Received: 20 November 2015 | Accepted: 7 June 2016

DOI: 10.1111/eva.12407

ORIGINAL ARTICLE

WILEY Evolutionary Applications Open Access

Single nucleotide polymorphisms to discriminate different classes of hybrid between wild Atlantic salmon and aquaculture escapees

Victoria L. Pritchard[1] | Jaakko Erkinaro[2] | Matthew P. Kent[3] | Eero Niemelä[2] | Panu Orell[2] | Sigbjørn Lien[3] | Craig R. Primmer[1]

[1]Department of Biology, University of Turku, Turku, Finland

[2]Natural Resources Institute Finland (Luke), Utsjoki, Finland

[3]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Aas, Norway

Correspondence
Victoria L. Pritchard, Department of Biology, University of Turku, Turku FI-20014, Finland.
Emails: victorialpritchard@gmail.com; craig.primmer@utu.fi

Abstract

Many wild Atlantic salmon (*Salmo salar*) populations are threatened by introgressive hybridization from domesticated fish that have escaped from aquaculture facilities. A detailed understanding of the hybridization dynamics between wild salmon and aquaculture escapees requires discrimination of different hybrid classes; however, markers currently available to discriminate the two types of parental genome have limited power to do this. Using a high-density Atlantic salmon single nucleotide polymorphism (SNP) array, in combination with pooled-sample allelotyping and an F_{st} outlier approach, we identified 200 SNPs that differentiated an important Atlantic salmon stock from the escapees potentially hybridizing with it. By simulating multiple generations of wild–escapee hybridization, involving wild populations in two major phylogeographic lineages and a genetically diverse set of escapees, we showed that both the complete set of SNPs and smaller subsets could reliably assign individuals to different hybrid classes up to the third hybrid (F3) generation. This set of markers will be a useful tool for investigating the genetic interactions between native wild fish and aquaculture escapees in many Atlantic salmon populations.

KEYWORDS

allelotyping, aquaculture escapee, Atlantic salmon, introgressive hybridization, *Salmo salar*, SNP array

1 | INTRODUCTION

In common with many other salmonid fishes (e.g. Katz, Moyle, Quiñones, Israel, & Purdy, 2013; Metcalf et al., 2007; Rand, 2013), wild Atlantic salmon (*Salmo salar*) have declined over the past two centuries as a result of overfishing and habitat loss. The species has been extirpated from half of the major European river basins and a third of the major American river basins in which it historically occurred, and many of the remaining populations are considered to be under threat (ICES 2015; Parrish, Behnke, Gephard, McCormick, & Reeves, 1998). In contrast, since artificial cultivation began in Norway in the late 1960s, the captive Atlantic salmon population has exploded. In Europe over 1,000 times more salmon is currently produced by the aquaculture industry than is caught in the wild (ICES 2015). Most of these domestic Atlantic salmon are reared in open net-pens in the marine environment, and escapes are frequent (Thorstad et al., 2008). In Norway, the world's biggest producer of farmed salmon, tens of thousands of aquaculture escapees, identified by body morphology and scale characteristics, are caught annually in the wild. Between 1989 and 2006, on average, 38% of the salmon catch of Norwegian coastal fisheries and 21% of the individuals sampled in spawning areas were aquaculture escapees (Thorstad et al., 2008).

Aquaculture salmon have undergone many generations of selection in the captive environment–both inadvertently, due to hatchery culture conditions, and deliberately. Some Norwegian aquaculture strains, for example, have been subject to artificial selection for multiple traits including speed of growth, weight and age at sexual maturity, and resistance to disease (Bicskei, Bron, Glover, & Taggart, 2014; Fleming, Agustsson, Finstad, Johnsson, & Björnsson, 2002; Gjøen, 1997). Correspondingly, farmed salmon have been shown to differ from their wild conspecifics in multiple traits including growth rate and age at maturation (Glover, Otterå, Olsen, & Slinde, 2009; Debes & Hutchings, 2014), migratory behaviour (Jonsson, Jonsson, & Hansen, 1991), heart rate and swimming endurance (Johnsson, Höjesjö, & Fleming, 2011), gene expression (Bicskei et al., 2014; Debes, Normandeau, Fraser, Bernatchez, & Hutchings, 2012) and response to stress (Solberg, Skaala, Nilsen, & Glover, 2013). These traits are expected to make aquaculture fish less well adapted to the natural environment, and multiple studies have demonstrated farmed salmon to have lower survival and reproductive fitness than native conspecifics in the wild (Fleming, Jonsson, Gross, & Lamberg, 1996; Jonsson & Jonsson, 2006; Naylor et al., 2005). Nevertheless, mature aquaculture escapees are frequently found in wild spawning areas (Erkinaro et al., 2010; Fiske, Lund, & Hansen, 2006) and can breed with wild fish (Clifford, McGinnity, & Ferguson, 1998). Introgressive hybridization – that is, the introduction of genetic material from the aquaculture escapees into the wild population via successive generations of interbreeding – poses a number of different threats to wild salmon, including the introduction of traits that are not locally adapted, a reduction in overall genetic diversity across populations, and the disruption of co-adapted gene complexes that have become established within populations over evolutionary time (Edmands, 2006). The genetic contribution of aquaculture escapees to wild populations can be considerable: studies in Norway (Glover et al., 2012, 2013), Ireland (Clifford et al., 1998) and North America (Bourret, O'Reilly, Carr, Berg, & Bernatchez, 2011) have observed recent genetic changes in many wild Atlantic salmon populations that can be attributed to the influence of aquaculture escapees. However, other populations have maintained their genetic integrity despite large numbers of escaped fish being observed in their natal rivers (Glover et al., 2012, 2013). It is largely unknown what demographic or environmental factors influence the vulnerability of wild Atlantic salmon populations to genetic invasion by aquaculture escapees, although population density may play a role (Heino, Svåsand, Wennevik, & Glover, 2015).

Few sets of markers that can reliably discriminate the genomes of wild Atlantic salmon and aquaculture strains have been described (but see Karlsson, Moen, Lien, Glover, & Hindar, 2011), and this limits research into the dynamics of hybridization between escapees and wild fish. This relative paucity of markers is partly a function of the genetic characteristics of the aquaculture fish. Norwegian aquaculture lines, which are also farmed in a number of other countries, have mixed origins. They derive primarily from Norwegian populations in the Atlantic evolutionary lineage of *S. salar* with small contributions from populations in the North Barents/White Sea and Baltic evolutionary lineages (Bourret et al., 2013; Gjedrem, Gjøen, & Gjerde, 1991). Various different lines have been maintained separately since their initiation and are genetically very different from one another (Gjøen, 1997; Karlsson et al., 2011). These lines have both given rise to further sublines and been combined to form new lines (Karlsson, Diserud, Moen, & Hindar, 2014), and a salmon farm may use a varying mix of lines (Gjedrem et al., 1991; Gjøen, 1997; Thorstad et al., 2008). Further, the genetic composition and diversity of these aquaculture lines will have changed over time as a result of selection and drift and the addition of new material. Thus, there may be no consistent genetic signature whereby aquaculture escapees may be discriminated from wild fish.

Using an array enabling the simultaneous genotyping of seven thousand (7K) single nucleotide polymorphism (SNP) markers in Atlantic salmon, Karlsson et al. (2011) identified a suite of 60 markers that discriminated major Norwegian aquaculture lines from wild Norwegian salmon. Collectively, these SNPs enabled identification of pure-bred wild and aquaculture individuals and their F1 hybrids. However, they have not been shown to allow discrimination of different classes of later generation hybrids. This is imperative for a full understanding of the genetic interactions between wild and aquaculture salmon, but is a much more challenging analytical task (Vähä & Primmer, 2006). Here, we use an Atlantic salmon SNP array that includes 220,000 (220K) mapped SNPs, and combine it with a cost-effective allelotyping approach, to identify a set of SNPs that can discriminate wild fish, a genetically diverse set of aquaculture escapees, and their first-, second- and third-generation hybrids. We focus on the Teno River of northern Finland and Norway, one of the world's largest and most diverse wild Atlantic salmon stocks (Vähä, Erkinaro, Niemelä, & Primmer, 2007; Erkinaro et al., 2010; Fig. 1). Reproductively mature aquaculture escapees have been caught throughout this river system since 1985 (Erkinaro et al., 2010). We further demonstrate that this SNP set performs well for hybrid class discrimination in additional populations, suggesting that it has general utility for examining wild–escapee hybridization over a wide geographical area.

2 | MATERIALS AND METHODS

2.1 | Scale samples & DNA extraction

Archived Teno River Atlantic salmon scales, collected as part of a four-decade monitoring program (Niemelä, Erkinaro, Julkunen, & Hassinen, 2005), were the primary source of genetic material for this study. All scales had been stored dry in envelopes prior to DNA extraction. As samples of aquaculture salmon, we used 240 aquaculture escapees captured in the Teno River between 1987 and 2010 (Erkinaro et al., 2010; hereafter referred to as 'Teno Escapees'), and 228 escapees captured in the adjacent coastal waters of Finnmark, Norway, in 2008

