Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (Salmo salar)

SUSAN E. JOHNSTON,*1 PANU ORELL,† VICTORIA L. PRITCHARD,* MATTHEW P. KENT,‡ SIGBJØRN LIEN,‡ EERO NIEMELÄ,† JAAKKO ERKINARO† and CRAIG R. PRIMMER*  
*Division of Genetics and Physiology, Department of Biology, University of Turku, Itäinen Pitkäkatu 4, Turku FIN-20520, Finland, †Finnish Game and Fisheries Research Institute, Utsjoki FIN-99980, Finland, ‡Centre for Integrative Genetics (CIGENE) and Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Aas N-1432, Norway

Abstract

Delaying sexual maturation can lead to larger body size and higher reproductive success, but carries an increased risk of death before reproducing. Classical life history theory predicts that trade-offs between reproductive success and survival should lead to the evolution of an optimal strategy in a given population. However, variation in mating strategies generally persists, and in general, there remains a poor understanding of genetic and physiological mechanisms underlying this variation. One extreme case of this is in the Atlantic salmon (Salmo salar), which can show variation in the age at which they return from their marine migration to spawn (i.e. their ‘sea age’). This results in large size differences between strategies, with direct implications for individual fitness. Here, we used an Illumina Infinium SNP array to identify regions of the genome associated with variation in sea age in a large population of Atlantic salmon in Northern Europe, implementing individual-based genome-wide association studies (GWAS) and population-based $F_{ST}$ outlier analyses. We identified several regions of the genome which vary in association with phenotype and/or selection between sea ages, with nearby genes having functions related to muscle development, metabolism, immune response and mate choice. In addition, we found that individuals of different sea ages belong to different, yet sympatric populations in this system, indicating that reproductive isolation may be driven by divergence between stable strategies. Overall, this study demonstrates how genome-wide methodologies can be integrated with samples collected from wild, structured populations to understand their ecology and evolution in a natural context.

Keywords: $F_{ST}$ outlier analysis, genome-wide association, life history variation, sea age, sexual maturity, trade-off

Received 13 March 2014; revision received 4 June 2014; accepted 6 June 2014

Introduction

The age and size at which an individual reaches sexual maturity is an important fitness trait in wild populations. Delayed maturation can lead to larger body size, higher fecundity and increased offspring survival, but longer generation times can carry an increased risk of death before maturity by prolonging the more vulnerable juvenile life stages (Stearns 2000). If age and size at maturity is heritable in a given system, then classical life history theory would predict that any trade-off between reproductive success and survival would lead to the evolution of an optimal strategy, as both traits are maximized (Roff 1992; Stearns 1992, 2000). This idea is of particular importance in many fish species,
particularly in those exhibiting external fertilization, yet a broad range of strategies such as sneaking, precocious maturity, migration, courting, drifting and/or territory defending have repeatedly evolved in these species, and variation in strategies within populations is common (Taborsky 1994; Mank & Avise 2006). A number of mechanisms have been proposed to explain this variation, such as frequency dependent selection, antagonistic pleiotropy and/or spatiotemporal environmental heterogeneity (Gross 1996; Roff & Fairbairn 2007), but there remains a poor general understanding of the genetic and physiological mechanisms underlying reproductive strategies (Stearns 2000; Roff & Fairbairn 2007; Roff 2011). If genomic regions determining age at maturity can be identified, we can begin to understand how and why variation is maintained (Aubin-Horth & Renn 2009; Edward & Chapman 2011).

Atlantic salmon (Salmo salar) show considerable plasticity in the age at which they reach sexual maturity (Hutchings & Jones 1998). They have an anadromous life cycle: reproduction takes place in freshwater, and most individuals reach sexual maturity after returning from their marine migration. This age at return (also known as age at maturity or ‘sea age’) can vary from 1 to 5 years within and between populations and strongly determines the size of the returning fish. For example, salmon returning after one winter at sea (called ‘one-sea-winter’ fish, or ‘griise’) can be 50–60 cm in length and weigh 1–2 kg, whereas those spending longer periods at sea (‘multi-sea winter’ fish) can reach sizes in excess of 1 m in length and 20–30 kg in weight (Hutchings & Jones 1998). These vast differences in size have important implications for fitness, as larger, multi-sea winter males typically dominate spawning grounds and have higher reproductive success (Fleming 1998; Garant et al. 2003) and multi-sea winter females have increased egg numbers and egg quality in comparison with one-sea-winter females (Heinimaa & Heinimaa 2004). The potential environmental and physiological mechanisms contributing to variation in sea age are relatively well studied (Salminen et al. 1995; Mather 1998; Friedland 2000; Kallio-Nyberg et al. 2006; Otero et al. 2012), yet the genetic basis of sea-age variation remains poorly understood. Sea-cage experiments in hatchery fish fed ad libitum indicated that the heritability of age at adult maturity may range from 0.04 to 0.48 (Gjerde 1984; Wild et al. 1994) and identified a genetic correlation between increased growth rate and earlier maturity (Gjerde & Gjedrem 1984; Gjerde et al. 1994), although the relevance of sea-cage experiments in the context of natural populations is questionable, due to the lack of a migration phase. Nevertheless, increased growth rate is associated with earlier sexual maturity in both juvenile (i.e. precocious maturity) and adult fish also in natural systems (Chadwick et al. 1986; Skilbrei 1989; Friedland & Haas 1996; Salminen 1997; Hutchings & Jones 1998; Garant et al. 2003; but see Jonsson & Jonsson 2007), and QTL analyses of sexual maturation and migration in aquaculture salmon imply that several regions of the genome are likely to contribute to sea-age variation (Pedersen et al. 2013; Gutierrez et al. 2014). Furthermore, modelling of population dynamics indicates that single- and multi-sea winter strategies are heritable and may be maintained in the wild by density-dependent selection (Gurney et al. 2012).

