Range-wide regional assignment of Atlantic salmon (Salmo salar) using genome wide single-nucleotide polymorphisms

Article in Fisheries Research · October 2018
DOI: 10.1016/j.fishres.2018.05.017

12 authors, including:

Nicholas W Jeffery
Bedford Institute of Oceanography
38 PUBLICATIONS 209 CITATIONS

Brendan F Wringe
Fisheries and Oceans Canada
44 PUBLICATIONS 134 CITATIONS

Ryan R.E. Stanley
Fisheries and Oceans Canada
35 PUBLICATIONS 120 CITATIONS

Louis Bernatchez
Laval University
787 PUBLICATIONS 23,985 CITATIONS

Some of the authors of this publication are also working on these related projects:

SNPFISK View project

Connectivity in the marine coastal environment: a multidisciplinary approach. View project
Range-wide regional assignment of Atlantic salmon (Salmo salar) using genome wide single-nucleotide polymorphisms


A R T I C L E   I N F O

Keywords:
Atlantic salmon
Salmo salar
Assignment
Mixed-stock fishery
Genetic stock identification

A B S T R A C T

The estimation of stock specific exploitation is imperative to fisheries management and the conservation of biodiversity, particularly in instances where fisheries simultaneously exploit mixtures of stocks. Mixed stock harvesting is particularly common in species that have extensive marine migrations, such as Atlantic and Pacific salmon. Here we develop a range-wide genetic baseline for Atlantic salmon (Salmo salar) from North American and European rivers to allow regional assignment of individuals targeted in international mixed stock fisheries. A combination of published data and additional genotyping was used to assemble a dataset of 96 SNPs for 285 range-wide Atlantic salmon populations for regional assignment. Clustering of baseline samples identified 20 North American and eight European reporting groups with mean individual assignment accuracy of 90% (range 79–100%). This baseline was applied to disentangle the stock composition of individuals in a subset of individuals from the West Greenland Atlantic salmon fishery. Genetic mixture analysis revealed that both European and North American individuals originated from multiple regions, with 92% of European individuals originating from the United Kingdom and Ireland, and North American individuals originating from three regions; Gulf of St. Lawrence (28%), Gaspe Peninsula (23%), and coastal Labrador (21%). The baseline represents a significant resource for the management of Atlantic salmon fisheries and the quantification of salmon migration patterns at sea.

1. Introduction

Mixed stock fisheries target a mixture of individuals that originate from several independent stocks thereby adding a layer of complexity to the management strategies of highly migratory marine and anadromous fish species (Begg et al., 1999; Hilborn, 1985; Seeb and Crane, 1999). As the conservation of stock diversity of exploited species is important to species and fisheries stability and persistence (Hilborn et al., 2003), the quantification of stock specific exploitation is central to successful management of mixed-stock fisheries (Begg et al., 1999; Carvalho and Hauser, 1994). Traditionally, approaches such as physical tagging (Reddin et al., 2012), morphometrics (Reddin and Friedland, 1999), and otolith chemistry or morphology (Friedland and Reddin, 1994) have been used to explore the composition of mixed-stock fisheries. More recently, genetic stock identification (GSI) has been utilized to identify fishery composition in several species including Atlantic herring (Bekkevold et al., 2011), Atlantic cod (Ruzzante et al., 2000), and various Pacific and Atlantic salmonids (Ackerman et al., 2011; Beacham et al., 2004; Bradbury et al., 2015b; Bradbury et al., 2016a, 2016b; Gauthier-Ouellet et al., 2009; Gilbey et al., 2017). Genetic stock
Identification has been dramatically facilitated with the development of large genomic datasets from which highly informative panels of genetic markers can be designed. Tens to hundreds of thousands of genome-wide markers can be screened, and targeted panels of informative loci can be designed for application to mixed-stock fishery analysis which may provide unprecedented resolution of stocks.

Atlantic salmon (*Salmo salar*) are an anadromous fish of high socioeconomic value that have been extensively studied across its native range in North America and Europe. Salmon show fine-scale population structure among rivers and regions due to their strong homing behaviour and local adaptation across a wide latitudinal and thermal range (e.g. Bourret et al., 2013b; King et al., 2001). Large-scale fisheries for Atlantic salmon in North America have largely ceased following the drastic decline of wild populations (COSEWIC, 2010), although several fisheries still exist in the Northwest Atlantic (Bradbury et al., 2016a, 2015b; King et al., 2001; Reddin and Friedland, 1999; Sheehan et al., 2010). In particular, the fishery off West Greenland is the largest intercepet fishery in the Northwest Atlantic and is comprised of multi sea-winter (MSW) fish from both North American and European populations, the relative proportions of which vary over time. The ability to assign individuals to their region of origin using genetic markers, both across continents (generally 100% success) and within a continent (70–90%), has been previously demonstrated (e.g. Bourret et al., 2013b; Bradbury et al., 2016a, 2015a; Gilbey et al., 2016, 2017). Most of these studies are regional in scope, and have relied on microsatellite loci, which require extensive standardization across laboratories for large-scale analyses (Ellis et al., 2011; Moran et al., 2006). Recently, Moore et al. (2014) demonstrated an improved regional and population-level assignment with SNPs for Canadian rivers, especially when using > 3000 SNPs, therefore supporting the hypothesis that large scale assignment is possible using an extensive and highly targeted SNP database.

