A microsatellite baseline for genetic stock identification of European Atlantic salmon
(Salmo salar L.).

John Gilbey¹*, Jamie Coughlan², Vidar Wennevik³, Paulo Prodöhl⁴, Jamie R. Stevens⁵, Carlos
Garcia de Leaniz⁶, Dennis Ensing⁷, Eef Cauwelier¹, Corrine Cherbonnel⁸, Sofia Consuegra⁹, Symon
Mark W. Coulson¹⁰, Tom F. Cross⁷, Walter Crozier⁷, Eileen Dillane⁵, Jonathan S. Ellis¹¹, Eva
García-Vázquez¹¹, Andrew M. Griffiths⁵, Sigurdur Gudjonsson¹², Kjetil Hindar¹³, Sten
Karlsson¹³, David Knox¹, Gonzalo Machado-Schiaffino¹¹, Dorte Meldrup¹⁴, Einar Nielsen¹⁴, Kristinn Ólafsson¹⁵, Craig R. Primmer¹⁶, Sergey Prusov¹⁷, Lee Stradmeyer¹, Juha-
Pekka Vähä¹⁸, Alexei Je. Veselov¹⁸, Lucy M.I. Webster¹⁰, Philip McGinnity² and Eric
Verspoor¹⁹

¹Marine Scotland, Freshwater Fisheries Laboratory, Faskally, Pitlochry PH16 4LB, UK
²Aquaculture & Fisheries Development Centre, School of Biological, Earth and Environmental
Sciences, University College, Cork, Ireland
³Institute of Marine Research, PO Box 1870 Nordnes, 5817 Bergen, Norway
⁴Institute for Global Food Security, School of Biological Sciences, Queen’s University, Belfast
⁵BT9 7BL, UK
⁶Department of Biosciences, Geoffrey Pope Building, University of Exeter, Stocker Road,
Exeter EX4 4QD, UK
⁷Department of Biosciences, Swansea University, Swansea, United Kingdom
⁸Agri-Food and Biosciences Institute Northern Ireland, Fisheries and Aquatic Ecosystems
Branch, Newforge Lane, Belfast BT9 5PX, UK
⁹GENINDEXE, 6 rue des Sports, 17000 La Rochelle, France
¹⁰IBERS, Aberystwyth University, Aberystwyth, United Kingdom
¹¹Rivers and Fisheries Trusts of Scotland (RAFTS), CBC House, 24 Canning Street, Edinburgh,
EH3 8EG, UK
¹²Departament of Functional Biology, Genetics, Universidad de Oviedo, C/Julian Claveria s/n,
33006 Oviedo, Spain
¹³Marine and Freshwater Research Institute, Skúlagata 4, 101 Reykjavik, Iceland
Running title: Stock identification of European Atlantic salmon
Abstract

Atlantic salmon *(Salmo salar* L.) populations of different river origins mix in the North Atlantic during the marine life stage. To facilitate stock identification in this environment, we developed a genetic baseline covering the European component of the species’ range from the Russian River Megra in the north-east, the Icelandic Elidaar in the west, and the Spanish Ulla in the south. Coverage extends over 3700 km North to South and over 2700 km East to West. The baseline encompasses 26,822 fish from 13 countries, 282 rivers and 467 sampling sites screened for 14 microsatellites. A hierarchical subdivision of regional genetic assignment units was defined using a combination of distance based and Bayesian clustering. A top level assignment level of three units was identified comprising Northern, Southern and Icelandic regions. A second assignment level was also defined composed of 18 regional units where individual assignments could be accurately performed and 29 units where accurate mixed stock estimates were possible. This baseline represents the most comprehensive population coverage for an Atlantic salmon genetic data-set, and constitutes a unique resource in the European marine fisheries context and is freely available to researchers to facilitate identification of the natal origin of European salmon.

Key words: Atlantic salmon, genetic stock identification, individual assignment, marine ecology, microsatellites

Introduction

The homing behaviour of Atlantic salmon together with other factors such as phylogeography, founder effects, isolation, selection and genetic drift, has resulted in local adaptations (Garcia de Leaniz et al., 2007) and significant population structuring at a hierarchy of levels from intra-river to inter-continental (King et al., 2001). Variation in marine migratory patterns of these various Atlantic salmon populations from different parts of the species range are known to occur but the full extent of differences among
populations and how this may be changing in response to shifting environmental conditions remains to be resolved (Jonsson et al., 2016). The study of population and stock-specific migration, distribution and feeding patterns, and their implications for marine mortality rates, and the impact of climate change, are hampered by a lack of information relating to the marine-phase of the lifecycle (Crozier et al., 2004). This makes it extremely difficult to mitigate anthropogenic influences on different stock components in mixed-stock fisheries or bycatches. Effective marine ecosystem management of Atlantic salmon could greatly benefit from accurate identification of genetically distinct populations and regional entities (MacKenzie et al., 2011) and the ability to discriminate the stock origins of fish in mixed feeding aggregations or during migratory phases.