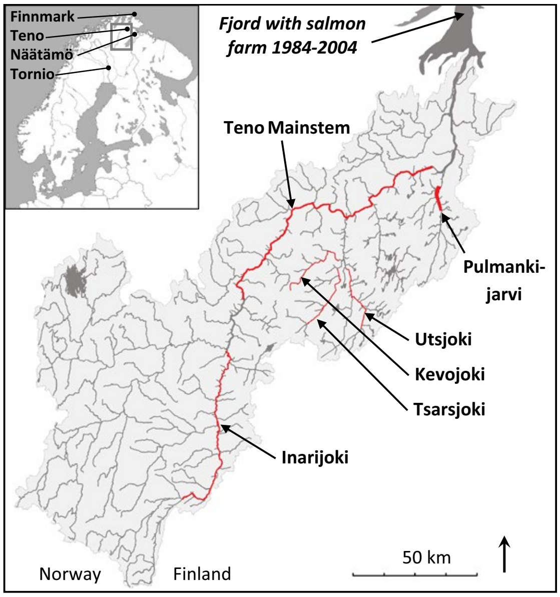

FIGURE 1 Sampling locations. Samples were collected by fisherman at multiple locations along the Teno mainstem, headwaters and tributaries

and 2009 (hereafter referred to as 'Finnmark Escapees'). All fish had been identified as escapees on the basis of morphological features and scale growth ring patterns consistent with previous captive rearing (Fiske, Lund, & Hansen, 2005). We expect these escapees to represent multiple different Norwegian strains of aquaculture salmon that are utilized in regional fish farms. As samples of wild Teno salmon unaffected by introgression from aquaculture escapees, we used individuals caught in the river between 1982 and 1987. Although a salmon farm was established near the mouth of the Teno in 1984, regional levels of salmon aquaculture were relatively low in the early 1980s, and the hybrid offspring of early escapes from this farm are not expected to return to the river before 1989 (Erkinaro et al., 2010). We chose not to use pre-1982 samples due to the decrease in SNP genotyping quality with sample age when using similar SNP arrays (Johnston et al., 2013). Individuals were collected from the Teno mainstem (n = 120, hereafter referred to as 'Old Teno Mainstem'), the Teno headwaters (Inarijoki, n = 114) and four tributaries: Kevojoki (n = 114), Pulmankijärvi (n = 114), Tsarsjoki (n = 114) and Utsjoki (n = 120) (Fig. 1). The salmon spawning in Kevojoki, Pulmankijärvi, Tsarsjoki and Utsjoki are known to be temporally stable, genetically distinct, populations (Vähä, Erkinaro, Niemelä, & Primmer, 2008; Vähä et al., 2007). The Teno mainstem has recently been shown to contain two overlapping subpopulations with low genetic divergence between them (Aykanat et al., 2015; Johnston et al., 2014); however, these were not separated for our study as the focus was on wild–aquaculture hybrid detection. The Inarijoki (headwater) population is genetically similar to the population in the upper Teno mainstem (Vähä et al., 2007). For the Old Teno Mainstem sample, we selected equal numbers of multi-seawinter (MSW, 3 years at sea) and one-seawinter (1SW, 1 year at sea) individuals (Johnston et al., 2014; Barson et al., 2015; sea age determined from scale growth ring patterns) from throughout the river. To obtain sufficient samples, we used scales from individuals that were caught from the last week of July onwards (6–8 weeks prior to spawning, Vähä et al., 2007) – thus, although the majority are expected to belong to the Teno mainstem populations, some may have been *en route* to other spawning locations. There are fewer multi-seawinter fish spawning in Inarijoki and Teno tributaries (Vähä et al., 2007), and we only sampled 1SW fish in these locations. We selected an equal number of males and females within each location or seawinter class.

We extracted DNA from 2 to 4 scales per individual using a QIAmp DNA mini kit (Qiagen), following the manufacturer's protocol and with an initial Proteinase K digestion step. Older scale samples are known to yield more degraded DNA (Johnston et al., 2013) and therefore filter tips were used with pre-2000 scales to minimize contamination risks. We assessed quality and concentration of all DNA extractions using a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific Inc.). The Nanodrop method is known to overestimate concentrations when DNA is partly degraded (Simbolo et al., 2013); therefore, we also measured DNA concentration in a subset of samples (n = 14) via fluorometric quantitation using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc.).

Pilot studies showed that efficacy of Affymetrix SNP genotyping for older scales declined with the DNA concentration in the initial extraction: therefore, only extracts with >150 ng DNA/µl (as quantified by Nanodrop) were used further. Old Teno Mainstem and Teno Escapees were genotyped individually. Individuals from Teno headwater and tributaries, and Finnmark Escapees, were pooled by population for allelotyping, which is a cost-effective approach for estimation of samplewide allele frequencies using SNP chips (Johnston et al., 2013; Ozerov et al., 2013; Sham, Bader, Craig, O'Donovan, & Owen, 2002). The following numbers of individuals were included in sample pools: Inarijoki 110, Kevojoki 102, Pulmankijärvi 83, Tsarsjoki 107, Utsjoki 108, Finnmark Escapees 227. Information on DNA concentration, estimated using Nanodrop, was used to equalize the amount of DNA contributed by each individual to the pool. To account for pipetting and allelotyping variability, six replicate pools were created (technical replicates) and two aliquots (subsamples) of each of these six pools were provided for analysis. For extracts from scales from the 1980s, there was an approximately linear relationship between DNA concentration estimated by Nanodrop and that estimated by Qubit, with Nanodrop estimates around five times higher than Qubit estimates. Therefore, for genotyping on the Affymetrix array we provided Old Teno Mainstem and Teno Escapee samples, and Teno pools, at a Nanodrop-estimated concentration of 70 ng/µl. The Finnmark Escapee pool, containing DNA extracted from recently collected scales, was provided close to the Affymetrix recommended concentration (10 ng/µl), at a Nanodrop-estimated concentration of 15 ng/µl.

For analysis and quality control purposes, we used data from three additional sets of scale samples: 530 Atlantic salmon individuals collected from the Teno mainstem between 2001 and 2003 as described in Johnston et al., 2013 & 2014 (hereafter referred to as 'New Teno Mainstem'); 240 individuals collected between 2006 and 2008 from the Näätämö River of Finland and Norway, which is adjacent to the Teno; and 120 individuals collected between 2005 and 2008 from the

Tornio River of Finland and Sweden, which flows into the Baltic Sea. New Teno Mainstem and Tornio fish were individually genotyped on the 220K array as part of the Barson et al. (2015) study. New Teno Mainstem individuals were also combined into four separate pools and allelotyped. Pooling of these samples was as described in Johnston et al., 2013: three pipetting replicates were generated per pool and three subsamples allelotyped per replicate. Näätämö samples were allelotyped: individuals were combined into four different pools each containing 60 individuals; four pipetting replicates were performed per pool, and each replicate allelotyped once.

2.2 | Genotyping and allelotyping

A custom 220K Affymetrix Axiom array was used to allelotype or genotype samples on a GeneTitan genotyping platform, according to manufacturer's instructions (Affymetrix, USA). The SNPs on this array were a subset of those included on the 930K XHD *Ssal* array developed by T. Moen and colleagues (unpublished data), and had been chosen for maximum informativeness on the basis of their SNPolisher performance (SNPolisher, V1.4, Affymetrix), minor allele frequency (MAF) in aquaculture samples and physical distribution. All of these SNPs have a known location on the NCBI RefSeq Atlantic salmon genome (Lien et al., 2016, available: http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Salmo_salar/100/). To ensure correct identification of genotype clusters, we applied the Affymetrix Best Practices Protocol for SNP calling simultaneously to a large data set that included the Old Teno Mainstem and Teno Escapee samples and all individuals genotyped for Barson et al. (2015). For allelotyping, pooled samples were subjected to the standard genotyping methodology, but normalized and summarized Allele A and Allele B probe intensities were returned instead of genotype calls.

2.3 | Quality control of individually genotyped samples

One hundred and one of 112 Old Teno Mainstem samples, 199 of 239 Teno Escapee samples, 526 of 530 New Teno Mainstem samples and 117 of 120 Tornio samples passed quality controls on the Affymetrix array. Subsequent quality control steps were performed using PLINK v.1.90 (Chang et al., 2015; Purcell et al., 2007). First, we removed 1,112 SNPs not mapped to an assembled *S. salar* chromosome, 35 SNPs known to have off-target variants, and 1,208 SNPs deviating from Hardy–Weinberg equilibrium at $p < .0001$ in the combined Old Teno Mainstem and New Teno Mainstem samples (indicative of technical genotype calling problems). Subsequently, we excluded 18,348 SNPs with >10% missing data or a MAF <10% in the combined Old Teno Mainstem and Teno Escapee data set. Finally, we excluded 17 individuals with >10% missing data. Following these quality control steps, 199,297 SNPs, genotyped in 94 Old Teno Mainstem, 192 Teno Escapee, 525 New Teno Mainstem and 115 Tornio individuals, were retained for analysis.

We examined genotyping repeatability in the Old Teno Mainstem samples by comparing genotype calls for four individuals that had been genotyped twice, using the *merge* function in PLINK; this was compared to the repeatability of five repeatedly genotyped New Teno Mainstem samples.

As an initial exploration of the genetic variation in the combined Old Teno Mainstem, New Teno Mainstem and Teno Escapee data set, we removed SNPs with >2% missing data, performed a linkage disequilibrium pruning step in PLINK (window size = 100, shift = 10, VIF = 2) and then used the *genome* function to calculate pairwise identity-by-state between all individuals, based on the remaining 48,375 SNPs. The presence of genotypic clusters was investigated by performing a two-dimensional multidimensional scaling analysis (MDS) on the genomic identity-by-state (IBS) matrix in PLINK and visualizing the output using ggplot2 in R 3.1.2 (Wickham, 2009; R Core Team 2015, Fig. 2).

2.4 | Estimation of allele frequency from pooled individuals

We estimated allele frequency at each SNP in each pool by calculating the relative intensity of the B-allele probe signal (=B-allele intensity/(A-allele intensity + B-allele intensity)), taking the median over all replicated pools, and applying a polynomial-based probe-specific (PPC) correction to account for differential hybridization efficiency of the two probes (Anantharaman & Chew, 2009; Brohede, Dunne, McKay, & Hannan, 2005). To obtain the PPC correction coefficients for each SNP, a second-order polynomial describing the relationship between relative B probe intensity and genotype call (AA, AB or BB) was fit over the 525 individually genotyped New Teno Mainstem samples, plus 86 samples from the Teno mainstem genotyped for a different study,

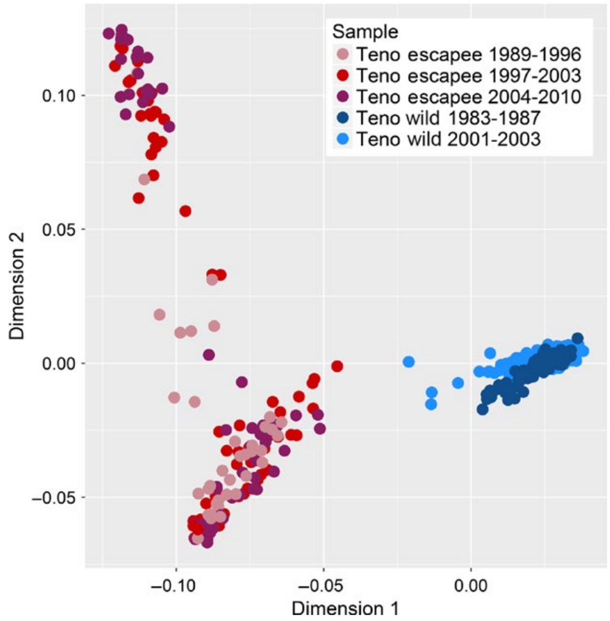

FIGURE 2 Multidimensional scaling analysis plot visualizing genomewide identity-by-state amongst Old Teno Mainstem, New Teno Mainstem and Teno Escapee samples. Each point represents an individually genotyped fish. Teno Escapee samples are colour-coded by collection period

using a custom script in R. The polynomial with these coefficients was then used to correct the estimated allele frequency for that SNP in all allelotyped pools. SNPs with a PPC-corrected allele frequency >1 or <0 were considered monomorphic, and allele frequencies were adjusted accordingly. The accuracy of the PPC-corrected allele frequency estimates for the four New Teno Mainstem pools was investigated by regressing them against the true allele frequencies obtained from individual genotyping. Linear regression was performed using the *lm* function in R, specifying the model as (PPC-corrected frequency) ~ 0 + (True frequency) and using default parameters.

