MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 2931–2940 doi: 10.1111/mec.12174

High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*

JUKKA CORANDER,* KERTTU K. MAJANDER,† LU CHENG* and JUHA MERILÄ†

Department of Mathematics and Statistics, University Helsinki, FI-00014, Helsinki, Finland, †Ecological Genetics Research Unit, Department of Biosciences, University Helsinki, FI-00014, Helsinki, Finland

Abstract

Marine fish species are characterized by a low degree of population differentiation at putatively neutral marker genes. This has been traditionally attributed to ecological homogeneity and a lack of obvious dispersal barriers in marine habitats, as well as to the large (effective) population sizes of most marine fish species. The herring (*Clupea harengus*) is a case in point – the levels of population differentiation at neutral markers, even across vast geographic areas, are typically very low ($F_{ST} \approx 0.005$). We used a RAD-sequencing approach to identify 5985 novel single-nucleotide polymorphism markers (SNPs) in herring and estimated genome-wide levels of divergence using pooled DNA samples between two Baltic Sea populations separated by 387 km. We found a total of 4756 divergent SNPs (79% of all SNPs) between the populations, of which 117 showed evidence of substantial divergence, corresponding to $F_{ST} = 0.128$ (0.125, 0.131) after accounting for possible biases due to minor alleles and uneven DNA amplification over the pooled samples. This estimate – based on screening many genomic polymorphisms – suggests the existence of hitherto unrecognized levels of genetic differentiation in this commercially important species, challenging the view of genetic homogeneity in marine fish species, and in that of the Baltic Sea herring in particular.

Keywords: fish, F_{ST}, population genomics, RAD-seq, Single-nucleotide polymorphism marker

Received 14 May 2012; revision received 1 November 2012; accepted 6 November 2012

Introduction

A general pattern of genetic population structuring in fishes is that populations of freshwater species are much more differentiated than those of marine species (DeWoody & Avise 2000; Ward 2004; Shikano *et al.* 2010a). This is understandable in the light of the fact that gene flow between effective population sizes and within freshwater habitats (viz. ponds, lakes, rivers) is reduced as compared to marine habitats. In contrast, our understanding of the low level of population structuring in marine fishes (e.g. Ward *et al.* 1994; Ward 2004) has been slower to develop. Traditionally, the low degree of differentiation in marine habitats has been attributed to their greater ecological homogeneity as compared to freshwater habitats, as well as to the lack of obvious dispersal barriers enabling extensive gene flow (e.g. Ward 2004; Conover *et al.* 2006; Cano *et al.* 2008). Also large (effective) population sizes of many marine fish species (but see Turner *et al.* 2002) have been identified as a possible explanation to the low degree of genetic differentiation in neutral markers (Cano *et al.* 2008).

In recent years, it has become clear that these low levels of genetic structuring in marine fish populations might not extend to genomic regions of functional importance, but that an extensive genomic heterogeneity in the degree of population differentiation is hidden in genomes of many organisms (e.g. Weir *et al.* 2005; Leinonen *et al.* 2008; Nosil *et al.* 2009), including marine fishes (e.g. Cano *et al.* 2008; Nielsen *et al.* 2009a; André *et al.* 2011; Shimada *et al.* 2011). In other words, despite the low degree of differentiation in neutral regions of genome, marine fish populations can also be locally adapted and highly differentiated in genomic regions

Correspondence: Juha Merilä, Fax: +358-9-19157694;
E-mail: juha.merila@helsinki.fi

under selection (Cano *et al.* 2008; Nielsen *et al.* 2009a). However, the evidence for this heterogenic differentiation is still scarce (reviewed in Cano *et al.* 2008; Nielsen *et al.* 2009a) and based on typically low-throughput approaches (e.g. André *et al.* 2011; Shimada *et al.* 2011; Westgaard & Fevolden 2007; Mäkinen *et al.* 2008; Nielsen *et al.* 2009b; but see: Hess *et al.* 2013).

The Baltic Sea herring (*Clupea harengus*) is a case in point: population genetic studies utilizing allozyme markers have found little differentiation across the Baltic ($F_{ST} \approx 0.001$–0.009; André *et al.* 2011; Ryman *et al.* 1984), and microsatellite-based studies have reached similar conclusions ($F_{ST} \approx 0.002$; André *et al.* 2011; Jørgensen *et al.* 2005). However, a single outlier microsatellite locus (Cpa112; $F_{ST} = 0.036$) has been identified (Larsson *et al.* 2007; André *et al.* 2011), but further efforts to probe additional divergence patterns have been modest in terms of the number of markers (<15) used. Consequently, the question of whether the Baltic herring can be considered as one large panmictic population, or a group of locally adapted populations connected by some gene flow, remains open. This question is also of practical importance: the herring is economically the most important species in the Baltic Sea fisheries, which has been deemed biologically and economically unsound (Kulmala *et al.* 2007; ICES 2011). Furthermore, given that the current management of herring fisheries in the Baltic Sea rests on quotas allocated to five different management units (ICES 2011), a mismatch between these units and possible cryptic population structure could add to the biologically unsound management of local stocks (Reiss *et al.* 2009).