Age at maturity is a key issue for the management and conservation of wild Atlantic salmon stocks as larger fish (i.e. multi-sea winter fish) are highly sought after in fisheries. However, the relative abundance of multi-sea winter fish has been decreasing in many populations (Hansen & Quinn 1998; Niemelä et al. 2006a; Friedland et al. 2009; Otero et al. 2011; Chaptut 2012). From both commercial and conservation perspectives, maintaining variation in sea age within a population is desirable: the proportion of multi-sea winter females within a population can be a predictor of genetic diversity, accounting for 80% of the observed variation in allelic richness in one large river system (Vähä et al. 2007); indeed, variability in reproductive strategies leading to overlapping generations is likely to play an important role in maintaining genetic diversity (Saunders & Schom 1985; Ellner & Hairston 1994) and ecosystem services more generally (Schindler et al. 2010). In addition, rapid anthropogenic change is affecting a number of aquatic ecosystems and commercially important wild salmon fisheries (Jennings & Kaiser 1998; Conover & Munch 2002; Kuparinen & Merilä 2007); if the variance in survival of fish at sea increases due to anthropogenic effects, the uncertainty of surviving longer periods at sea may also increase (Charlesworth 1994). Therefore, understanding which regions of the genome are associated with sea-age variation is of both fundamental and applied importance in wild salmon populations, particularly for providing guidance on how to maintain any underlying genetic variation within wild populations. Additionally, the ability to identify individuals with a higher likelihood of adopting a particular life history strategy based on genotype information alone may pave the way for reintroduction of the multi-sea winter life history strategy in populations where it has disappeared.

One approach to address this problem in wild populations is to use high throughput genetic markers, such as single nucleotide polymorphisms (SNPs), to identify regions of the genome associated with phenotypic variation (Gilad et al. 2009; Slate et al. 2010). Two common approaches are (i) genome-wide association studies (GWAS), which examine the association between SNPs
and phenotypes throughout the genome (Lewis 2002; McCarthy et al. 2008), and (ii) examination of $F_{ST}$ outliers between populations to detect signatures of adaptive selection at specific genomic regions (Stinchcombe & Hoeskstra 2007; Helyar et al. 2010). However, determining the genetic architecture in wild populations can be challenging, as the ability to characterize genetic variation may be limited by lower numbers of genetic markers, low linkage disequilibrium and small sample sizes (Balding 2006). Furthermore, wild populations may be subject to environmental heterogeneity, population structuring (e.g. due to reproductive isolation and/or cryptic relatedness), differential introgression of alleles (Gosset & Bierne 2013) and historical bottlenecks (Storz 2005), all of which may lead to systematic differences in allele frequencies between cases and controls for the phenotype in question (Huizinga et al. 2014), risking spurious phenotype–genotype associations (Platt et al. 2010; Price et al. 2010) and/or false signatures of selection (Bierne et al. 2011, 2013); the latter is particularly acute in one-dimensional habitats, such as rivers (Fourcade et al. 2013). Studies in the wild have more commonly used the $F_{ST}$ outlier approach between distinct populations (Namroud et al. 2008; Hohenlohe et al. 2010; Bourret et al. 2013; Bruneaux et al. 2013), whereas GWAS in nonmodel organisms have generally been limited to pedigreed populations (Johnston et al. 2011; Santure et al. 2013) or several unstructured, polymorphic populations (Parchman et al. 2012).

The Teno river system in Northern Europe supports the largest wild Atlantic salmon fishery in Europe, accounting for 15–22% of the European riverine salmon catch in a given year (ICES 2013) and has been subject to widespread sampling of scales from harvested salmon for age determination and population monitoring purposes. These scales provide a reliable resource for genetic studies utilizing high throughput SNP technology (Johnston et al. 2013). In addition, there is extensive within and between subpopulation variation in age at smoltification (i.e. age of outward migration) and sea age in both sexes, which range from 2–7 years and from one-sea winter to five-sea winters, respectively (Niemelä et al. 2000), resulting in largely overlapping generations. In this study, we sought to determine whether sea age in Atlantic salmon has a genetic basis, and if so, to identify the specific genetic mechanisms maintaining variation in sea age in the Teno population. Using genomic data from a 6K SNP array (Lien et al. 2011; Bourret et al. 2013), we characterized population structure and demography and used both GWAS and $F_{ST}$ approaches to identify regions of the genome associated with differences in sea age in the Teno river population.

**Materials and methods**

**Study site and sample collection**

The Teno river system (Norwegian: Tana, Sami: Deatnu) is located in Northern Europe (68–70°N, 25–27°E) and runs between Finland and Norway, draining into the Tana fjord at the Barents Sea. The stocking of reared fish or eggs is strictly forbidden in this system, and so salmon stocks can be considered as fully wild. Scales were collected from anadromous adult Atlantic salmon harvested by local fishermen using a variety of methods during August (the end of the fishing season and after the peak migration period; Niemelä et al. 2006b) from 2001 to 2003. The fishermen measured the length and weight of the fish and recorded the date and location of the catch. Scales were dried and archived in paper envelopes at room temperature by the Finnish Game and Fisheries Research Institute (Utsjoki, Finland). The sea age of each individual was determined by counting the growth rings on scales, following the internationally agreed guidelines for Atlantic salmon scale reading (ICES 2011), and confirmed using length and weight measures. We predict that sea age is likely to be a threshold trait with an underlying quantitative distribution ( Falconer & Mackay 1996; Hutchings & Jones 1998); by selecting individuals that are likely to deviate substantially from the mean of the phenotypic distribution (i.e. from either end of the distribution), it is possible maximize the power of detecting loci associated with sea age, given a limited number of genotyping arrays (Lander & Botstein 1989; Manolio et al. 2009; Li et al. 2011). Therefore, we restricted our analysis to virgin fish returning after one-sea winter (hereafter ‘1SW’) or after three or more consecutive sea winters (hereafter ‘3SW’). To minimize differences in the genetic background between the sea-age categories, we ‘category matched’ as many samples as possible by selecting by (i) fisherman-assigned sex, (ii) sea age and (iii) sampling location, to ensure that enough individuals were selected from the same mating population. Scales from a total of 534 individuals, covering a 129 km stretch of the Teno mainstem, were selected for genotyping (Fig. 1).

