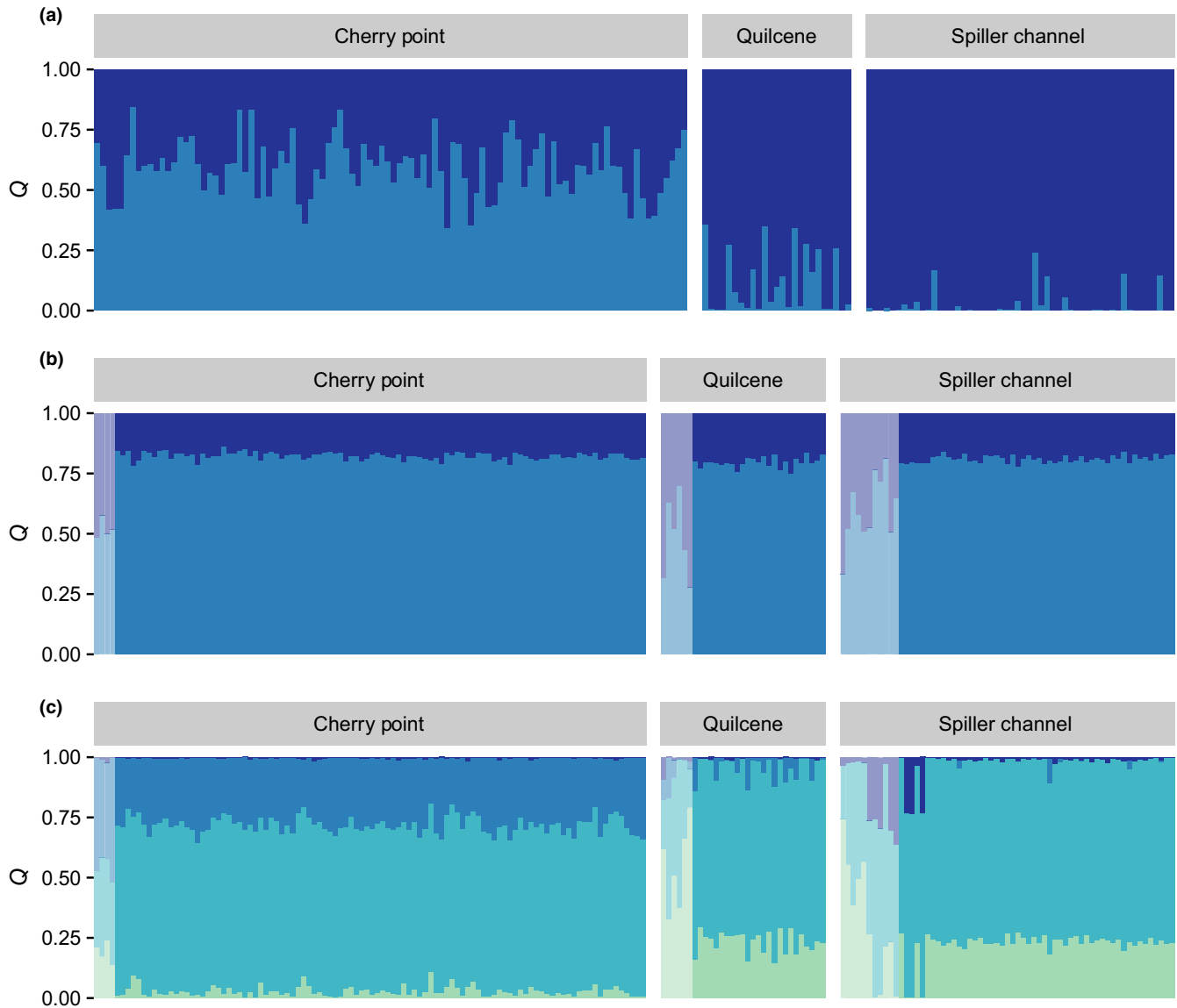


FIGURE 4 Population structure estimated using *Structure*. Each sample is portrayed by a vertical line which consists of coloured segments, representing the estimated fraction of an individual's ancestry (Q) belonging to k clusters. Individuals represented by transparent bars are contaminated adult samples. (a) *Structure* analysis using only bleached samples and no H_I outliers; $\text{LnP}(D)$ and ΔK unambiguously identify $K = 2$ as the most likely number of clusters. These clusters correspond to the major known spawning phenotypes of Pacific herring ("late spawners" and "primary spawners"). (b) *Structure* analysis using all samples and $K = 2$. The presence of contaminated samples alters the values of $\text{LnP}(D)$ and ΔK , compared to the clean data set. (c) *Structure* analysis using $K = 4$ and all samples [Colour figure can be viewed at wileyonlinelibrary.com]