The aim of this study was to take advantage of an affordable next generation sequencing technology that has provided unprecedented opportunities to conduct genome-wide studies of differentiation in organisms previously lacking extensive genomic resources (e.g. Hohenlohe *et al.* 2010, 2011). To this end, we used RAD-sequencing (Miller *et al.* 2007; Davey *et al.* 2011) to characterize genetic differentiation in 5985 SNP loci among herring caught from two Baltic Sea sites separated by 387 km. While earlier studies conducted on this geographical scale have uncovered no (or very little) genetic differentiation in mitochondrial and nuclear DNA in this species (e.g. Jørgensen *et al.* 2005; André *et al.* 2011; Limborg *et al.* 2012), we reasoned that given the accumulating evidence for extensive genetic heterogeneity in levels of differentiation in various organisms (e.g. Weir *et al.* 2005; Nosil *et al.* 2009), a screen of genome-wide polymorphisms might also uncover cryptic structuring in this species for which three decades of population genetic studies have failed to discern no or little population structuring.

Methods

The samples for this study were collected from two sites: Eckerö, Torp, Finland (Åland Islands; 60°11′ 19.34′ ′N, 19°36′47.87″E) in May 2009, and from Jūrmalciems, Lipeaja, Latvia (56°47′32.68″N, 21°2′43.83″E) in April 2010. The two locations belong to two different fisheries divisions [28–2 and 29, respectively (ICES 2011)], which are treated as one management unit in the current herring fisheries (Fig. 8.3.2.1 in ICES 2011). Adult herring were captured from spawning sites at the peak of the local spawning season to obtain tissue samples used for DNA extraction. Fin clips were collected from the Åland fish and preserved in 70% ethanol, whereas muscle tissue samples were collected from the Latvian fish.

Total genomic DNA was extracted using Qiagen DNeasy kit (Qiagen, Finland) following manufacturer's instructions. Quality and concentrations of individual DNA samples were checked on 1% agarose gels and measured using a Nanodrop 2000 (Thermo Scientific) spectrophotometer. Six individuals per population were pooled for final concentration of approximately 50 ng/μL genomic DNA, and there was ca. 4000 ng of DNA per pool. RAD libraries were generated by FLORAGENEX (Eugene, OR, USA), using the methods outlined by Baird *et al.* (2008), Hohenlohe *et al.* (2010) and Emerson *et al.* (2010). Briefly, Illumina sequencing adaptors and population-specific barcodes were ligated to digested (using SbfI enzyme) to pooled, total genomic DNA. Barcoded RAD samples were then sequenced on the Illumina HiSeq2000 platform with single-end 1 × 100 bp chemistry at University of Oregon (HT-Seq; Eugene).

De novo assembly of consensus sequences

The Åland library was used as a reference in the *de novo* assembly with FLORAGENEX unitag assembler v2.0. The total number of sequence reads submitted to the assembler pipeline was 5 023 398, of which 4 087 625 (81%) were used in the initial unique Tag sequence (unitag) clustering. The number of initial unitags assembled equalled 104 071, and the final number retained was 63 742 according to the standard FLORAGENEX pipeline criteria (see below). RAD sequence analysis followed the methods as in Pfender *et al.* (2011) with slight modifications. Briefly, using the FLORAGENEX software, sequence reads from one population were first grouped into clusters of identical sequences (RAD tags) and clusters with <5 or >500 sequences were discarded, allowing up to three mismatches in calling identical RAD tags. Consensus sequences were used as a reference for downstream alignment and variant calling.

Read alignment and SNP discovery

The total number of sequence reads produced from the Latvian library was 4 526 378. Reads were aligned using the Bowtie alignment algorithm (Langmead *et al.* 2009), with three mismatches allowed per alignment, with a maximum of a single permissible alignment per read. The median sequencing depth was 16 and Illumina sequencing pipeline 1.3 quality score threshold was set to 20 for single-nucleotide polymorphism markers (SNP) discovery.

SNP determination and sensitivity analysis

As the sequencing was performed using pooled DNA samples, we used as conservative approach as possible to ensure that the conclusions would not be biased towards detecting divergence. Of the total 7228 variant loci (SNP) detected between the two samples, 6442 passed the quality filter threshold, that is, the average sequence quality score for the locus was larger than 20. In addition, all loci with missing data for either population were excluded, leaving a total of 5985 loci for the analysis of genetic diversity and differentiation. As the sample size was relatively small for both locations, we conducted a simulation study to assess the effect of minor allele frequencies (MAF) in the populations to the SNP calling results. Also, to quantify the impact of possible variation in DNA amplification levels within each pooled sample (cf. amplification effective only for a subsample), we conducted another simulation study to assess the combined effect of MAF and uneven DNA amplification on the results. Details of the performed simulations are as follows.