The objective of our study was to develop a highly informative and cost-effective panel of SNPs for accurate genetic assignment and analysis of mixed stocks across the native range of Atlantic salmon. We demonstrate the utility of this panel using a cross range reference baseline to identify regional contributions to a subset sample of the mixed stock fishery from West Greenland. We build upon previously developed Atlantic salmon baselines based on microsatellites (Bradbury et al., 2015b; Moore et al., 2014) and SNPs (Bourret et al., 2013b; Moore et al., 2014) to extend our North American coverage and further refine our fine-scale regional clustering of rivers and compare the assignment power of our SNP baseline to a microsatellite baseline from Bradbury et al. (2015b). The development of a small panel (< 100) of SNPs will allow for range-wide, rapid regional assignment of salmon to inform fisheries management and aid in the conservation of at-risk populations in both North America and Europe.

2. Methods

2.1. Outline of methods

Panel development and application was divided into the following eight steps: 1) We used genotype data from Moore et al. (2014), Bradbury et al. (2015b) and Sylvester et al. (2018) for SNP panel identification. 2) Samples from preliminary reporting groups based on those from (Bradbury et al., 2015b) were split into training and hold-out groups for SNP panel selection and testing, each consisting of 50% of the individuals. 3) Using the training set of individuals, we tested six methods for panel design of 288 SNPs to further refine into a final panel of 96 SNPs. While we aimed to produce a panel of 96 SNPs for mixture analyses, 288 were initially chosen in case a second panel of 96 was required for desired levels of accuracy, and to provide redundant SNPs for assay design and testing since this would not be possible for each of the top 96 chosen. 4) Additional genotype data were added to our baseline, including individuals from populations from Bourret et al. (2013b), Mäkinen et al. (2015), Barson et al. (2015), and Gilbey et al. (2016). 5) Fish from 100 additional North American rivers were genotyped using our panel of 96 SNPs for assignment and incorporated into our baseline. 6) Final reporting groups were evaluated using clustering analyses and geographic proximity of 285 rivers from North America and Europe (Fig. 1). 7) Baseline power was assessed using self-assignment and a leave-one-out approach, as well as simulations in the R package *rubias* (Anderson et al., 2008; Hasselman et al., 2015). This resulted in a final baseline of 28 reporting groups. 8) 280 individuals from a mixed-stock fishery in West Greenland were genotyped using...
these 96 SNPs and assigned to their North American or European regions of origin (Supplementary Fig. S1).

2.2. Initial baseline samples

We initially combined 6 K SNP array (CIGENE; Bourret et al., 2013b, Lien et al., 2011) data from 75 North American rivers available from Moore et al. (2014) and Bradbury et al. (2015b), as well as individuals genotyped using a 220 K SNP array from Lake Melville, Labrador, available from Sylvester et al. (2018), including two aquaculture populations originally derived from the St. John River, NB (codes CKA and NLA; Supplementary Table S1). 2760 overlapping SNPs from the 220 K and the 6 K arrays were subset and merged using the R package genePopedit (Stanley et al., 2017). Multi-site variants (MSVs) were not included in this panel, as these were flagged and removed in the original studies (Bourret et al., 2013a). Further filtering for minor allele frequency < 0.05 and missing data resulted in 1874 loci to design our panel from. The populations were combined into 13 reporting groups, which were based on 12 from Bradbury et al. (2015b) plus Lake Melville (see Sylvester et al., 2018), for initial testing. We did not remove potentialsiblings from our data prior to downstream analysis due to the potentially harmful effects if the pedigree is inferred incorrectly, such as reduced precision in the estimation of $F_{ST}$ and population allele frequencies (see Waples and Anderson, 2017). We used gsi_sim (Anderson et al., 2008) to remove individuals with 100% identical genotypes due to some overlap among combined data sets.