**DNA extraction, sex determination and genotyping**

DNA was extracted from either one or two scales per individual using a QIAamp DNA mini kit (Qiagen Inc. Valencia, CA, USA). The DNA concentration of each sample was determined using a Nanodrop ND-1000 Spectrophotometer (Thermo Scientific), and samples were normalized to 100 ng/µl. All samples were

© 2014 John Wiley & Sons Ltd
genotyped at 5568 SNP loci using a custom-designed Illumina® iSelect SNP array (Lien et al. 2011; Bourret et al. 2013), and sample order was randomized to remove possible bias due to variation between genotyping runs. Individual genotypes were scored using the clustering algorithm implemented in the Illumina® GenomeStudio Genotyping Analysis Module v2011.1. SNPs were selected for further analysis based on the protocol outlined in Johnston et al. (2013): clustering was carried out excluding samples where the raw call rate before quality control was below 0.95; of the remaining loci, genotypes with an Illumina GenCall score of less than 0.05, SNPs with a call rate of less than 0.99, a minor allele frequency of less than 0.01 and/or a heterozygote excess/deficit of more than 0.1 were removed. Samples matching at more than 90% of loci were removed from the study. As the salmon genome is still in the assembly process, all SNP positions are given in centiMorgans (cM) from the female linkage map published in Lien et al. (2011). Linkage disequilibrium (LD) measures were calculated between loci in sliding windows of 20 SNPs using a Spearman’s rank correlation ($r^2$) and $D’$ in the R library genetics v1.3.8 (Warnes et al. 2012) in R v2.15.3 (R Core Team 2013). Samples were also genotyped on an optimized panel of 14 microsatellites known to be highly polymorphic within Teno Atlantic salmon using the protocol outlined in Våhå et al. (2008). The sex of each fish was verified using a sex marker genotyping protocol outlined in Yano et al. (2013). The final samples (after confirming sex) included the following: 1SW females = 59, 1SW males = 194, 3SW females = 162, 3SW males = 119.

**Analysis of population structure**

A matrix of genomic kinship was determined for all pairs of genotyped samples across all SNPs using the software GENABEL (Aulchenko et al. 2007) implemented in R v2.15.3. Classical multidimensional scaling (MDS) was carried out to identify whether there was clustering of genetically similar individuals, and the number of clusters was determined using hierarchical clustering in R library mclust (Fraley & Raftery 2002). The association between sea age, year of capture, river age, sex and location on cluster assignment was examined using a $\chi^2$ test to determine possible sources of stratification. In the case where individuals appeared to cluster into more than one population, global $F_{ST}$ between MDS clusters was calculated; to achieve unbiased estimates, we created a second data set including only SNPs that had a pairwise $r^2$ of 0.01 (hereafter ‘unlinked SNPs’). Unlinked SNPs were identified using PLINK v1.07 (Purcell et al. 2007), with a sliding window of 50 SNPs, with a window shift of 5 SNPs and a variance inflation factor of 1.01. Global $F_{ST}$ was estimated using ARLEQUIN v3.5.1.3 (Excoffier & Lischer 2010). The degree of relatedness between all pairs of individuals was calculated using microsatellite data in the software ML-RELATE (Kalinowski et al. 2006).

**Genome-wide association study (GWAS) of sea age**

The appropriate genome-wide and chromosome-wide thresholds for multiple testing after accounting for linkage disequilibrium (LD) between all SNP loci were calculated using the software KESS VSEP 2007 (Moskvina & Schmidt 2008), using a sliding window of 50 SNPs. To determine the model structure for the GWAS, general linear models (GLMs) of sea age (modelled as a binary trait with a binomial error structure; 1SW = 0, 3SW = 1) were fitted in R to identify additional covariates associated with sea age; these included age at smoltification, sex, year of collection and cumulative distance within the river. Genome-wide association between sea age and each SNP genotype was then tested in all individuals using the gtscore function in GENABEL; sea age was again modelled as a binary trait, with sex, year of capture and cumulative distance within the river included as fixed effects in all association models. To account for sources of population stratification, three different GWAS approaches were used, depending on the nature of the structure:

1. Genomic control ($\lambda$): If all individuals appeared to be members of the same population after hierarchical clustering (and not first-degree relatives), test statistics were divided by $\lambda_{GC}$ to reduce inflation,
where $\lambda_{GC}$ was the median $\chi^2$ statistic (with 1 d.f.) association statistic across all SNPs, divided by its theoretical mean under the null distribution. A value of $\lambda_{GC} \approx 1$ means that there is no stratification, whereas $\lambda_{GC} > 1$ indicates stratification or other confounding effects, such as family structure or cryptic relatedness (Price et al. 2010).

2 Structured association: If samples clustered in more than one MDS cluster (i.e. there is more than one population within the sample) and individuals within these clusters were considered not to be first-degree relatives, GWAS was conducted independently within each population cluster and estimates summed to determine the global significance. This method was implemented by providing the MDS cluster classification for each individual in the strata argument within the egscore function in GENABEL.

3 EIGENSTRAT (Price et al. 2006): This method is a form of structured association, where in the case of ambiguity in defining individual populations, GWAS were repeated fitting principle components derived from the genomic kinship matrix, to account for any differences in ancestry between the different sea-age classes. This method was implemented using the egscore function in GENABEL.

Analysis of population demography and $F_{ST}$ outlier analysis

Bottleneck detection and effective population size estimation. The programme BOTTLENECK v 1.2.02 (Piry et al. 1999) was used to test for recent population bottlenecks on the microsatellite data grouped by MDS cluster. A Wilcoxon’s test under a two-phase mutation (TPM) model (Di Rienzo et al. 1994) was used to determine whether population clusters had undergone a recent bottleneck. The effective population size ($N_e$) for each MDS cluster was estimated using microsatellite data in the software NEESTIMATOR v2 (Do et al. 2013).

