SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*)

VINCENT BOURRET,*1 MATTHEW P. KENT,† CRAIG R. PRIMMER,‡ ANTI VASEMÄGI,‡§ STEN KARLSSON,¶ KJETIL HINDAR,¶ PHILIP MCGINNITY,** ERIC VERSPOOR,†† LOUIS BERNATCHEZ* and SIGBJØRN LIEN†

*Département de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, 1030 avenue de la Médecine, Québec, Canada G1V 0A6, †Department of Animal and Aquacultural Sciences, Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, PO Box 5003, 1432, Aas, Norway, ‡Division of Genetics and Physiology, Department of Biology, University of Turku, 20014, Turku, Finland, §Department of Aquaculture, Institute of Veterinary Medicine and Animal Science, Estonian University of Life Sciences, 51014, Tartu, Estonia, ¶Norwegian Institute for Nature Research (NINA), 7485, Trondheim, Norway, **School of Biological, Earth & Environmental Sciences, University College Cork, Cork, Ireland, ††Rivers and Lochs Institute Inverness College, University of Highlands and Islands, Longman Building Inverness, Scotland IV1 1SA, UK

Abstract

Atlantic salmon (*Salmo salar*) is one of the most extensively studied fish species in the world due to its significance in aquaculture, fisheries and ongoing conservation efforts to protect declining populations. Yet, limited genomic resources have hampered our understanding of genetic architecture in the species and the genetic basis of adaptation to the wide range of natural and artificial environments it occupies. In this study, we describe the development of a medium-density Atlantic salmon single nucleotide polymorphism (SNP) array based on expressed sequence tags (ESTs) and genomic sequencing. The array was used in the most extensive assessment of population genetic structure performed to date in this species. A total of 6176 informative SNPs were successfully genotyped in 38 anadromous and freshwater wild populations distributed across the species natural range. Principal component analysis clearly differentiated European and North American populations, and within Europe, three major regional genetic groups were identified for the first time in a single analysis. We assessed the potential for the array to disentangle neutral and putative adaptive divergence of SNP allele frequencies across populations and among regional groups. In Europe, secondary contact zones were identified between major clusters where endogenous and exogenous barriers could be associated, rendering the interpretation of environmental influence on potentially adaptive divergence equivocal. A small number of markers highly divergent in allele frequencies (outliers) were observed between (multiple) freshwater and anadromous populations, between northern and southern latitudes, and when comparing Baltic populations to all others. We also discuss the potential future applications of the SNP array for conservation, management and aquaculture.

Keywords: aquaculture, conservation, population genetics, salmonids, selection, single nucleotide polymorphism

Received 27 February 2012; revision received 17 May 2012; accepted 7 June 2012

[1]Correspondence: Vincent Bourret, Fax: 418-656-7176; E-mail: vincent.bourret.1@ulaval.ca
Vincent Bourret and Matthew P. Kent equally share first co-authorship.
Louis Bernatchez and Sigbjørn Lien equally share senior co-authorship.

Introduction

Since their divergence from other salmonids, Atlantic salmon (*Salmo salar* L.) populations have been subjected to strong natural and artificial selection throughout their evolutionary history. These included pronounced habitat shifts associated with glaciations, 'landlocking' of fjords or lakes previously open to the sea, as well as various sources of anthropogenic changes (Verspoor *et al.* 2007). The greatest divergence between Atlantic salmon populations can be seen when comparing the European and North American lineages, which diverged approximately 600 000 years before present (ybp) (King *et al.* 2007). A more recent fundamental division is seen in populations of nonanadromous fish on both continents, which have evolved independently following 'landlocking' of their rivers or fjords (Tessier & Bernatchez 2000; King *et al.* 2007). In addition, Atlantic salmon inhabit a wide range of habitats with different temperatures, from Spain to the high Arctic (Aas *et al.* 2011). Atlantic salmon thus represents an ideal species to disentangle the genomic basis of potential parallel evolution at both historic and recent scales.

More recently, there has been a dramatic decline of many wild Atlantic salmon populations due to overfishing, habitat destruction and possibly other indirect causes such as impacts from aquaculture or climatic change (Friedland *et al.* 2003; Carr *et al.* 2004; Ferguson *et al.* 2007; Ford & Myers 2008; Todd *et al.* 2008). Consequently, wild Atlantic salmon populations have been the focus of international conservation efforts, where population genetic studies have played an important role by genetically inferring phylogeographic patterns and defining evolutionary significant units (ESU) as well as management units (e.g. Vasemägi *et al.* 2005; Palstra *et al.* 2007; Dionne *et al.* 2008; Tonteri *et al.* 2009). However, the limited resolution of the available genetic tools has meant that the genomic basis of local adaptation shaped by natural selection is still largely unknown for Atlantic salmon (Fraser *et al.* 2011). Single nucleotide polymorphism (SNP) arrays represent excellent tools for studying population structure and the effect of natural and artificial selection at the genome level (e.g. The Bovine HapMap Consortium, Gibbs *et al.* 2009; Willing *et al.* 2010; many others). For example, for Atlantic cod (*Gadus morhua*), Bradbury *et al.* (2010) were able to show parallel clinal association between SNP allele frequencies and water temperatures across the species range, in both North America and Europe. A SNP array for salmon would add substantially to previous efforts by providing managers with an enhanced tool for the definition of ESUs on the basis of neutral and adaptive divergence (Kohn *et al.* 2006; Primmer 2009). Moreover, it could also improve the resolution of differentiation of genetic stocks and the efficiency of mixed-stock analysis (Gauthier-Ouellet *et al.* 2009; Griffiths *et al.* 2010).

Aquaculture of Atlantic salmon is practised on a large scale in both the Northern and Southern hemispheres, and today, it is the most cultivated fish in the western world and the most farmed salmonid worldwide. This expanding industry is recognizing the potential benefits arising from development and application of high-density SNP technologies (Dominik *et al.* 2010). Indeed, similar resources have been developed for other production species and are proving invaluable for the detection of quantitative trait loci (QTL) underlying economically important traits (Dekkers & Hospital 2002; Daetwyler *et al.* 2008), strain identification (Suekawa *et al.* 2010) and to implement marker-assisted selection and genomic selection (Soller 1990; Meuwissen *et al.* 2001; Dekkers & van der Werf, 2007). Moreover, accidental release and/or the escape of farmed fish into the surrounding wild environment has raised substantial concerns (McGinnity *et al.* 2003) over the potential genetic impact these fish can have on local wild populations at the genome level (e.g. Hindar *et al.* 1991; Roberge *et al.* 2006; Normandeau *et al.* 2009; Bourret *et al.* 2011). High-density SNP arrays represent a powerful tool to assess the impacts of escapes on wild populations (i.e. introgression), to trace the source of escapees and to characterize the current status of at-risk wild populations (Karlsson *et al.* 2011).

The first goal of this study is to describe the development and application of a dense SNP array for Atlantic salmon. Second, we present the most genetically detailed population study of Atlantic salmon whereby 6176 SNPs were genotyped in 1430 individuals from 38 naturally occurring anadromous and landlocked populations collected across the species range. While the main emphasis concerns a description of the population genetic structure of European populations, several North American populations are also included in the assessment. Third, we aim to identify markers and chromosomal regions potentially under divergent selection among populations between and within lineages as well as between anadromous and landlocked populations. Finally, we consider whether potential adaptive divergence is associated with key biological pathways, and environmental patterns of variation.

Materials and methods

Detection of SNPs in EST databases and genome complexity reduction (GCR)

Expressed sequence tags (EST) mining was performed according to Hayes *et al.* (2007), and using the EST libraries described in this study, a total of 9240 putative SNPs were identified. Limited availability of

DNA excluded the possibility of preparing a Reduced Representation Library (RRL; Altshuler *et al.* 2000); instead, we describe an approach called genome complexity reduction (GCR) that includes a PCR step to isolate a distributed subfraction of the genome. Genomic DNA was extracted from adipose-fin clips collected from *Salmo salar* using a commercial extraction kit. Samples were digested to completion using Xba1 which, compared to other six-cutter restriction enzymes (RE), produced the most uniform smear of DNA fragments. Synthetic double-stranded DNA adapters containing a PCR-binding site were ligated to the resulting sticky-ended fragments using T4DNA ligase. The resulting product was amplified under standard PCR conditions using a 15-s extension time, which promotes the amplification of fragments up to 1200 bp. The combination of RE cut frequency and amplicon length means that a random fraction of the genome is represented in the final product. Gel analysis of the PCR product revealed an expected smear of fragments and a single region of high intensity, most likely resulting from overamplification of a repeated DNA element. Because repeated DNA can produce false-positive SNPs and consume sequencing capacity, we developed an optional step [reduced GCR (rGCR)] whereby PCR product is separated on a 1% agarose gel and a fragment range (e.g. 500–700 bp), which avoids brighter (repeat) regions being isolated. The typically low DNA recovery requires an additional PCR amplification, which is performed as described above. For SNP discovery, GCR libraries were prepared for individual haploid ($n = 2$) and diploid ($n = 6$) samples; in addition, rGCR libraries were prepared for the haploid individuals ($n = 2$) and a single pool sample containing all diploid samples ($n = 1$). Haploid fish were produced by fertilizing eggs with irradiated sperm according to Refstie (1983). All biological material originated from Norwegian commercial aquaculture strains (Aqua Gen).

SNP discovery

The diploid and haploid GCR libraries were each sequenced twice, while the rGCR diploid pool and haploid samples were sequenced just once using a Genome Sequencer FLX system (Roche). A total of 6 927 968 reads were generated across 19 runs with an average read length of 182 bp. Reads were filtered for repeats and adapter sequence using Lucy (http://lucy.ceh.uvic.ca/repeatmasker/cbr_repeatmasker.py); approximately, 30% of reads were eliminated from the data set because they were identified as repeat sequences. Atlas overlapper (Havlak *et al.* 2004) was then used to group 1 404 933 reads into 53 769 bins

(minimum bin size = 5 reads; average = 26, overlap criteria: 92% identity, minimum read length = 40 bp, min overlap = 36 bp, MaxOverlapSeed = 50). Each set of binned reads was processed by Phrap (Gordon *et al.* 1998) to generate 77 858 contigs with an N50 contig length of 349bp. Putative SNP discovery was made by aligning individual binned reads back to their matching contigs using Cross_Match (Gordon *et al.* 1998). Criteria for calling a SNP were a minimum read coverage of 2, and minimum SNP coverage/total coverage ≥ 0.2. The possibility of these SNPs ($n = 17 844$) being located within unknown (and therefore unmasked) repeats was addressed by blasting the 40 bp flanking each SNP against the contig database. Sequences found to match with 100% sequence identity at any position other than their source were discarded. SNPs were then ranked into categories ($n = 3$) according to the genotypes of the individual diploid samples sequenced. Category-1 corresponded to SNPs for which there was evidence of homozygous and heterozygous (i.e. AA, BB and AB) allelotypes, category-2 to those with the occurrence of both homozygote allelotypes (AA and BB), and category-3 to those with heterozygote and one homozygote (AB and AA or BB).