Methods for identifying the region or river/tributary of origin of salmonids using DNA profiling have advanced significantly over recent decades and are now widely applied to stock management of Pacific salmon (Oncorhynchus spp.) (e.g. Shaklee et al., 1999; Beacham et al., 2004; Beacham et al., 2006; Shedd et al., 2016). Their application to Atlantic salmon stock management has been less extensive but has provided valuable insights into stock mixing at several scales including intercontinental (e.g. North American and European stocks in the West Greenland fishery: Gauthier-Ouellet et al., 2009), regional (e.g. stock composition in Canadian gill-net fisheries: Bradbury et al., 2016) and river level (e.g. population structuring in the river Teno: Vähä et al., 2016).

The assignment of salmon to continent of origin is now routine and increasingly cost-effective (Gauthier-Ouellet et al., 2009). On the Western side of the Atlantic, several genetic baselines are available (e.g. Bradbury et al., 2015; Sheehan et al., 2010) including a recently developed fine scale range-wide North American microsatellite baseline (Bradbury et al., 2016). Together, these allow the within-region identification of fish originating from
Western Atlantic populations at high geographic resolution. Although partial baselines have been developed on the Eastern side of the Atlantic (e.g. Griffiths et al., 2010; Verspoor et al., 2012; Ensing et al., 2013; Gilbey et al., 2016a; Vähä et al., 2016) there is no high resolution resource covering the entire species’ (non-Baltic) eastern Atlantic range. A DNA-based approach to the GSI of marine samples from the Eastern Atlantic would, in conjunction with ecological studies, provides a basis for advancing understanding of the migration and distribution patterns of Atlantic salmon. This would help to improve our knowledge of factors influencing marine mortality, and facilitate the implementation of more effective management programmes (Crozier et al., 2004).

Advances in DNA profiling over recent years have allowed for the development of genetic stock identification (GSI) using various genetic markers. Allozymes (Koljonen and McKinnell, 1996) and mitochondrial DNA (Moriya et al., 2007) have both been successfully used for stock identification in salmonid species including Atlantic salmon. However, the levels of resolution achieved with such markers have been insufficient for intra-regional discrimination. Historically microsatellites have been the genetic marker most widely used with Atlantic salmon. Various studies have screened numerous populations of salmon over many years resulting in high resolution coverage of many parts of the species’ range. As such, even with the development of other markers such as Single Nucleotide Polymorphisms, the large body of microsatellite data available provides a powerful resource for GSI. The use of microsatellite data does, however, come with certain problems (reviewed in Moran et al., 2006) which include: laboratories using different sets of markers; variations in allele-calling with different size markers or allele-size bins; different screening platforms; differences in chemistry; differences in the fluorophore; whether the forward or reverse primer is labelled; and differences in primer sizes. All of these can result in
inconsistent allele-size designations. Nevertheless, evidence from large-scale standardisation projects with these marker types among Pacific salmonid species such as *Oncorhynchus mykiss* (Stephenson et al., 2009) and *Oncorhynchus tshawytscha* (Seeb et al., 2007), together with previous studies in Atlantic salmon (e.g. Ellis et al., 2011), suggest that these issues can be overcome to construct comprehensive integrated genetic baselines (Moran et al., 2006).

This study builds on existing national and international microsatellite screening programmes to develop a comprehensive database of microsatellite variation. It contains data for a common set of 14 microsatellite loci for a geographically extensive range of rivers, spanning the species’ Eastern Atlantic European range from the Russian river Megra in the north-east (66.151 N, 41.484 W), to the Icelandic Ellidaar in the west (64.117 N, 21.833 E) and the Spanish Ulla river in the south (42.639 N, 8.761 E). Samples encompass rivers responsible for about ~85% of wild-salmon production in the eastern Atlantic (estimate based on rod-catch data from numerous sources). Baltic salmon populations are excluded from the baseline, as they do not migrate outside the Baltic Sea (Karlsson and Karlstrom, 1994; Torniainen et al., 2013), though one Baltic sample was included as a genetic out-group to represent this region. Existing and new data supplied by partners in a multi-laboratory trans-European consortium were calibrated (Ellis et al., 2011), subjected to stringent quality control and integrated to form the baseline.