Identifying loci under selection. We examined the pairwise genetic differentiation between MDS clusters for each individual SNP locus. This allowed us to identify loci which were likely to differ between MDS clusters as a result of selection, rather than genetic drift. To avoid false positive results, two approaches were used to detect loci under selection (Vasemägi & Primmer 2005). First, $F_{ST}$ was calculated using a Bayesian approach to directly estimate the posterior probabilities of each locus being subject to selection. This was implemented in BAYESCAN v2.1 (Foll & Gaggiotti 2008) for a total of 100 000 iterations, with a burn in of 50 000 iterations, a thinning period of 10 iterations and a sample size of 5000. Second, $F_{ST}$ was calculated with 200 000 simulations using the hierarchical island model (Beaumont & Nichols 1996) implemented in ARLEQUIN v3.5.1.3. The results were then compared to determine loci which were significantly under selection in both studies at an experiment-wide threshold of $P = 0.001$.

Determination of candidate genes

The Atlantic salmon genome project was ongoing during this study, but a number of assembled contiguous sequences (hereafter 'contigs') could be accessed via the cGRASP salmon database on ASalBase (available at: http://asalbase.org/, accessed June 2013), which included information on SNP positions; we retrieved contigs containing the most highly associated SNPs from both the GWAS and $F_{ST}$ outlier analyses. We also obtained sequence scaffolds containing the most highly associated SNPs from the salmon genome sequence DRAFT v3.6 (available at http://www.ncbi.nlm.nih.gov/nuccore/655945504). All contigs are hereafter referred to as 'reference contigs'. The location of the SNP flanking sequences on the reference contigs was determined using blastn in BLAST v 2.2.27+ (Altschul et al. 1990). To determine the relative locations of the significant SNPs relative to coding regions, we obtained nucleotide sequences, expressed sequence tags and proteins from the NCBI database (available at: http://www.ncbi.nlm.nih.gov/, accessed: June 2013). Two approaches were used to determine the location of genes in the reference contigs:

1 blastn (nucleotide) of all Salmo NCBI sequences ($N = 60140$) with a word size of 15 and e-value of 1e-10. Alignments with <100 matches and/or <95% similarity between the NCBI sequences and reference contigs were discarded. Hits to transposases, reverse transcriptases, BAC clone sequence and microsatellites were also discarded. Sequence segments aligning to several independent regions of the reference contigs (i.e. likely repetitive sequences) were also removed. In addition, segments of the reference contigs with multiple alignments to more than five different sequences were examined; alignments were removed if indicative of a repetitive sequence.

2 tblastn (proteins against translated nucleotides) of all Salmo ($N = 22766$) and Oncorhynchus (Pacific salmon, $N = 21401$) NCBI proteins, excluding transposases and reverse transcriptases, with a word size of 6. Alignments with <10 matches and/or <70% similarity between NCBI proteins and the reference contigs were discarded. As in (1), segments aligning to several independent regions of the reference contigs were removed.

© 2014 John Wiley & Sons Ltd
Gene ontology information

To gain an insight into the potential function of genes occurring in regions of the genome significantly associated with sea age, gene ontology (GO) information was extracted for all loci occurring on reference contigs with significantly associated SNPs. GO terms were obtained from databases for human (Homo sapiens), rat (Rattus norvegicus), mouse (Mus musculus) and zebrafish (Danio rerio), obtained from the web resource (http://www.geneontology.org/) in July 2013. No enrichment test was conducted due to the relatively small number of significant loci identified.

Retrospective assignment of individuals to sea-age class

To test the ability of SNP genotypes in the study to assign individuals to sea-age classes or population clusters, we tested the assignment of genotyped individuals to the correct grouping (i.e. sea age or population cluster) using the Rannala & Mountain method (Rannala & Mountain 1997) implemented in the software GENECLASS2 V2.0.h (Piry et al. 2004). To determine the optimum number of SNPs required for group assignment, SNPs were ranked in order of their association with grouping: for assignments to a particular sea age, we used the GWAS results corrected with genomic control; for assignment to a particular MDS population cluster, we used the individual locus $F_{ST}$ estimates across all four populations. To determine the optimal number of SNPs for group assignment, we first used all individuals as reference population to assign all individuals to the correct groups with subsets of the top 10, 20, 30, 50, 100, 200, 300, 400, 500 and 1000 SNPs with an assignment threshold of 0.05. The proportion of individuals assigned to the correct grouping was calculated, and the optimal number of SNPs was taken as the smallest number at which this proportion appeared to asymptote. Using this method alone means that population assignment may be biased, as the same individuals are used as both the reference population and assigned individuals. To account for this, we determined the accuracy of group assignment using a twofold cross-validation approach with the optimal number of SNPs. Briefly, the data set was divided into two groups, A and B, each comprised of half of the individuals in the total data set; individuals were assigned to either data set A or B at random. Data sets A and B were then both used as a reference data set, with individuals of the opposite data set being assigned, meaning that all individuals were used as both reference and assigned samples. This process was repeated 10 times, and the mean proportion of correctly assigned individuals was calculated over all replicates ($N = 20$).

Results

A total of 503 individuals and 4353 SNP loci passed quality control, with a median intramarker distance of 0.1 cM (mean $= 0.597$ cM). Using sliding windows of 20 SNPs, the median values of linkage disequilibrium were 0.054 for $r^2$ and 0.126 for D’ (means $= 0.123$ and 0.253, respectively).

Analysis of population structure and demography

Four clusters of genetically similar individuals were identified (classical multidimensional scaling, hereafter ‘MDS’, Fig. 2a; Table 1); cluster classification was strongly associated with sea age ($\chi^2 = 168.25, P = 3.04 \times 10^{-10}$; Fig. 2b, Table 1); and more weakly with year of capture ($\chi^2 = 23.504, P = 6.44 \times 10^{-4}$). The mean relatedness coefficient across all individuals was low (0.047), with around 10% of individuals within each cluster estimated to be related as half-sibs and 90% estimated as unrelated (Table S1, Supporting information). The global $F_{ST}$ over all MDS clusters was 0.0103
Genome-wide association study of sea age