2.3. Identification of informative SNPs

We first created training and hold-out data sets each consisting of 50% randomly selected individuals from the initial reporting units to prevent upward bias of accuracy in order to develop an informative panel of SNPs for baseline establishment and mixed-stock analysis (see Anderson, 2010). The training set was used to select the SNP selection methods, while the hold-out individuals were used solely to test the self-assignment accuracy of the panels. We used the Toolbox for Ranking and Evaluation of SNPs (TRES; Kavvakiotsis et al., 2015) to evaluate the informativeness of various SNP panels for assignment using the built in file conversion tool. We selected six panels of 288 SNPs to test for assignment accuracy based on a number of parameters, including informativeness for assignment (Rosenberg et al., 2003); delta, the absolute differences in allele frequency (Shriver et al., 1997); and Wright’s $F_{ST}$ (Wright, 1949) (Fig. S2). Additionally, we used random forest (RF) and regularized random forest (RRF) ranking of locus importance using 5000 trees in the R packages randomForest (Liaw and Wiener, 2002) and RRF (Deng and Runger, 2013), which have been shown to be effective for individual assignment in salmonids (Sylvestre et al., 2018). Finally, we obtained the top 288 high $F_{ST}$ loci with low linkage disequilibrium (LD) values ($r^2 < 0.25$) using the genopop_to_toploci function in the R package genopopedit (Stanley et al., 2017). 288 SNPs were initially chosen in case a second panel of 96 SNPs was necessary for the reporting group resolution we aimed to achieve, and also to have redundant SNPs to design assays on the Fluidigm platform, assuming that not all of the top SNPs would pass SNPtype validation.

We used the R package assigner (Gosselin et al., 2017) which is based on gsi_sim (Anderson, 2010; Anderson et al., 2008) or the discriminant analysis of principal components (DAPC) function in the R package adegenet (Jombart, 2008) to conduct self-assignment analysis for each of the six SNP panels. We tested each panel of 288 loci using both the gsi_sim and DAPC analyses and the ‘random’ method using all individuals in the hold-out dataset within assigner, which implements a leave-one-out (LOO) approach in assigning individuals to their region of origin. Preliminary analyses revealed that for our data, using DAPC within assigner always yielded worse results than gsi_sim (by an average of 5% lower overall accuracy), and so all analyses within assigner used gsi_sim. Panel assessments were based on both individual reporting group self-assignment and average assignment accuracy of all of the reporting groups. Once we tested each of the six methods, we identified the best panel of 288 loci (i.e. that with the highest accuracy) and designed SNPtype assays for the top 96 SNPs which passed validation (see Section 2.4 Additional data and genotyping for details).

2.4. Additional data and genotyping

Following SNP panel selection, additional publicly available data including Mäkinen et al. (2015) for the Saint John River, Canada, Barson et al. (2015) for Norwegian rivers, and United Kingdom rivers from Gilby et al. (2016), which used the same SNPs from Bourret et al. (2013b) were also added to our baseline following filtering for the top 96 common loci using genopopedit (Table S2). All genotypes were inspected for strand-flips, and neighbour-joining trees were constructed to compare overlapping British/Scottish (Gilby et al., 2016) and Norwegian (Barson et al., 2015) river genotypes with those in Bourret et al. (2013) to ensure the SNP calls were the same between each published data set. All rivers which overlapped were sister-taxon on the NJ trees as would be expected if genotype calls were the same among studies.

SNP genotyping was performed using SNPtype assays (Fluidigm) per the manufacturer’s protocols, without the STA step, using 96.96 genotyping IFCs (Fluidigm) and read on an EP1 (Fluidigm) and analyzed using SNP Genotyping Analysis software (Fluidigm). Each 96 well plate setup included 10 redundant samples that were repeated on the plate to detect processing errors (row or plate reversal) and ensure consistent clustering interpretation; positive controls (see above for details) and the required negative controls. Any samples with > 9 failed loci were removed from the final data set (35 samples in total, 1.03%). SNPtype assays were designed using the D3 Assay Design application (https://d3.fluidigm.com) (Fluidigm, San Francisco, CA, USA) based on sequences for each locus obtained from dbSNP (GenBank www.ncbi.nlm.nih.gov). Assays were evaluated for inclusion in the final panel based on: obtaining correct genotypes with known samples and positive controls (see below); reproducible genotypes across multiple Integrated Fluidic Circuit (IFC) runs; the ranking of the target SNP in the prioritized list; assay performance without the STA (Specific Target Amplification). To calculate the genotype error rate, 11.5% of the samples were reanalyzed from the original tissue where tissue samples permitted. The finalized panel consisted of 96 loci (Table S2).

Synthesized double stranded DNA (gBlocks; Integrated DNA Technologies, Coralville, IA, USA) was prepared for use as positive controls (Richards-Hrdlicka, 2014). Finally, 3406 additional North American fish from 100 rivers were genotyped at these 96 informative loci on a Fluidigm EP1 platform at the Aquatic Biotechnology Laboratory at the Bedford Institute of Oceanography, Dartmouth, Nova Scotia. 280 salmon from the West Greenland fishery of known European (n = 140) and North American (n = 140) origin from Bradbury et al. (2016b) were also genotyped using the 96 SNP panel for mixture analysis.