The baseline was constructed to identify the region of origin of marine-phase salmon in the Eastern Atlantic, and a hierarchical approach was used to partition the baseline into genetically distinctive regional assignment units. The power and accuracy of assignment to these units were assessed using both simulations and test samples constructed by removing
fish from the dataset, and the utility of the baseline for regional assignment of salmon from the marine environment was evaluated.

Methods

Baseline samples

Samples were collected from 32,888 Atlantic salmon from 551 sites representing 325 rivers in 13 countries across Europe (Denmark, England, Finland [two rivers with outlets in Norway], France, Iceland, Ireland, Northern Ireland, Norway, Russia, Scotland, Spain, Sweden and Wales) (Fig. 1, Table 1, Supplementary data S1 & S2), including one Baltic River acting as a genetic out-group. Sampled sites spanned the entire eastern Atlantic range of the species covering a range of 3737 km from North to South and 2717 km from East to West.

Samples were collected from 1994 to 2010, with the majority collected in 2008-2009. In general, they were from juvenile fish, mostly parr and fry, but in some cases from smolts or mature salmon, sampled when returning to fresh water to spawn. Numbers sampled at a site ranged from 11 to 300 with a mean of 58, and rivers were characterised by 1 to 12 sites, depending largely on river size, with a mean number of sample sites per river of 1.7. Full details of sites are given in the Supplementary material (S1 & S2).

Genotyping

Microsatellite data were obtained from DNA extracted from tissue samples (typically fin clips or scales) screened by a consortium of 11 laboratories located across Europe (Table 1) for 14 of the 15 loci identified by a consortium of researchers and described by Olafsson et al. (2010). SsaD486 (King et al., 2005) was excluded from the analysis due to its lack of variation over much of the European range. The panel of 14 loci used here were SsaF43
(Sanchez et al., 1996), Ssa14, Ssa289 (McConnell et al., 1995), Ssa171, Ssa197, Ssa202

(O'Reilly et al., 1996) SSsp1605, SSsp2201, SSsp2210, SSsp2216, SSspG7 (Paterson et al.,

2004), SsaD144, SsaD157 (King et al., 2005) and SSsp3016 (unpublished, GenBank number

AY37820).

PCR conditions, thermocyclers and multiplexes varied across laboratories, as did
genotyping platforms, size standards and other chemistry employed. Genotyping details and
standardisation of genotype assignments among laboratories appear in Ellis et al. (2011). In
summary, two 96-well ‘control plates’ were prepared containing template DNA extracted
from samples representing the widest coverage of the range of S. salar as was practicable
and which covered sites from both the Eastern and Western Atlantic (Matis, Iceland). These
were subsampled and typed by each laboratory. Genotypes were submitted by each
member of the consortium to a single depository (Exeter University) where conversion
algorithms and standardised nomenclature were applied. For each locus, lists of allele
counts and sizes for each laboratory were aligned and cross-referenced for the sample
genotypes in the control plates. Standard allele scores were designated for each locus and
size differences between allele lists from each laboratory were determined, which allowed
laboratory specific standardisation rules to be defined. It should be noted that using this
approach not every possible allele was screened, but the approach did allow the individual
microsatellite bin ladders to be defined at each location. It cannot be ruled out therefore
that rare alleles or alleles affected by regional idles may be have been missed using such an
approach, although the coherence of the reference baseline produced (see below) suggests
this is unlikely to have been a major influencing factor.

Based on the standardisation rules, all data generated for baseline sites was converted
to the standard size ranges and stored in a single bespoke database for further analysis (see
Sib-ship analysis among individuals in each sample was investigated using \textit{COLONY} (Jones and Wang, 2010) and used to exclude all but one fish from each full-sib family in each sample prior to inclusion in the database. Fish with less than 10 microsatellite loci genotyped were removed from further analysis due to concerns with DNA and genotype quality. Sites with more than half of the loci out of Hardy-Weinberg equilibrium (examined in \textit{GENEPOP} 4.2.2; Rousset, 2008) (potentially not representative of a single population), or had less than 70\% of fish scored at all loci (potentially poor quality DNA and genotypes), or consisted of less than 30 individuals after quality control checks listed above (potential failure to provide accurate estimates of allele frequencies), were also removed. We estimated descriptive statistics with \textit{GenAlEx} 6 (Peakall and Smouse, 2006).