Given the small sample sizes, MAF in the two locations can have a substantial effect on the SNP discovery. Given L loci, the sampling scenarios described below yield a probability distribution over the number of loci where the shared allelic variation is not detected, and consequently, the loci would be considered as fixed at different alleles in the two pooled samples. To quantify this effect, we considered two different scenarios in both of which the MAF had a uniform distribution over the range 0.01–0.40 at any given locus out of L independent loci. Under the first scenario, a locus was assumed to be tri-allelic (with alleles A, B, C), such that alleles A, B, were present in one location and alleles A, C in the other location. In both locations, A was the minor allele and the random value of MAF for a single locus was assumed to be the same (P). We calculated the corresponding probability for the event that the minor allele is detected in neither of the two pooled samples given the sample sizes, which is the probability to observe 0 successes in two independent Binomial($2n$,P) experiments. In the second scenario, we assumed that a locus is bi-allelic (with alleles A, B), such that A is the minor allele in one location and B in the other location. MAF was again assumed to equal P in both cases. The probability to simultaneously fail to observe A from location 1 and B from location 2 is given by the same Binomial expression as in the previous scenario; thus, it is not necessary to determine explicitly which loci are bi-allelic and which tri-allelic.).

As an uneven amplification of DNA from individuals over the genome can also influence the outcome of the genotyping process of pooled samples, we further extended the simulations to assess the combined effect of the two factors. For both locations, it was assumed that successful DNA amplification had the probability 0.80 (chosen to make our approach conservative) independently for each individual. This leads to a high level of variation in the number of detectable alleles per locus, and we used a compound stochastic process (e.g. Casella & Berger 2001) over the loci by randomly deleting the genotypes of as many individuals as indicated by the corresponding binomial probabilities. The result is again a probability distribution over the number of loci where the shared allelic variation is not detected in the two samples, but with an inflated level of variation compared with the case where only the effect of MAF was considered. As the effects of MAF and uneven DNA amplification are expected to be very much larger than fixed loci emerging due to sequencing errors under the standard quality filters applied (for details see above), we did not include this factor in the simulation study.

Population genetic analyses

The allele frequencies were estimated for both locations using the standard Bayesian method with conjugate Beta-priors as used in BAPS software (Corander *et al.* 2003; Corander & Marttinen 2006), such that at every locus only a single copy of each allele in the detected genotype was used in the likelihood. This approach is conservative and minimizes the baseline information content to the same level over all loci, because it cannot be known at which loci there may have been problems with uneven DNA amplification, and thereby failures to capture minor alleles among the sampled individuals. Given the weak likelihood, the used priors will automatically imply considerable uncertainty about allele frequencies in the posterior distributions.

To quantify the genetic divergence, we estimated F_{ST} using the Bayesian Monte Carlo method described in Corander *et al.* (2003), based on the formula derived in Nei (1977), creating 10 000 independent posterior samples for each locus. Again, to employ a more

conservative approach, we preferred an alternative Bayesian estimate based on averaging the locus-wise posterior means, to minimize the effect of the prior distribution under a weak likelihood. Note that as the likelihood contributions of the observed alleles were minimized at all loci, only three distinct values of the posterior mean of F_{ST} are possible.

Additionally, to invoke the results from the sensitivity analysis of SNP calling, the final estimates of F_{ST} were down-weighted based on the probability distribution of the number of loci (Z) where the shared allele would be erroneously missing in both samples. Given that L loci were observed to be fixed at distinct alleles in the two locations, we assigned $K < L$ of them to be due to MAF and uneven DNA amplification using the threshold 0.01 on the probability $P(Z > K)$ according to the distribution obtained as described above (see: SNP determination and sensitivity analysis). For the L-K remaining loci, we used the Bayesian estimate of F_{ST} as described above. For the former K loci, the estimate was down-weighted by assigning it a uniform distribution between zero, and the estimate that was used for the L-K 'unaffected' loci, ensuring that the resulting locus-wise estimates would always be smaller and on average half of the other values. This procedure was adopted for simplicity as it is challenging to model explicitly the effect of MAF and the uneven DNA amplification under the conservative approach to minimize the likelihood contributions.

To test the hypothesis of equal allele frequencies in the two populations against distinct allele frequencies, we used the 'test' option in BAPS software (Corander *et al*. 2008) with equal prior probabilities for the two hypotheses.

BLAST analyses

After initial blast-screen of all 7228 variable RAD tag reads with liberal quality criteria (E-value < 0.1; similarity >70%), the resulting 943 RAD tag reads were used as query sequences in more stringent blast searches conducted with software BLAST2GO (http://www.blast2go.com/b2ghome) and Embster (http://chipster.csc.fi/embster/) to search for homologous sequences, genes and gene annotations. Minimal E-value of 1.0E−3 was set for BLASTN analysis with BLAST2GO and regarded as threshold when considering results with Embster too. In Embster, both BLASTN and the MEGABLAST analyses against the nonredundant database selection (nr nucleotide database maintained by NCBI, a composite of GenBank, GenBank updates and EMBL updates) were conducted. In BLAST2GO, a similar nr–blast was conducted, which runs against all GenBank, EMBL, DDBJ and PDB sequences (but no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS sequences). Although the nr-type search uses a collection of genomes including other species as well, our focus was on the fish genomes. Most of the hits yielded by blast analyses occurred against the genomes of *Danio rerio*, *Takifugu rubripes*, *Gasterosteus aculeatus*, *Salmo salar*, *Oncorhynchus mykiss*, *Oreochromis niloticus* and *Oryzias latipes*, and less commonly against *Osmerux mordax* and *Astotilapia burtoni*. All the used analyses yielded highly congruent results. Below, we report a representative selection hits with UniProtKB classification for biological and molecular functionalities. The 30 overall best-quality hits (75% similarity and E-value < 1.0E−5) are also included in the Supporting information (Table S1), as are the identities of the 133 divergent RAD sequences of 943 most strongly divergent RAD sequences used in BLAST analyses (Table S2).