Sea-age variation was significantly associated with sex (GLM; \( P < 0.001 \)), year of capture (\( P = 0.024 \)) and cumulative distance within the river (\( P = 0.033 \)), but not with age of smoltification (\( P = 0.62 \)) in this data set; significant effects were fitted in all subsequent genome-wide association models. The significance threshold for genome-wide association was \( P = 1.23 \times 10^{-5} \) (i.e. equivalent to an experiment-wide significance threshold of \( P = 0.05 \) with \( 4060.278 \) independent tests). After correction for population stratification using genomic control (\( \lambda_{GC} = 3.24 \)), two markers at 50.5 cM on chromosome 9 passed the threshold for genome-wide significance (GCR_cBin3754_Ctg1_259 and GCR_cBin3754_Ctg1_594, \( P < 0.05 \), Fig. 3a), and eight further markers on chromosomes 9, 12 and 16 were significant at the chromosome-wide level (\( P < 0.05 \); Fig. 3a; Table 2). As analysis of population structure indicated that structure was associated with sea age, the GWAS was repeated using structured association (Fig. 3b) and EIGENSTRAT (Fig. 3c) methods. There were no significantly associated loci with either approach, with the exception a single marker on chromosome 27 which reached chromosome-wide significance with the structured association approach (Fig. 3b). Full results for all approaches are provided in Table S3 (Supporting information).

Detection of loci under selection

Outlier analyses using the methods implemented in BAYESCAN and ARLEQUIN indicated that 14 of the 4353 loci (0.32%) were significantly divergent between the four MDS clusters (\( P < 0.001 \); Fig. 4, Table 2). Six of the most significant loci were located on the same region of chromosome 9 (50.5–52.7 cM). Other highly divergent markers grouped on chromosomes 12 and 16 as in the GWAS results (Table 2, Fig. 4), with additional markers identified on chromosomes 9, 1 and one unknown location. At the most divergent loci, observed \( F_{ST} \) was 0.264, which was more than 25 times greater than the global \( F_{ST} \) of 0.0103. Full results for all loci using both approaches are provided in Table S3 (Supporting information).

Identification and characterization of candidate genes

Five of the outliers had been derived from expressed sequence tags (ESTs; Table 2), and the most highly associated SNP occurred on contigs with a median length of \( \sim 10.6 \) kB (mean \( \sim 18.6 \) kB). In most cases, only one SNP was present on each contig (see Tables S4 and S5, Supporting information); in cases where two or more significant outlier SNPs mapped to the same contig (cM), LD between those SNPs was very high (\( r^2 > 0.817, D' > 0.999 \)). However, three CIGENE contigs associated with chromosome 9 were very large (1.68 MB, 1.84 MB and 3.94 MB respectively); given this large discrepancy in contig sizes for different associated regions, we restricted GO annotation and candidate gene selection to regions directly spanning or flanking the associated SNP locus.

The region at 50.5 cM on chromosome 9 had two SNPs significantly associated with differences in sea age, with a further two SNPs showing a suggestive association in this region (spanning a region of \( \sim 0.5 \) kB). Nucleotide and protein BLAST indicated that these SNPs are not situated within a known gene coding region, but were flanked by likely protein coding regions for ubiquitin thioesterase (OTUB2) and...
inositol-tetrakisphosphate 1-kinase (ITPK1, Table 2); ITPK1 may have functions related to ATP binding. Further loci in this region included 60 kDa lyso phospholipase (LLP60) and charged multivesicular body protein 3 (CHMP3); LLP60 is involved in metabolism of lipids, whereas both LLP60 and CHMP3 are associated with intracellular protein transport. Further regions with suggestive associations on chromosome 9 were at 51.3, 52.7 and 88.2 cM, situated closely to protein coding regions for kinesin light chain 1 (KLC1) and creatine kinase (CK), yippee-like 5 (YPEL5), and histone H3 (Hist3), respectively (Table 2). Interestingly, high CK levels in salmon have been postulated as an indicator of cold tolerance in zebrafish (Wu et al. 2011), as well as muscle degeneration (Rodger et al. 1991) and heart and skeletal muscle inflammation (Yousaf & Powell 2012).
Table 2 SNPs identified as significant outliers using genome-wide association studies (GWAS) (genomic control) and $F_{ST}$ outlier approaches. MAF = minor allele frequency; $H_{OBS}$ = observed heterozygosity. The approximate distance to the nearest coding region is stated if the SNP does not correspond to a coding sequence. Contig-wide BLAST results are provided in Tables S4 and S5 (Supporting information).