\textbf{Assignment units}

Assignment units were defined in an iterative way similar to that employed by Gilbey et al. (2016a). Initial units were first defined by a combination of distance based and Bayesian clustering. Individual assignment accuracies using these units were then examined and units where accuracies did not meet a predefined threshold were combined with units which saw reciprocal misassignments until all units had accuracies at or above the threshold level.

The distance based approach was based on a neighbour-joining tree (Saitou and Nei, 1987) constructed using Nei’s genetic distance $D_s$ (Nei et al., 1983) calculated in \textit{POPTREE2} (Takezaki et al., 2010) and visualised in \textit{MEGA7} (Kumar et al., 2016). The clustering approach was carried out in \textit{STRUCTURE} (Pritchard et al., 2000), using a burn-in of 100,000 and a run phase of 300,000 iterations during each application. Three replicates for each cluster number ($K$) were run with values of $K$ from 1 to 10. $K = 10$ was chosen as an upper limit after examination of the results of the runs while they were underway which showed in each case estimates true $K$ at the level under analysis had been identified by this point (see results).
Prior site information was incorporated into the analysis using the LOCPRIOR option. The smallest $K$ capturing the major structure in the dataset was defined by the $\Delta K$ method of Evanno et al. (2005), which was calculated using STRUCTURE HARVESTER (Earl and vonHoldt, 2012). Replicate membership coefficients were combined with CLUMPP (Jakobsson and Rosenberg, 2007) using the Full Search method.

We used a hierarchical approach, starting with the full dataset. Evanno et al. (2005) showed that STRUCTURE tends to capture the major structure in a reference dataset but that more fine scale structure may become evident if a hierarchical analysis is performed. In the current analysis, at each hierarchical level a STRUCTURE analysis was performed and the minimum best $K$ identified. The data were then split up into the cluster units and further STRUCTURE analysis performed on each one independently, as above. This was repeated at each hierarchical split until either single-river structuring was observed or geographical coherence of the clusters was lost.

Once both the distance-based and clustering analysis had been performed the degree to which the assignment units identified by each technique corresponded was examined. Where the same units were identified these were incorporated into the initial assignment unit panel. Where the two approaches had identified different units the smallest unit from either approach was incorporated into the initial assignment unit panel (for example if one technique had identified a single unit where the other had identified sub-units with this then the sub-units were added to the initial panel). In this way the smallest units identified by one or both technique were incorporated into the initial assignment unit test panel.

Once the initial assignment unit panel had been identified individual assignment accuracy was then calculated for each of these units (see below). If accuracy to a unit was at or above 80% the unit was retained in the panel. If accuracy was below this level the unit
was combined with other units to which reciprocal misassignments were taking place.

Accuracies were tested again and the process repeated until all units in the panel had individual assignment accuracies at or above the 80% level. Nei’s genetic distance $D_A$ (Nei et al., 1983) was calculated for all pairwise final assignment combinations using the Populations 1.2.3 software package (Langella, 1999).

**Assignment analysis**

**Individual assignment**

Individual assignment accuracy was calculated using Maximum likelihood based mixture analyses carried out using ONCOR (Kalinowski et al., 2007) with mixture proportions estimated using the EM algorithm and genotype probabilities calculated by the method of Rannala and Mountain (1997). Accuracies were based on fish randomly removed from the reference baseline and combined into a mixture file. A random 10% of fish were removed from each of the three top level assignment units identified (see results) resulting in a total of 2682 fish in the mixture file. For each fish the most likely assignment unit of origin and associated assignment probability was calculated. Fish with assignment probabilities below 0.8 were classified as unassigned and excluded from the analysis. Accuracy to the assignment units was then calculated with the remaining fish. Using such a cut-off means that fish whose origin is difficult to determine (low probability) are removed from the analysis and so potential accuracy can be increased (Gilbey et al., 2016a; Bekkevold et al., 2015). However, the application of cut-off scores also increases the proportion of unassigned fish (Gilbey et al., 2016a) and will thus influence apparent stock proportions if calculated from the individual assignments, and so this should never be performed. In order to estimate accurate stock proportions a Mixed Stock Analysis approach was therefore utilised (see below).
Simulated fishery mixtures were analysed in ONCOR and comprised sets of 100% simulated samples of fish from each assignment unit. Genotypic frequencies for each locus in each unit were re-sampled following Anderson et al. (2008). The 100% simulations were based on 1000 simulations of 200 fish per hierarchical assignment unit and the same simulated reference sample sizes as in the actual dataset.

**Mixed stock analysis**

Mixed stock proportions were calculated for each assignment unit. The same set of removed fish was used and mixture proportions estimated in ONCOR using conditional maximum likelihood (Millar, 1987) with confidence intervals calculated based on 1000 bootstraps.