Because category-1 SNPs and SNPs with higher read-depth displayed the greatest proportion of polymorphic loci (data not shown), these SNPs were preferentially selected for the array. SNP-assay designability was assessed using Illumina's online Assay Design Tool (ADT; www.illumina.com) with a required minimum ADT score of 0.7. The final array order included 7021 SNPs originating from GCR, 9240 SNPs from the alignment of EST reads from both European and Canadian material, 58 SNPs from public databases, 169 SNPs detected through BAC-end sequencing (Lorenz *et al.* 2010) and 63 SNPs detected by mitochondrial DNA resequencing (Karlsson *et al.* 2010). More details regarding the array development strategy and SNP discovery can be found in supplementary methods (Appendix S1).

DNA samples

A total of 1431 extracted DNA samples sampled between 1977 and 2008 from 38 sample sites (31 from Europe and seven from North America, between 20 and 72 samples per location) were included in this study (Fig. 1; Table 1), thus covering the natural geographical distribution of Atlantic salmon. Five locations in Europe and one in North America represented landlocked populations, while other populations were anadromous.
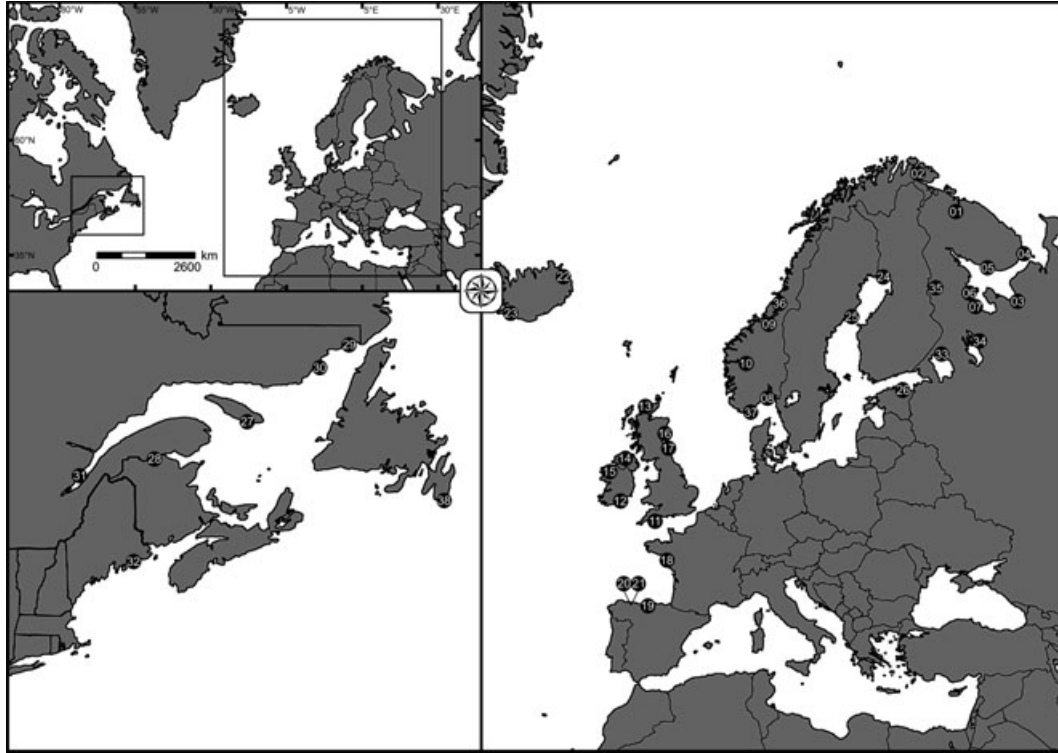
Fig. 1 Map showing sample sites in Europe and North America. Populations are linked to the numbers in Table 1.

Genotyping and quality control

Genotyping was performed according to the manufacturer's instructions using the Illumina infinium assay (Illumina, San Diego, CA, USA). The assay conversion rate was 92% with the final array containing 15 225 SNP assays. Using Illumina's Genotyping Module software, it was possible to examine each SNPs cluster pattern using data from a set of pedigree samples ($n = 3297$) provided by a Norwegian aquaculture programme (Aqua Gen AS, Norway) and data from this study ($n = 1430$). Visual inspection allowed for the classification of SNPs into different categories [i.e. single-locus SNPs, and the more complex paralogous sequence variants (PSVs), and multisite variants (MSVs), arising from genome duplication]. Classification was objectively supported using the program described by Gidskehaug *et al.* (2011) and is reported in supplementary methods of Lien *et al.* (2011). Following this (see also Results), a total of 6176 SNP markers were retained in the subsequent analyses of the 1431 population samples, 55% of these were derived from EST, 43% from GCR, and the remaining 2% from the other SNP sources. Because European fish were used in GCR and rGCR, we compared mean observed heterozygosity in both EST and GCR SNPs for European and North American populations to assess potential ascertainment bias. We also compared pairwise population measure of differentiation (F_{ST}) for both types of markers. Moreover, within each identified regional genetic groups, we pooled the per population minor allele frequency distribution across loci and compared each regional pooled distribution to each other using the Kolmogorov–Smirnov test in order to further evaluate potential bias in minor allele distribution. Note that we systematically excluded within-population monomorphic markers from the distributions.

Population structure and differentiation

We measured global and per SNP observed and expected heterozygosity within each population. Pairwise genetic differentiation between populations was estimated by θ (Weir & Cockerham 1984) using ARLEQUIN 3.5 (Excoffier & Lischer 2010) with 10 000 permutations to determine statistical significance. Based on previous mtDNA analysis (King *et al.* 2007) and intrinsic regional life history characteristics (e.g. migration, winter feeding areas), we determined the regional clustering of populations to avoid confounding hierarchical stratification in subsequent analyses. Thus, regional grouping among European populations was assessed first by calculating pairwise genetic distances using the D_A distance (Nei 1977). The resulting genetic distance

Table 1 Description of regional groupings and parameters associated with sample sites composing the groups: latitude and longitude, number of individuals genotyped (N_{GEN}), number of individuals with call rate superior to 0.85 ($N_{>85}$), average call rate per population (CR) and average expected (H_E) and observed (H_O) heterozygosities per population

Regional groups			Population ID	Code	Number	Years	Latitude	Longitude	N_{GEN}	$N_{>85}$	CR	H_E	H_O
Anadromous	Europe	Barents-White	Tuloma	TUL	1	2000	68.88	33.00	40	40	0.980	0.290	0.286
			Tana	TAN	2	1989 & 1995	70.50	28.38	55	29	0.929	0.296	0.303
			Emtsa	EMT	3	2001	64.57	40.48	40	40	0.986	0.255	0.249
			Lebyazhya	LEB	4	2001	66.95	41.28	40	40	0.982	0.275	0.275
			Yapoma	YAP	5	2000	66.27	36.85	40	40	0.985	0.263	0.261
			Pongoma	PON	6	1999	65.00	34.63	40	40	0.984	0.253	0.257
			Suma	SUM	7	2001	64.28	35.40	40	40	0.985	0.219	0.230
		Atlantic	Numedalslågen	NUM	8	1989	59.04	10.05	50	43	0.934	0.294	0.302
			Gaula	GAU	9	1989–1990	63.34	10.23	43	43	0.964	0.303	0.302
			Lærdalselva	LAR	10	1977–1997	61.10	7.47	72	25	0.968	0.297	0.298
			Dart	DAR	11	2005	50.40	−3.63	40	35	0.978	0.285	0.284
			Blackwater	BLW	12	2006	51.94	−7.84	40	40	0.989	0.287	0.289
			Dionard	DIO	13	2006	58.52	−4.80	40	40	0.984	0.287	0.287
			Foyle	FOY	14	2006	55.05	−7.25	40	40	0.990	0.276	0.273
			Moy	MOY	15	2004	54.10	−9.15	40	40	0.988	0.276	0.274
			North Esk	NES	16	2005	56.75	−2.42	40	40	0.989	0.288	0.289
			Tweed	TWE	17	1996	55.75	−1.98	40	37	0.981	0.284	0.283
			Loire	LOI	18	2006–2007	47.27	−2.18	40	39	0.985	0.238	0.239
			Cares	CAR	19	2004	43.39	−4.51	40	40	0.982	0.199	0.268
			Piguena	PIG	20	2004	43.57	−6.08	40	20	0.983	0.238	0.241
			Narcea	NAC	21	2004	43.57	−6.08	20	20	0.970	0.219	0.265
			Selá	SEL	22	2004–2005	65.83	−14.81	32	32	0.988	0.236	0.238
			Ölfusá	OLF	23	2004–2005	63.93	−21.23	32	32	0.991	0.226	0.228
		Baltic	Tornionjoki	TOR	24	1997	65.81	24.15	40	40	0.988	0.215	0.213
			Vindelälven	VIN	25	2005	63.75	20.32	40	40	0.984	0.212	0.211
			Kunda	KUN	26	2005	59.52	26.53	40	40	0.990	0.195	0.198
	North America	Anticosti	Chaloupe	CHA	27	2004	49.14	−62.54	23	23	0.971	0.166	0.165
		Southern Québec	Matapédia	MAP	28	2004	47.97	−66.93	25	25	0.967	0.170	0.167
		Labrador	Saint-Paul	STP	29	2004	51.45	−57.70	25	25	0.982	0.155	0.151
		Lower North Shore	Gros Mécatina	MEC	30	2004	50.77	−59.09	25	25	0.981	0.139	0.138
		Québec city	Du Gouffre	DGO	31	2004	47.43	−70.48	25	25	0.982	0.148	0.148
		USA	Narraguagus	NAR	32	2001–2004	44.60	−67.92	25	25	0.978	0.123	0.123
Landlocked	Europe	Baltic	Sysky	LL_SYS	33	1999	61.65	31.27	32	32	0.992	0.165	0.170
			Pyalma	LL_PYA	34	2004	62.40	35.87	40	40	0.983	0.139	0.141
		Barents-White	Pisto	LL_PIS	35	1999	65.26	30.56	40	40	0.988	0.144	0.148
		Atlantic	Namsen-Småblank	LL_NAM	36	2005–2008	64.46	11.52	46	40	0.985	0.119	0.113
			Otra-Byglandsbleka	LL_BYG	37	2004	58.14	8.01	40	40	0.988	0.103	0.104
	North America	Newfoundland	Bristol Cove	LL_BCR	38	1997	46.63	−53.19	41	40	0.975	0.145	0.146

matrix was used to construct a neighbour-joining (N-J) phylogram, and confidence estimates on tree topology were obtained by re-sampling over loci with 1000 bootstrap replicates. The genetic distance estimation and bootstrapping procedures were carried out using POWER-MARKER (Liu & Muse 2005). Based on the N-J tree (Fig. 2), European wild populations were clustered in three regional groups: Atlantic, Baltic Sea and Barents + White Seas (hereafter referred as Atlantic, Baltic and Barents–White). The composition of each subcontinental group is presented in Table 1. To confirm the relevance of hierarchical groupings (intercontinental and between regions within Europe), analyses of molecular variance (AMOVAS) were performed using ARLEQUIN 3.5. Seven North American populations were included mainly for estimating the overall divergence between continents, each of these having been chosen to represent one of the seven regional genetic groups defined by Dionne *et al.* (2008). A principal component approach was also used to identify the axes of greatest genetic differentiation between all populations. The principal components were constructed by decomposing the genomic relationship matrix among all individuals where the genomic relationship matrix was defined as in Yang *et al.* (2010). Principal components were fitted using R (R core development team).