population. Subsamples containing highly contaminated individuals were outliers in the nMDS analysis (Figure 3d–f). Both the "dirty cocktail" and the unbleached adult subsamples exhibited high differentiation from bleached subsamples taken from the same population (Figure 3d, $F_{ST} = 0.015$ – 0.070 , Table S1). After bleaching, adult herring subsamples taken from the same population were less differentiated from each other (Figure 3f, $F_{ST} = -0.009$ – 0.019 , Table S1), although subsamples containing H_I outliers exhibited higher differentiation (Figure 3e, $F_{ST} = 0.016$ – 0.028 , Table S1). The lowest F_{ST} values were observed between the bleached and unbleached replicate subsamples of the same juvenile individuals (Table S1).

Hierarchical AMOVAs demonstrated that contamination can inflate underlying genetic population differentiation (Table 2). When contaminated individuals were included in comparisons of population and treatment (Table 2, AMOVA 1), the differentiation between treatment groups from the same population (F_{SC}) was greater than the differentiation observed between distinct populations (F_{CT}). When contaminated individuals were included in an AMOVA using subsamples of individuals (Table 2, AMOVA 2), contamination inflated the overall F_{ST} . Contamination also increased the differentiation between population groups (F_{CT}) as well as the differentiation among subsamples within a population (F_{SC}). Adding individual-level analyses into the

TABLE 2 AMOVA results using two different hierarchical groupings. In AMOVA 1, groups are defined by population (Cherry Point; Quilcene Bay; Spiller Channel) and subgroups consist of the two different treatments (bleach, null). In AMOVA 2, groups are defined by population and subgroups consist of subsamples of individuals ($N = 4-6$); different iterations of this AMOVA were conducted excluding untreated individuals and H_i outliers. F_{SC} is the differentiation among subsamples within a group, while F_{CT} represents the differentiation among groups (i.e., among the three populations). Bold formatting: $p < 0.001$, no formatting: $p > 0.05$

	Without individual level			With individual level		
	F_{ST}	F_{SC}	F_{CT}	F_{IS}	F_{SC}	F_{CT}
AMOVA 1						
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

AMOVA did not change these trends, although the presence of contaminated samples was clearly indicated by more negative F_{IS} values.

4 | DISCUSSION

4.1 | Effects of contamination

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

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

These findings underscore the utility of using clean samples to estimate empirical distributions of H_i . A modest number of clean reference samples can be used to construct a baseline for comparison with potentially contaminated samples using the simple metric of H_i . Furthermore, H_i is a standard metric that is commonly reported in population genetic studies (Hoffman et al., 2014; Kjeldsen et al., 2016; Tarpey et al., 2017). To our knowledge, this is the one of the first studies of wild populations to examine patterns of H_i as a quality-control measure, even though a related metric (ratio of

heterozygous/nonreference homozygous sites) is commonly used in the quality control of human genomic data (Wang, Raskin, Samuels, Shyr, & Guo, 2015). We recommend that researchers examine the distribution of H_i in their data across individuals and populations, and carefully consider whether outlier samples could be caused by intraspecific DNA contamination.

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

Marine species are characterized by weak population differentiation that is sensitive to sampling errors (Waples, 1998). A possible consequence of contamination would be that "noise" introduced into a data set through contaminating alleles would overwhelm faint signals of genetic differentiation between populations. Indeed, this hypothesis was confirmed by our results; contaminated samples appeared as outliers in every analysis and led to inflated estimates of population differentiation (F_{ST}) and differentiation among subsamples within a population (F_{SC}) in an

AMOVA framework. Clustering approaches were also strongly affected by contamination: heavily contaminated individuals and population subsamples were outliers in PCA and nMDS analyses, and may thus impact the interpretation from such approaches. *Structure* results were also dominated by contaminated samples, and $\ln P(D)$ and ΔK did not converge on the same value of K when these contaminated samples were included in the data. Without contaminated samples, *Structure* detected subtle but clear population structure. Contamination can therefore distort true population structure, which is especially problematic in the context of conservation genetics and resource management, as genetic data are often used to help delineate conservation or management units (Funk, McKay, Hohenlohe, & Allendorf, 2012; Palsbøll, Bérubé, & Allendorf, 2007; Scribner et al., 2016). Thus, it is possible that contaminated genotypes could lead to the erroneous designation of management units and the accidental overexploitation of harvested populations.