**Equal proportions**

Mixed stock proportions were calculated for each assignment unit using simulated fishery mixtures with equal proportions of fish at each assignment unit in ONCOR. 100 fish were simulated for each unit and confidence intervals of the estimates calculated using 1000 bootstraps.

**Baseline coverage analysis – River removal**

A baseline rarely completely covers all possible source populations, and so some fish in fishery mixtures may be from populations either not sampled or included in the baseline. Hence, simulation analysis may overestimate the success rates of assignments of fish in an actual fishery due to being based only on samples from sites and rivers contained in the baseline (Waples et al., 2008). This issue was addressed using a further test panel and associated test baseline. A random 10% of the rivers in each assignment unit were removed from the baseline and used as test mixtures which were then assigned back to the
reconstructed baseline. All assignment units comprising more than one river had at least one river randomly removed (see Supplementary material S1 for details of sites and rivers removed). Fish in these ‘unrepresented’ mixture panels were thus from sites and rivers not included in the reconstructed baseline. In this way, we tested the capability of the baseline to reflect the regional signal of each assignment unit and to assign fish from sites and rivers not included in the baseline but from the assignment unit. This procedure was repeated at both assignment unit levels again using ONCOR with confidence intervals calculated based on 1000 bootstraps.

Results

Baseline QC

From a total of 551 sites sampled, 84 sites were removed, leaving 467 sites containing 26,822 fish representing 282 rivers in the final baseline (Table 1). From the 551 sites, 17 sites were removed as genotypes were not in H-W proportions, 51 had <70% of fish screened at all loci, and 15 had <30 individuals representing the site after correction for full-siblings and individual fish for which <10 loci could be reliable genotyped. A further site (a sample of adult rod-caught fish from the Norwegian river Flekkeelva in 2007) was removed due to extreme outlier behaviour in the STRUCTURE analysis (data not shown). Full details of sites are contained in Fig. 1, Table 1 and Supplementary data S1 & S2. Most loci across sites were highly variable with allele numbers ranging from 10 for Ssa14 to 46 for SsaD157 (mean 29.9). Additional descriptive and diversity estimates for each locus and site appear in Supplementary material S3.

Definition of initial assignment regions

13
A neighbour-joining tree of Nei’s $D_A$ is summarised in Fig. 2 with an expanded version with all nodes labelled detailed in Supplementary data S4 and full site level $D_A$ matrix in Supplementary data S5. A plot of $\Delta K$, and a map showing the geographic positioning of the clusters at each hierarchical STRUCTURE level are shown in Fig. 3. Assignment units as defined by POPTREE and STRUCTURE are compared in Supplementary data S6.

Both POPTREE and STRUCTURE identified three large regional groupings of sites covering the Northern, Southern and Icelandic regions and these will henceforth be referred to as the Level 1 assignment units. In general there was also good agreement between the two techniques in the lowest level units identified. POPTREE identified 26 distinct units (coloured differently in Fig. 2) and STRUCTURE identified 22 (lowest level splits on Fig. 3) and in 17 cases the same units were identified by both techniques (Supplementary data S6).

Using the lowest level split in from each technique a total of 29 units were identified for the initial Level 2 assignment accuracy tests (column 1 in Table 2, Supplementary data S6).

**Assignment analysis**

**Initial assignment accuracy**

Using the 2682 fish removed from the baseline, individual assignments were performed at Level 1 and at the initially defined Level 2 assignment units. At Level 1 accuracy of all fish to the Northern, Southern and Icelandic unit respectively was 90.8%, 92.7% and 99.5%. Using a probability cut-off score ≥ 0.8 this increased to 94.2%, 95.5% and 100% with 86.8%, 90.2% and 99.5% of fish being assigned using such a cut-off.

Assignment accuracy of fish with probability scores ≥ 0.8 to the Level 2 units was ≥ 80% in 19 of the 29 units (Table 2; for full breakdown of assignments at each Level 2 iterative level see Supplementary data S7). Assignment units which contained reciprocal misassignments were then combined resulting in 21 assignment units and accuracy
recalculated resulting in 18 of the 21 achieving accuracies ≥ 80%. A final round of assignment unit combination resulted in 18 assignment units at which assignment accuracies were all ≥ 80% (Table 2, Supplementary data S7). Initial assignment units at both levels are mapped in Fig. 1 with $D_x$ matrices detailed in Supplementary data S8.