<table>
<thead>
<tr>
<th>SNP name</th>
<th>$F_{ST}$ Outlier Rank</th>
<th>GWAS rank</th>
<th>Chr Female map pos. (cM)</th>
<th>MAF</th>
<th>$H_{OBS}$</th>
<th>ARLEQUIN $F_{ST}$ estimate</th>
<th>ARLEQUIN hierarchical Island model P-Value</th>
<th>BAYESCAN $F_{ST}$ estimate</th>
<th>BAYESCAN P-Value</th>
<th>Nearest coding regions (Abbreviation, distance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCR_cBin3754_Ctg1_259</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>50.5</td>
<td>0.154</td>
<td>0.288</td>
<td>0.264</td>
<td>1.00E-07</td>
<td>0.0657</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>GCR_cBin3754_Ctg1_594</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>50.5</td>
<td>0.154</td>
<td>0.288</td>
<td>0.264</td>
<td>1.00E-07</td>
<td>0.0651</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>GCR_cBin3754_Ctg1_228</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>50.5</td>
<td>0.184</td>
<td>0.328</td>
<td>0.236</td>
<td>1.00E-07</td>
<td>0.0630</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>GCR_cBin3754_Ctg1_41</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>50.5</td>
<td>0.184</td>
<td>0.328</td>
<td>0.236</td>
<td>1.00E-07</td>
<td>0.0625</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>ESTNV_34703_149</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>52.7</td>
<td>0.481</td>
<td>0.542</td>
<td>0.221</td>
<td>1.00E-07</td>
<td>0.0830</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>GCR_cBin30825_Ctg1_213</td>
<td>1</td>
<td>6</td>
<td>12</td>
<td>106.5</td>
<td>0.291</td>
<td>0.448</td>
<td>0.221</td>
<td>1.00E-07</td>
<td>0.0758</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>ESTNV_22419_604</td>
<td>1</td>
<td>7</td>
<td>16</td>
<td>60.9</td>
<td>0.378</td>
<td>0.495</td>
<td>0.160</td>
<td>8.54E-06</td>
<td>0.0609</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>ESTNV_30244_819</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>51.3</td>
<td>0.441</td>
<td>0.495</td>
<td>0.188</td>
<td>1.00E-07</td>
<td>0.0713</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>ESTV_16605_658</td>
<td>9</td>
<td>8</td>
<td>16</td>
<td>60.9</td>
<td>0.378</td>
<td>0.496</td>
<td>0.144</td>
<td>5.08E-05</td>
<td>0.0541</td>
<td>2.73E-04</td>
</tr>
<tr>
<td>AY388603_a</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>106.5</td>
<td>0.386</td>
<td>0.499</td>
<td>0.137</td>
<td>9.17E-05</td>
<td>0.0537</td>
<td>2.73E-04</td>
</tr>
<tr>
<td>ESTNV_34364_237</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>88.2</td>
<td>0.348</td>
<td>0.48</td>
<td>0.151</td>
<td>2.28E-06</td>
<td>0.0556</td>
<td>2.73E-04</td>
</tr>
<tr>
<td>GCR_cBin25404_Ctg1_420</td>
<td>12</td>
<td>22</td>
<td>16</td>
<td>61.2</td>
<td>0.308</td>
<td>0.453</td>
<td>0.162</td>
<td>1.00E-07</td>
<td>0.0547</td>
<td>3.83E-04</td>
</tr>
<tr>
<td>GCR_cBin7997_Ctg1_181_V2_NullAll</td>
<td>13</td>
<td>23</td>
<td>—</td>
<td>—</td>
<td>0.220</td>
<td>0.363</td>
<td>0.153</td>
<td>1.00E-07</td>
<td>0.0526</td>
<td>4.92E-04</td>
</tr>
<tr>
<td>ESTNV_35486_240</td>
<td>14</td>
<td>16</td>
<td>1</td>
<td>103.1</td>
<td>0.330</td>
<td>0.459</td>
<td>0.099</td>
<td>5.33E-04</td>
<td>0.0524</td>
<td>8.43E-04</td>
</tr>
</tbody>
</table>

$OTUB2$, Ubiquitin thioesterase; $ITPK1$, Inositol-tetrakisphosphate 1-kinase; $YPEL5$, yippee-like 5; $STK16$, serine/threonine kinase 16; $CK$, creatine kinase; $KLC1$, kinesin light chain 1; $TUBA$, tubulin alpha chain; $MHC$ class II, major histocompatibility complex II; $Hist3$, Histone H3.3; $GLTSCR2$, Glioma tumour suppressor candidate region gene 2; $EBP50$, Ezrin-radixin-moesin-binding phosphoprotein 50.

*SNP loci derived from expressed sequence tags (ESTs).
†The locus was not significantly associated with sea age in the GWAS study.
2012), and its expression likely to be an important disease marker in salmon populations (Braceland et al. 2013); yippee-like 3 has been shown to be upregulated in zebrafish during fasting (Amaral & Johnston 2011). Two markers showed a suggestive association at 106.5 cM on chromosome 12, and the same markers were among the top significant F<sub>ST</sub> outliers. No flanking genes could be identified for one of these loci, but the second ranked locus was flanked on both sides by protein coding regions for major histocompatibility complex class II alpha (MHC2). Finally, two SNPs showed a suggestive association with sea age at 60.9 cM on chromosome 16 and were also significant F<sub>ST</sub> outliers; both were flanked by protein coding regions for tubulin alpha chain (TUBA), associated with cytoskeleton structure and serine/threonine protein kinase 16 (STK16), which is associated with ATP metabolism. Full results are in Tables S4 (Supporting information) (nucleotide BLAST), S5 (protein BLAST) and S6 (GO terms associated with significant nucleotide and protein BLAST hits).

**Discussion**

In this study, two complementary methods, GWAS and F<sub>ST</sub> outlier analyses, were used to identify regions of the genome associated with differences in sea age in Teno Atlantic salmon. We identified three regions, on chromosomes 9, 12 and 16, exhibiting significant signatures of selection and significant and suggestive genome-wide associations. An additional, unexpected finding was the strong genetic structuring of the population by sea age, indicating that possible reproductive isolation between sympatric populations is maintaining life history variation in this population. Here, we discuss the possible functions of outlier regions and the interpretations of our findings, as well as implications for the conservation and management of sea-age variation within Atlantic salmon populations.

**Genomic regions associated with sea-age variation**

Clues to the nature of phenotypic differentiation between 1SW and 3SW fish may be provided by genes occurring in the significant outlier regions detected in this study, although it should be emphasized that direct inferences are speculative, and follow-up studies are required to strengthen these claims. Several of these genes had functions related to metabolism, intracellular protein transport and muscle development (see Results, Table 2, Table S6, Supporting information). Such functions are viable candidates for being under differential selection between, or indeed, causal variants for, differences in sea age, as it has been speculated that 1SW and 3SW fish may utilize different marine feeding grounds (Jonsson & Jonsson 2011) and vary in their growth rate (see Introduction) resulting in different selection pressures on cold tolerance, muscle development and metabolism. With the exception of the MHC locus (see below), there was no correspondence of our associated loci with previously identified QTL.
associated with sexual maturation in aquaculture Atlantic salmon (Pedersen et al. 2013; Gutierrez et al. 2014). In addition, none of the associated regions are likely to be sex-linked based on the current understanding of sex determination in Atlantic salmon (Eisbrenner et al. 2013; Yano et al. 2013).