Results

Population genetic analyses

The fraction of polymorphic sites over the total covered sequence length was approximately 1 × 10E−3 per base (5985/6 055 490). Of the 5985 variable loci without missing data, 1567 (26%) were fixed for different alleles in the two populations (however, this number was pruned in the subsequent calculations as explained in Methods). In contrast, 1229 exhibited no difference in the allele frequencies, and the remaining 3246 loci were intermediate between these two extremes. In total, 4756 loci (79%) showed a pattern of non-zero divergence ($F_{ST} > 0$). A summary count of putative SNP loci and final counts of candidate SNPs after different filtering steps is given in Table 1.

Results of the sensitivity analyses are shown in Fig. 1. Figure 1a shows variation in the probability distribution of the number of loci (Z) where the shared allele failed to become detected in both locations. Choosing the upper 95% confidence limit of the underlying distribution and given that 1567 loci were considered, the probability of Z exceeding 1224 equals the threshold 0.01 (see Methods). Thus, here K = 1224 and the remaining L-K = 286 loci would be interpreted to represent – with high probability – polymorphic sites where MAF did not cause the observed fixation.

Correspondingly, Fig. 1b shows how the effect of uneven DNA amplification combines with the effect of MAF. Here, at the upper 95% confidence limit the probability of Z exceeding 1393 equals the threshold 0.01, which leads to K = 1393 and L-K = 117. The latter value can thus be considered as a conservative estimate of the number of sites where the two locations display a substantial signal of divergence.

Table 1 Counts putative loci in herring after different filtering steps and final counts of candidate single-nucleotide polymorphism markers (SNPs) after filtering

	Count	% of total
Filtering step		
Total variable RAD tag loci	7228	100.0
Loci with average quality score > 10	6935	95.9
Loci with average quality score > 20	6442	89.1
Loci with no missing data for either pooled sample	5985	82.8
Candidate SNPs		
Total	5985	100.0
Fixed between samples	1567	26.2
Fixed between samples after pruning by 95% c.i.	117	2.0
No difference in allele frequencies between samples	1229	20.5
Intermediate difference in allele frequencies between samples	3246	54.2

The posterior mean estimate of F_{ST} was equal to 0.315 (95% credible interval: 0.308, 0.322). However, the preferred estimate is to use the posterior mean of allele frequencies at every locus (see Methods) and then correct the estimate by down-weighting according to the results of the sensitivity analysis. The uncorrected mean F_{ST} equalled 0.179 (95% credible interval: 0.172, 0.182), and the corrected final estimate was 0.128 (95% credible interval: 0.125, 0.131).

Bayesian hypothesis testing resulted in the posterior probability $P = 1.000$ for the hypothesis of unequal allele frequencies for the two locations, which further supports the quantitative finding in terms of F_{ST}.

BLAST *analyses of the divergent loci*

The blast analyses result in a total of 133 hits with *E*-value smaller than 0.001 (Table S1, Supporting information). The 14 best-quality hits ($E < 10^{-5}$; <75% similarity) are shown in Table 2. These hits occurred against various fish genomes and involved biological functions related to immunological functions (e.g. MHC1), hypoxia (Hif1a) and stress responses (e.g. HSP; Table 2).

Discussion

The most significant finding of this study was the high degree of population differentiation between two herring samples collected from the Baltic Sea over a geographic distance of <400 km. The observed (minimum) level of differentiation ($F_{ST} = 0.128$) is much larger than the typical estimates uncovered in earlier studies of herring over much larger geographic areas (Ryman *et al.* 1984; Jørgensen *et al.* 2005; Larsson *et al.* 2007; André *et al.* 2011; Limborg *et al.* 2012). This finding is not entirely unexpected, as Baltic herring are known to exhibit substantial morphological and life-history differentiation (e.g. autumn and spring spawning stocks; Ryman *et al.* 1984). Furthermore, recent studies of other species have started to uncover genetic