The 100% simulations for each assignment unit showed robust estimates of stocks proportions at both assignment levels (Fig. 4). At Level 1, the mean estimates matched the estimated proportions extremely well with a maximum difference of just 0.3% between the actual and estimated values and all upper CI at 100%. At the initial Level 2 assignment units again showed relatively accurate estimates with an average difference between the estimated and actual mean proportions of 4.5%. However individual units did perform below this with a maximum difference with the West and Central Scotland level of 17.6%. At the first round of assignment unit combinations accuracies are seen to improve as expected with average and maximum differences between the estimated and actual mean proportions of 4.5% and 9.0%. These levels reduced to 1.9% and 8.0% respectively at the final assignment unit combination round.

Mixed stock analysis

The results of the MSA using the 2682 fish removed from the baseline and used as a fishery mixture are shown in Fig. 5A. At all assignment units within both assignment levels estimated proportions matched actual proportions (were within the CI bands) apart from a single unit in Level 2, South France/Spain where the upper CI was just 0.19 below the actual value. The estimates are also seen to be very precise with average CI bands of just 2.2 and a maximum of 4.7. Considering the high accuracy of the mixed stock estimates at this initial
assignment unit composition, no further assignment unit combinations are presented for mixed stock analysis.

**Equal proportions**

As with the previous analysis the equal proportion simulation shows excellent agreement between the actual and estimated proportions in the mixture (Fig. 5B). At Level 1 there is an average difference between actual and estimated of just 0.06% and a maximum of 0.09% and at Level 2 these two differences only rises to a mean difference of 0.4 and a maximum of 1.1%.

**Baseline coverage analysis – River removal**

The most demanding test of assignment capabilities of the baseline was the “river removal” test in which entire river systems were removed from the baseline and its fish assigned to region using the remaining rivers. However, even here relatively high levels of accuracy were obtained (Fig. 5C). There is an average difference between actual and estimated mixture proportions of just 1.9% and a maximum of 2.3% at Level 1 and 1.3% and 2.9% respectively at Level 2. At no time were significant proportions assigned to any of the six assignment units consisting of a single river which did not therefore have representatives in the mixture file (lower CI at zero in these units).

**Discussion**

The study presented here represents the largest analysis of Atlantic salmon population structure. The results demonstrate the utility of microsatellites to successfully assign Atlantic salmon from the NE Atlantic to regions of origin. The genetic baseline developed represents a powerful resource to better understand the biology of Atlantic salmon in the marine environment. The use of this resource may help to understand the causes of
differential mortality among salmon stocks and inform a more efficient management of
Atlantic salmon fisheries (Crozier et al., 2004).

Distance based and cluster analysis both revealed substantial hierarchical sub-
structuring of river populations of European and Icelandic salmon. At the highest level,
structure is related to large-scale geographical discontinuities between Scandinavia-Russia,
Iceland, and the southern region (Britain-Ireland-France-Denmark-Spain) populations. These
units are similar to those identified in previous analyses of population structure in salmon.
For example, King et al. (2001) showed with microsatellites an unambiguous separation of
Iceland, Norway and Scotland-Ireland-Spain (their Fig. 3), and Verspoor et al. (2005)
identified an Icelandic group together with a southern British Isles-Northern France group
using allozymes (although a more complex pattern was apparent in their analysis among the
more central range groups).

At the second level, two assignment units shared the largest average degree of
distinctiveness to other units and this is reflected in the fact that both are on the extreme
end of the neighbour-joining tree (Fig. 2). The Baltic unit had a mean $D_A$ of 0.236 to other
units (Supplementary data S8) and this significant differentiation to other European rivers
has been seen in previous studies (Bourret et al., 2013). This divergence is consistent with
the restricted migration of Baltic salmon (Karlsson and Karlstrom, 1994) and the Baltic Sea's
long history of geographical isolation (Bourret et al., 2013). Interestingly, a second
assignment unit, the English Chalk streams also shared this same very high mean $D_A$ of
0.236. Griffiths et al. (2010) and Ikediashi et al. (2018) also found these rivers highly
differentiated in comparison to others in the southern part of the European range but it is
perhaps unexpected that the degree of differentiation matches that of the Baltic when the
entire European and Icelandic range is examined.
Salmon populations in the Icelandic region segregated into two distinct units. This division into Northern and Western Icelandic units was also reported by Olafsson et al. (2014) and may reflect the patterns of recolonisation after the Last Glacial Maximum.