One interesting association was that of the major histocompatibility complex (MHC) class II alpha on chromosome 12. This locus is associated with immune response (Grimholt et al. 2003) and parasite susceptibility in Atlantic salmon (Glover et al. 2007) and is highly polymorphic in Atlantic salmon populations (Landry & Bernatchez 2001), and variation at the related MHC class I locus has previously been implicated in differences in migration strategies in rainbow trout (Oncorhynchus mykiss) (Hale et al. 2013). Again, salmon

Fig. 5 Predictive capability of top-ranked SNP loci for sea age and multidimensional scaling (MDS) population cluster assignment. (a) Proportion of individuals with sea age correctly assigned relative to the number of SNPs used for assignment. (b) Proportion of individuals with sea age correctly assigned after using twofold cross-validation using 20 sampled data sets with the top 300 GWAS SNPs (genomic control). (c) Proportion of individuals with MDS population cluster correctly assigned relative to the number of SNPs used for assignment. (d) Proportion of individuals with MDS population cluster correctly assigned after using twofold cross-validation using 20 sampled data sets with the top 500 $F_{ST}$ SNPs (between MDS clusters).
spending different time periods at sea, possibly in differing locations, may have different parasite and infection loads and exposure; indeed, MHC variation has been shown to be correlated with latitude and temperature (Dionne et al. 2007; Tonteri et al. 2010). In addition, MHC loci have also been implicated in mate choice in many vertebrate species (Milinski 2006) including Atlantic salmon (Landry et al. 2001; Evans et al. 2012) and have been shown to influence kin discrimination (Rajakaruna et al. 2006). Therefore, it is possible that selection at MHC may be driven by differences in immunity requirements, but it also may be a driver leading to (or preventing further development of) the level of population structuring observed in our data set.

**Challenges and interpretations of trait mapping in a structured system**

Our study highlighted several challenges for carrying out GWAS in an unpedigreed, wild population. The original study design selected 1SW and 3SW fish from the same locations throughout the Teno mainstem to minimize population structuring via spatial isolation. However, the unexpected high level of structuring by the sea-age phenotype meant that genome-wide associations corrected using genomic control alone had a higher risk of being false positive. Moreover, GWAS methods accounting for structure had low power to detect regions of the genome associated with sea age: structured association had a skewed number of cases and controls within each MDS cluster (see Table 1), effectively reducing the sample size; and the primary principle component in EIGENSTRAT was strongly associated with sea age (see Fig. 2b). Therefore, we extended our analysis to include a population-based approach (i.e. the $F_{ST}$ outlier study) to identify loci undergoing differential selection between the different population clusters. Both the GWAS and $F_{ST}$ outlier approaches identified the same regions of the genome as being associated with differences in sea ages and different selection between populations dominated by a particular sea age. It should be noted that the two approaches are not necessarily independent, as both use information on variation in allele frequencies between cases and controls (GWAS) or between populations ($F_{ST}$ outliers); however, the latter approach, particularly when considering several different models of selection, identifies loci which are significant outliers in relation to a neutral distribution of allele frequencies in a scenario where no selection is taking place. Furthermore, as the effective population sizes were large and the population clusters had not undergone significant bottlenecks in their recent history, we are confident that significant outlier loci identified by two different outlier tests in this study are indeed a result of differential selection between sea ages, and not merely false positives generated by population drift.

One further issue affecting the interpretation of the results was that 4353 loci are unlikely to capture all of the genetic variation within this data set. A recent GWAS study in a wild bird population (Santure et al. 2013) indicated that the power to detect trait loci of even a moderate effect is likely to be low when levels of linkage disequilibrium between markers are low (as is the case in the Teno population: median $r^2 = 0.054$). This is because a low density SNP array will fail to capture a sizeable proportion of genome-wide genetic variation; this effect is expected to be acute if trait variation is determined by many loci of small effect and/or rare variants (Spencer et al. 2009; Visscher et al. 2012). Given the low levels of linkage disequilibrium and high effective population sizes observed in the Teno data set, there is a risk that further loci associated with sea age cannot be detected, as they were not tagged by any of the SNPs within our study (i.e. false-negative findings).

Our study appears to go against this trend, however, as we have detected several genomic regions significantly associated with differences in sea age. However, this may be because these genomic regions do not directly lead to differences in sea age per se, but are rather the result of differences in selection between 1SW and 3SW fish during their lifetime. This is because an inevitable constraint in the study of sea age is that it can only be determined in surviving adults. Therefore, it is not possible to distinguish between hypotheses that (i) significant associations in this study are due to causative effects of loci, or (ii) are a result of differences in selection regimes between 1SW and 3SW during juvenile stages and their marine migration. Differences in diet, migration routes, parasites, predation and growth rates between sea ages (Salminen et al. 1995; Mather 1998; Friedland 2000; Kallio-Nyberg et al. 2006) may have led to fitness differences between particular genotypes within each sea-age phenotype, resulting in a change in allele frequencies within the adult population that was not necessarily present in the population before migration. Nevertheless, future studies of the genetic architecture of sea age will require much higher densities of SNP markers, and if possible, larger numbers of case and control individuals, as well as replication within other polymorphic systems.

**Population structuring by sea age in the Teno salmon**

Although population structuring by phenotype may prove inconvenient for effective trait mapping, this
finding does provide an important contribution to our understanding of the genetic contribution to salmonid life history variation. Our findings not only compliment previous studies showing that genetic variation underlying migration strategies can persist within Atlantic salmon families and populations (Gjerde 1984; Gutiérrez et al. 2014), but also that population structuring between migration strategies indicates that this phenotype may be an important driver of ecological divergence (e.g. as with anadromy in the rainbow trout, *O. mykiss*; see Hale et al. 2013 and references therein).