Fig. 2 Genetic relationships in Atlantic salmon as resolved by neighbour-joining tree constructed using *DA* distance. Population code and numbers are as in Table 1 (populations with LL are landlocked). Nodes marked with a dot were supported by bootstrap support of <70% of 1000 replicates, and others are supported by more than 70% of 1000 replicates.

Outlier markers detection

To identify the most divergent markers among populations, which may potentially comprise loci under divergent selection (but see Bierne *et al.* 2011), we used hierarchical Fdist (Excoffier *et al.* 2009), a genome scan analysis implemented in ARLEQUIN 3.5 (Excoffier & Lischer 2010). The finite island model that was used in Fdist has been shown to lead to large fraction of false positives in some systems; therefore, ARLEQUIN 3.5 proposes to use a hierarchical island model where migration rates among groups are different than migration rates among populations within groups. Thus, this hierarchical method can detect outlier loci among groups of populations (F_{CT}). In cases where no hierarchical structure was present, outliers were detected using F_{ST}, and therefore, the method was identical to the Fdist test of Beaumont & Nichols (1996).

Although regional groups were represented in North American samples, a regular Fdist was used because only one population per group was genotyped. Overall, five hierarchical tests were carried out, including an intercontinental comparison between anadromous European and North American populations, and one intra-Europe test contrasting the three regional groups defined (see Results section). Three additional hierarchical tests were performed to contrast landlocked and anadromous populations within each group (respectively, Atlantic anadromous: NUM, GAU and LAR; and landlocked: LL_NAM and LL_BYG; Baltic anadromous: KUN, VIN and TOR; and landlocked: LL_SYS and LL_PYA; and Barents anadromous: YAP, LEB, PON, EMT and SUM; and landlocked: LL_PIS) (See Table 1 for abbreviations definition). Hierarchical and nonhierarchical Fdist were run three times each in ARLEQUIN 3.5. Only loci detected as outliers at the significance level of 0.01 in all three runs were reported as potentially under the effect of selection. Outliers from each test were compared to examine possible parallelism between continents or between landlocked vs. anadromous populations from different regions. Evidence for parallelism would strengthen support for interpreting putative outliers as being under the effect of divergent selection.

Candidate genomic regions affected by selection

As an alternative to single-locus outlier tests, for European populations, we combined Atlantic salmon genetic map information (Lien *et al.* 2011) with a kernel-smoothing moving average approach (Hohenlohe *et al.* 2010) to generate genome-wide distributions of the divergence estimates (measured F_{CT}) from hierarchical and nonhierarchical Fdist (Excoffier *et al.* 2009). Only SNPs mapped

by Lien *et al.* (2011) were used in this analysis. To identify genomic regions with an unexpectedly high proportion of SNPs showing increased or decreased divergence indicative of divergent and balancing selection, respectively, we performed 30 000–1 000 000 permutations to estimate local *P*-values. We tested multiple smoothing parameter values and selected 2 cm bandwidth that identified relatively narrow genomic regions of interest while being large enough to reduce sampling variance. For linkage groups (LGs) with relatively low SNP coverage, larger bandwidth (SSA08 = 5 cm; SSA26 = 4 cm; and SSA29 = 3 cm) was used. Despite the fact that the choice of bandwidth can have a strong effect on kernel density estimation, different smoothing parameters did not change the position of the major peaks (data not shown).

Gene ontology and SNP annotation

Blast2go (Gotz *et al.* 2008) was used to associate gene ontology (GO) annotation terms with all 6176 SNPs. Homology searching was first realized through a BLAST search of the available flanking sequences for each SNP on the NCBI nr public database with the *e*-value threshold set to 1×10^{-10}. Blast2go then retrieved GO terms associated with the obtained BLAST hits. The output GO annotation was then classified in multilevel biological processes, molecular functions and cellular components from the most general (level 2) terms of each category to more specific (upper levels) terms. In order to determine whether the biological processes, molecular functions or cellular components of the outliers herein identified were over-, equally or under-represented when compared to the 6176 analysed SNPs, we performed an enrichment analysis using Fisher's exact test corrected for multiple tests by applying a false discovery rate (FDR) (Benjamini & Hochberg 1995).

Clinal variation among outliers

Because latitude is integrative of many correlated environmental variables along a South–North coastal line, we tested for the presence of latitudinal clines in allele frequency at outlier SNPs among European populations. Given that the allele frequency of many outlier markers (34 of 52) showed significant correlations with latitude (see Results), we visually compared the cline pattern at those markers with the 18 nonsignificant ones. When excluding the Barents–White populations from the regression analysis, we observed that the latter markers presented more significant cline patterns driven by shifts in Baltic populations. Therefore, instead of using latitude, we tested for clinal allelic frequency distribution using generalized linear models of coastal distances along two different continua: (i) from northeastern

populations to southern populations excluding the Baltic populations and (ii) using one of the Baltic population as the point of origin and measuring the distance away from this population excluding populations from the Barents–White group. For each continuum, respectively called the White–Barents and Baltic continua hereafter, we compared the outliers' patterns with 52 randomly selected markers among the assumedly neutral (nonoutlier) markers (excluding European monomorphic markers). Finally, we used CFIT-6 (Gay *et al.* 2008) to test whether the observed clines presented a common centre or width. We thus compared four models: (i) no constraint, (ii) centre constrained, slope not constrained, (iii) centre not constrained, slope constrained and (iv) centre and slope constrained. Finally, outliers were positioned using the genetic map information from Lien *et al.* (2011).

Results

Genotyping and quality control

After initial quality control and classification of genotypes obtained from 1431 samples (Table 1; Fig. 1 for site locations), we classified (i) 5436 markers of 15 225 SNP features on the array as single-locus and polymorphic SNPs (i.e. diploid SNPs), (ii) 1725 markers as being represented at two homologous loci and polymorphic (i.e. MSV), (iii) 930 markers being represented at two homologous loci but displaying no polymorphism (i.e. PSV), (iv) 1853 failed assays and (v) 5281 nonpolymorphic markers (i–iv detailed in Lien *et al.* 2011). Mitochondrial SNPs were excluded from further analysis. Among diploid and MSV SNPs, 6112 markers showed an overall minor allele frequency (MAF) over 0.01. These, along with 64 other markers with an overall MAF < 0.01, but with a MAF > 0.05 in at least one population, made for a total of 6176 SNPs used in all analysis except stated otherwise. From the initial 1431 samples, 95 were discarded as they generated an average call rate <0.85 (proportion of SNPs genotyped), which was chosen as a threshold differentiating acceptable and unacceptable genotyping data. Table 1 shows summary data for call rates (CR), observed heterozygosity (H_O) and expected heterozygosity (H_E) across populations. A significantly higher overall observed heterozygosity (Wilcoxon's rank sum test $P < 0.001$) for European compared to North American populations (0.241 and 0.148, respectively) indicated an ascertainment bias towards European populations or alternatively a naturally lower diversity in North American populations. This was anticipated as a large proportion of the SNPs on the array were detected by sequencing European aquaculture strains. GCR-derived

markers were 2.4 times more heterozygous among European populations than among North Americans (average observed heterozygosity = 0.290 vs. 0.122). However, it should be noted that a few GCR markers were monomorphic across European populations. EST-derived markers, which should be less affected by ascertainment bias, were also more heterozygous in European samples albeit with a smaller difference (0.203 vs. 0.168) (Table S1, Supporting information). Pairwise F_{ST} values were not significantly different between GCR- and EST-derived markers for comparisons of populations within Europe or within North American (respective Wilcoxon's tests $P = 0.640$ and 0.881), but significantly different for intercontinental comparisons (mean F_{ST}: GCR = 0.463 and EST = 0.603; Wilcoxon's $P < 0.001$). According to the Kolmogorov–Smirnov tests, within-region pooled population distribution of MAF showed that the Atlantic group had a significantly different distribution compared to all other groups ($P < 0.005$). The Baltic, North American and landlocked groups were similar to each other in MAF distribution. The Barents–White group presented a significantly different distribution from all other groups, while intermediate between the Atlantic groups and the others (Fig. 3).

Genetic diversity and population differentiation

Within-population observed and expected heterozygosities varied across populations and markers (Tables 1 and S2, Supporting information). North American populations showed lower observed heterozygosity (range: 0.123 NAR to 0.167 MAP) than European populations (range: 0.198 KUN to 0.303 TAN), while landlocked populations demonstrated the lowest values (0.104 LL_BYG to 0.170 LL_SYS). All samples showed similar observed and expected heterozygosities except for the Spanish Narcea population (NAC), which had higher observed heterozygosity than expected, probably due to stocking of genetically divergent non-native fish from northern European rivers (Ayllon *et al.* 2006). All pairwise comparisons of genetic differentiation between populations were highly significant ($P < 0.001$, Table S3, Supporting information). Overall, F_{ST} varied between 0.011 (NES vs. DIO and NES vs. TWE) and 0.758 (NAR vs. LL_BYG). The within-Europe comparison varied between 0.011 and 0.347 (CAR vs. KUN), while within-North American F_{ST} varied between 0.048 (CHA vs. MAP) and 0.151 (NAR vs. STP). Intercontinental differentiations along with intra-Europe landlocked comparisons showed the highest F_{ST} with respective ranges of 0.439 (CHA vs. TUL and MAP vs. TUL) to 0.636 (CAR vs. NAR) and 0.320 (LL_PYA vs. LL_SYS) to 0.630 (LL_BYG vs. LL_NAM).

An AMOVA considering both continents as regional groups and another carried out at the intra-Europe level considering the clusters identified previously with the phylogram as regional groups (Fig. 2) showed significant genetic variation among groups (Table 2). In the intercontinental analysis, 41% of variation was found among continental groupings (Europe and North America), while within Europe, 7.55% of variation was found among regional groupings (Atlantic, Baltic and Barents–White). In a principal component analysis (PC) of the genomic relationships among individuals, seven PC factors individually determined at least 1% of variation and together explained 55.1% of the total genetic variation among populations. Principal components 1–4 accounted for 38.4%, 9%, 2.1% and 1.7%, respectively. PC1, PC2 and PC3 differentiated the European populations from the North American populations (Fig. 4a), while PC3 separated the populations within Europe almost along a northeast–southwest axis, with Spanish populations at one end and most Russian populations and Baltic populations at the other end (Fig. 4b).