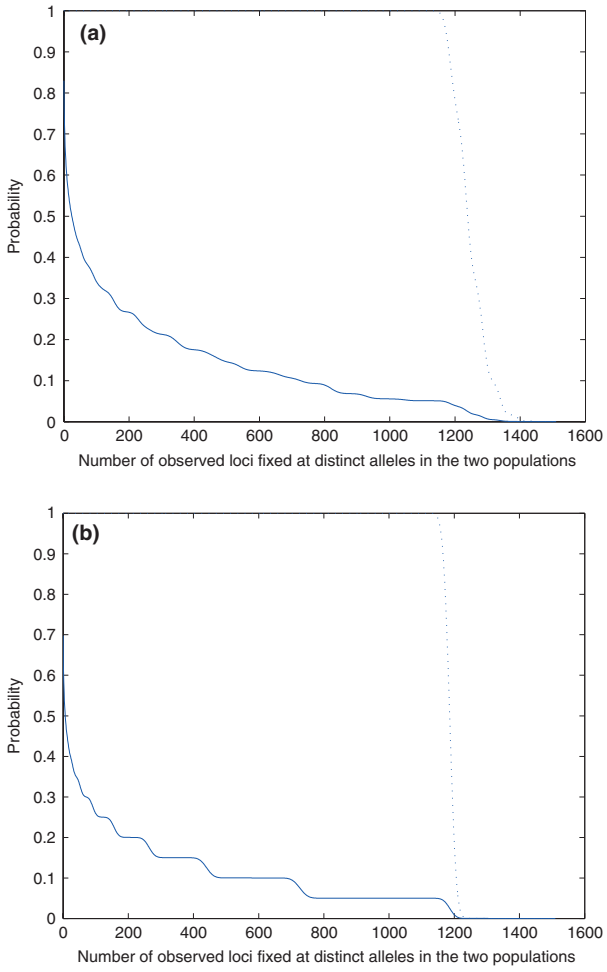

Fig. 1 Probability distribution over the number of loci (Z) where the shared minor allele is not detected in genotyping in our simulation study. (a) The value of the vertical axis equals the probability $P(Z > z)$ that Z exceeds any given value z on the horizontal axis. The unbroken curve represents the mean of the probability over the distribution of minor allele frequency (MAF) values across loci and the dotted curve equals the 95% upper confidence limit (lower limit not visible due to scaling of the axes). (b) Same as the previous except that the distributions were obtained using a compound process in which also the DNA amplification success varied randomly over loci, in addition to the sampling effect due to MAF.

differentiation that has gone undetected in studies employing only neutral marker loci (e.g. Westgaard & Fevolden 2007; Nielsen *et al.* 2009b; Russello *et al.* 2011; Shimada *et al.* 2011). Nonetheless, it is still surprising to uncover such a high degree of differentiation in a species otherwise shown to be genetically homogenous in earlier studies and even in a more recent study which revealed little differentiation using 281 SNPs (Limborg *et al.* 2012). Our results are particularly interesting given that most marine organisms tend to show low levels of genetic differentiation within the Baltic Sea, likely owing to its young age (Johannesson & André 2006; Johannesson *et al.* 2011).

As no reference genome is available for herring, the genomic positions of the SNPs used in this study are mostly unknown. Hence, it is not possible to determine whether the divergent SNPs detected in study reside within or close to coding or regulatory regions of functional genes that might be subject to locally varying selection. Nevertheless, our BLAST analyses revealed that many of the highly divergent SNPs reside in genomic regions that contain physiologically and/or immunologically important genes in other fish species. This finding suggests the possibility that those SNPs may also have corresponding roles in herring. For instance, ionic channel activity-related genes (e.g. kcnn1, ryr2, DDX43 and KCNH8) likely to be involved with osmoregulation and electrolyte homeostasis (e.g. Tse *et al.* 2007; Inokuchi *et al.* 2008; Norman *et al.* 2011; Shimada *et al.* 2011) were among the top hits. Likewise, growth factors (e.g. pdgrfa), and heat shock and immunogenic responses mediating factors (e.g. MHC1, HSPA14, UBA and UBB families), as well as genes suggested to be tissue specific for morphogenesis such as gill and ureteric tract development (Igr4, wdr55), regulation of peripheral blood cell formation (plg), oxygen transportation (hif1a, HBB) and urea transportation (ut1) were identified as potential candidates of divergence (see: Marshall & Grosell 2006; Shikano *et al.* 2010b; for similar findings). Nevertheless, given that only a conservative approach to F_{ST} estimation is possible under the current study design, it is not meaningful to try to statistically identify outlier loci that would have putatively been under selection pressure or experienced hitch-hiking with such loci (e.g. Foll & Gaggiotti 2008). The reason is that the derived information content is identical among all the 1567 loci that showed substantial evidence of divergence, such that they would appear equally outlying in terms of any applicable statistical test. Consequently, it is not reasonable to ask which of them could actually be explicitly considered as outliers. In the future, with the aid of a sequenced reference genome for this species (or if individual rather than pooled samples are used; for example Limborg *et al.* 2012), it will be feasible to perform detailed genome scans and map divergent loci to gain insights towards the genomic regions on which selection is acting.