Initially the Northern Level 2 unit was subdivided into eleven geographically coherent second-level genetic clusters that match well with previously reported structure in this region. Bourret et al. (2013), using SNP markers, found separation of northern Norway and Russian rivers from the Norwegian and Swedish Atlantic coast rivers, and Kjærner-Semb et al. (2016) found separation of northern and southern Norwegian groupings. Within the northern Norway-Russian complex, Vähä et al. (2016) also found the same North Kola, Northern Norway and Russia-White sea units, as reported here. Their use of 33 microsatellites and a more comprehensive baseline coverage allowed them to define structure at further hierarchical levels within these groups, which was not apparent with the 14 microsatellites and site coverage used here.

The population structuring of rivers from across the part of the range covered by the Southern Level 1 unit into an initial sixteen Level 2 units is coherent with that reported by Griffiths et al. (2010). They used 12 microsatellites, 11 of which were also contained in the panel used here, on fish sampled from 57 rivers across the Southern region, but excluding rivers from the East coast of Scotland and Northern Ireland. They reported similar geographic patterns of genetic structure as in this study (their Fig. 2). Similar assignment units in France and Northern Spain appeared in both analyses and also broadly reflected allozyme-based regional differentiation (Verspoor et al., 2005).

Despite these similarities, differences were seen with some of the units and between the two methods identifying the assignment units. Griffiths et al. (2010) identified groupings stretching across both Scotland and Ireland (see their Fig. 2) and similar groups were
identified here using the STRUCTURE based approach (Fig. 3). However in the distance
based approach the various Scottish and Irish units were clearly separated (Fig. 2) and this
was also reflected in the final assignment units where accuracy of assignment could be
made between the two geographic areas. Nevertheless, misassignments were still evident
between the Irish and Scottish units (Supplementary data S7) suggesting a degree of
homology between the units.

Accurate assignments to the initial 32 Level 2 units was not possible using individual
assignments but was achieved when using a mixed stock fishery approach. Acceptable levels
of individual assignments could be made to some units using the initial split but some areas
proved problematic at this scale particularly within the UK/Ireland areas. This observation
reflects the differing power of the two techniques (Manel et al., 2005) and suggests that
when using the baseline for a particular purpose the required levels of both accuracy and
resolution should be defined \textit{a priori} and this will depend on the specific questions being
examined and the tools being utilised.

Overall, the two levels of genetic structure are geographically coherent and in basic
agreement with regional groups previously reported. This agreement over many studies and
across marker types suggests the higher level regional structuring identified is likely to be
ageographically and temporally robust. However, differentiation between the identified
regional units at the finer geographic scales may be influenced by human activities, such as
the transport and escape of fish from aquaculture facilities, stocking, habitat alteration,
fisheries-induced evolution, and indirect genetic changes from disease and ecological
disturbances.

Potential genetic changes resulting from these contemporary influences mean that
the temporal stability of contributing populations will require future monitoring. In a
previous examination of temporal stability on assignment of Atlantic salmon in the species’ southern European range (Griffiths et al., 2010), test samples collected 20 years before the baseline samples still showed predominant allocation back to region of origin. This finding suggests at least at the larger regional levels temporal stability may be temporarily stable, however this should not be assumed to be the case for all units, and a program of resampling should be incorporated into future developments using these reference baseline populations.

At Level 1 and the final Level 2 assignment units all tests of assignment power suggest high accuracies can be achieved with both individual assignments and mixed stock analysis. The use of an assignment probability cut-off of 0.8 for individual assignments will always improve assignment accuracies; however, this comes at the cost of the proportion of fish assigned. The actual cut-off used will thus depend on the situation under investigation and will be a decision for the investigator/manager at the time. Further the actual assignment units may also be varied. If reduced accuracies to some of the combined units are acceptable these may also be used in specific circumstances.

Based on the various assignment tests, the baseline described can be exploited to investigate patterns of ocean utilisation and associated differences in marine mortality at the regional stock level, however important quantitative variation linked to how individual population components use the ocean and which may affect mortality rates, also exists at the level of individual rivers within regions and among river tributaries (Barson et al., 2015).

Evaluation of river-specific problems in some contexts will require information at the individual river level, for which the current baseline may have limited usefulness. However, even if river level identification is problematic, identification of region of origin may allow further analysis using region specific baselines of higher resolution.
The identification of intra-regional population contributions in mixed samples will be facilitated by further increases in the coverage and resolution of the baseline. This is an ongoing process; for example, higher resolution is already being achieved in selected areas covered by the baseline reported here using other markers (Gilbey et al., 2016a; Ozerov et al., 2017; Vähä et al., 2016). Future baseline development will likely increase the coverage of the baseline reported here towards the estimated 2000 rivers in the North-East Atlantic Commission area. However, diminishing returns will apply given that the rivers currently in the baseline represent an estimated ~85% of the non-Baltic European adult salmon production.