2.5 | Selection of SNPs for hybrid class discrimination

To select SNPs potentially informative for hybrid class discrimination, we focused on regions of the genome that are unusually divergent between aquaculture escapees and wild Teno fish that were collected prior to aquaculture influence. To identify these regions, we used two different genome scan approaches which take population allele counts as input and use F statistics to identify outlying loci: Fdist2 (Beaumont & Nichols, 1996, available: http://www.maths.bris.ac.uk/~mamab/software/) and Bayescan (Foll & Gaggiotti, 2008). The source code for Fdist2 was slightly modified to allow a larger number of markers and compiled under Linux. Genomes scans were applied to the entire data set of 199,297 SNPs. Although our wild and aquaculture populations are expected to violate some of the model assumptions underlying these approaches (e.g. island migration model; independent divergence from a common ancestor), as our primary aim was to identify loci that could be used to discriminate between them rather than make inferences about differential selection we considered their use justified.

We converted allele frequency estimates from pooled samples into allele counts assuming 200 allele copies per locus per pool. For Fdist2 and Bayescan analyses, the 'All Escapee' sample was the combined allele counts of the individually genotyped Teno Escapees and the allelotyped Finnmark Escapee pool. Initially, this sample was compared to the combined 'All Wild' sample (allele counts from individually genotyped Old Teno Mainstem fish plus allele counts from pooled Inarijoki, Kevojoki, Tsarsjoki, Utsjoki and Pulmankijärvi). Subsequently, we compared the All Escapee sample each of the six Teno subpopulations separately. We ran Bayescan using default parameters. To apply Fdist2, we used the package function *datacal* to calculate observed heterozygosity and F_{st} for each locus. We then simulated the expected null distribution of heterozygosity and F_{st}, assuming two demes/populations of 100 individuals, over 10,000,000 loci, using the function *fdist2*. In order to identify the model input value for *Expected* F_{st} that would generate a mean simulated F_{st} approximating the mean observed F_{st} between each escapee–wild comparison, we made pilot runs of 100,000 simulations iteratively changing *Expected* F_{st} until we obtained the desired values. Input values for *Expected* F_{st} were as follows: All Wild/All Escapee: 0.143; Inarjoki/All Escapee: 0.172; Kevojoki/All Escapee: 0.200; Pulmankijärvi/All Escapee: 0.207; Old Teno Mainstem/All Escapee: 0.115; Tsarsjoki/All Escapee: 0.361; Utsjoki/All Escapee: 0.228. To obtain the distribution of F_{st} values simulated by *fdist2* (within each H_e bin of 0.04), calculate empirical probabilities of our observed F_{st} values on the basis of these distributions, and convert these into *q* values, we used the R functions in *getPvalues.R*, written by Lotterhos and Whitlock (2014) and available in the Dryad repository (doi: 10.5061/dryad.v8d05).

All outlying loci identified by Bayescan or Fdist2 in all comparisons were examined for intrachromosome linkage disequilibrium using the r^2 function in Plink, applied to the combined Old Teno Mainstem and New Teno Mainstem data sets. From these results, we identified clusters of physically adjacent linked SNPs: from observation of the data, a SNP was arbitrarily considered to be within a linked cluster when its pairwise r^2 with any other SNP within the cluster was >0.2.

We initially selected a set of 200 SNPs for use in hybrid discrimination, as pilot studies demonstrated no substantial increase in assignment efficacy using additional markers (data not shown), and because roughly this number of SNPs can be conveniently analysed on two 96-well plates. We first chose SNPs located in linked clusters that were identified as outliers by both Bayescan and Fdist2 in the All Wild/All Escapee comparison. We chose one SNP per cluster, selecting the one that had the strongest outlying pattern over all the population comparisons. We supplemented these with SNPs in different linkage clusters that were identified as outliers by either Bayescan or Fdist2 in at least four comparisons between different Teno subpopulations and All Escapees.

2.6 | Population genetic characteristics

For each population, mean expected heterozygosity was estimated from allele frequencies of the 199,297 SNPs using R. Pairwise unbiased F_{st} values between all populations (weighted by heterozygosity, Cockerham & Weir, 1993; Weir & Cockerham, 1984) were calculated from estimated allele counts using *datacal*. To examine whether use of pooled samples biased our estimation of F_{st} we calculated pairwise F_{st} between the All Escapee sample and each of the four New Teno Mainstem pools, first using allele frequencies obtained from individual genotyping and then using allele frequencies estimated using allelotyping. We also repeated all pairwise F_{st} comparisons using the subsets of 200, 160, 120, 80 and 40 loci selected for hybrid discrimination.

2.7 | Simulation of hybrid populations

To test whether the 200 SNPs could in combination be used to discriminate hybrid classes, we used the Python package simuPOP (Peng & Kimmel, 2005) to simulate populations with allele frequencies at these SNPs approximating those observed in our genotyped and/or allelotyped samples (All Escapee, Old Teno Mainstem, Inarijoki, Kevojoki, Pulmankijärvi, Tsarsjoki & Utsjoki, plus the combined All Wild sample; see further details below), and ran them through three generations of hybridization. SimuPOP enables the user to simulate physical linkage between markers on the same chromosome by providing linkage distances between markers. Based on total physical length of the 29 assembled *S. salar* chromosomes (≈2200 million bp) and mean total linkage map distance estimated by Gonen et al., 2014 (2190 cM), we used the approximation 1 million bp = 1 cM.

Several sources of bias could lead us to overestimate the efficacy of our 200 SNPs to discriminate the genomes of wild fish and aquaculture escapees. First, we observed that using allele frequencies obtained via allelotyping caused F_{st} between the All Escapee and New Teno Mainstem pools to be overestimated by up to 7% (Table S1, see below). Second, the estimate of pairwise F_{st} from population samples is in general expected to be higher than the true population-level value due to sampling variance (Anderson, Waples, & Kalinowski, 2008). In order to minimize these biases, we deliberately adjusted the allele frequencies in our simulated wild and escapee populations to be closer to one another. Initial B-allele frequencies for the simulated populations were those estimated from genotyping/allelotyping the real population samples. First, based on a comparison of actual allele frequencies in the New Teno Mainstem pools to those estimated by allelotyping (Fig. S1), we adjusted allele frequencies of 0 or 1 to 0.025 or 0.975, respectively. For each locus in each population, we then generated a new B-allele frequency by making random draws from the beta-distribution with parameters (*f**200, (1-*f*)*200), where *f* is the B-allele frequency following the previous adjustment step. Draws were performed using the R function *rbeta()*. For simulated Inarijoki, Kevojoki, Tsarsjoki, Pulmankijarvi, Utsjoki and All Wild populations, we chose the first value that adjusted the B-allele frequency closer to the All Escapee value; for the simulated All Escapee population we chose the first value closer to the All Wild value. As Old Teno Mainstem allele frequencies were estimated from individual genotyping only, we did not adjust these values for the simulated Old Teno Mainstem population.

To create Generation 0 in simuPOP, we used the adjusted allele frequencies to initialize an escapee and a wild population each containing 500 individuals of each sex. We estimated pairwise F_{st} between the simulated Generation 0 populations by converting the adjusted allele frequencies into allele counts assuming 1,000 alleles per population and using these as input into the Fdist2 function *datacal*. We then simulated three generations of hybridization in simuPOP as follows: 100 individuals migrated from the escapee population into the wild population each generation; subsequently, 500 offspring of each sex were generated by random mating within each population, with parents sampled with replacement and each pair producing a single offspring. Individual IDs, parental IDs and genotypes for all individuals in all Generations (0 to 3) were output to a single file in (PED) format. We used a custom R script to calculate hybrid class for each individual by tracing its lineage from Generation 0. To facilitate comparison of hybrid class assignments amongst simulations with different ancestral wild populations, assignments were performed on standardized subsets of 400 individuals for each simulation. These subsets were generated by harvesting the same number of individuals of each hybrid class for each simulation. The number of individuals to be harvested was determined from the observed frequency of that hybrid class in the relevant generation over all simulations. For example, in Generation 2 overall frequencies of Pure Wild and Wild Backcross approximated 0.505 and 0.215, respectively; therefore, the test subset of 400 Generation 2 individuals for each wild ancestral population included 202 Pure Wild and 86 Wild Backcross individuals. Three possible Generation 3 hybrid classes (Escapee Backcross X Escapee Backcross, Escapee Backcross X F2, and F2 X F2) were excluded due to the small proportional representation of these classes. We also generated an independent set of Generation 0 individuals for each simulation (50 wild, 50 escapee) using the same initial allele frequencies. These individuals were created as wild and escapee reference samples and did not contribute to the hybridizing populations.

2.8 | Hybrid class discrimination in simulated populations

We first used a Bayesian approach implemented in the program NewHybrids (Anderson & Thompson, 2002) to assign simulated individuals to user-defined pure or hybrid classes (six possible classes for Generation 2; 21 possible classes for Generation 3, of which 18 were present in the data set and 14 could potentially be discriminated by NewHybrids). We provided the 100 reference individuals described above as reference samples: these individuals were not considered to be part of the hybrid test population. We additionally ran the Generation 2 analysis without a reference sample. We used the command-line version of NewHybrids, compiled under Linux, and ran the analysis twice for each data set using different random seeds, with a burn-in of 10,000 followed by 50,000 sweeps and default values for all other parameters. Number of sweeps was determined a priori by performing pilot runs in the graphical version of NewHybrids, with one of the Generation 3 data sets, using the same default parameters, and visually following the progress of the analysis. An individual was considered to be correctly classified when it was assigned to its own class with probability >0.5.