2.5. Identification of reporting groups

We constructed a neighbour-joining (NJ) tree of all 285 populations based on Cavalli-Sforza and Edwards (1967) chord distances between sites in POPULATIONS 1.2.33 (Langella, 2015) with 1000 bootstrap pseudoreplicates to define our final reporting groups for individual baseline assignment and mixture analysis. Second, we conducted a discriminant analysis of principal components (DAPC) in the R package adegenet (Jombart, 2008). Initially, we retained all principal components and selected the best value for K based on the Bayesian information criterion using the find.clusters function. A DAPC was then conducted based on this optimal number of clusters and the probabilities that an individual belonged to a cluster were used to assign individuals to reporting groups.
Fig. 2. A) A neighbour-joining tree for all baseline rivers genotyped at the top 96 SNPs and based on Cavalli-Sforza and Edwards (1967) chord distances and 1000 bootstrap replicates. Each colour corresponds to a reporting group. B) Discriminant analysis of principal components (DAPC) of European rivers and C) DAPC of North American rivers using the same 96 SNPs.
Using these two methods, and additional modifications to the baseline based on geographic proximity of rivers, we tested reporting group iterations using the leave-one-out approach in the R package *rubias* (*Anderson et al.*, 2008). *rubias* is based on gsi_sim and attempts to control for biases caused by differences in the number of populations among reporting groups to maximize overall assignment success, which was defined as each reporting group having a minimum self-assignment success of 70%, and an average accuracy > 85% across all reporting groups. We report both accuracy (i.e. the proportion of correctly assigned fish) and efficiency (i.e. the proportion of fish of all the samples which assigned at a probability > 70%) (*Vähä et al.*, 2011; *Vähä and Primmer*, 2006) for our final 28 reporting groups.

We additionally compared the power of our SNP panel to a previously reported microsatellite baseline by *Moore et al.* (2014) and *Bradbury et al.* (2015b) by conducting 100% simulations and a leave-one-out approach for the same reporting groups from these publications, and compared the accuracy of each genetic marker type.

### 2.6. Simulations

Next we conducted mixture simulations based on our 28 defined reporting groups to test the power of our 96 SNP panel baseline on mixed stocks in *rubias*. We first conducted 100% simulations, in which mixtures are simulated where 100% of the individuals come from a single reporting unit. We ran 600 simulations of 200 fish per simulated mixture, with a flat Dirichlet distribution and determined the assignment accuracy per reporting unit. Second, we conducted more realistic fisheries mixtures, first using equal proportions of all reporting units (n = 100 individuals from each of 28 reporting units) and 500 simulations, and then comparing a range of actual to simulated proportions across 500 replicates of simulated mixtures each consisting of 500 individuals.

### 2.7. Mixed-stock analysis

Mixed-stock analysis was performed on 280 Atlantic salmon from the coast of West Greenland during the 2014 fishery (Table S3), using a Bayesian approach with parametric bootstrapping (n = 100 iterations) in *rubias* (*Anderson et al.*, 2008) and a maximum likelihood approach in ONCOR (*Kalinowski et al.*, 2008). We genotyped equal numbers of fish of known North American (n = 140) and European (n = 140) origin to test how well our panel could assign individuals to their region of origin. We first combined all North American and European populations into continental reporting groups to ensure 100% identification of continent of origin. We used 5000 MCMC iterations in *rubias*, with the first 500 iterations discarded as burn-in to obtain the individual assignments and overall proportions of each mixed-stock fishery at the continental level. We then analyzed the proportions of these mixed-stock fish on our regional reporting units using the same approaches in both *rubias* and ONCOR.

### 3. Results

#### 3.1. Initial baseline and loci assessment

75 initial North American populations were genotyped at 1874 loci after being filtered for SNPs common to the Atlantic Salmon 6 K SNP array and Affymetrix × 220 K array and minor allele frequency (> 0.05) and grouped into 13 reporting groups based on previously published reporting units (*Bradbury et al.*, 2015b; *Moore et al.*, 2014; *Sylvester et al.*, 2018). The aquaculture-raised fish (CKA and NLA) could not be distinguished from wild populations in the Saint John River, and so were grouped with these rivers into the SJR reporting group. Overall assignment accuracy of the top 288 loci from six methods of locus detection ranged from 92.95 ± 1.8 to 95.48 ± 2.03% (mean ± sd) (Fig. S2). The top 288 high FST unlinked loci from *genepopedit* yielded the highest overall assignment success (Fig. S3). While these differences are minimal, this selected panel of 288 loci consistently outperformed the next highest methods (random forest and informativeness for assignment) by 0.5–1.0%. The top 96 SNPs for which we were able to design assays from this panel of 288 were found to be evenly distributed across the genome (Fig. S4), and were associated with both non-coding regions and a wide variety of genes, including rhd8 (retinol dehydrogenase), hsp90ab1 (heat shock protein 90-beta), tieg3 (transforming growth factor), and other genes involved in growth and development, metabolism, and homeostasis (Table S2).