4.2 | Efficacy of bleach treatment

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

Once contaminated individuals were removed from the data set, subsamples of individuals taken from the same location produced very concordant estimates of F_{ST} , even though subsample sizes were tiny ($N = 4-7$). However, it has been shown that reliable F_{ST} estimates can be obtained from very few individuals if loci can be sampled without bias (Willing et al., 2012). Population genetic analyses using individuals or subsamples of individuals confirmed previous genetic studies of Pacific herring, which found that Cherry Point herring were reproductively isolated from other populations due to differences in their spawn timing (Beacham et al., 2008; Mitchell, 2006; Small et al., 2005). Additionally, we found that nMDS and AMOVA analyses using subsamples of individuals detected subtle but significant genetic differentiation between herring populations from Quilcene Bay and Spiller Channel, which spawn at similar times of year (Table 1). This result indicates that analyses based on small subsamples of individuals may be more powerful than those based on full samples, as suggested by Nielsen et al. (2012).

Previous research has shown that RAD sequencing requires very high-quality DNA as input; otherwise, there is a significant reduction

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

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

In conclusion, we show that intraspecific DNA contamination can affect subtle patterns of population structure that are characteristic of many marine fish. We verified that treatment 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 Briec and Charles D. Waters contributed to 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 subagencies. 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.

DATA ACCESSIBILITY

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

AUTHOR CONTRIBUTIONS

E.L.P., L.H., R.K., D.L., M.M. and D.Y. designed research. E.L.P. and D.D. performed research and analysed the data. E.L.P., L.H. and D.D. wrote the paper.

ORCID

Eleni L. Petrou  <https://orcid.org/0000-0001-7811-9288>

REFERENCES

- Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, 11(10), 697–709. <https://doi.org/10.1038/nrg2844>.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17(2), 81–92. <https://doi.org/10.1038/nrg.2015.28>.
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10), e3376. <https://doi.org/10.1371/journal.pone.0003376>.
- Balloux, F., Amos, W., & Coulson, T. (2004). Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, 13(10), 3021–3031. <https://doi.org/10.1111/j.1365-294X.2004.02318.x>.
- Beacham, T. D., Schweigert, J. F., MacConnachie, C., Le, K. D., & Flostrand, L. (2008). Use of microsatellites to determine population structure and migration of Pacific herring in British Columbia and adjacent regions. *Transactions of the American Fisheries Society*, 137(6), 1795–1811. <https://doi.org/10.1577/T08-033.1>.
- Briec, M. S. O., Waters, C. D., Seeb, J. E., & Naish, K. A. (2014). A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence after an ancestral whole genome duplication event. *G3: Genes|genomes|genetics*, 4(3), 447–460. <https://doi.org/10.1534/g3.113.009316>.
- Bucholtz, R. H., Tomkiewicz, J., & Dalskov, J. (2008). *Manual to determine gonadal maturity of herring (Clupea harengus L.)*. Charlottenlund, Denmark: DTU Aqua, National Institute of Aquatic Resources.
- Cale, C. M., Earll, M. E., Latham, K. E., & Bush, G. L. (2016). Could secondary DNA transfer falsely place someone at the scene of a crime? *Journal of Forensic Sciences*, 61(1), 196–203. <https://doi.org/10.1111/1556-4029.12894>.
- Campana, M. G., Robles García, N., Rühli, F. J., & Tuross, N. (2014). False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. *BMC Research Notes*, 7(1), 111. <https://doi.org/10.1186/1756-0500-7-111>.
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140. <https://doi.org/10.1111/mec.12354>.
- Clarke, K. R., & Gorley, R. N. (2006). *PRIMER v6: User Manual/Tutorial*. Plymouth: PRIMER-E.
- Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4(2), 359–361. <https://doi.org/10.1007/s12686-011-9548-7>.
- Eklom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1–15. <https://doi.org/10.1038/hdy.2010.152>.
- Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., & Cresko, W. A. (2012). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In V. Orgogozo & M. Rockman (Eds.), *Molecular methods for evolutionary genetics. Methods in molecular biology (Methods and protocols)* (Vol. 772, pp. 157–178). New York, NY: Humana Press.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, 14(8), 2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>.
- Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3), 564–567. <https://doi.org/10.1111/j.1755-0998.2010.02847.x>.
- Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M., & Kang, H. M. (2015). Correcting for sample contamination in genotype calling of DNA sequence data. *The American Journal of Human Genetics*, 97(2), 284–290. <https://doi.org/10.1016/j.ajhg.2015.07.002>.
- Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Molecular Ecology Resources*, 16(4), 966–978. <https://doi.org/10.1111/1755-0998.12519>.
- Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, 27(9), 489–496. <https://doi.org/10.1016/j.tree.2012.05.012>.
- Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., ... Somers, C. M. (2015). Impacts of degraded DNA on

- restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources*, 15(6), 1304–1315. <https://doi.org/10.1111/1755-0998.12404>.
- Greenstone, M. H., Weber, D. C., Coudron, T. C., & Payton, M. E. (2011). Unnecessary roughness? Testing the hypothesis that predators destined for molecular gut-content analysis must be hand-collected to avoid cross-contamination. *Molecular Ecology Resources*, 11(2), 286–293. <https://doi.org/10.1111/j.1755-0998.2010.02922.x>.
- Greenstone, M. H., Weber, D. C., Coudron, T. A., Payton, M. E., & Hu, J. S. (2012). Removing external DNA contamination from arthropod predators destined for molecular gut-content analysis. *Molecular Ecology Resources*, 12(3), 464–469. <https://doi.org/10.1111/j.1755-0998.2012.03112.x>.
- Hoffman, J. I., Simpson, F., David, P., Rijks, J. M., Kuiken, T., Thorne, M. A. S., ... Dasmahapatra, K. K. (2014). High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences*, 111(10), 3775–3780. <https://doi.org/10.1073/pnas.1318945111>
- Hourston, A. S., & Rosenthal, H. (1976). Sperm density during active spawning of Pacific herring (*Clupea harengus pallasii*). *Journal of the Fisheries Research Board of Canada*, 33(8), 1788–1790. <https://doi.org/10.1139/f76-226>.
- Hubisz, M. J., Falush, D., Stephens, M., & Pritchard Jonathan, K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9(5), 1322–1332. <https://doi.org/10.1111/j.1755-0998.2009.02591.x>.
- Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., ... Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *The American Journal of Human Genetics*, 91(5), 839–848. <https://doi.org/10.1016/j.ajhg.2012.09.004>.
- Kemp, B. M., & Smith, D. G. (2005). Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. *Forensic Science International*, 154(1), 53–61. <https://doi.org/10.1016/j.forcint.2004.11.017>.
- King, R. A., Davey, J. S., Bell, J. R., Read, D. S., Bohan, D. A., & Symondson, W. O. C. (2011). Suction sampling as a significant source of error in molecular analysis of predator diets. *Bulletin of Entomological Research*, 102(3), 261–266. <https://doi.org/10.1017/S0007485311000575>.
- Kjeldsen, S. R., Zenger, K. R., Leigh, K., Ellis, W., Tobey, J., Phalen, D., ... Raadsma, H. W. (2016). Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population variability across its range. *Conservation Genetics*, 17(2), 337–353. <https://doi.org/10.1007/s10592-015-0784-3>.
- Knutsen, H., Olsen, E. M., Jorde, P. E., Espeland, S. H., André, C., & Stenseth, N. C. (2011). Are low but statistically significant levels of genetic differentiation in marine fishes 'biologically meaningful'? A case study of coastal Atlantic cod. *Molecular Ecology*, 20(4), 768–783. <https://doi.org/10.1111/j.1365-294X.2010.04979.x>
- Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., ... Blaxter, M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of Sciences*, 113(18), 5053–5058. <https://doi.org/10.1073/pnas.1600338113>.
- Lamichaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., ... Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences*, 114(17), E3452–E3461. <https://doi.org/10.1073/pnas.1617728114>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Meth*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25. <https://doi.org/10.1186/gb-2009-10-3-r25>.
- Limborg, M. T., Helyar, S. J., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R. O. B., ... Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, 21(15), 3686–3703. <https://doi.org/10.1111/j.1365-294X.2012.05639.x>.
- Longo, M. S., O'Neill, M. J., & O'Neill, R. J. (2011). Abundant human DNA contamination identified in non-primate genome databases. *PLoS ONE*, 6(2), e16410. <https://doi.org/10.1371/journal.pone.0016410>.
- Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15(1), 28–41. <https://doi.org/10.1111/1755-0998.12291>.
- Miller, K. M., Laberee, K., Schulze, A. D., & Kaukinen, K. H. (2001). Development of microsatellite loci in Pacific herring (*Clupea pallasii*). *Molecular Ecology Notes*, 1(3), 131–132. <https://doi.org/10.1046/j.1471-8278.2001.00048.x>.
- Mitchell, D. M. (2006). Biocomplexity and metapopulation dynamics of Pacific herring (*Clupea pallasii*) in Puget Sound, Washington. (Master of Science), University of Washington, Seattle, WA.
- Mitchell, D., McAllister, P., Stick, K., & Hauser, L. (2008). Sperm contamination in archived and contemporary herring samples. *Molecular Ecology Resources*, 8(1), 50–55. <https://doi.org/10.1111/j.1471-8286.2007.01840.x>.
- Nielsen, E. E., Cariani, A., Aoidh, E. M., Maes, G. E., Milano, I., Ogden, R., ... Carvalho, G. R. (2012). Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*, 3, 851. <https://doi.org/10.1038/ncomms1845>
- Olsen, J. B., Lewis, C. J., Kretschmer, E. J., Wilson, S. L., & Seeb, J. E. (2002). Characterization of 14 tetranucleotide microsatellite loci derived from Pacific herring. *Molecular Ecology Notes*, 2(2), 101–103. <https://doi.org/10.1046/j.1471-8286.2002.00160.x>
- Palsbøll, P. J., Bérubé, M., & Allendorf, F. W. (2007). Identification of management units using population genetic data. *Trends in Ecology & Evolution*, 22(1), 11–16. <https://doi.org/10.1016/j.tree.2006.09.003>.
- Paradis, E. (2010). pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26(3), 419–420. <https://doi.org/10.1093/bioinformatics/btp696>.
- Paris, J., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: A road map for stacks. *Methods in Ecology and Evolution*, 8(10), 1360–1373. <https://doi.org/10.1111/2041-210X.12775>.
- Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, 28(19), 2537–2539. <https://doi.org/10.1093/bioinformatics/bts460>.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine fish: Mismatch between biological and fisheries management units. *Fish and Fisheries*, 10(4), 361–395. <https://doi.org/10.1111/j.1467-2979.2008.00324.x>.
- Rousset, F. (2008). Genepop'007: A complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources*, 8, 103–106. <https://doi.org/10.1111/j.1471-8286.2007.01931.x>.
- Schmieder, R., & Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE*, 6(3), e17288. <https://doi.org/10.1371/journal.pone.0017288>.