To examine the efficacy of hybrid class discrimination using fewer SNPs, we repeated the NewHybrid analyses with subsets of the initial 200 loci. Loci were ranked by allele frequency difference between the All Wild and All Escapee samples, and those with the smallest difference removed first. Results from analyses with $n = 160$, 120, 80 and 40 SNPs were compared to those from the full set of 200. We quantified the ability of different numbers of SNPS to correctly assign individuals to different hybrid classes, following Vähä and Primmer (2006), by defining the following measures. *Efficiency* is the proportion of individuals in a certain hybrid class that were actually assigned to that class by NewHybrids, for example (Total number of simulated F1 assigned to the F1 class)/(Total number of simulated F1). *Accuracy* is the proportion of individuals assigned to a class by NewHybrids that actually belong to that class, for example (Total number of simulated F1 assigned to the F1 class)/(Total number of all individuals assigned to the F1 class). *Overall performance* is the mean of *Efficiency* multiplied by *Accuracy* for each hybrid class over all populations.

As a supporting analysis, we also estimated the proportion of wild and escapee ancestry of each simulated individual using the command-line version of the program Structure (Pritchard, Stephens,

& Donnelly, 2000). Again, we provided reference samples of 50 wild fish and 50 escapees and defined them using 'POPDATA', 'POPFLAG' and 'USEPOPINFO' with MIGRPRIOR = 0.0001. We used *k* = 2, a burn-in of 20,000 followed by 200,000 MCMC steps, recorded 95% confidence intervals of estimated ancestry, and retained default values for all other parameters.

From Structure results, five classes (pure wild, pure escapee, F1 or F2 hybrids, and backcrosses in each direction) in Generation 2, and nine classes in Generation 3, could be discriminated on the basis of expected admixture proportions. We assigned individuals to classes by examining their proportion of ancestry from the wild cluster. Individuals were considered to be assigned to a class when the 95% confidence intervals of their estimated wild ancestry did not overlap the expected mean ancestry of adjacent classes (e.g. an F1/F2 individual has an upper CL of wild ancestry <0.75 and a lower CL of wild ancestry >0.25; pure wild or pure escapee individuals had the 95% CL of their wild ancestry overlapping 1.0 and 0.0, respectively). We examined the assignment performance of different sets of SNPs as described above.

2.9 | Hybrid discrimination in additional wild populations

As the discriminatory SNPs were selected based on allele frequency differences between the same populations that were subsequently used to test them, our results may overestimate the efficacy of SNPs to discriminate wild–escapee hybrids in other populations ('high grading bias', Anderson, 2010; Waples, 2010). We therefore tested the ability of these 200 SNPs to discriminate different classes of simulated hybrids using two additional wild populations. The Näätämö Atlantic salmon population belongs to same evolutionary lineage (North Barents/White Sea) as the Teno River, while the Tornio population is within a genetically divergent lineage (Baltic, Bourret et al., 2013, Fig. S9). For Näätämö, B-allele frequency at the 200 SNPs was estimated by taking the median over all 16 allelotyped pools and applying the PPC correction; for Tornio, it was estimated from all 115 individually genotyped fish. Allele frequencies for Näätämö and Tornio were not further adjusted. Although we did not have an independent escapee sample, we made our analysis more conservative by estimating Generation 0 allele frequencies for our simulated escapee population from a randomly chosen subset of 90 Teno Escapees rather than the entire All Escapee sample. The remaining 99 Teno Escapee were used to estimate allele frequencies for the escapee reference sample. As before, three generations of hybridization were simulated using simuPOP, and NewHybrids used to assign simulated Generation 2 and Generation 3 individuals to different hybrid classes as described above.

3 | RESULTS

3.1 | Initial exploration

Overall, Old Teno Mainstem samples, collected in the 1980s, failed Affymetrix quality controls more frequently and exhibited more missing genotypes than New Teno Mainstem samples, collected in the 2000s. Mean genotype repeatability over the four replicate Old Teno Mainstem samples was 98.0%, as compared to 99.3% over the five replicate New Teno Mainstem samples.

Preliminary exploration of genomewide patterns of identity-by-state revealed clear genetic differentiation between the individually genotyped Teno Mainstem and Teno Escapee samples (Fig. 2). Further, the escapee sample comprised two, largely discrete, genetic clusters. These clusters were not related to the time period in which the escapees were collected, and overall, we observed no clear temporal patterns in the genomic composition of the escapee or Teno samples. Three individuals classified as 'aquaculture escapees' clustered with the wild Teno individuals and were considered misidentified and removed (Fig. S9). No individuals classified as wild Teno fish clustered with the aquaculture escapees; however, several New Teno Mainstem individuals located between the Teno and Escapee clusters were considered potential wild–escapee hybrids (Fig. 2).

3.2 | Accuracy of allele frequency estimation by allelotyping

For the 199,297 SNPs that remained after filtering, population allele frequencies estimated from the four New Teno Mainstem pools with the PPC correction were closely linearly related to the true frequencies estimated by individually genotyping (Fig. S1; each pool: r^2 = .996, residual SE = 0.037–0.039).

3.3 | Population genetic parameters

Mean H_e within the sampled populations, calculated over all 199,297 SNPs, was as follows: Teno Escapees, 0.415; Finnmark Escapees, 0.348; Old Teno Mainstem, 0.381; New Teno Mainstem 0.376: Inarijoki, 0.340; Kevojoki, 0.377; Pulmankijarvi, 0.349; Tsarsjoki, 0.276; Utsjoki, 0.313; Naatamo, 0.300; Tornio, 0.280. Pairwise F_{st} values amongst the samples (All Escapee plus eight wild populations) are shown in Table 1. Mean pairwise F_{st} between aquaculture escapees and the six Teno populations was 0.097, compared to a mean pairwise F_{st} of 0.083 amongst the Teno populations. Relatively high pairwise F_{st} between Tsarsjoki and most other locations reflected the reduced genetic diversity in this population. With Tsarsjoki removed, mean aquaculture–wild pairwise F_{st} and mean amongst-wild pairwise F_{st} were 0.084 and 0.067, respectively. The high pairwise F_{st} between Tornio and all the other populations reflected the phylogeographic distinctness of this population.

As expected, pairwise F_{st} between the All Escapee and wild populations was much higher when calculated only from the discriminatory subset of SNPs (Table S1, mean F_{st} across Teno/All Escapee comparisons = 0.464). Pairwise F_{st} values estimated between the New Teno Mainstem pools and the All Escapee sample were slightly higher when allele frequencies had been estimated by allelotyping than when they had been calculated by individual genotyping (Table S1, mean F_{st} from 199,297 loci = 0.067 vs. 0.058; mean F_{st} from 200 loci = 0.324 vs. 0.306).

TABLE 1 Pairwise F_{st} between samples based on all 199,297 SNPs (above diagonal) and 200 discriminatory SNPS (below diagonal)

	All escapee	Inarijoki	Kevojoki	Pulmanki järvi	Teno Old Mainstem	Teno New Mainstem	Tsarsjoki	Utsjoki	Näätämö	Tornio
All escapee		0.078	0.091	0.094	0.054	0.056	0.163	0.103	0.095	0.170
Inarijoki	0.470		0.074	0.072	0.018	0.027	0.141	0.066	0.057	0.203
Kevojoki	0.466	0.087		0.076	0.057	0.078	0.114	0.069	0.131	0.217
Pulmankijärvi	0.471	0.096	0.084		0.062	0.074	0.170	0.105	0.111	0.220
Teno Old Mainstem	0.333	0.075	0.103	0.110		0.006	0.125	0.057	0.049	0.172
Teno New Mainstem	0.307	0.110	0.130	0.134	0.010		0.130	0.061	0.039	0.171
Tsarsjoki	0.529	0.141	0.090	0.140	0.161	0.191		0.034	0.162	0.288
Utsjoki	0.518	0.101	0.077	0.109	0.137	0.168	0.037		0.080	0.227
Näätämö	0.401	0.135	0.184	0.178	0.067	0.063	0.252	0.215		0.213
Tornio	0.272	0.417	0.410	0.424	0.259	0.232	0.487	0.471	0.319	

3.4 | Outlier analysis

In the comparison between all wild fish and all aquaculture escapees (All Wild/All Escapee), we found 227 outlying loci ($q < .05$) with Fdist2 and 1,112 ($q < .05$) with Bayescan, of which 183 overlapped between the two methodologies (Fig. S2; Table S2). Outlier loci were distributed across all 29 chromosomes (Fig. S2). More than 95% of these 1,156 loci were also identified as outliers in one or more of the comparisons between escapees and fish from different Teno locations (Kevojoki, Inarijoki, Pulmankijärvi, Tsarsjoki, Utsjoki or Old Teno Mainstem; 78% of loci were outliers in at least two comparisons, Table S2). Across the genome, 67 linked clusters of SNPs plus 60 single SNPs were identified by both Bayescan ($q < .05$) and Fdist2 ($q < .05$) as outliers in the All Wild/Escapee comparison (Table S2). The set of 200 SNPs selected for use in hybrid discrimination were distributed over 28 of the 29 Atlantic salmon chromosomes (Fig. S2, Tables S2, S3). None of these loci overlapped with those described by Karlsson et al. (2011). Estimated B-allele frequencies for these SNPs and adjusted frequencies used for simulations are provided in Table S3. Pairwise F_{st} values between simulated escapee and wild populations at Generation 0 were lower than the unbiased pairwise F_{st} values calculated from estimated allele frequencies of the real samples (Table S1).

3.5 | Assignment of simulated individuals to different hybrid classes

Replicate runs of NewHybrids for the same data set gave almost identical outcomes, and in all cases, we present results from the first run. For the second generation of hybridization between wild fish and aquaculture fish (Generation 2), the full set of 200 SNPs exhibited a high *overall performance*, assigning all but 24 individuals (1%) to the correct hybrid class over all six simulated Teno subpopulations, irrespective of whether or not a reference sample was provided (Fig. 3, Fig. 5, Tables S4, S5). This performance barely changed when number of SNPs was reduced to 160 (Fig. 5, Fig. S3, Tables S4, S5: 1.4% misassignment). Further reductions in SNP number led to a decline in *overall performance*, particularly when assigning F2 hybrids and escapee backcrosses (Fig. 5, Fig. S3; Tables S4, S5). However, even using as few as 40 SNPs, only 12 hybrid individuals (all wild backcross, 1% of all hybrids) were erroneously identified as pure wild fish.