#### 3.2. Final baseline and reporting groups

As additional data were made publicly available, we added populations and newly genotyped individuals to our baseline. We genotyped 3406 additional individuals using our final 96 SNPs, which showed a genotype success rate of 98.9% and a genotype error rate of 0.32% (based on *Pompanon et al.*, 2005), consistent with what has been reported in other studies on fish (e.g. *Hess et al.*, 2015; *Jones et al.*, 2015; *Larson et al.*, 2014; *Petrou et al.*, 2014). Our final baseline was thus
Table 1
Baseline self-assignment accuracy using a Bayesian leave-one-out approach in rubias showing accuracy, the proportion of correctly assigned fish, and efficiency, the proportion of individuals that assigned at a high probability (> 70%), of 28 reporting groups defined in the present study. The largest misidentification, where fish from a known reporting group were incorrectly assigned, is also reported per group.

<table>
<thead>
<tr>
<th>Reporting Group</th>
<th>Code</th>
<th>Accuracy (%)</th>
<th>Efficiency (%)</th>
<th>Largest Misidentification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>North America</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anticosti</td>
<td>ANT</td>
<td>95.8</td>
<td>95.1</td>
<td>WNF (1.4%)</td>
</tr>
<tr>
<td>Avalon Peninsula</td>
<td>ARA</td>
<td>95.2</td>
<td>96.3</td>
<td>NIF1 (2.7%)</td>
</tr>
<tr>
<td>Eastern Nova Scotia</td>
<td>ENS</td>
<td>94.0</td>
<td>93.1</td>
<td>USA (2.8%)</td>
</tr>
<tr>
<td>Gaspé</td>
<td>GAS</td>
<td>90.3</td>
<td>82.7</td>
<td>QUE (5.3%)</td>
</tr>
<tr>
<td>Gulf of St. Lawrence</td>
<td>GUL</td>
<td>91.5</td>
<td>77.3</td>
<td>IFN (9.6%)</td>
</tr>
<tr>
<td><strong>Europe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Baltic Sea</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baltic Sea</td>
<td>BAL</td>
<td>100</td>
<td>100</td>
<td>NA</td>
</tr>
<tr>
<td>Barents-White Seas</td>
<td>BAR</td>
<td>99.0</td>
<td>99.6</td>
<td>NOR (2.2%)</td>
</tr>
<tr>
<td>United Kingdom/ Ireland</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>France</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>FRN</td>
<td>92.3</td>
<td>100</td>
<td>BRI (5.1%)</td>
</tr>
<tr>
<td><strong>Iceland</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iceland</td>
<td>ICE</td>
<td>98.4</td>
<td>100</td>
<td>BRI (1.6%)</td>
</tr>
<tr>
<td><strong>Northern Norway</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern Norway</td>
<td>NNO</td>
<td>86.7</td>
<td>83.9</td>
<td>SNO (8.8%)</td>
</tr>
<tr>
<td><strong>Southern Norway</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southern Norway</td>
<td>SNO</td>
<td>86.4</td>
<td>97.9</td>
<td>BRI (1.8%)</td>
</tr>
<tr>
<td>Spain</td>
<td>SPN</td>
<td>100</td>
<td>100</td>
<td>NA</td>
</tr>
</tbody>
</table>

The baseline self-assignment accuracy using a Bayesian leave-one-out approach in rubias showed high accuracy, ranging from 79% in the NLS reporting group to 100% in the BAL, SPN, and FRN reporting groups ($x^2 = 97.3 \pm 0.01\%$) (Fig. 5A).

3.3. Simulations and mixture analysis

100% simulations of mixtures conducted in rubias showed high accuracy, ranging from 79% in the NLS reporting group to 100% in the BAL, SPN, and FRN reporting groups ($x^2 = 97.3 \pm 0.01\%$) (Fig. 5A).

A PCA of our 280 mixture individuals from West Greenland revealed clear division at the continental scale along the first principal component, which explained 28% of the variance (Fig. 7A). Assignment to the continent of origin was 99.7 ± 0.1% using both rubias and ONCOR. Mixture proportions at the regional scale for individuals of known European origin were 92.4% fish of UK/Irish origin, and ~2–4% southern Norwegian, Spanish and French, with negligible proportions from Iceland, Norway, the Baltic Sea, and Barents-White Sea. The fish of North American origin consisted of 21% coastal Labrador, 23.3% Gaspé, and 28% Gulf of St. Lawrence fish, with ~6% contributions each from Ungava Bay, Lake Melville (central Labrador) and the St. Lawrence lower shore, and minor (< 3%) contributions from the other reporting groups (Fig. 7B). However, when only considering individuals which assigned to a reporting unit with greater than 80% probability, no fish assigned to Lake Melville (Fig. 7C). Bootstrap corrected reporting unit proportions showed no appreciable difference from the non-corrected proportions (Fig. S5). Mixture proportions at the reporting group level did not differ between ONCOR and rubias using a two-tailed t-test ($p = 0.95$).