- Scribner, K. T., Lowe, W. H., Landguth, E., Luikart, G., Infante, D. M., Whelan, G. E., & Muhlfeld, C. C. (2016). Applications of genetic data to improve management and conservation of river fishes and their habitats. *Fisheries*, 41(4), 174–188. <https://doi.org/10.1080/03632415.2016.1150838>.
- Sehn, J. K., Spencer, D. H., Pfeifer, J. D., Bredemeyer, A. J., Cottrell, C. E., Abel, H. J., & Duncavage, E. J. (2015). Occult specimen contamination in routine clinical next-generation sequencing testing. *American Journal of Clinical Pathology*, 144(4), 667–674. <https://doi.org/10.1309/ajcpr88wdjjldmbn>.
- Small, M. P., Loxterman, J. L., Frye, A. E., Von Bargen, J. F., Bowman, C., & Young, S. F. (2005). Temporal and spatial genetic structure among some pacific herring populations in Puget Sound and the Southern Strait of Georgia. *Transactions of the American Fisheries Society*, 134(5), 1329–1341. <https://doi.org/10.1577/T05-050.1>.
- Tarpey, C. M., Seeb, J. E., McKinney, G. J., Templin, W. D., Bugaev, A. V., Sato, S., & Seeb, L. W. (2017). SNP data describe contemporary population structure and diversity in allochronic lineages of pink salmon (*Oncorhynchus gorbuscha*). *Canadian Journal of Fisheries and Aquatic Sciences*, 75(6), 987–997. <https://doi.org/10.1139/cjfas-2017-0023>.
- Vandewoestyne, M., Van Hoofstat, D., De Groote, S., Van Thuyne, N., Haerinck, S., Haerinck, S., Deforce, D. (2011). Sources of DNA contamination and decontamination procedures in the forensic laboratory. *Journal of Forensic Research*, S2(001). <https://doi.org/10.4172/2157-7145.S2-001>.
- Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, 31(3), 318–323. <https://doi.org/10.1093/bioinformatics/btu668>.
- Waples, R. S. (1998). Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, 89(5), 438–450. <https://doi.org/10.1093/jhered/89.5.438>
- Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358–1370.
- Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE*, 7(8), e42649. <https://doi.org/10.1371/journal.pone.0042649>.
- Yang, D. Y., & Watt, K. (2005). Contamination controls when preparing archaeological remains for ancient DNA analysis. *Journal of Archaeological Science*, 32(3), 331–336. <https://doi.org/10.1016/j.jas.2004.09.008>.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Petrou EL, Drinan DP, Kopperl R, et al. Intraspecific DNA contamination distorts subtle population structure in a marine fish: Decontamination of herring samples before restriction-site associated sequencing and its effects on population genetic statistics. *Mol Ecol Resour*. 2019;19:1131–1143. <https://doi.org/10.1111/1755-0998.12978>