The analytical approach we have used provides a very conservative estimate of F_{ST}. Yet, the considerable level of divergence – as compared to earlier studies based on limited numbers of putatively neutral markers – accords with the contention that at least part of the observed divergence is likely to have been caused by directional selection. For example, we estimated that approximately 7.5% (117/1567) of the substantially divergent loci do in fact represent genuine variation that may be linked to directional selection. Similar results were shown by Roberts *et al.* (2012), in which 6% (6/96) of the loci (a subset of 10 993 SNPs screened) demonstrated substantial population differentiation ($F_{ST} > 0.1$) in the Pacific herring (*Clupea pallasii*). Likewise, 5.7% (16) of the 281 SNP loci screened in the global study of the Atlantic herring were found to show significant divergence (Limborg *et al.* 2012). Much like the herring in the Baltic Sea, earlier microsatellite studies of Pacific herring also failed to find similar levels of differentiation. Hence, one possible explanation for the lack of correspondence with these earlier studies is the larger number of loci ($n = 5985$) screened in our study. Yet, a recent study that screened 281 SNPs in five Baltic Sea herring populations did not detect any significant differentiation (Limborg *et al.* 2012). It is also possible that some of the divergent SNPs in our data are tightly linked, but as we used pooled data, we were unable to estimate the linkage disequilibrium among these loci. However, in the light of the BLAST results, it seems unlikely that linkage alone would explain the large number of divergent loci in our results. Whatever the explanation, we wish to emphasize that the derived F_{ST} estimate should not be interpreted too literally as a direct measure of gene flow, given both the restriction to only two locations within the Baltic Sea and the limitations imposed by the study design. Nevertheless, given our limited sampling and conservative approach, it is possible that we have even underestimated the general degree of divergence among Baltic Sea herring populations.

From the point of herring management in the Baltic Sea, the results underline the need for further studies and possible refinements in management practices. The two populations in the focus of our study both belong to the same herring fishery management unit (see Table 8.3.2.1 in ICES 2011), in which fishing mortality in relation to that producing maximum sustainable yield (F_{MSY}) is considered to be above target level. In addition, harvesting is considered to be unsustainable in relation to precautionary limits (ICES 2011) in this fishery management unit. If one considers the high degree

Table 2 A selection and characterization of 14 high-quality BLAST matches obtained in comparison of *Clupea harengus* RAD-sequence reads against various fish genomes

RAD_id	NCBI_reference	Gene name	*E*-value	Hit length (bp)	Similarity	Species	Found also in species	Uniprot GO
RADid_182_depth_21	NM_001165271.1	ut1	7.11005E−10	1667	77%	*Salmo salar*	*Danio rerio*, *Takifugu rubripes*, *Oreochromis niloticus*	Urea/water transmembrane transporter activity
RADid_342_depth_153	NM_001045199.2	kcnn1	1.05429E−26	1206	89%	*D. rerio*	*O. niloticus*	Small conductance calcium-activated potassium channel activity
RADid_471_depth_41	EF375485.1	MHC1	8.66181E−9	163200	77%	*Gasterosteus aculeatus*		Antigen processing and presentation, immune response
RADid_538_depth_150	XM_682092.5	lgr4	1.12667E−13	2894	83%	*D. rerio* (predicted)	*O. niloticus* (predicted)	Immunoglobulin-like receptor; positive regulation of branching involved in ureteric bud morphogenesis
RADid_823_depth_114	DQ386648.1	pdgfrba	3.93246E−13	97628	82%	*Astotilapia burtoni*		Vascular endothelial growth factor receptor signalling pathway
RADid_823_depth_114	AB270897.1	UBA1, UBA2, UAA1	2.03706E−10	183264	81%	*O. niloticus*		Antigen processing and presentation, immune response
RADid_2363_depth_98	NM_001124391.1	plg	4.79071E−12	2778	91%	*Oncorhynchus mykiss*	*E. coioides*, *D. rerio*	Blood clogging, peripheral blood cell formation
RADid_2753_depth_158	NM_001003871.1	wdr55	2.48165E−9	1916	81%	*D. rerio*	*Oryzias latipes*, *O. niloticus* (predicted)	Organogenesis, pharyngeal arch formation, eyes and swim bladder
RADid_2801_depth_27	BT075462.1	HBB	8.66181E−9	1218	85%	*Osmerus mordax*		Involved in oxygen transport from gills to the various peripheral tissues
RADid_2801_depth_27	AL808019.6	KCNH8	1.28553E−6	159698	83%	*D. rerio* (predicted)	*O. niloticus*, *T. rubripes*, *G. aculeatus*	Pore-forming (alpha) subunit of voltage-gated potassium channel, two-component sensor activity.
RADid_3247_depth_88	XM_001921102.1	ryr2	5.4662E−5	14848	81%	*D. rerio*		Ryanodine-sensitive calcium-release channel activity
RADid_3514_depth_107	BX255914.3	hif1a	1.20295E−19	217957	84%	*D. rerio*	*O. mykiss*	Hypoxia-inducible factor 1-alpha; functions as a master transcriptional regulator of the adaptive response to hypoxia
RADid_4024_depth_12	NM_001045076.1	hspa14	1.46549E−18	2242	83%	*D. rerio*	*O. niloticus* (predicted)	Stress and virus immune response, heat shock
RADid_5829_depth_133	BX890626.7	DDX43	1.67212E−11	20707	78%	*D. rerio*	*G. aculeatus*, *O. latipes*, *T. rubripes*	Similar to vertebrate DEAD; ATP-binding, helicase activity

of genetic heterogeneity within this management unit, it implies that it is not necessarily a demographically and biologically coherent unit, but in fact, perhaps a heterogeneous collection of divergent stocks. If so, there is a clear mismatch between fisheries units and genetic population structure, which can lead to biologically unsound management of local stocks (Reiss *et al.* 2009).