4. Discussion

Mixed stock fisheries challenge fisheries management (Begg et al., 1999; Hilborn, 1985; Seeb and Crane, 1999), threatening fisheries and species stability and persistence (Hilborn et al., 2003). Although several assignment methods, including tagging, otolith chemistry, and microsatellite genotyping have traditionally been used for stock identification, range-wide highly-accurate and transferable methods for genetic stock identification have been elusive. Here we develop the first-targeted panel of genome-wide SNPs for range-wide assignment of Atlantic salmon. This small cost-efficient panel can successfully assign salmon from both the East and West Atlantic to their European country or sea of origin, and fine-scale region of origin in North America, ranging from Maine, USA, to Ungava Bay, northern Quebec. We build on previous genetic baselines increasing the North American baseline reporting groups from 12 (Bradbury et al., 2015b; Moore et al., 2014) to 20, which include the 16 Designatable Units of Atlantic salmon within Canada (COSEWIC, 2010). The main additional reporting groups were in Newfoundland and Labrador which is the northern portion of the North American range. The ability to perform range-wide accurate assignments will be directly applicable to fisheries such as the West Greenland mixed-stock fishery which consists of individuals native to both Europe and North America and whose relative contributions have fluctuated over time (Bradbury et al., 2016b; Reddin and Friedland, 1999).

4.1. Marker selection

Although a variety of approaches exist for panel design (see Sylvester et al., 2018), in this context we found that optimal panel
design was based on ranking by high $F_{ST}$ values and low levels of linkage disequilibrium (LD). All six methods examined for SNP selection yielded highly accurate results, yet the low-linkage high-$F_{ST}$ panel consistently outperformed the other panels in terms of assignment accuracy by $\sim 1\%$. Interestingly, perhaps the most commonly used approach of simply ranking by $F_{ST}$ (e.g., Gilbey et al., 2016; McKinney et al., 2017; Moore et al., 2014) consistently yielded the worst results. It seems selecting the highest $F_{ST}$ markers that are not in LD will allow for greater baseline refinement by reducing redundancy in the dataset (e.g., Sinclair-Waters et al. in review). Random forest and variations of random forest are promising avenues for ranking SNPs by their overall importance (Sylvester et al., 2018), but importance rankings may be skewed when SNPs are in LD (Meng et al., 2009). As such, pre-filtering prior to $F_{ST}$ ranking or random forest analysis seems to maximize accuracy in panel design. Further refinement of regional groups using SNPs could potentially be conducted using larger panels of unlinked SNPs or amplicon based multi-SNP haplotypes. For example, McKinney et al. (2017) report that they were able to use half as many haplotypes relative to single SNPs to achieve $> 90\%$ assignment accuracy in Chinook salmon. However, this means that haplotypes are composed of SNPs in LD due to their physical proximity in the genome, and while this does provide more alleles for assignment, further in-depth analyses comparing high $F_{ST}$ low LD SNPs to multi-SNP haplotypes will provide insight into the true performance of these types of markers for assignment purposes (e.g. Morin et al., 2009; Seeb et al., 2011). Additionally, the retention of MSVs could improve the accuracy of our assignments or potentially increase the resolution of reporting units (Waples et al., 2016). Gilbey et al. (2016) found that MSVs ranked significantly higher relative to regular SNPs based on river-level $F_{ST}$, possibly because of neo-functionalisation and their potentially adaptive nature. MSV loci were not included in our study, as they can be difficult to genotype across different platforms (Dufresne, 2016). However, their potential for fine-scale regional assignment could be investigated in future work.

4.2. Comparison to previous baselines

Previous attempts to resolve regional groups of Atlantic salmon using microsatellite loci or SNPs had identified similar, though larger, clusters of populations. Direct comparison using the 12 reporting groups used by previous microsatellite based analyses (Bradbury et al., 2015b), demonstrates that our SNP panel performed similarly in terms of accuracy and efficiency, showing no significant difference for these reporting groups. The advantage here is an increase in the number of reporting groups defined by Bradbury et al. (2015b) from 12 to 20 North American groups, and the inclusion of eight European groups resulting from the increased number of baseline populations included as well as the increased resolution provided by the SNP panel. Most notably, our baseline increases the number of reporting groups in the northern portion of the North American range, such as in Newfoundland from two (Newfoundland and the Avalon peninsula) to six and in Labrador from three to four. Our SNP panel is also able to differentiate the Inner Bay of Fundy, western Nova Scotia, and the St. John River watershed at a high level of accuracy which are all geographically proximate rivers. Despite our SNP panel being originally developed using only North American populations, assignment accuracy and efficiency was generally higher in Europe, likely due to the broader-scale coverage of the reporting groups across a greater geographic region, and higher genetic distances among European rivers. The Baltic and Spain regions, in particular, had $100\%$ accuracy and efficiency in self-assignment tests and for the other six regions there was high accuracy and efficiency of $> 87\%$, therefore suggesting that this panel of SNPs could be used to identify the country of origin for other European mixed stock fisheries such as those in the Faroe Islands (Gilbey et al., 2017). Simulations of each of the reporting groups showed high accuracy as well, and none of our reporting groups appeared to show bias based on population structure across a wide range of simulated proportions. Admittedly, the baseline is currently biased towards North American populations and improved spatial resolution should be possible with the inclusion of additional European salmon populations.