In our study, the unexpected strong structuring of the populations by sea age could be explained by two alternative scenarios. First, as noted in the previous section of the Discussion, differences in selection pressures between 1SW and 3SW fish may have driven divergence at allele frequencies throughout the genome during the lifetimes of the fish. Second, 1SW and 3SW fish are subject to some degree of reproductive isolation within the Teno mainstem, despite being sympatric populations. Although the fish within each MDS cluster were predominately of one sea age (Table 1, Fig. 1b), the presence of both 1SW and 3SW fish in each cluster indicated that the clusters are genetically differentiated populations (albeit to a low degree). Fish were selected for this study in such a way as to minimize spatial isolation between samples, that is, well after the peak return migration, so that fish should be sampled from their most likely spawning ground. Therefore, the mechanisms potentially driving reproductive isolation between the 1SW and 3SW fish remain unclear. One mechanism may be that reproductive isolation is due to assortative mating and/or temporal variation in spawning time, rather than spatial isolation. This scenario was assumed to be unlikely, as both 1SW and 3SW spawn at the same time (Karppinen & Erkinaro 2009; E.N, J.E. & P.O, personal observations with underwater camera monitoring). Furthermore, parentage assignment in the large Utsjoki tributary of the Teno river indicates that matings do occur between 1SW and 3SW fish (Mikko Ellmén & C.R.P., unpublished). In addition, the observed association with MHC loci indicates a potential mechanism by which assortative mating may arise. An alternative explanation is that the populations may be currently sympatric. Overfishing in tributaries, particularly of the more sought-after 3SW individuals, may have led to population disturbance causing a movement of individuals into the river mainstem. This movement may have been historical, leading to an admixture event, or is temporary, with larger fish returning to their spawning grounds after the end of the fishing season (Aug 31st). Further studies using temporal sampling of the Teno salmon population, and the inclusion of individuals with intermediate life history strategies (i.e. 2SW fish), will be required to shed light on the exact mechanisms by which this reproductive isolation has arisen.

**Retrospective assignment of sea age and consequences for population management**

The SNP array used in the current study also highlighted potential future applications for population monitoring purposes in the Teno population, particularly in prediction of age structure in future populations, and relating changes in genetic variation over time to climatic and other environmental variables. The top 300 SNPs associated with sea age using GWAS (with genomic control) had around 80% retrospective assignment success in determining sea age in the adult fish, indicating the likelihood of a fish returning at a particular age, should it survive until adulthood. It is unlikely that perfect assignment success can be achieved, as the trait is unlikely to be completely heritable: our data suggest that environmental factors are still important in determining sea age. Nevertheless, some improvement may be possible if associations with SNPs in additional genome regions can be identified. Given that different sea ages appear to exist as different populations in the Teno river, one important management outcome of the study is that they may require different management strategies to maintain both age classes within the river system. In particular, overfishing of the 3SW fish could result in permanent changes to their frequency within the system – or even their loss from the population.

**Conclusions**

The results presented here support the idea that sea-age variation in Atlantic salmon is associated with several regions of the genome, possibly with functions related to muscle development, metabolism, immune response and mate choice, although further studies would be required to confirm this. Although we cannot rule out that genetic differentiation between sea ages is due to selection rather than causal variants, we have still identified loci which may play an important role in driving or maintaining variation in this important life history trait. Ultimately, the causal relationship between genetic variants and sea-age phenotype cannot be definitively determined without functional studies in Atlantic salmon, which would be costly, time-consuming and difficult to implement, given the long generation time of these animals. However, future studies examining a greater diversity of populations and temporal sampling, as well as utilizing a higher density of markers to capture a greater proportion of genomic variation, are
likely to shed light on the evolution of sea-age variation in Atlantic salmon populations. In particular, this would enable examination of the extent to which sea-age variation is heritable in Atlantic salmon populations, if the same genomic regions are important in other populations, and why variation in life history strategies are maintained within wild populations.

Acknowledgements

This work would not have been possible without the help and cooperation of the fishermen on the Teno River who contributed scales and phenotypic information to the Finnish Game and Fisheries Research Institute. Scale age measurements were carried out by Jari Haantie. The samples were prepared for SNP genotyping by Katja Salminen, Karin Söstari and Terhi Pajula with guidance from Meri Lindqvist. Data analysis was greatly improved by discussion and feedback from Jarrod Hadfield, Jisca Huisman, Konrad Lobse, Lewis Spurgin, Graham Stone, Silva Uusi-Heikkilä, Shihab Hasan, Hannu Mäkinen and participants of the ERC Wild Animal Genomics workshop at The Burn (Eskdale, Scotland). The comments provided by three anonymous reviewers are also acknowledged. This study was funded by an Academy of Finland professorship awarded to C.R.P.

References


Chadwick EMP, Randall RG, Luger C (1986) Ovarian development of Atlantic salmon (Salmo salar) smolts and age at first maturity. Canadian Special Publication of Fisheries and Aquatic Sciences, 89, 15–23.


Fleming IA (1998) Pattern and variability in the breeding system of Atlantic salmon (Salmo salar), with comparisons to other salmonids. Canadian Journal of Fisheries and Aquatic Sciences, 55 (Suppl.), 59–76.


Purcell S, Neale B, Todd-Brown K et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*, 559–575.


© 2014 John Wiley & Sons Ltd
Yano A, Nicol B, Jouanno E et al. (2013) The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. Evolutionary Applications, 6, 486–496.

J.E., E.N. and P.O. organized the collection of samples. M.P.K. and S.L. developed the SNP chip and genotyped individuals. C.R.P., J.E., P.O. and S.E.J. designed the study. S.E.J. analysed the data. V.L.P. conducted preliminary BLAST analyses. S.E.J and C.R.P. wrote the first version of the paper. All authors contributed significantly to revisions.

Data accessibility
GWAS, F<sub>ST</sub> and BLAST results uploaded as Supporting information.
Sampling locations, phenotype data, SNP genotypes and DNA sequences uploaded to Dryad (doi:10.5061/dryad.fr43s).

Supporting information
Additional supporting information may be found in the online version of this article.
Table S1 Proportions of the four clusters identified to have different levels of relatedness.
Table S2 F<sub>ST</sub> between MDS clusters based on unlinked SNP markers (N = 521).
Table S3 Genotyping summary and genome-wide association study (GWAS) and F<sub>ST</sub> outlier results for all SNPs included in the current study.
Table S4 Nucleotide BLAST (blasn) results of all Salmo nucleotide sequences (N = 60 140) accessed from NCBI in June 2013. Sequences were aligned to reference contigs containing significant SNPs.
Table S5 Protein to nucleotide BLAST (tblastn) results of all Salmo and Oncorhynchus protein sequences (N = 22 766 and 21 401, respectively) accessed from NCBI in June 2013.
Table S6 Gene ontology (GO) information for genes spanning or flanking SNPs significantly associated with sea age variation.