Finally, although pooling of samples is shown to be a sound approach to population genetic inference (e.g. Baird *et al.* 2008; Futschik & Schlötterer 2010; Davey *et al.* 2011), there is always a risk that some error is introduced, for example in the case where some individuals within the pool failed to amplify. This possibility, combined with the effect of MAF, motivated our sensitivity analyses to assess what fraction of the divergent loci should be discarded to avoid inflating the false-positive rate. Despite the relatively small (but comparable to that used in earlier studies; e.g. Jones *et al.* 2011; Feulner *et al.* 2012) number of sampled individuals per population, our analytical approach suggests that the large number of loci that are derivable using the RAD technology enables discoveries about biologically important variation both within and between species, which has not been possible using the traditional genotyping methods. That said, the SNP density recovered in this study (ca. 1 SNP per kb) is substantially lower than that discovered for instance in pooled samples of nine-spined sticklebacks (*Pungitius pungitius*; 1.93 SNP per kb; Bruneaux *et al.* 2012) or in three-spined sticklebacks (*Gasterosteus aculeatus*; 21.8 SNP per kb; Hohenlohe *et al.* 2010). However, these differences are likely to reflect both biological and technical sources of variation. For instance, the number of individuals and populations screened in our study was substantially smaller than that in the two above mentioned studies. Nevertheless, although even small number of individuals may suffice to get good estimates of degree of population differentiation when the number of screened markers is large (Willing *et al.* 2012), further studies with larger sample sizes are clearly warranted to get more confidence on the levels of genetic variability and differentiation in the Baltic Sea herring. Likewise, we note that studies aiming to utilize SNPs discovered in study should consider the fact that as only two populations were used for discovery, ascertainment bias might ensue.

In conclusion, our results provide evidence for extensive heterogeneity in the levels of genetic differentiation among Baltic Sea herring populations and suggest the existence of hitherto unrecognized cryptic population structuring within this species. This cryptic differentiation is particularly noteworthy, as the two study populations belong to the same herring fisheries management unit and are separated only by a relatively short geographic distance. In general, our results highlight the potential utility of second-generation sequencing technologies in identifying hidden structuring in populations of marine fishes, which have traditionally been viewed as genetically homogenous. When genome sequences become more generally available for such organisms, this will open numerous opportunities to continue unravelling patterns of local adaptation and divergence, offering valuable information both for management and conservation purposes in the future.

Acknowledgements

We thank the three anonymous reviewers for helpful comments and suggestions leading to a significant improvement over an earlier version of the article, Jacquelin DeFaveri, Baocheng Guo and Susan Johnston for the critical comments, Janis Birzaks for help in obtaining samples from Latvia, and for Jason Boone (FLORAGENEX) and Amber Teacher for their help with the laboratory work. Our research was supported by the Academy of Finland (grants 129662 and 134728 to JM; grant 251170 to JC), ERC grant 239784 to JC, and the European Community's Seventh Framework Programme (FP/2007-2013) under grant agreement no. 217246 made with the joint Baltic Sea research and development programme BONUS (to JM).

References

André C, Larsson LC, Laikre L *et al.* (2011) Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity*, **106**, 270–280.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Bruneaux M, Johnston SE, Herczeg G, Merilä J, Primmer CR, Vasemägi A (2012) Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology*, in press. doi:10.1111/j.1365-294X.2012.05749.x

Cano JM, Shikano T, Kuparinen A, Merilä J (2008) Genetic differentiation, effective population size and gene flow in marine fishes: implications for stock management. *Journal of Integrative Field Biology*, **5**, 1–10.

Casella G, Berger RL (2001) *Statistical Inference*, 2nd edn. Duxbury Press, Pacific Grove, California.

Conover DO, Clarke LM, Munich SB, Wagner GN (2006) Spatial and temporal scales of adaptive divergence in marine fishes and the implications for conservation. *Journal of Fish Biology*, **69**, 21–47.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multi-locus molecular markers. *Molecular Ecology*, **15**, 2833–2843.

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

DeWoody JA, Avise JC (2000) Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *Journal of Fish Biology*, **56**, 461–473.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA*, **107**, 16196–16200.

Feulner PGD, Chain FJJ, Panchal M *et al.* (2012) Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molecular Ecology*, in press. DOI: 10.1111/j.1365-294X.2012.05680.x.

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.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Hess J, Campbell N, Close D, Docker M, Narum S (2013) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, **22**, 2898–2916.

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

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

ICES (2011) *Report of the ICES Advisory Committee, 2011*. ICES Advice, 2011. Book 8, pp. 119.

Inokuchi M, Hiroib J, Watanabe S, Lee K, Kaneko T (2008) Gene expression and morphological localization of NHE3, NCC and NKCC1a in branchial mitochondria-rich cells of Mozambique tilapia (*Oreochromis mossambicus*) acclimated to a wide range of salinities. *Comparative Biochemistry and Physiology*, **151**, 151–158.