4.3. Mixed fishery proportions

Our 96 SNP panel and baseline provided extremely accurate individual assignment at multiple spatial scales. For continent of origin, assignment success was comparable to continent of origin assignment by King et al. (2001) using microsatellite loci. In this study we genotyped equal proportions of individuals with known continent of origin from Bradbury et al. (2016b) to provide as a demonstration of baseline utility with large groups of east and west Atlantic salmon. Currently, the composition of the mixed stock fishery at West Greenland estimated at approximately $20\%$ and $80\%$ European and North American, respectively (Bradbury et al., 2016b; Sheehan et al., 2010). The panel and
baseline developed here can provide accurate estimates of continent of origin in fisheries such as this with the additional benefit of regional assignment power.

At smaller spatial scales, the proportions of European and North American reporting groups contributing to the West Greenland fishery were again similar to those previously reported using microsatellites. The North American contributions were similar to those reported by Bradbury et al. (2016b) and Gauthier-Ouellet et al. (2009), with ~21% contributions from Labrador, 23–28% from both the St. Lawrence Gulf and Gaspé Peninsula, and ~6% contributions from Lake Melville in central Labrador, Ungava Bay, and the St. Lawrence lower north shore. These regions are characterised by multi-sea winter individuals (Chaput et al., 2006; Jensen, 1990), contrasting southern Newfoundland or Avalon regions, which were absent from the fishery and are generally

---

Fig. 5. A) The results of 100% simulations performed in rubias. The dashed line indicates the mean accuracy of the simulations (97%). B) Mixed fishery simulations based on equal proportions of all reporting units performed in rubias, where the dashed line indicates the mean mixture proportion when including equal proportions of all reporting units.
single sea winter fish (i.e. grilse). Our results for the European contributions to the West Greenland fishery were consistent with those previously reported with microsatellite analyses, with 92% of individuals originating from the UK and Ireland (ICES, 2015). Southern Norway, Spain and France contributed \( \sim 6\% \) of this fishery, while northern Europe contributed \( \sim 3\% \) of the fishery which is also comparable to previously reported contributions (ICES, 2015), though the confidence intervals around these minor contributions are high and include 0%. Though individual assignments were well supported for the Spanish and French individuals (i.e. > 80%), sampling was sparse in this region and it is possible that fish only assigned to these rivers due to unsampled “ghost” rivers, which can thus lead to spurious assignments to other baseline rivers (Slatkin, 2005; Waples and Gaggiotti, 2006). Alternatively, individuals from the mixed stock could assign to the UK/Irish reporting unit due to ‘large group attraction’ where they spurious best assign to a region with a large sample size because of higher genetic diversity within the reporting unit. We do not think this is the case for our results, as no individuals assigned to southern Norway with high confidence, which has a similar size. Additionally, parametric bootstrapping conducted in rubias showed similar mixture proportions to the non-bootstrapped results, suggesting the reporting units are well resolved. Sampling to fill in rivers south of the UK will allow us to determine if French and Spanish rivers are contributing to the West Greenland fishery, or whether the European contributions are actually composed of nearly 100% United Kingdom and Irish fish. Very few individuals fish from the mixed fishery assigned to northern Europe as expected as northern Europe’s multi-sea winter fish do not migrate to West Greenland, but instead migrate to the Norwegian Sea near the Faroe Islands (Gilbey et al., 2017).

### 4.4. Ascertainment bias and other limitations

Given the design of the study, ascertainment bias is likely influencing the spatial resolution reported as observed elsewhere (e.g.}

---

**Fig. 6.** Simulations of mixed fisheries using varying proportions of each reporting unit show a near 1:1 relationship between the true proportion and the simulated proportion performed in rubias. Points above or below the line represent upward or downward bias, respectively (sensu Hasselman et al., 2015). Note that panels are arranged from southern to northern latitudes with European followed by North American reporting units.
Bradbury et al., 2011; Seeb et al., 2011), and evident in the fact that a much larger number of reporting groups were identified in North America (20) than Europe (8). Two sources of ascertainment bias may be present here. Firstly, the 6 K SNP array was originally developed on Norwegian Atlantic salmon (Lien et al., 2011), suggesting there may be an ascertainment bias for European genotypes in the initial SNP discovery process (Albrechtsen et al., 2010). Secondly, as our initial SNP panel selection process only used North American populations, there is
almost certainly ascertainment bias enhancing the detection of fine-scale reporting units in this continent relative to Europe. This is a certainly a limitation of this panel which could be addressed with additional panel development in the future. Nonetheless, the regional groups identified in Europe (Gilbey et al., 2016; King et al., 2001; King et al., 2005; Moore et al., 2014), and the mixture results for West Greenland (Bradbury et al., 2016b; ICES, 2015, 2016) are similar to those reported using microsatellites supporting the utility of this panel range wide. Additionally genotyping with microsatellite Gilbey et al. (2016) or SNP panels targeted to Europe may be used to increase resolution of reporting Groups within the East Atlantic.