Outlier markers detection

Six genome scans (Fig. 5) were performed using 6176 SNPs with a varying number of markers potentially under divergent (range between 26 and 139) and balancing selection (range between 20 and 101) at the 0.01 significance level (Table S4, Supporting information). The hierarchical genome scan performed over all anadromous populations revealed the highest number of loci potentially under divergent selection (139) while also yielding the lowest number of markers potentially under balancing selection (20) in spite of an average F_{CT} across loci of 0.311 (ranging from −0.055 to 0.997). European (mean $F_{CT} = 0.063$; ranging from −0.052 to 0.732) and North American (mean $F_{ST} = 0.081$; ranging from −0.024 to 0.780) genome scans showed similar numbers of potentially selected loci with, respectively, 52 and 51 divergent outliers and 61 and 46 markers under balancing selection. Four outliers were common to both European and North American genome scans. Three showed similar trends as two were detected as potentially under divergent selection (ESTNV_28701_207 and ESTNV_23580_687) and one under balancing selection (ESTNV_28516_389). However, one showed opposite trends (ESTNV_32552_113) as it was potentially under divergent selection in the European scan and under balancing selection in the North American scan.

All three hierarchical genome scans also revealed SNPs under potential divergent and balancing selection when anadromous populations were compared with landlocked populations from the same geographical
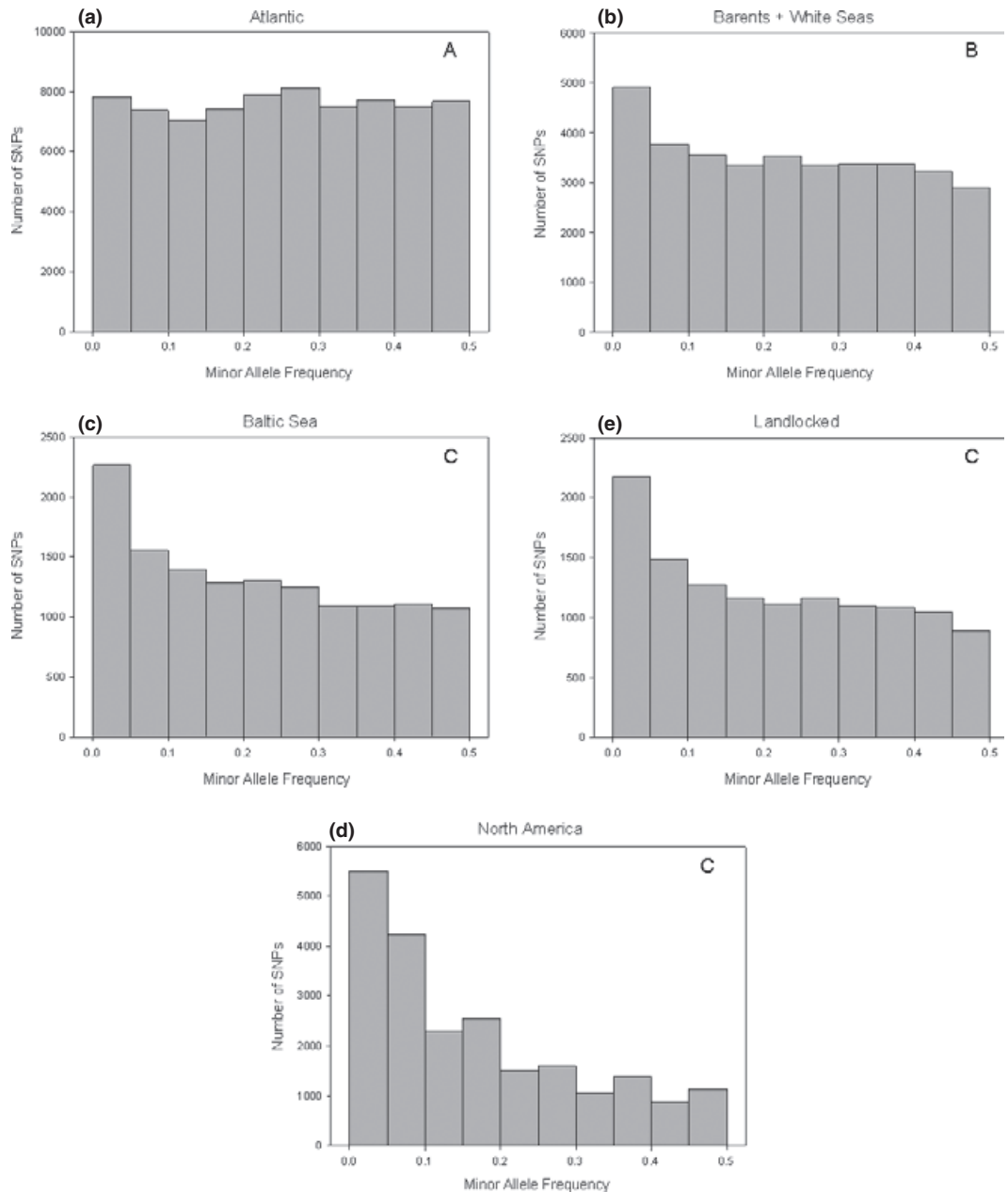
Fig. 3 Within-region pooled population distribution of minor allele frequency (MAF) for (a) Atlantic populations, (b) Barents–White populations, (c) Baltic populations, (d) North America populations and (e) landlocked populations. Panels sharing letters have distributions not significantly different according to the Kolmogorov–Smirnov test. Markers with MAF = 0 are not shown.

region (respectively, N divergent = 40, 49 and 40; N balancing = 31, 41 and 38; mean F_{CT} = 0.057, 0.060 and 0.100; maximal values 0.876, 0.883 and 0.983). Three outliers were common to landlocked vs. anadromous genome scans. Marker ESTNV_33891_846 was found to be under balancing selection among Barents–White and Baltic populations, GCR_cBin31530_Ctg1_84 under balancing selection among Baltic populations while under divergent selection in the Atlantic populations and

GCR_cBin14325_Ctg1_429 was under divergent selection in Barents–White and Atlantic populations. None were common to all three genome scans.

Candidate genomic regions affected by selection

To further gain insight into the chromosomal patterns of divergence (measured as F_{CT}) along the mapped linkage groups, we identified genomic regions showing
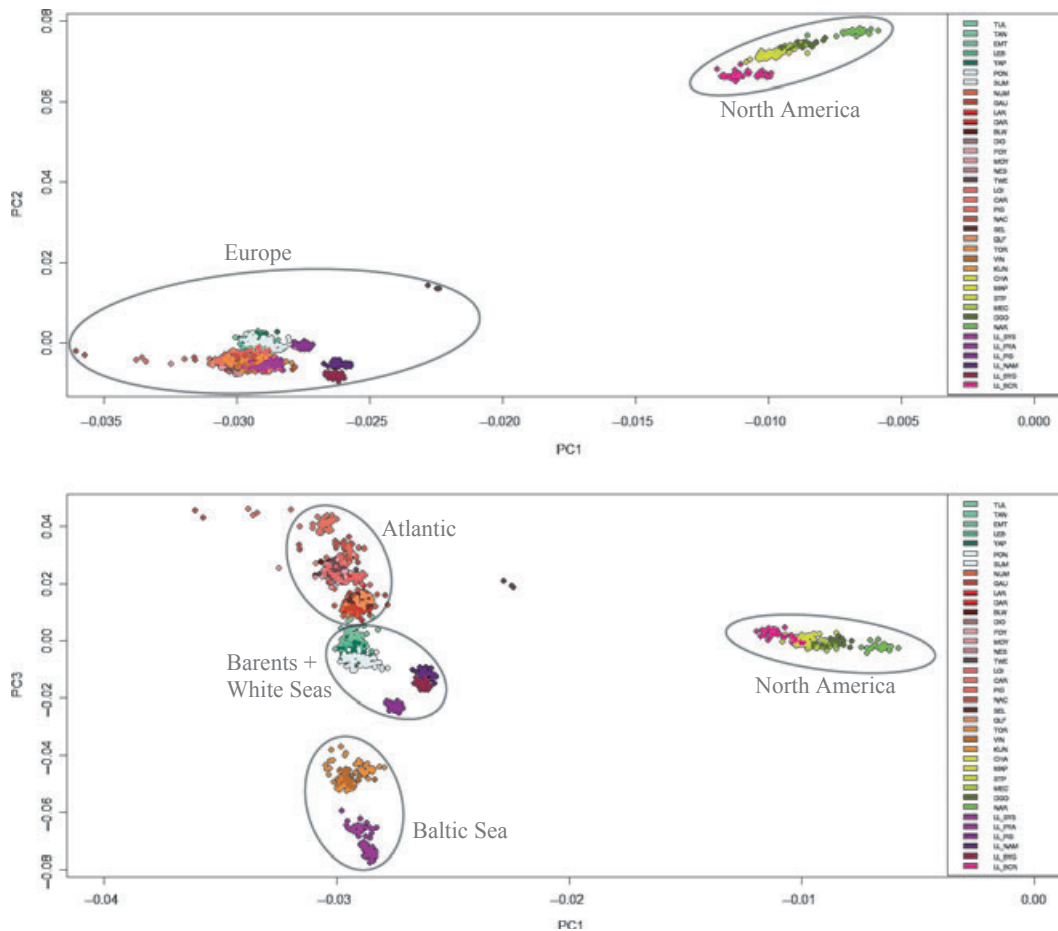
Fig. 4 Principal components analysis of genetic differentiation among populations based on 6176 single nucleotide polymorphism markers (each point represents one individual) with (a) principal component 1 (PC1: 38.4% of variance) against PC2 (9.0% of variance); (b) PC1 against PC3 (2.1% of variance). Each population shows a different colour, and colour grading relates in part to geographical similarity.

unexpectedly high or low divergence between different European regional groups. Altogether, 19 genomic regions showed increased differentiation ($P < 0.01$) in

Table 2 Analysis of molecular variance (AMOVA) in two hierarchical groupings: (a) groups defined as North America and Europe and (b) groups defined as regional groups intra-Europe as identified in Table 1

Source of variation	d.f.	Percentage of variation
(a) Continental groups		
Among groups	1	41.01*
Among populations within groups	30	8.84*
Within populations	2174	50.15*
(b) European groups		
Among groups	2	7.55*
Among populations within groups	23	10.39*
Within populations	1884	82.06*

*P-value < 0.001.

European comparisons (Fig. 6a; Table S5, Supporting information). A slightly larger number of genomic regions ($n = 25$) exhibited reduced levels of differentiation ($P < 0.01$) (Fig. 6a; Table S5, Supporting information). When compared with the single-locus outlier tests, both analyses often revealed similar candidate regions potentially under divergent selection (Figs S1–S29, Supporting information e.g. LG: SSA01, SSA05, SSA10, SSA12, SSA15-SSA17, SSA18, SSA20). The strongest evidence ($P < 10^{-5}$) for increased or reduced differentiation was found in LG SSA08 and SSA16 (Fig. 6bc).