Assignment of individuals to the many possible hybrid classes created by three generations of wild–escapee hybridization (Generation 3) is a much more difficult analytical problem, and we focus our discussion on the five most frequent classes generated in our scenario of 10% escapees per generation: Pure Wild, Pure Escapee, F1 Hybrids, Wild Backcross and Wild Backcross X Wild. Using the full set of 200 SNPs, 99.1% of Pure Wild individuals, 96.4% of Pure Escapees, 94.8% of F1 Hybrids, 83.8% of Wild Backcross X Wild and 73.2% of Wild Backcross were assigned correctly (Fig. 4, Fig. 5, Table S6). However, rarer hybrid classes were correctly assigned with much lower success (Fig. 4, Table S6). Misassigned individuals were almost invariably assigned to hybrid classes with similar proportions of wild/escapee ancestry. As expected, performance declined with decreasing numbers of SNPs (Fig. 5, Fig S4, Table S6). Once again, misassignment of hybrid individuals to the Pure Wild class was rare and almost entirely limited to Wild Backcross X Wild individuals; even with 80 SNPs, only 4% of hybrids were incorrectly assigned as Pure Wild individuals.

Replicate runs of Structure for the same data set gave congruent results, and we report results from the first run. In general, Structure was a less effective analytical approach than NewHybrids for assigning simulated individuals to hybrid classes, particularly with fewer SNPs and after three generations of hybridization (Fig. 5, Figs S5, S6, Tables S7, S8). Nevertheless, at Generation 2, and with 120 or more SNPs, Structure and NewHybrids assigned individuals to the correct class with similar high levels of accuracy (Fig. 5, Fig. S5, Table S7).

3.6 | Hybrid discrimination in additional wild populations

Simulated hybrids between aquaculture escapees and fish from Näätämö or Tornio were assigned to the correct hybrid class with

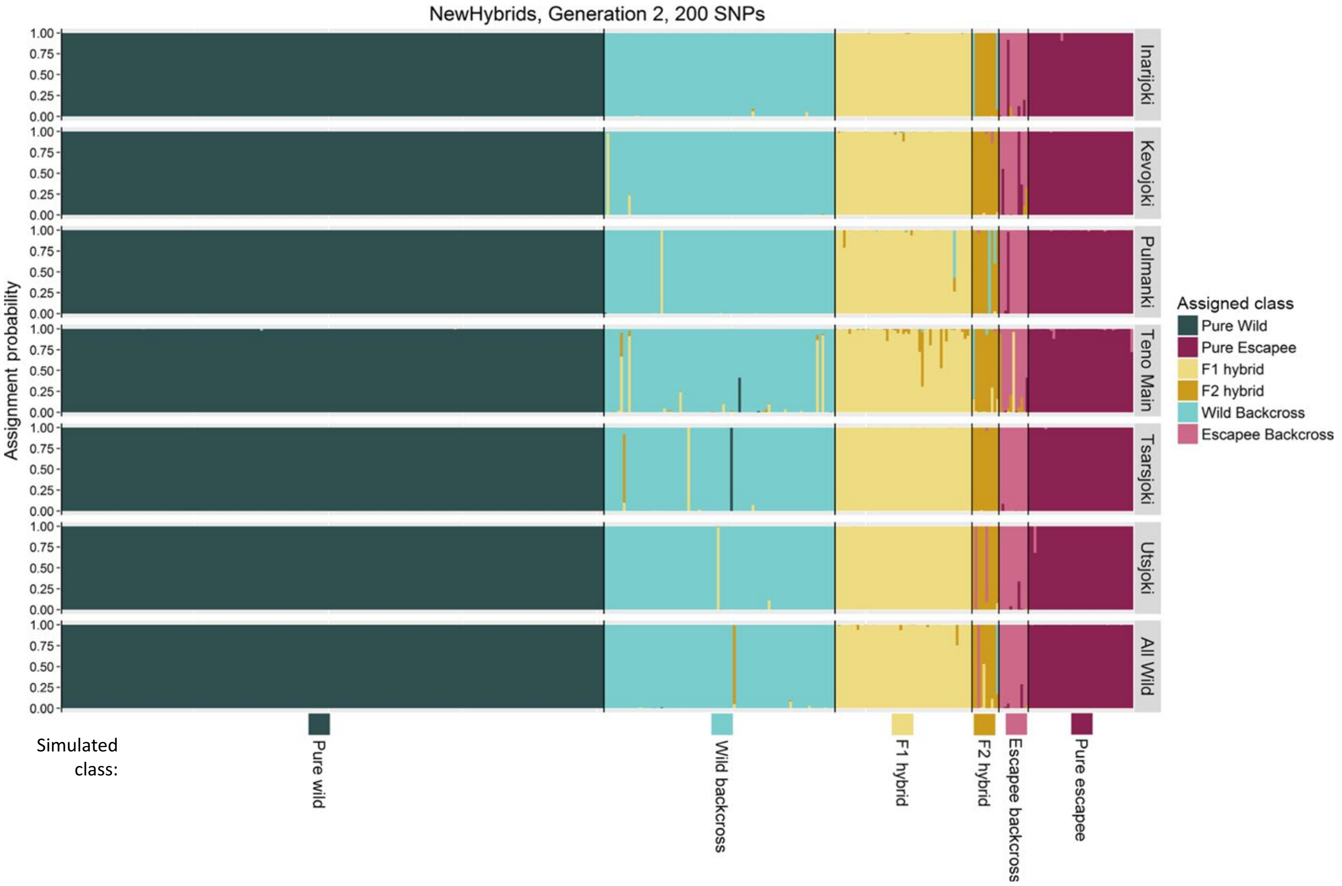

FIGURE 3 Results of NewHybrids analysis for 400 individuals produced by two generations of simulated hybridization between aquaculture escapees (10% of the population) and wild fish from different Teno subpopulations. Each individual is represented by a vertical bar. Individuals are arranged along the *x*-axis by simulated hybrid class, with different hybrid classes bounded by black lines. *Y*-axis indicates the probability, returned by New Hybrids, that an individual belongs to one of the six possible hybrid classes ('Assignment probability'). The different possible hybrid classes are indicated by different colours. For 'All Wild', the wild population was simulated using the average allele frequencies over all seven subpopulations

similar accuracy to that observed in the simulations involving Teno populations (Fig. 6, Fig. S8, Tables S4, S5 & S6), especially using larger numbers of SNPs. Again, misassignment of simulated hybrids as pure wild fish was rare. With 200 SNPs, at Generation 2, there were no such misassignments for either Tornio or Näätämö. At Generation 3, no hybrids were misidentified as wild fish for Näätämö and <1.9% of hybrids were misidentified for Tornio, all of which were Wild Backcross X Wild.

4 | DISCUSSION

Here, we have shown that a suite of 200 SNPs can collectively discriminate advanced-generation classes of hybrid between wild fish from a genetically diverse river and the aquaculture escapees that may be reproducing in that river. Our assessment of the efficacy of these markers is expected to be conservative because we deliberately simulated hybridizing populations with a lower level of genetic divergence between them than was estimated from our real population sample. For example, our simulated wild populations contained no monomorphic loci, while many of these loci may truly be monomorphic in several of the Teno subpopulations. For two generations of hybridization, these markers assign simulated individuals to their correct hybrid class with a very high level of accuracy. Even after three generations of hybridization, many hybrid classes can be reliably identified. Importantly, individuals with hybrid ancestry are rarely identified as pure wild fish in either of the hybrid generations examined. This is a much more accurate level of hybrid class identification than has previously been shown using genetic markers that discriminate aquaculture and wild fish (Karlsson et al., 2011). Further, we observe similarly good discrimination of hybrid classes when simulating hybrids from parental populations that were not originally used to identify the SNPs, including a population in a highly divergent evolutionary lineage, which suggests this marker set may be useful across a range of populations. The rather low number of 200 SNPs can nowadays be assayed relatively cheaply through genotyping via sequencing (Campbell, Harmon, & Narum, 2014) or similar approaches. Moreover, we have shown that a smaller subset of these SNPs enables hybrid class discrimination to a level of accuracy that may be sufficient in many scenarios.