Despite high assignment accuracy to regions on both sides of the Atlantic this SNP panel did have several additional limitations. Generally, the panel was unable to distinguish wild and aquaculture individuals, and at present any aquaculture individual would be assigned to the source regional group. For example, salmon collected from the Saint John River are indistinguishable from Canadian aquaculture-raised salmon, which are themselves originally derived from the Saint John River. Similarly, in Europe, aquaculture salmon could not be distinguished from Norwegian wild fish, from which they were derived. However, targeted cost-effective SNP panels are available and being developed to distinguish farmed fish from wild fish (Karlsson et al., 2011; Mäkinen et al., 2015; Wringle et al. in review a, b), and can be used downstream of our developed panel should this resolution be desired. Finer scale geographic assignment of wild salmon will likely require more alleles either through additional SNP genotyping (Moore et al., 2014; Sylvester et al., 2018), direct amplicon sequencing (Beacham et al., 2017), or sequencing of large microsatellite panels (Bradbury et al., 2018).

Interestingly, in several instances rivers did not cluster as expected based on geography. For example, the Hunt River in northern Labrador consistently clustered within Lake Melville in our NJ trees, but is >200 km north of the Melville system. Similarly, Cap Chat River, which is in the Gaspé Peninsula but clusters genetically with the Gulf reporting group, was included in the Gulf reporting unit for assignment purposes. The Québec and Upper North Shore (QUE and NLS) reporting groups had the lowest self-assignment accuracies and simulation power, and could be combined into a single reporting group to increase their assignment accuracy depending on the desired resolution of the study; however, we chose to keep these groups separate as they both self-assigned at our minimum success threshold (>70%) and were distinct (but genetically proximate) on our neighbour-joining tree.

Currently, most of our European data were for Scottish rivers from Gilbey et al. (2016) and Norwegian rivers from Barson et al. (2015); however, our results are consistent with those previously reported from microsatellites and we do not expect this high concentration of regional rivers to bias our overall assignment for the mixed fishery. As mentioned previously, unsampled "ghost" rivers may lead to some inaccurate assignments in Europe due to the sparsity of sampled rivers relative to North America, and additional data for rivers from Spain and France would be beneficial to increase the resolution of our European baseline and make it comparable to the North American baseline.

5. Conclusions

Here we have developed the first highly targeted panel of SNPs for range-wide, fine-scale assignment of Atlantic salmon as a tool for fisheries management, as well as understanding migration patterns in this species. Our panel of 96 SNPs and baseline performed similarly to previously published microsatellite baselines in terms of accuracy, but significantly outperforms microsatellites in the resolution of regional groups, as we were able to resolve 28 range-wide reporting groups. The ability to perform both range-wide and fine scale regional genetic assignment in North America represents a significant advance and will directly inform management of this species. This panel of SNPs can be used to detect changes in the proportions of these mixed-stock fisheries over time, and will be beneficial for the long-term monitoring and management of wild populations in North America and Europe.

Data Statement

Genetic baseline and mixed stock genotype files along with meta-data will be deposited in the Dryad Digital Repository upon acceptance of the manuscript. Individual river information and the SNP panel used is available in the supplementary materials.

Acknowledgements

The authors thank staff of the Newfoundland DFO Salmonids section, the Nunatsiavut Government, the Sivunivut Inuit Community Corporation, the Innu Nation, the Labrador Hunting and Fishing Association and fishers from Rigolet, North West River and Sheshatshiu (Labrador) for their support and active participation in this project, and the staff of the Aquatic Biotechnology Lab at the Bedford Institute of Oceanography for genotyping North American samples used in this study. We also thank the Centre for Integrative Genetics (CIGENE) for genotyping the range-wide data used in the present study, Hannu Mäkinen for providing additional genotype data for individuals from the Saint John River, Sigbjørn Lien for providing data from Norwegian rivers, and Sarah J. Lehnert for reading and providing helpful comments on an early draft of the manuscript. This study was supported by an Atlantic Canada Opportunities Agency and Department of Tourism, Culture, Industry and Innovation grants allocated to the Labrador Institute (MC), Atlantic Salmon Conservation Foundation grant allocated to the Nunatsiavut Government, and an Atlantic Salmon Federation Olin Fellowship allocated to Jonathan Pearce who collected samples, as well as a Genomics Research and Development Initiative (GRDI) Grant and a National Sciences and Engineering Research Council (NSERC) Discovery Grant and Strategic Project Grant to IRB. NWJ is supported by an NSERC visiting fellowship in a government laboratory.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.fishres.2018.05.017.

References


Bourret, V., Dionne, M., Kent, M.P., Lien, S., Bernatchez, L., 2013a. Landscape genomics over time, and will be beneficial for the long-term monitoring and management of wild populations in North America and Europe.

Fisheries Research 206 (2018) 163–175