Annotation of outlier SNPs

The BLAST (Altschul *et al.* 1990) yielded 2691 SNPs with significant hits (e-value < 1×10^{-10}, Table S6, Supporting information). From these results, a total of 17 701 GO terms were associated with SNPs. GO terms segregated into many levels of biological processes,
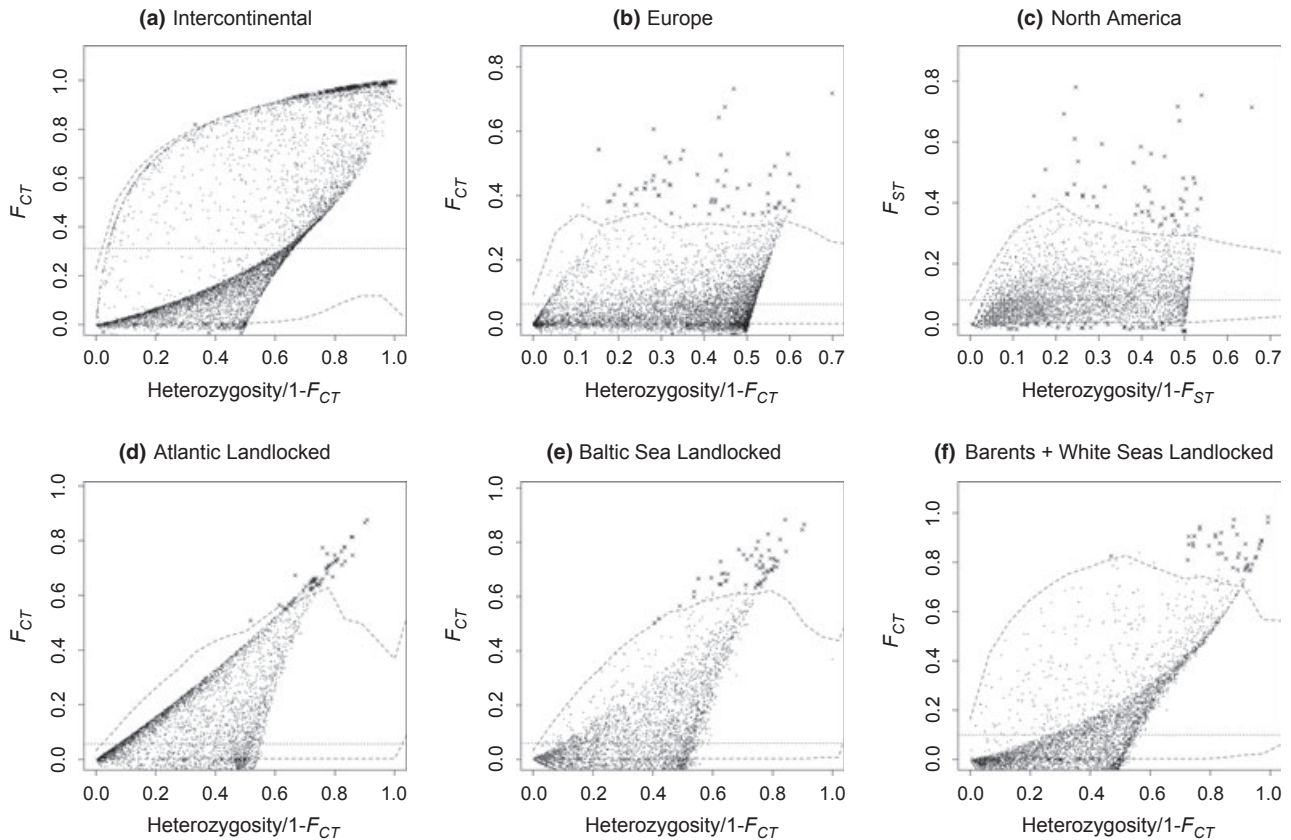
Fig. 5 Differentiation (F_{CT} or F_{ST}) as a function of *heterozygosity/1-differentiation* as calculated by ARLEQUIN 3.5 for 6 genome scans: (a) comparison between European and North American anadromous populations (hierarchical test), (b) comparison between three regional groups (hierarchical test), (c) comparison among North American anadromous populations (nonhierarchical test), (d–f) comparison of landlocked vs. anadromous populations within each of the three European groupings (Atlantic, Baltic and Barents–White) (hierarchical test). In each panel, outliers markers ($P < 0.01$) are marked by X, dashed lines represent upper and lower 99% confidence level, and dotted line indicates the average F_{CT} or F_{ST} across loci.

molecular functions and cellular components with a mean level of 5.819. Overall, 19-level 2 biological pathways associated with cellular process, metabolic process and biological regulation were highlighted, and together, these represented 48% of GO annotations. After correction for multiple testing, an enrichment analysis across both categories of outliers (divergent and balancing) did not indicate significant over- or under-representation of any biological pathway in the outliers identified. Among common outliers in European vs. North American comparison, ESTNV_28701_207 (potentially under divergent selection) was associated with a B-cell receptor (CD22-like), which prevents overactivation of the immune system (Hatta *et al.* 1999), while marker ESTNV_32552_113 (divergent in Europe, balancing in North America) has a close relationship with a calcium ion–binding protein. Common outliers among anadromous and landlocked populations revealed one marker, ESTNV_33891_846 (under balancing selection in Barents–White and Baltic), which has an association with coagulation factor V.

Clinal variation among outliers

Among the 52 outliers potentially under divergent selection among all anadromous European populations, the population allelic frequencies of 34 markers were significantly correlated with latitude (R^2 ranging from 0.177 to 0.498, mean = 0.308). The generalized linear models applied to each of two different continuums, one starting in White Sea (excluding the Baltic) and the other in the Baltic Sea, revealed that 23 and 18 outliers present only in the White–Barents and Baltic continua, respectively, and 11 outliers showing clinal patterns common to both continua. Therefore, all 52 markers showed a clinal pattern for one or both continua. The congruence of outlier markers was striking when contrasted with randomly selected neutral markers (Fig. 7). In both continua, the model with no constraint was always the best one (Table S7, Supporting information), meaning that there was no common centre or width associated with observed clines. However, the average
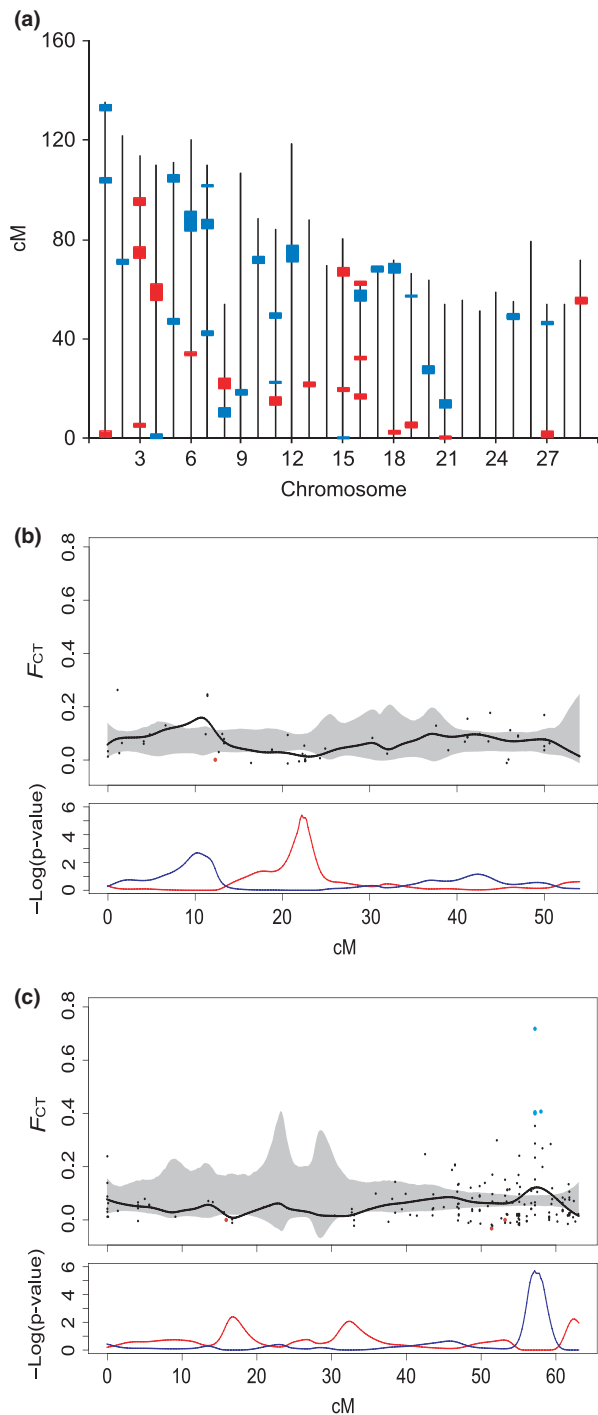
Fig. 6 Genetic linkage map showing the distribution of genomic regions showing increased or decreased divergence (P < 0.01) marked as blue and red bars, respectively, in (a) European populations. The distribution of F_{CT} along the linkage groups: (b) SSA08 and (c) SSA16. Grey areas correspond to 95% confidence interval of the smoothed kernel curve (black line). Individual outlier loci showing increased or decreased divergence are marked as blue and red bars. Estimated significance levels (−log10-transformed *P*-value) for region of increased or decreased divergence are shown as blue and red curves.

localizations of the centres were 1092 and 1009 km from the point of origin for the White Sea and Baltic continua, respectively. These averages closely coincide with the boundaries between the Barents Sea vs. Atlantic and Baltic Sea vs. Atlantic. This also closely matches the genetic boundaries inferred from the three genetic groupings defined earlier. Finally, the outliers showed a random distribution on the genetic map with no apparent clustering on specific linkage groups (Fig. 8).

Discussion

This study represents the single most extensive population genetic study on Atlantic salmon performed to date. With 1360 individuals successfully genotyped for over 6000 SNP markers, it also stands as one of the most comprehensive population genetics studies of wild populations in a nonmodel species. An important outcome of our study, particularly for population management, is the robust confirmation of three major regional genetic groups occurring in Europe. Moreover, based on strong clinal patterns observed for identified outlier markers, we propose that these groups are bordered by secondary contact zones where highly divergent markers are associated with endogenous (intrinsic genetic incompatibilities) and possibly exogenous (environmental or ecological) barriers (Bierne *et al.* 2011).

Ascertainment bias

Among the 6176 SNPs retained on the SNP array, 3383 were detected from available EST reads from both European and Canadian materials. On the other hand, the GCR markers were discovered using only Norwegian commercial aquaculture fish, meaning that these markers are potentially more prone to an eastern Atlantic European bias when used for genotyping more geographically dispersed samples. The sampling bias towards the Atlantic populations and to a lesser extent in the Barents–White groups could potentially bias our differentiation estimates. Indeed, Albrechtsen *et al.* (2010) observed a small upward bias in F_{ST} estimates using ascertained populations. They also showed that the intensity of the bias depends on the genetic distance of the populations being compared relative to the population within which the markers were originally developed, suggesting that F_{ST}, when two populations outside of the ascertained regional group are compared to each other, may be less affected by ascertainment bias.