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

Johannesson K, Smolarz K, Grahn M, André C (2011) The future of Baltic Sea populations: local extinction or evolutionary rescue? *Ambio*, **40**, 179–190.

Jones FC, Chan YF, Schmutz J *et al.* (2011) A genome-wide SNP genotyping array reveals patterns of global and repeated species pair divergence in sticklebacks. *Current Biology*, **22**, 83–90.

Jørgensen HBH, Hansen MM, Bekkevold D, Ruzzante DE, Loeschcke V (2005) Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. *Molecular Ecology*, **14**, 3219–3234.

Kulmala S, Peltomäki H, Lindroos M, Söderkultalahti P, Kuikka S (2007) Individual transferable quotas in the Baltic Sea herring fishery: a socio-bioeconomic analysis. *Fisheries Research*, **84**, 368–377.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome, Biology*, **10**, R25.

Larsson LC, Laikre L, Palm S, André C, Carbalho GR, Ryman N (2007) Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. *Molecular Ecology*, **16**, 1135–1147.

Leinonen T, O'Hara RB, Cano JM, Merilä J (2008) Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology*, **21**, 1–17.

Limborg MT, Helyar SJ, De Bruyn M *et al.* (2012) Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, **21**, 3686–3703.

Marshall WS, Grosell M (2006) Ion transport, osmoregulation, and acid-base balance. In: *The physiology of fishes*, (eds Evans DH, Claiborne JB), pp. 177–230. CRC Press, Boca Ranton, Florida.

Mäkinen HS, Cano JM, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. *Molecular Ecology*, **17**, 3565–3582.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

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

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009a) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, **18**, 3128–3150.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009b) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.

Norman J, Danzmann R, Glebe B, Ferguson M (2011) The genetic basis of salinity tolerance traits in Arctic charr (*Salvelinus alpinus*). *BMC genetics*, **12**, 81.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Pfender WF, Saha MC, Johnson EA, Slabaugh MB (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theoretical and Applied Genetics*, **122**, 1467–1480.

Reiss H, Hoarau G, Dickey-Collas M, Wolff WJ (2009) Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish and Fisheries*, **10**, 361–395.

Roberts SB, Hauset L, Seeb LW, Seeb JE (2012) Development of genomic resources for Pacific herring through targeted transcriptome pyrosequencing. *PLoS One*, **7**, e30908.

Russello MA, Kirk SL, Frazer KK, Askey PJ (2011) Detection of outlier loci and their utility for fisheries management. *Evolutionary Applications*, **5**, 39–52.

Ryman N, Lagercrantz U, Andersson L, Chakraborty R, Rosenberg R (1984) Lack of correspondence between genetic and morphologic variabilty patterns in Atlantic herring (*Clupea harengus*). *Heredity*, **53**, 687–704.

Shikano T, Shimada Y, Herczeg G, Merilä J (2010a) History vs. habitat type: explaining the genetic structure of European nine-spined stickleback (*Pungitius pungitius*) populations. *Molecular Ecology*, **19**, 1147–1161.

Shikano T, Ramadevi J, Merilä J (2010b) Identification of local- and habitat-dependent selection: scanning functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*). *Molecular Biology and Evolution*, **27**, 2775–2789.

Shimada Y, Shikano T, Merilä J (2011) A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. *Molecular Biology and Evolution*, **28**, 181–193.

Tse W, Au D, Wong C (2007) Effect of osmotic shrinkage and hormones on the expression of Na+/H+ exchanger-1, Na+/K+/2Cl– cotransporter and Na+/K+-ATPase in gill pavement cells of freshwater adapted Japanese eel, Anguilla japonica. *Journal of Experimental Biology*, **210**, 2113–2120.

Turner TF, Wares JP, Gold JR (2002) Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish (*Sciaenops ocellatus*). *Genetics*, **162**, 1329–1339.

Ward RD (2004) Genetics of fish populations. In: *Handbook of Fish Biology and Fisheries* (eds Hart JB, Reynolds JD), vol. 1, 2nd edn, pp. 200–224. Blackwell Science Ltd, Massachusetts, USA.

Ward RD, Woodwark M, Skibinski DOF (1994) A comparison of genetic diversity levels in marine, freshwater and anadromous fishes. *Journal of Fish Biology*, **44**, 213–232.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15**, 1468–1476.

Westgaard JI, Fevolden S-E (2007) Atlantic cod (*Gadus morhua* L.) in inner and outer coastal zones of northern Norway display divergent genetic signature at non-neutral loci. *Fisheries Research*, **85**, 306–315.

Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLoS One*, **7**, e42649.

J.C. and L.C. are statisticians interested on challenging population genomic problems. K.M. is a bioinformatician with particular interest on population genomic data. J.M. is an evolutionary biologist interested on ultimate and proximate causes of population differentiation.

Data accessibility

Raw sequence data and pileups, details concerning the BLAST analyses, and information on all SNPs and fixed SNPs can be found in Dryad (doi:10.5061/dryad.jr56h).

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 A list of the 30 best-quality blast matches together with their identifiers, hit lengths, *E*-values and similarity values.

Table S2 A list 133 best-match blast hits.