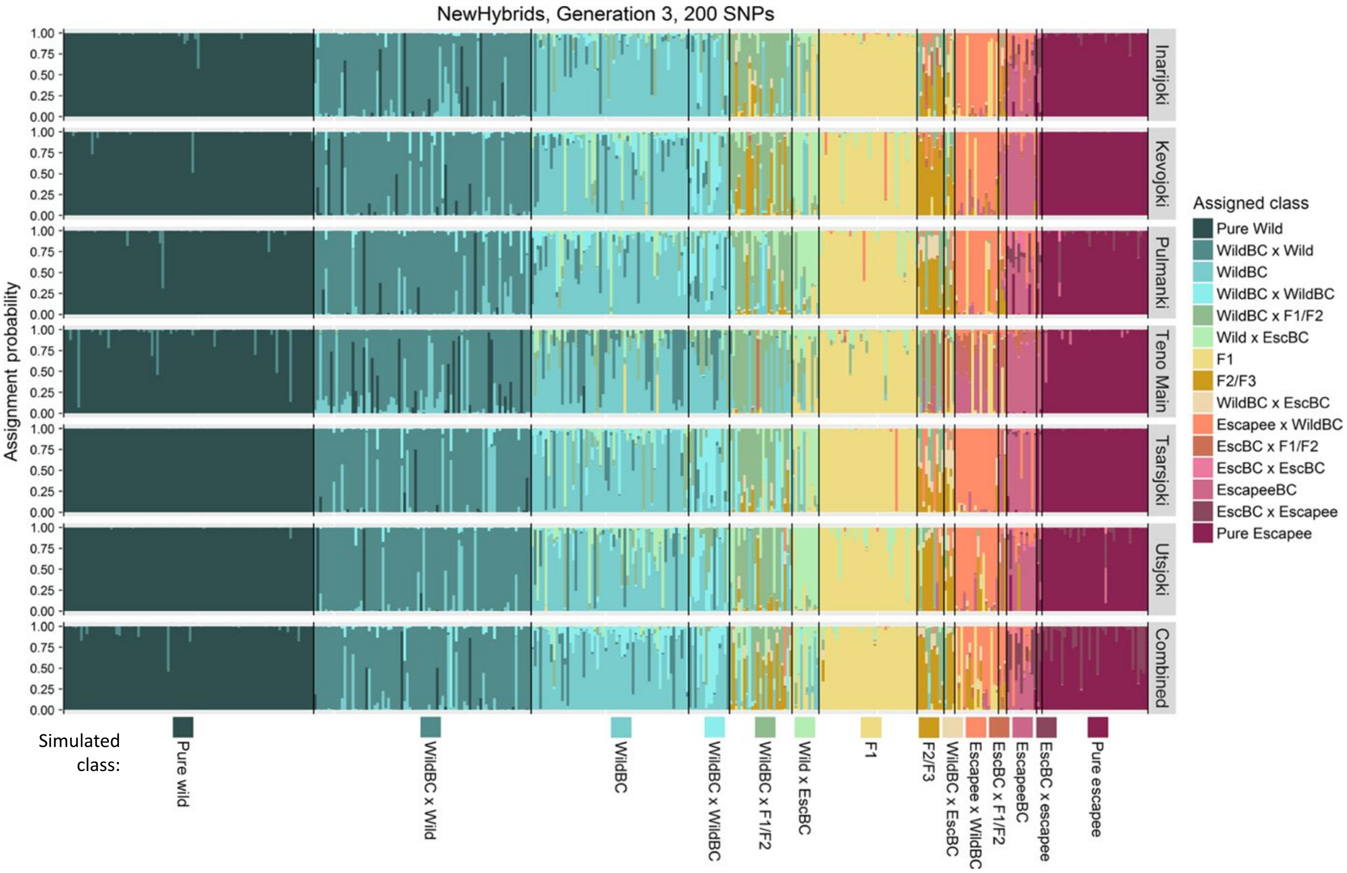

FIGURE 4 Results of NewHybrids analysis for 400 individuals produced by three generations of simulated hybridization between aquaculture escapees and wild fish. Fifteen different hybrid classes can potentially be discriminated by NewHybrids, of which one (EscBC X EscBC) is not present in the mixture due to low frequency. ‘X’: crossed with; ‘Esc’: escapee; ‘BC’: backcross; ‘F1/F2’: F1 hybrid or F2 hybrid; ‘F2/F3’: F2 hybrid or F3 hybrid. For further details see Fig. 3

To examine SNP allele frequencies in the majority of our populations, we used a cost-effective allelotyping approach with the 220K Affymetrix SNP array. Comparison of allele frequencies estimated from individually genotyped fish and pools of the same individuals demonstrated that allelotyping, followed by the PPC correction, well approximated true allele frequencies. We note, however, that the PPC correction coefficients were largely estimated from the same set of samples that were subsequently used to test the accuracy of allele frequency estimates from allelotyping. The performance of these PPC correction coefficients in an independent set of samples has not yet been examined. In particular, in the sample used to generate the PPC correction, 2.3% of SNPs were missing one or more of the three possible genotypes, meaning that the coefficients for these SNPs were unlikely to be estimated correctly (Anantharaman & Chew, 2009). Further, the increased difficulty of estimating DNA concentration when it is partly degraded (Simbolo et al., 2013) may cause allelotyping accuracy to be lower for older scale samples because individuals contribute unequally to the pools. Nevertheless, many of the same outlier regions were identified when comparing the individually genotyped Old Teno Mainstem sample with the aquaculture escapees and when comparing the allelotyped Teno tributaries with the same escapees, suggesting that allelotyping did not bias the results of our analyses. Estimated levels of differentiation between All Escapee and New Teno Mainstem samples were higher when using allele frequencies estimated from allelotyping than when using those obtained from individual genotyping. This bias may partly be caused by rare shared alleles remaining undetected when allelotyping, and should be born in mind when combining results obtained from allelotyping and genotyping in other studies.

Utilizing escapees caught in the wild as our sample of aquaculture salmon enabled us both to capture a large range of aquaculture salmon genetic diversity (including variability over time), and directly sample individuals that were likely to be entering the spawning areas of the Teno River. Visualization of genomic differentiation amongst individually genotyped fish revealed a clear distinction between wild Teno Mainstem individuals (sampled both before and during aquaculture influence) and aquaculture escapees caught in the same river. Moreover, the aquaculture escapees fell into two distinct clusters, which were not strongly related to period of collection, and are likely to represent different aquaculture lines. Norwegian Atlantic salmon aquaculture strains have mixed ancestry from multiple wild populations and are known to be genetically diverse (Glover, Hansen, & Skaala, 2009; Karlsson et al., 2011). This is reflected in the relatively high heterozygosity that we observe within our aquaculture escapee

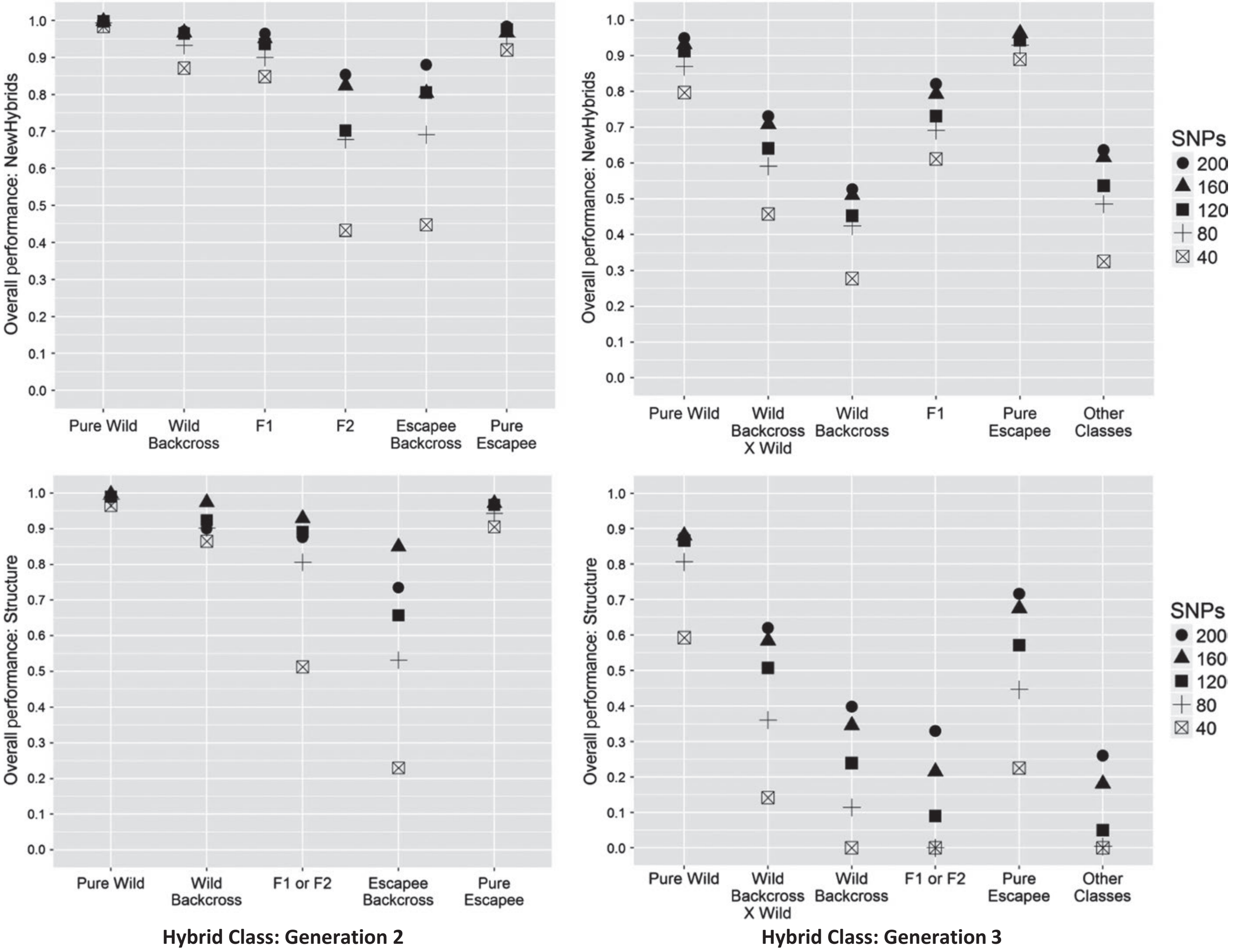

FIGURE 5 Overall performance of different numbers of SNPs for correctly assigning different hybrid classes generated by two or three generations of hybridization and using two different assignment methods. Top: NewHybrids; bottom: Structure

sample, although estimated H_e is also expected to be biased higher in escapees due to the preferential inclusion of SNPs that are polymorphic in aquaculture lines in the 220K array. Previous microsatellite genotyping of the Teno Escapees also found them to be genetically diverse (Erkinaro et al., 2010).

We selected discriminatory SNPs from regions of the genome that are unusually divergent between wild Teno River salmon and aquaculture escapees. This F_{st} outlier approach has previously been used successfully to identify SNPs for stock discrimination in *Oncorhynchus nerka* (Ackerman, Habicht, & Seeb, 2011; Russello, Kirk, Frazer, & Askey, 2012) and wild *S. salar* (Ozerov et al., 2013). Outlying regions are expected to harbour loci under differential selection between the studied groups. Some of the regions that we identified here may contain loci under convergent directional selection in aquaculture strains compared to wild fish. However, given that Norwegian aquaculture strains are primarily derived from a different evolutionary lineage to Teno salmon (Bourret et al., 2013), it is likely that a proportion of these outlying regions reflect long-term evolutionary divergence between the Atlantic (escapee) and North Barents/White Sea (Teno) lineages. Disentangling these two sources of genomic variation between wild Teno fish and aquaculture escapees, and identifying the genes potentially under selection, requires assessment of additional samples from populations in the Atlantic and North Barents/White Sea lineages and is outside the realm of this study.

The fact that the 220K Affymetrix SNP array that we used is optimized for SNPs variable within Norwegian aquaculture lines limited our ability to identify truly discriminatory markers. Ninety per cent of our 200 ‘discriminatory’ SNPs had minor allele frequencies >0.3 in the All Escapee sample. In contrast, 93% of these SNPs had minor allele frequencies <0.1 in the All Wild sample (Table S3). Identification of a set of SNPs with fixed differences between wild salmon and aquaculture lines would greatly increase the accuracy of later generation hybrid identification, with fewer markers being required (Vähä & Primmer, 2006). The outlying genomic regions that we have identified with the genome scan approach are likely to be enriched for such diagnostic SNPs, and targeted sequencing of these regions is a promising approach to find such markers in the future.