A significantly higher genetic diversity (H_O) in European samples was observed for both EST- and GCR-based markers (Table S1, Supporting information). Contrasting estimates of genetic diversity between the continents are partly explained by the ascertainment
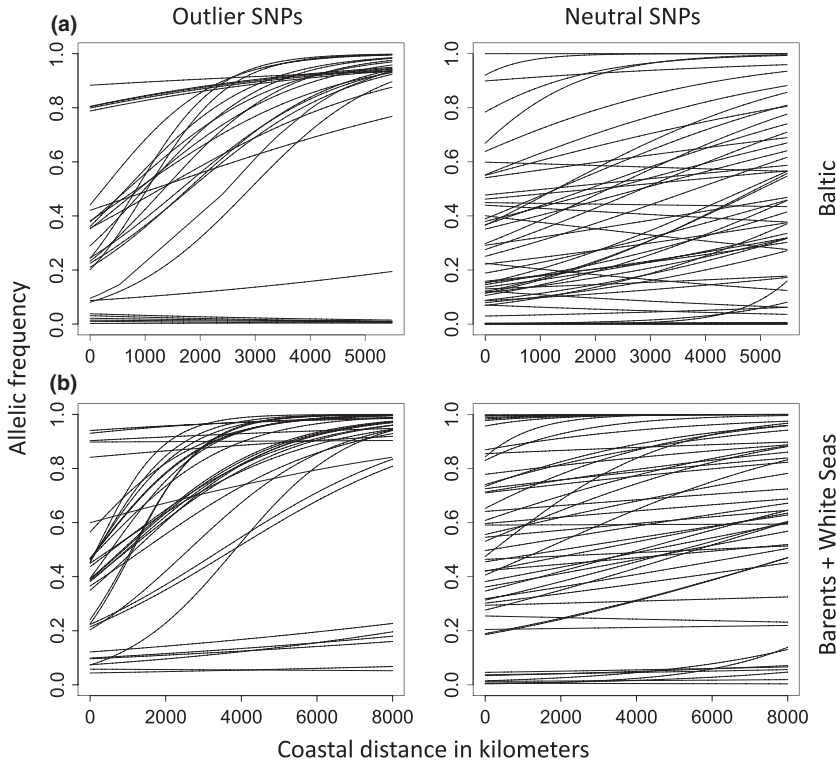
Fig. 7 Generalized linear models illustrating latitudinal clines (or lack thereof) for 52 European outliers (left panels) and 52 randomly selected neutral (nonoutlier) markers (right panels) shown along two different coastal distance continuum: (a) the Baltic and (b) the White–Barents.
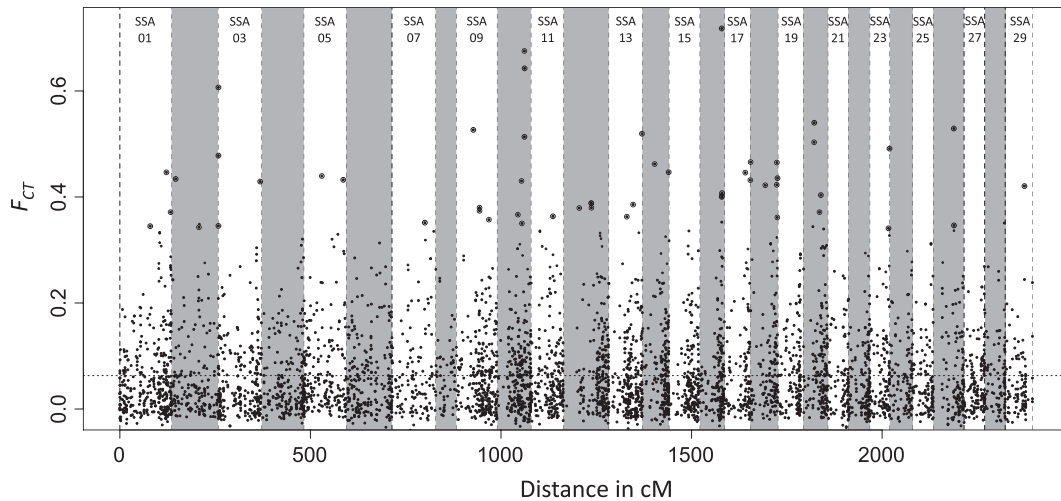

Fig. 8 European regional differentiation (F_{CT}) of each single nucleotide polymorphism marker mapped on the European Atlantic salmon genetic map. Grey and white rectangles separated by vertical dashed lines represent separate linkage groups (named SSA–). Large black dots indicate outlier markers, and the horizontal dotted line indicates the average F_{CT} among markers (0.063).

bias and are of concern for estimating and interpreting other genetic parameters. Indeed, we found that intercontinental population differentiations are significantly higher when measured with EST-based markers than with GCR-based markers. Therefore, this bias should be taken into consideration in any intercontinental comparisons using the array. That said, it is noteworthy that in a survey of microsatellite diversity, King *et al.* (2001)

also found that genetic diversity was reduced in North American populations compared with European populations using markers developed with material from both origins. It is thus possible that different demographic history partly explains the overall differences in genetic diversity between continents.

In contrast, no significant difference in intracontinental population differentiation was observed when estimates

were calculated with EST and GCR markers separately (data not shown). However, within Europe, genetic diversity is also regionally heterogeneous with lower genetic diversity associated with Baltic and landlocked populations. Here again, this regional pattern of variation was observed in previous studies using microsatellites and was attributed to the phylogeographic history of these populations (Säisä *et al.* 2005; Tonteri *et al.* 2007). Overall then, it appears that variation in genetic diversity between regions results from both historical contingency and ascertainment bias. We thus advise that absolute values in intercontinental comparisons should be interpreted carefully. However, the bias effect is less of a concern for within continental comparisons, especially for interpreting broad-scale patterns of differentiation.

Population structure

While Atlantic salmon population structure has been extensively investigated (Ståhl 1987; King *et al.* 2001; Verspoor *et al.* 2005; Dionne *et al.* 2008), this study is the first to examine the structure of population over its entire range with extensive genomic coverage in a single original study. Overall, the results provided by the SNP markers show similar patterns as those previously observed, but the level of differentiation is amplified with F_{ST} values up to twice as high as previously reported, may be due to the lower level of polymorphism in SNPs compared to microsatellites (Hedrick 1999). As an example, a within-continent comparison of pairwise F_{ST}, Dionne *et al.* (2008) reported a F_{ST} value of 0.048 based on 12 microsatellites, compared to the value of 0.112 we observed with SNPs for the same comparisons. This enhanced structure definition is observed on every scale examined. Indeed, the range-wide analysis suggested that 41% of the genetic variation occurred between the continents, which contrasts with the 21.9% reported by King *et al.* (2001). Although we did not observe any continent-specific alleles as previously observed for microsatellite markers (King *et al.* 2001; Wennevik *et al.* 2004), we did identify over 100 SNP markers with $F_{CT} > 0.95$, that is, nearly diagnostic between continents.

Following the pioneering study of Stahl (1987) who first identified two distinct genetic groups of Atlantic salmon within Europe, namely Eastern Atlantic and Baltic Sea, many studies subsequently focused on the colonization history of northern Europe following the last ice age and its impact on regional genetic groupings of Atlantic salmon populations (Kazakov & Titov 1991; Skaala *et al.* 1998; Koljonen 2001; Wennevik *et al.* 2004; Makhrov *et al.* 2005; Saisa *et al.* 2005; Tonteri *et al.* 2005). Most of these studies were congruent in differen-

tiating the Baltic Sea populations and regrouping other populations into an Atlantic + Barents Sea + White Sea group. Verspoor *et al.* (2005) reported that more groups could be defined in Europe, especially within the previously named Atlantic group. More recently, Tonteri *et al.* (2009) demonstrated the genetic distinctiveness of the Eastern Barents Sea and White Sea populations from the Western Barents and Eastern Atlantic ones. Our results confirm this partitioning of northern populations and further suggest the contribution of at least three glacial refugia in colonizing European populations. Essentially, we observed three major clusters corresponding to an Atlantic group, a Baltic Sea group and a Barents–White Seas group. A relatively high proportion (7.55%) of the genetic variation occurred among these groups, although a higher proportion (10.4%) was also attributed to differences among populations within each of them. These three groups are consistent with the main postglacial colonization routes previously identified as the West Atlantic, the Baltic Ice Lake and Eastern Barents Sea (Tonteri *et al.* 2007, 2009; and references therein).

Landlocked populations exhibited a structure also consistent with previous studies. However, this study allowed a finer understanding regarding the origins of these populations from anadromous ancestors. Tonteri *et al.* (2005) suggested that freshwater populations from the Baltic and White Sea basins originated from different populations, explaining their segregation into two different clusters based on 14 microsatellite markers. Here, we observed approximately 1.5 times more differentiation when contrasting landlocked populations from these two basins with any anadromous populations of their respective basins. Additionally, freshwater Baltic populations clearly clustered with anadromous Baltic populations, while the nonanadromous population from the White Sea basin clustered with the Barents–White group. Based on their geographical location, Norwegian populations were expected to cluster as Atlantic populations, but showed a high degree of divergence from all three regional genetic groups. Because their closest neighbours on the tree are Norwegian or Western European populations, it is highly possible that their differentiation from the Atlantic group stems from a very pronounced drift effect resulting from a founder event, which could have occurred during land upheaval following the last deglaciation in Norway (Berg 1985).

Signatures of selection

The detection of loci under selection using different genetic differentiation methods shares several known caveats, particularly the potential for false positives with such a high degree of multiple testing (Foll &

Gaggiotti 2008). Nonetheless, these methods have been successfully used for identifying candidate genes and QTL for local adaptation to environmental conditions (Schmidt *et al.* 2008; Storz & Wheat 2010; Gagnaire *et al.* 2012b), localizing genomic regions under adaptive divergence (Rogers & Bernatchez 2007; Flori *et al.* 2009; Hohenlohe *et al.* 2010; Gagnaire *et al.* 2012a) or identifying markers to investigate introgression (Hohenlohe *et al.* 2011; Karlsson *et al.* 2011; Lamaze *et al.* 2012). While a main objective of this study was to assess the extent of potential adaptive divergence among populations throughout the native range of Atlantic salmon, we were not able, in some comparisons, to distinguish the effect of selection from that of pronounced drift. In other comparisons, using multiple populations subject to the same putative selection reduced the chance that the outliers identified were mainly the result of drift.

Markers identified as outliers when contrasting genetic variation on both continents genetic compositions were almost differentially fixed within each continent. Given the estimated time of 600 000 years since continental interdivergence (King *et al.* 2007) and very modest contemporary migration rates between the two continental lineages, we were expecting many more strictly differentially fixed markers than were observed. Thus, the absence of fixed markers could mean that the migration rate between both lineages was previously underestimated. However, we cannot exclude that ascertainment bias, the method of SNP selection and the unbalanced coverage of genetic variation towards Europe might also contribute to this pattern.