Considerable value could be added by combining the European baseline reported here with North American information to provide a trans-ocean baseline to enable oceanic scale investigations. This has already begun in a limited way using a reduced set of microsatellite markers and shows promise in the ability to assign fish from the entire species' range (Gilbey et al., 2016b). A trans-Atlantic baseline is likely to benefit from identification of strategic, level specific, diagnostic markers for continental, regional and intra-river groupings.

**Supplementary data**

Supplementary material is available at the ICESJMS online version of the manuscript.

**Acknowledgments**
This work forms part of the SALSEA-Merge research project (Project No. 212529) and was funded by the European Union under theme six of the 7th Framework programme. It was also co-sponsored by the Atlantic Salmon Trust and the Total Foundation, who we thank for financial support. PMcG and JC were partly supported by the Beaufort Marine Research Award in Fish Population Genetics funded by the Irish Government under the Sea Change Programme. The work was also supported under financial support of the program of fundamental research of Presidium of RAS “Searching fundamental scientific investigations in the interests of development of the Arctic zone of Russian Federation”. The authors also thank the numerous people responsible across Europe whom helped collect samples and assisted with the laboratory analyses. The manuscript benefited greatly from editorial and reviewer comments on an earlier draft and the authors wish to express thanks for their time and comprehensive inputs.

References


divergence across the natural range of Atlantic salmon (Salmo salar). Molecular Ecology, 22: 19.


Rousset, F. 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources, 8: 103-106.


Table 1. Sample baseline coverage pre and post genotype quality control (see text for details).

<table>
<thead>
<tr>
<th>Country</th>
<th>Pre-QC</th>
<th></th>
<th>Post-QC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rivers</td>
<td>Sites</td>
<td>Fish</td>
<td>Rivers</td>
</tr>
<tr>
<td>Denmark(^1)</td>
<td>3</td>
<td>6</td>
<td>253</td>
<td>2</td>
</tr>
<tr>
<td>England(^2,3)</td>
<td>24</td>
<td>38</td>
<td>1652</td>
<td>23</td>
</tr>
<tr>
<td>Finland(^4)</td>
<td>2</td>
<td>5</td>
<td>395</td>
<td>2</td>
</tr>
<tr>
<td>France(^2,3,5,6)</td>
<td>13</td>
<td>16</td>
<td>759</td>
<td>9</td>
</tr>
<tr>
<td>Iceland(^7)</td>
<td>17</td>
<td>25</td>
<td>2352</td>
<td>16</td>
</tr>
<tr>
<td>Ireland(^8)</td>
<td>29</td>
<td>45</td>
<td>2345</td>
<td>29</td>
</tr>
<tr>
<td>Northern Ireland(^9)</td>
<td>9</td>
<td>20</td>
<td>1469</td>
<td>7</td>
</tr>
<tr>
<td>Norway(^4,10,11)</td>
<td>90</td>
<td>109</td>
<td>7749</td>
<td>81</td>
</tr>
<tr>
<td>Russia(^4,12)</td>
<td>33</td>
<td>36</td>
<td>2506</td>
<td>30</td>
</tr>
<tr>
<td>Scotland(^3)</td>
<td>87</td>
<td>230</td>
<td>11625</td>
<td>69</td>
</tr>
<tr>
<td>Spain(^6)</td>
<td>7</td>
<td>7</td>
<td>342</td>
<td>4</td>
</tr>
<tr>
<td>Sweden(^1,4)</td>
<td>4</td>
<td>4</td>
<td>180</td>
<td>4</td>
</tr>
<tr>
<td>Wales(^2)</td>
<td>7</td>
<td>10</td>
<td>375</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>325</td>
<td>551</td>
<td>32002</td>
<td>282</td>
</tr>
</tbody>
</table>

Institutions contributing data: \(^1\) Danish Institute for Fisheries Research, Denmark; \(^2\) University of Exeter, England; \(^3\) Marine Scotland Science, Scotland; \(^4\) University of Turku, Finland; \(^5\) Geneindex, France; \(^6\) University of Oviedo, Spain; \(^7\) Marine and Freshwater Research Institute, Iceland; \(^8\) University College Cork, Ireland; \(^9\) Queen’s University Belfast & Agri-Food and Biosciences Institute Northern Ireland, Northern Ireland; \(^10\) Institute of Marine Research, Norway; \(^11\) Norwegian Institute for Nature Research, Norway; \(^12\) Knipovich Polar Research Institute of Marine Fisheries & Oceanography, Russia.