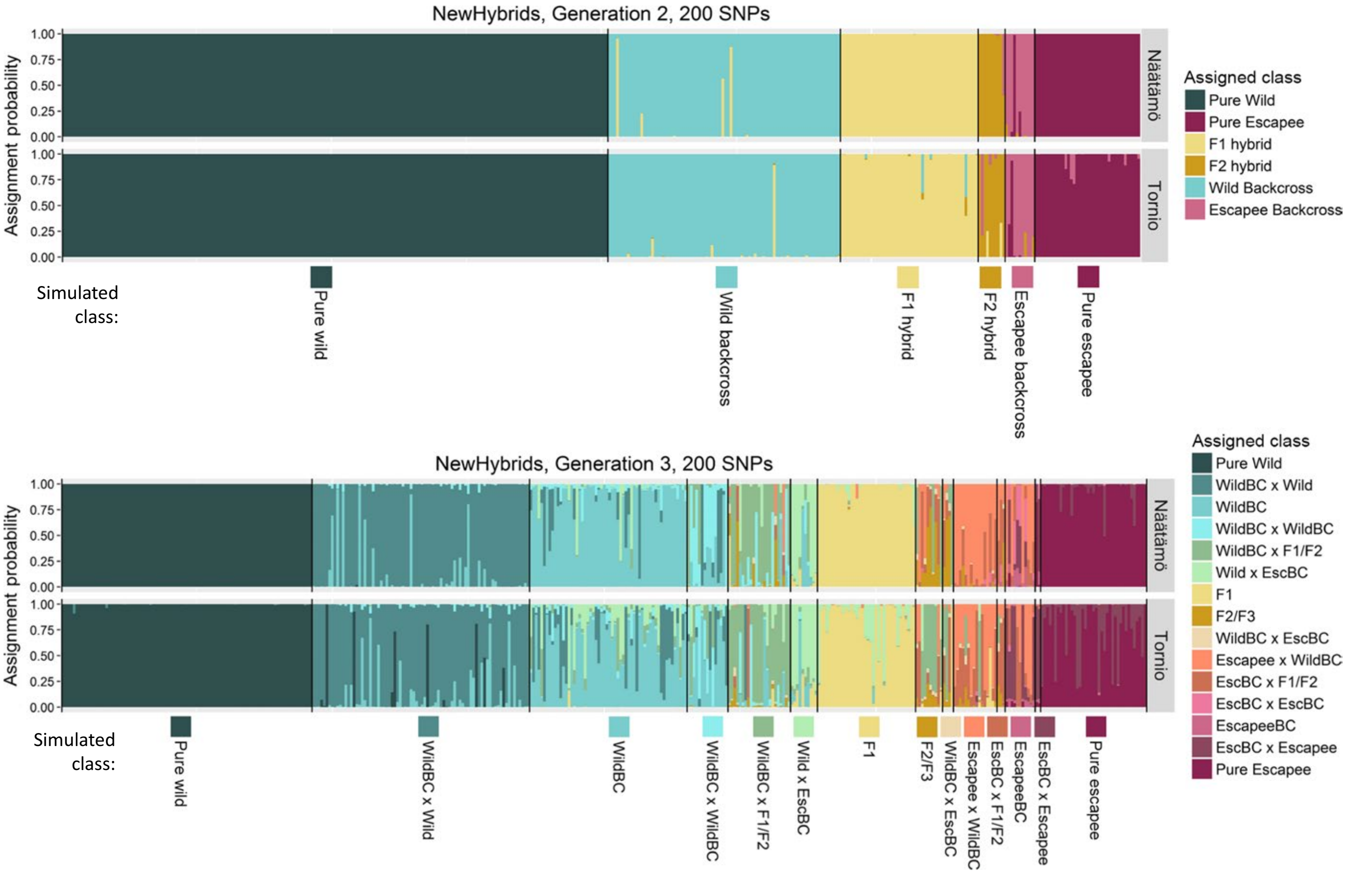

FIGURE 6 Results of NewHybrids analysis for 400 individuals produced by two and three generations of simulated hybridization between aquaculture escapees (10% of the population) and wild fish from Näätämö and Tornio. See Fig. 3 and Fig. 4 for further details

We expect the set of 200 SNPs that we have identified in this study to be useful in investigating wild–aquaculture introgression in all Atlantic salmon belonging to the North Barents/White Sea evolutionary lineage, which includes important populations in northern Norway, Finland and Russia (Bourret et al. 2013). Farming of Norwegian aquaculture salmon is continuing to expand in this region, particularly in Russia (ICES 2015), increasing the introgression threat to these northern populations. This set of 200 SNPs also proved effective at discriminating different classes of simulated hybrids between aquaculture escapees and wild salmon from the Tornio River, which is within the highly divergent Baltic evolutionary lineage (Bourret et al. 2013). Examination of allele frequencies within our samples (Table S3) suggests that our SNP selection procedure has pinpointed many loci with alleles at relatively high frequency within Norwegian aquaculture lines but at low frequency or absent in other lineages. Baltic Atlantic salmon populations are not currently at risk for hybridization with aquaculture escapees, due to environmental conditions (including the presence of the parasite *Gyrodactylus salaris*, Zueva et al., 2014) that are unsuitable for commercial aquaculture strains. However, the strong performance of our SNP set in discriminating different classes of simulated hybrid between aquaculture escapees and the Tornio population suggests they might perform similarly well in other regions where the native wild Atlantic salmon populations are genetically distinct from the Norwegian ancestors of the aquaculture lines. Although several countries have developed their own aquaculture lines from local salmon populations (e.g. Bourret et al., 2011), Norwegian aquaculture strains are also utilized worldwide (e.g. Clifford et al., 1998).

Even if the set of SNPs described here do not perform well in other regions, our results suggest that, using our approach, it may be straightforward to find similar markers that identify hybrids between Norwegian aquaculture strains and other genetically divergent populations. This will be facilitated by the recent development of several other high-density Atlantic salmon SNP arrays (Houston et al., 2014; Yáñez, Houston, & Newman, 2014). Identifying SNPs to discriminate Norwegian aquaculture strains and the wild central and southern Norwegian populations from which they are largely derived is expected to be a more difficult task. The work of Karlsson et al. (2011) was directed at finding markers to study the important problem of wild–escapee hybridization within this evolutionary lineage. Although the Karlsson et al. (2011) SNPs have not been tested for their usefulness at discriminating hybrids beyond the F1, they are expected to be less effective for the identification of later generation hybrid classes simply because of the much lower genetic divergence between the hybridizing lineages compared to our study. The genome scan approach that we have applied here provides opportunities to identify additional discriminatory markers in genomic regions that have diverged between aquaculture fish and their wild ancestors due to shared directional selection amongst domestic lines.

In summary, we have demonstrated that a set of 200 SNPs enables discrimination of different classes of hybrid between wild Atlantic salmon and aquaculture escapees for up to three generations of hybridization. This suite of SNPs will allow detailed examination of the hybridization dynamics between aquaculture escapees and native fish in an important wild Atlantic salmon population. We expect these SNPs, and/or the analytical approaches that we have used to identify them, to also be useful when investigating escapee–wild hybridization throughout a much wider geographic range.

ACKNOWLEDGEMENTS

Jorma Kuusela, Tuomas Metsäniemi & Lauri Rintamäki (Luke); Jani Aaltonen, Tutku Aykanat, Susan Johnston, Katja Salminen & Silva Uusi-Heikkala (University of Turku); and Silje Karoliussen (CIGENE) assisted with scale subsampling, DNA extraction, pooling, and/or genotyping. Tutku Aykanat wrote the R code for the PPC correction. Computational resources were provided by the CSC—*IT Center for Science, Finland*. This study was funded by an Academy of Finland grant (284941, 272836) to CRP and a Finnish Cultural Foundation grant to VLP. We thank the many fishermen who contributed scales to the Teno archive. Comments from four anonymous referees greatly improved the manuscript.

DATA ARCHIVING STATEMENT

Raw data and code used in analyses are archived in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.dg8f3.

LITERATURE CITED

Ackerman, M., Habicht, C., & Seeb, L. (2011). Single-nucleotide polymorphisms (SNPs) under diversifying selection provide increased accuracy and precision in mixed-stock analyses of sockeye salmon from the Copper River, Alaska. *Transactions of the American Fisheries Society*, *140*, 865–881.

Anantharaman, R., & Chew, F. T. (2009). Validation of pooled genotyping on the Affymetrix 500K and SNP 6.0 genotyping platforms using the polynomial-based probe-specific correction. *BMC Genetics*, *10*, 82.

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, *10*, 701–710.

Anderson, E. C., & Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, *160*, 1217–1229.

Anderson, E. C., Waples, R. S., & Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*, 1475–1486.

Aykanat, T., Johnston, S. E., Orell, P., Niemelä, E., Erkinaro, J., & Primmer, C. R. (2015). Low but significant genetic differentiation underlies biologically meaningful phenotypic divergence in a large Atlantic salmon population. *Molecular Ecology*, *24*, 5158–5174.

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., Jacq, C., Jensen, A. J., Johnston, S. E., Karlsson, S., Kent, M., Moen, T., Niemelä, E., Nome, T., Næsje, T. F., Orell, P., Romakkaniemi, A., Sægrov, H., Urdal, K., Erkinaro, J., Lien, S., & Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in Atlantic salmon. *Nature*, *528*, 405–408.

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B: Biological Sciences*, *263*, 1619–1626.

Bicskei, B., Bron, J. E., Glover, K. A., & Taggart, J. B. (2014). A comparison of gene transcription profiles of domesticated and wild Atlantic salmon (*Salmo salar* L.) at early life stages, reared under controlled conditions. *BMC Genomics*, *15*, 884.

Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor, E., Bernatchez, L., & Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, *22*, 532–551.

Bourret, V., O'Reilly, P. T., Carr, J. W., Berg, P. R., & Bernatchez, L. (2011). Temporal change in genetic integrity suggests loss of local adaptation in a wild Atlantic salmon (*Salmo salar*) population following introgression by farmed escapees. *Heredity*, *106*, 500–510.

Brohede, J., Dunne, R., McKay, J. D., & Hannan, G. N. (2005). PPC: an algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of DNA using data from high density microarrays. *Nucleic Acids Research*, *33*, e142.

Campbell, N. R., Harmon, S., & Narum, S. R. (2014). Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*, 855–867.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation Plink: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7.

Clifford, S. L., McGinnity, P., & Ferguson, A. (1998). Genetic changes in Atlantic salmon (*Salmo salar*) populations of northwest Irish rivers resulting from escapes of adult farm salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, *52*, 118–127.

Cockerham, C. C., & Weir, B. S. (1993). Estimation of gene flow from F-statistics. *Evolution*, *47*, 855–863.