Although environmental gradients occur along the geographical distribution of Atlantic salmon, contrasting the three major clusters identified failed to isolate markers that could be strictly under the influence of environmental selection. Instead, our results revealed that the majority of outlier markers showed allele frequencies that correlated with geographical clines (compared to nonoutlier markers) in both the Baltic vs. Atlantic and Barents–White vs. Atlantic comparisons. This could be a result of historical colonization and persisting tension zones following secondary contact between once geographically isolated lineages. We did not identify a common centre or width for these clinal outliers (Table S7, Supporting information). However, all markers showing clinal variation were strictly outliers, they showed an overall congruence in geographical patterns of variation (Fig. 7), and that they were randomly distributed in the genome (Figs 6 and 8). This strongly suggests that these markers are likely to reveal endogenous genetic barriers (genetic incompatibilities) occurring in zones of secondary contact between distinct evolutionary lineages, rather than genetic–environment associations (exogenous barriers). As demonstrated by Bierne *et al.* (2011), endogenous

genetic barriers can easily be coupled with exogenous genetic barriers associated with environmental gradients. Here, environmentally driven selection could be responsible for some of the observed clines since environmental heterogeneity, mainly in terms of temperature, and salinity also occurs along these zones. However, our data and analysis do not allow us to tease apart markers potentially under the influence of environmental selection vs. others. For instance, the Öresund and Danish Belts delimitate an important environmental discontinuity between the Baltic Sea and the Atlantic Coast characterized by an abrupt change in salinity (among other parameters) that have been identified as a selective agent in many species (Johannesson & Andre 2006; Gaggiotti *et al.* 2009; Limborg *et al.* 2009). Here, we found that this geographical region is associated with numerous outlier markers showing clinal variation. We do not challenge the environmental selection acting in this area and certainly acknowledge previous studies identifying this possible hotspot for genetic–environment associations. However, our results suggest that careful interpretation of outlier markers is needed, especially in known secondary contact zones. Namely, as it is likely the case for the Baltic, hybrid zone theory predicts that tension zones will be trapped by natural barriers (Barton 1979; Barton & Hewitt 1985; Hewitt 1988; Bierne *et al.* 2011). Should this be the case in the Barents Sea as well, characterization of the putative environmental barrier operating there might reveal possible functional targets of selection. Overall, we emphasize that identifying genomic regions underlying a true genetic–environment association first needs functional support and/or convincing rejection of historical contingency on top of detailed environmental characterization. Therefore, in the case of Atlantic salmon, we suggest that future research should focus on disentangling the interplay between historical and selective forces.

As for the North American populations analysed in this study, only cautious interpretations can be drawn given that results are still equivocal. Previous studies of Western Atlantic salmon provided evidence for a hierarchical genetic structure. However, depending on the markers used and geographical coverage, defined regional groups differed among studies (King *et al.* 2007 and references therein; Dionne *et al.* 2008). Interestingly, there is evidence that North America was colonized by at least two major refuges given the presence of European mtDNA haplotypes in many landlocked populations (Knox *et al.* 2002; King *et al.* 2007). Nonetheless, the regional structure of anadromous populations was never associated with postglacial colonization from multiple source populations, as in Europe. Instead, recent landscape genetics studies have shown that regional differences are likely driven by distinct environmental conditions, which would imply a regional scale of local adaptation (Dionne

et al. 2007, 2008), which may be more common in anadromous salmonids than previously thought (Fraser *et al.* 2011). In this context, we argue that identified outliers could represent or be associated with actual targets of environmental selection. However, the limited coverage of North American populations hampered the possibility to pinpoint any specific biological function or process among annotated outliers that could have established potentially functional targets of selection and thus useful candidate genes. Clearly, further investigation of adaptive divergence among North American populations will necessitate deeper sampling coverage within and among regional groups along with a detailed landscape genomics approach.

Contrasting landlocked versus anadromous populations in Europe identified markers that were strikingly differentiated between the two life history strategies. Such differences may indicate ecological selection and/ or random genetic difference due to genetic drift. None of the outliers that were detected showed parallel patterns of divergence among the three regional genome scans we performed. However, we found three markers identified as outliers in two of three scans. Although all three showed different patterns (divergent selection in one group, balancing selection in opposite directions in the other groups), and only one was successfully annotated, these exploratory scans suggest that there is sufficient divergence between landlocked and anadromous populations to justify further examination. We propose three main steps to undertake in a follow-up study that would aim specifically at elucidating the adaptive divergence between these populations. First, comparing regional replicates holds the key to distinguishing selection from drift. Second, a less stringent significance threshold could be used to detect outliers in order to allow more markers to be further examined and screened using ecological and functional context. Finally, particular attention should be paid to candidate genes involved in osmoregulation and immune-related functions that are possible targets for selection given the contrasting environmental conditions encountered by anadromous and landlocked populations. The upcoming genome sequence (Davidson *et al.* 2010) of Atlantic salmon promises to improve the functional context and further contribute to disentangling historical from adaptive divergence.

Implications and perspectives

In addition to the recent demographic decline of most anadromous populations, related concerns such as the relative genetic contribution of individual populations to fisheries or bycatch, translocation, reintroduction strategies and assessments of farmed escapees' impacts

are all management issues that could benefit from the improved genetic information derived from the SNP array. Information from the array allows finer, more precise definition of populations, management and ESU. In this study, using the SNP array certainly enhanced the resolution of such units by revealing increased level of differentiation estimates at every geographical scale relative to previous studies. Our study suggests that escapees from aquaculture (the majority of which stem from Norwegian aquaculture populations of the Atlantic group in Europe) could lead to different genetic consequences whether introgressing into wild populations of the same (Atlantic) or different (Barents–White) phylogeographical groupings. The array should contribute in resolving the continuing controversy surrounding the potential impacts and level of introgression of captive bred fish in wild populations. For instance, the SNP array developed here has recently been used by Karlsson *et al.* (2011) to develop a subpanel of markers that can discriminate between wild populations and the major strains of domestic Atlantic salmon used for farming in Europe, enabling the assessment of the impacts of farmed escapees on wild populations. Furthermore, linking genetic and environmental divergence could highlight biological processes that evolve under the effect of natural selection and identify the actual selective agents. Although further investigation is needed to delineate the relative contribution of adaptive divergence in observed patterns of genetic differentiation, the ubiquitous occurrence of highly divergent markers representing a diversity of biological functions holds the potential for adding information about the adaptive nature of divergence in defining significant units of management and conservation. Being aware of historical contingency effect on observed patterns of differentiation, we are now better suited to grasp the real contribution of environmental selection in shaping population divergence in Atlantic salmon. Overall then, this SNP array and subsequent versions of it should bring considerable benefits to Atlantic salmon management and conservation community as well as for aquaculture applications.

inputs to the manuscript and analyses. Research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to VB, from the Canadian Research Chair in genomics and conservation of aquatic resources to LB, the Academy of Finland to CRP and AV, Estonian Science Foundation to AV and the Beaufort Marine Research Award in Fish Population Genetics funded by the Irish Government under the Sea Change Programme to PMcG.

References

Aas Ø, Einum S, Klemetsen A, Skurdal J (2011) Atlantic Salmon Ecology. Wiley-Blackwell, Oxford.

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, **27**, 2534–2547.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Ayllon F, Martinez JL, Garcia-Vazquez E (2006) Loss of regional population structure in Atlantic salmon, *Salmo salar* L., following stocking. *Ices Journal of Marine Science*, **63**, 1269–1273.

Barton NH (1979) Dynamics of hybrid zones. *Heredity*, **43**, 341–359.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **263**, 1619–1626.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289–300.

Berg OK (1985) The formation of non-anadromous populations of Atlantic salmon, *Salmo salar* L., in Europe. *Journal of Fish Biology*, **27**, 805–815.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bourret V, O'Reilly PT, Carr JW, Berg PR, 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.

Bradbury IR, Hubert S, Higgins B *et al.* (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **277**, 3725–3734.

Carr JW, Whoriskey F, O'Reilly P (2004) Efficacy of releasing captive reared broodstock into an imperilled wild Atlantic salmon population as a recovery strategy. *Journal of Fish Biology*, **65**, 38–54.

Daetwyler HD, Schenkel FS, Sargolzaei M, Robinson JAB (2008) A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. *Journal of Dairy Science*, **91**, 3225–3236.

Davidson WS, Koop BF, Jones SJM *et al.* (2010) Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biology*, **11**, 403–409.

Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, **3**, 22–32.

Dekkers JCM, Van der Werf JHJ (2007) Chapter 10: strategies, limitations, and opportunities for marker-assisted selection in livestock. In: Marker-assisted Selection—Current Status and Future Perspectives in Crops, Livestock, Forestry and Fish (eds Guimarães EP, Ruane J, Scherf BD, Sonnino A, Dargie JD), pp. 167–184. FAO, Rome.

Dionne M, Miller KM, Dodson JJ, Caron F, Bernatchez L (2007) Clinal variation in MHC diversity with temperature: evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution*, **61**, 2154–2164.

Dionne M, Caron F, Dodson JJ, Bernatchez L (2008) Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology*, **17**, 2382–2396.

Dominik S, Henshall JM, Kube PD *et al.* (2010) Evaluation of an Atlantic salmon SNP chip as a genomic tool for the application in a Tasmanian Atlantic salmon (*Salmo salar*) breeding population. *Aquaculture*, **308**, S56–S61.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Ferguson A, Fleming IA, Hindar K *et al.* (2007) Farm escapes. In: The Atlantic Salmon: Genetics, Conservation and Management (eds Verspoor E, Stradmeyer L, Nielsen JL), pp. 357–398. Blackwell Publishing, Oxford.

Flori L, Fritz S, Jaffrezic F *et al.* (2009) The genome response to artificial selection: a case study in dairy cattle. *PLoS ONE*, **4**, e6595.

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.

Ford JS, Myers RA (2008) A global assessment of salmon aquaculture impacts on wild salmonids. *PLoS Biology*, **6**, 0411–0417.

Fraser DJ, Weir LK, Bernatchez L, Hansen MM, Taylor EB (2011) Extent and scale of local adaptation in salmonid fishes: review and meta-analysis. *Heredity*, **106**, 404–420.

Friedland KD, Reddin DG, Castonguay M (2003) Ocean thermal conditions in the post-smolt nursery of North American Atlantic salmon. *ICES Journal of Marine Science*, **60**, 343–355.

Gaggiotti OE, Bekkevold D, Jorgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, **63**, 2939–2951.

Gagnaire PA, Normandeau E, Bernatchez L (2012a) Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and

American eels. *Molecular Biology and Evolution*, Online early. doi: 10.1093/molbev/mss076

Gagnaire PA, Normandeau E, Côté C, Hansen MM, Bernatchez L (2012b) The effect of spatially varying selection on locally adaptive alleles in the panmictic American eel (*Anguilla rostrata*). *Genetics*, **190**, 725–736.

Gauthier-Ouellet M, Dionne M, Caron F, King TL, Bernatchez L (2009) Spatiotemporal dynamics of the Atlantic salmon (*Salmo salar*) Greenland fishery inferred from mixed-stock analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 2040–2051.

Gay L, Crochet PA, Bell DA, Lenormand T (2008) Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution*, **62**, 2789–2806.

Gibbs RA, Taylor JF, Van Tassell CP *et al.* (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, **324**, 528–532.

Gidskehaug L, Kent M, Hayes BJ, Lien S (2011) Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics*, **27**, 303–310.

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.

Gotz S, Garcia-Gomez JM, Terol J *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420–3435.

Griffiths AM, Machado-Schiaffino G, Dillane E *et al.* (2010) Genetic stock identification of Atlantic salmon (*Salmo salar*) populations in the southern part of the European range. *BMC Genetics*, **11**, 31.

Hatta Y, Tsuchiya N, Matsushita M, Shiota M, Hagiwara K, Tokunaga K (1999) Identification of the gene variations in human CD22. *Immunogenetics*, **49**, 280–286.

Havlak P, Chen R, Durbin KJ *et al.* (2004) The atlas genome assembly system. *Genome Research*, **14**, 721–732.

Hayes B, Laerdahl JK, Lien S *et al.* (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

Hedrick PW (1999) Highly variable loci and their interpretation in evolution and conservation. *Evolution*, **53**, 313–318.

Hewitt GM (1988) Hybrid zones—natural laboratories for evolutionary studies. *Trends in Ecology & Evolution*, **3**, 158–167.

Hindar K, Ryman N, Utter F (1991) Genetic effects of cultured fish on natural fish populations. *Canadian Journal of Fisheries and Aquatic Sciences*, **48**, 945–957.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Johannesson K, Andre C (2006) Life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. *Molecular Ecology*, **15**, 2013–2029.

Karlsson S, Moen T, Hindar K (2010) An extended panel of single nucleotide polymorphisms in Atlantic salmon (*Salmo salar* L.) mitochondrial DNA. *Conservation Genetics*, **11**, 1171–1175.

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

Kazakov RV, Titov SF (1991) Geographical patterns in the population-genetics of Atlantic salmon, *Salmo-salar* L, on USSR territory, as evidence for colonization routes. *Journal of Fish Biology*, **39**, 1–6.

King TL, Kalinowski ST, Schill WB, Spidle AP, Lubinski BA (2001) Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation. *Molecular Ecology*, **10**, 807–821.

King TL, Verspoor E, Spidle AP *et al.* (2007) Biodiversity and population structure. In: The Atlantic salmon—Genetics, Conservation and Management (eds Verspoor E, Stradmeyer L, Nielsen JL), pp. 117–166. Blackwell Publishing, Oxford.

Knox D, Lehman K, Reddin DG, Verspoor E (2002) Genotyping of archival Atlantic salmon scales from northern Québec & West Greenland using novel PCR primers for degraded DNA. *Journal of Fish Biology*, **60**, 266–270.

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology & Evolution*, **21**, 629–637.

Koljonen ML (2001) Conservation goals and fisheries management units for Atlantic salmon in the Baltic Sea area. *Journal of Fish Biology*, **59**, 269–288.

Lamaze FC, Sauvage C, Marie A, Garant D, Bernatchez L (2012) Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (*Salvelinus fontinalis*). *Molecular Ecology*, **21**, 2877–2895.

Lien S, Gidskehaug L, Moen T *et al.* (2011) A dense SNP-based linkage map of Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics*, **12**, 615.

Limborg MT, Pedersen JS, Hemmer-Hansen J, Tomkiewicz J, Bekkevold D (2009) Genetic population structure of European sprat *Sprattus sprattus*: differentiation across a steep environmental gradient in a small pelagic fish. *Marine Ecology Progress Series*, **379**, 213–224.

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, **21**, 2128–2129.

Lorenz S, Brenna-Hansen S, Moen T *et al.* (2010) BAC-based upgrading and physical integration of a genetic SNP map in Atlantic salmon. *Animal Genetics*, **41**, 48–54.

Makhrov AA, Verspoor E, Artamonova VS, O'Sullivan M (2005) Atlantic salmon colonization of the Russian Arctic coast: pioneers from North America. *Journal of Fish Biology*, **67**, 68–79.

McGinnity P, Prodöhl P, Ferguson A *et al.* (2003) Fitness reduction and potential extinction of wild populations of Atlantic salmon, *Salmo salar*, as a result of interactions with escaped farm salmon. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **270**, 2443–2450.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Nei M (1977) F-Statistics and analysis of gene diversity in sub-divided populations. *Annals of Human Genetics*, **41**, 225–233.

Normandeau E, Hutchings JA, Fraser DJ, Bernatchez L (2009) Population-specific gene expression responses to hybridization between farm and wild Atlantic salmon. *Evolutionary Applications*, **2**, 489–503.

Palstra FP, O'Connell MF, Ruzzante DE (2007) Population structure and gene flow reversals in Atlantic salmon (*Salmo salar*) over contemporary and long-term temporal scales: effects of population size and life history. *Molecular Ecology*, **16**, 4504–4522.

Primmer CR (2009) From conservation genetics to conservation genomics. In: Year in Ecology and Conservation Biology 2009 (eds Ostfeld RS, Schlesinger WH), pp. 357–368. Blackwell Publishing, Oxford.

Refstie T (1983) Induction of diploid gynogenesis in Atlantic salmon and rainbow trout using irradiated sperm and heat shock. *Canadian Journal of Zoology*, **61**, 2411–2416.

Roberge C, Einum S, Guderley H, Bernatchez L (2006) Rapid parallel evolutionary changes of gene transcription profiles in farmed Atlantic salmon. *Molecular Ecology*, **15**, 9–20.

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (Coregonus sp Salmonidae) species pairs. *Molecular Biology and Evolution*, **24**, 1423–1438.

Säisä M, Koljonen ML, Gross R et al. (2005) Population genetic structure and postglacial colonization of Atlantic salmon (*Salmo salar*) in the Baltic Sea area based on microsatellite DNA variation. *Canadian Journal of Fisheries and Aquatic Sciences*, **62**, 1887–1904.

Schmidt PS, Zhu CT, Das J, Batavia M, Yang L, Eanes WF (2008) An amino acid polymorphism in the couch potato gene forms the basis for climatic adaptation in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 16207–16211.

Skaala O, Makhrov AA, Karlsen T et al. (1998) Genetic comparison of salmon from the White Sea and north-western Atlantic Ocean. *Journal of Fish Biology*, **53**, 569–580.

Soller M (1990) Genetic-mapping of the bovine genome using deoxyribonucleic acid-level markers to identify loci affecting quantitative traits of economic importance. *Journal of Dairy Science*, **73**, 2628–2646.

Ståhl G (1987) Genetic population structure of Atlantic salmon. In: Population Genetics and Fisheries Management (eds Ryman N, Utter F), pp. 212–240. University of Washington Press, Seattle.

Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution*, **64**, 2489–2509.

Suekawa Y, Aihara H, Araki M, Hosokawa D, Mannen H, Sasazaki S (2010) Development of breed identification markers based on a bovine 50K SNP array. *Meat Science*, **85**, 285–288.

Tessier N, Bernatchez L (2000) A mtDNA and microsatellite assessment of single *vs* multiple origins of landlocked Atlantic salmon (*Salmo salar* L.) from Lake Saint-Jean, Québec, Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 797–804.

Todd CD, Hughes SL, Marshall CT, Maclean JC, Lonergan ME, Biuw EM (2008) Detrimental effects of recent ocean surface warming on growth condition of Atlantic salmon. *Global Change Biology*, **14**, 958–970.

Tonteri A, Titov S, Veselov A et al. (2005) Phylogeography of anadromous and non-anadromous Atlantic salmon (*Salmo salar*) from northern Europe. *Annales Zoologici Fennici*, **42**, 1–22.

Tonteri A, Veselov AJ, Titov S, Lumme J, Primmer CR (2007) The effect of migratory behaviour on genetic diversity and population divergence: a comparison of anadromous and freshwater Atlantic salmon *Salmo salar*. *Journal of Fish Biology*, **70**, 381–398.

Tonteri A, Veselov AJ, Zubchenko AV, Lumme J, Primmer CR (2009) Microsatellites reveal clear genetic boundaries among Atlantic salmon (*Salmo salar*) populations from the Barents and White seas, northwest Russia. *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 717–735.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Verspoor E, Beardmore JA, Consuegra S et al. (2005) Population structure in the Atlantic salmon: insights from 40 years of research into genetic protein variation. *Journal of Fish Biology*, **67**, 3–54.

Verspoor E, Stradmeyer L, Nielsen JL (2007) The Atlantic Salmon: Genetics, Conservation, and Management. Blackwell Publishing Ltd, Oxford.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution*, **38**, 1358–1370.

Wennevik V, Skaala O, Titov SF, Studyonov I, Naevdal G (2004) Microsatellite variation in populations of Atlantic salmon from North Europe. *Environmental Biology of Fishes*, **69**, 143–152.

Willing EM, Bentzen P, van Oosterhout C et al. (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology*, **19**, 968–984.

Yang JA, Benyamin B, McEvoy BP et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–571.

into the wild. E.V.'s main interest is the application of genetics to the management of freshwater biodiversity with a particular interest in the evolution of population structuring Atlantic salmon and other salmonid fishes. L.B'.s research focuses on understanding the patterns and processes of molecular and organismal evolution as well as their significance to conservation. S.L.'s scientific interests are the mapping and molecular dissection of complex traits in cattle, pig, salmon, cod and wheat, as well as studying the mechanisms of rediploidization in salmonids.

Data accessibility

Individual sampling location and 6176 SNP data: DRYAD entry doi: 10.5061/dryad.gm367.

Supporting information

Additional Supporting Information may be found in the online version of this article.

Table S1 Observed heterozygosity (HO) among GCR and EST based SNPs in both continents.

Table S2 Single nucleotide polymorphism (SNPs) markers observed (HO) and expected (HE) heterozygosities per population.

Table S3 Pairwise measures of genetic differentiation (F_{ST}).

Table S4 Summary of the detection of markers potentially under selection following Fdist (North America) and hierarchical Fdist (Intercontinental, Europe, Russian Landlocked and Norwegian landlocked) genome scans implemented in AR-LEQUIN 3.5 (Excoffier & Lischer 2010).

Table S5 Summary of the detection of genomic regions showing elevated or decreased divergence in European (measured as FCT) salmon populations.

Table S6 Blast results from BLAST2GO with blast e-value treshold of 1×10^{-3} and gene ontology (GO) terms annotation for blast of e-value inferior to 1×10^{-10}.

Table S7 Likelihood of models tested to best explain the geographical positions of clines in (a) the Baltic continuum and (b) the Barents continuum.

Fig. S1–S29 The distribution of European F_{CT} along the linkage groups [number of Figure corresponds to linkage group number (e.g. Figure S1 corresponds to linkage group SSA01)].

Appendix S1 Supplementary methods.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.