Debes, P. V., & Hutchings, J. A. (2014). Effects of domestication on parr maturity, growth, and vulnerability to predation in Atlantic salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, *71*, 1371–1384.

Debes, P. V., Normandeau, E., Fraser, D. J., Bernatchez, L., & Hutchings, J. A. (2012). Differences in transcription levels among wild, domesticated, and hybrid Atlantic salmon (*Salmo salar*) from two environments. *Molecular Ecology*, *21*, 2574–2587.

Edmands, S. (2006). Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding for conservation and management. *Molecular Ecology*, *16*, 463–475.

Erkinaro, J., Niemelä, E., Vähä, J.-P., Primmer, C. R., Brørs, S., & Hassinen, E. (2010). Distribution and biological characteristics of escaped farmed salmon in a major subarctic wild salmon river: implications for monitoring. *Canadian Journal of Fisheries and Aquatic Sciences*, *67*, 130–142.

Fiske, P., Lund, R. A., & Hansen, L. P. (2005). Identifying fish farm escapees. In S. X. Cadrin, K. D. Friedland, & J. R. Waldeman (Eds.), *Stock Identification Methods, Applications in Fishery Science*. Amsterdam: Elsevier.

Fiske, P., Lund, R., & Hansen, L. (2006). Relationships between the frequency of farmed Atlantic salmon, *Salmo salar* L., in wild salmon populations and fish farming activity in Norway, 1989–2004. *ICES Journal of Marine Science*, *63*, 1182–1189.

Fleming, I. A., Agustsson, T., Finstad, B., Johnsson, J. I., & Björnsson, B. T. (2002). Effects of domestication on growth physiology and endocrinology of Atlantic salmon (*Salmo salar*). *Canadian Journal of Fisheries and Aquatic Sciences*, *59*, 1323–1330.

Fleming, I. A., Jonsson, B., Gross, M. R., & Lamberg, A. (1996). An experimental study of the reproductive behaviour and success of farmed and wild salmon (*Salmo salar*). *Journal of Applied Ecology*, *33*, 893–905.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*, 977–993.

Gjedrem, T., Gjøen, H., & Gjerde, B. (1991). Genetic origin of Norwegian farmed Atlantic salmon. *Aquaculture*, *98*, 41–50.

Gjøen, H. (1997). Past, present, and future of genetic improvement in salmon aquaculture. *ICES Journal of Marine Science*, *54*, 1009–1014.

Glover, K., Hansen, M., & Skaala, Ø. (2009). Identifying the source of farmed escaped Atlantic salmon (*Salmo salar*): Bayesian clustering analysis increases accuracy of assignment. *Aquaculture*, *290*, 37–46.

Glover, K., Otterå, H., Olsen, R., & Slinde, E. (2009). A comparison of farmed, wild and hybrid Atlantic salmon (*Salmo salar* l.) reared under farming conditions. *Aquaculture*, *286*, 203–210.

Glover, K. A., Pertoldi, C., Besnier, F., Wennevik, V., Kent, M., & Skaala, Ø. (2013). Atlantic salmon populations invaded by farmed escapees: quantifying genetic introgression with a Bayesian approach and SNPs. *BMC Genetics*, *14*, 74.

Glover, K. A., Quintela, M., Wennevik, V., Besnier, F., Sørvik, A. G. E., & Skaala, Ø. (2012). Three decades of farmed escapees in the wild: a spatio-temporal analysis of Atlantic salmon population genetic structure throughout Norway. *PLoS One*, *7*, e43129.

Gonen, S., Lowe, N. R., Cezard, T., Gharbi, K., Bishop, S. C., & Houston, R. D. (2014). Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. *BMC Genomics*, *15*, 166.

Heino, M., Svåsand, T., Wennevik, V., & Glover, K. A. (2015). Genetic introgression of farmed salmon in native populations: quantifying the relative influence of population size and frequency of escapees. *Aquaculture Environment Interactions*, *6*, 185–190.

Houston, R. D., Taggart, J., Cezard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R. T., Bishop, S. C., Archibald, A. L., Bron, J., Penman, D., Davassi, A., Brew, F., Tinch, A. E., Gharbi, K., & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). *BMC Genomics*, *15*, 90.

ICES (2015). *Report of the Working Group on North Atlantic Salmon (WGNAS)*, 17–26 March, Moncton, Canada: ICES. 332 pp.

Johnsson, J. I., Höjesjö, J., & Fleming, I. A. (2011). Behavioural and heart rate responses to predation risk in wild and domesticated Atlantic salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, *58*, 788–794.

Johnston, S. E., Lindqvist, M., Niemelä, E., Orell, P., Erkinaro, J., Kent, M. P., Lien, S., Vähä, J-. P., Vasemägi, A., & Primmer, C. R. (2013). Fish scales and SNP chips: SNP genotyping and allele frequency estimation in individual and pooled DNA from historical samples of Atlantic salmon (*Salmo salar*). *BMC Genomics*, *14*, 439.

Johnston, S. E., Orell, P., Pritchard, V. L., Kent, M. P., Lien, S., Niemelä, E., Erkinaro, J., & Primmer, C. R. (2014). Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, *23*, 3452–3468.

Jonsson, B., & Jonsson, N. (2006). Cultured Atlantic salmon in nature: a review of their ecology and interaction with wild fish. *ICES Journal of Marine Science*, *63*, 1162–1181.

Jonsson, B., Jonsson, N., & Hansen, L. (1991). Differences in life history and migratory behaviour between wild and hatchery-reared Atlantic salmon in nature. *Aquaculture*, *98*, 69–78.

Karlsson, S., Diserud, O. H., Moen, T., & Hindar, K. (2014). A standardized method for quantifying unidirectional genetic introgression. *Ecology and Evolution*, *4*, 3256–3263.

Karlsson, S., Moen, T., Lien, S., Glover, K. A., & Hindar, K. (2011). Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, *11*(Suppl 1), 247–253.

Katz, J., Moyle, P. B., Quiñones, R. M., Israel, J., & Purdy, S. (2013). Impending extinction of salmon, steelhead, and trout (Salmonidae) in California. *Environmental Biology of Fishes*, *96*, 1169–1186.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., vonSchalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Våge, D. I., deJong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang, X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J., Jonassen, I., Maass, A., Omholt, S. W., & Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *553*, 200–205.

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of Fst outlier tests. *Molecular Ecology*, *23*, 2178–2192.

Metcalf, J. L., Pritchard, V. L., Silvestri, S. M., Jenkins, J. B., Wood, J. S., Cowley, D. E., Evans, R. P., Shiozawa, D. K., & Martin, A. P. (2007). Across the great divide: genetic forensics reveals misidentification of endangered cutthroat trout populations. *Molecular Ecology*, *16*, 4445–4454.

Naylor, R., Hindar, K., Fleming, I. A., Goldburg, R., Williams, S., Volpe, J., Whoriskey, F., et al. (2005). Fugitive salmon: assessing the risks of escaped fish from net-pen aquaculture. *BioScience*, *55*, 427.

Niemelä, E., Erkinaro, J., Julkunen, M., & Hassinen, E. (2005). Is juvenile salmon abundance related to subsequent and preceding catches? Perspectives from a long-term monitoring programme *ICES Journal of Marine Science*, *62*, 1617–1629.

Ozerov, M., Vasemägi, A., Wennevik, V., Diaz-Fernandez, R., Kent, M., Gilbey, J., Prusov, S., Niemelä, E., & Vähä, J. P. (2013). Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. *PLoS One*, *8*, e82434.

Ozerov, M., Vasemägi, A., Wennevik, V., Niemelä, E., Prusov, S., Kent, M., & Vähä, J.-P. (2013). Cost-effective genome-wide estimation of allele frequencies from pooled DNA in Atlantic salmon (*Salmo salar* L.). *BMC Genomics*, *14*, 12.

Parrish, D. L., Behnke, R. J., Gephard, S. R., McCormick, S. D., & Reeves, G. H. (1998). Why aren't there more Atlantic salmon (*Salmo salar*)? *Canadian Journal of Fisheries and Aquatic Sciences*, *55*, 281–287.

Peng, B., & Kimmel, M. (2005). simuPOP: a forward time population genetics simulation environment. *Bioinformatics*, *21*, 3686–3687.

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

Purcell, S., Neale, B., Todd-Brown, K., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., deBakker, P. I. W., Daly, M. J., & Sham, P. K. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*.Vienna, Austria: R Foundation for Statistical Computing.

Rand, P. S. (2013). Current global status of taimen and the need to implement aggressive conservation measures to avoid population and species-level extinction. *Archives of Polish Fisheries*, *21*, 119–128.

Russello, M. A., Kirk, S. L., Frazer, K. K., & Askey, P. J. (2012). Detection of outlier loci and their utility for fisheries management. *Evolutionary Applications*, *5*, 39–52.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., & Owen, M. (2002). DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, *3*, 862–871.

Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., Lawlor, R. T., et al. (2013). DNA qualification workflow for next generation sequencing of histopathological samples. *PLoS One*, *8*, e62692.

Solberg, M., Skaala, Ø., Nilsen, F., & Glover, K. (2013). Does domestication cause changes in growth reaction norms? a study of farmed, wild and hybrid Atlantic salmon families exposed to environmental stress. *PLoS One*, *8*, e54469.

Thorstad, E. B., Fleming, I. A., McGinnity, P., Soto, D., Wennevik, V., & Whoriskey, F. (2008). *Incidence and Impacts of Escaped Farmed Atlantic Salmon Salmo salar in Nature*. NINA Special Report 36. 110 pp.

Vähä, J.-P., Erkinaro, J., Niemelä, E., & Primmer, C. R. (2007). Life-history and habitat features influence the within-river genetic structure of Atlantic salmon. *Molecular Ecology*, *16*, 2638–2654.

Vähä, J.-P., Erkinaro, J., Niemelä, E., & Primmer, C. R. (2008). Temporally stable genetic structure and low migration in an Atlantic salmon population complex: implications for conservation and management. *Evolutionary Applications*, *1*, 137–154.

Vähä, J.-P., & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, *15*, 63–72.

Waples, R. S. (2010). High grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology*, *19*, 2599–2601.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*, 1358–1370.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Yáñez, J. M., Houston, R. D., & Newman, S. (2014). Genetics and genomics of disease resistance in salmonid species. *Frontiers in Genetics*, *5*, 415.

Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P., Lien, S., & Primmer, C. R. (2014). Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? *PLoS One*, *9*, e91672.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.