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

Gilbey, John; Coughlan, Jamie; Wennevik, Vidar; Pro Kohl, Paulo; Stevens, Jamie; de Leaniz, Carlos Garcia; Ensing, Dennis; Cauweliers, Eef; Cherbonnel, Corrine; Consuegra, Sofia; Coulson, Mark; Cross, Tom; Crozier, Walter; Dillane, E.; Ellis, J. S.; Garcia-Vazquez, E.; Griffiths, A. M.; Gudjonsson, Sigurdur; Hindar, Kjetil; Karlsson, Sten

Published in:
ICES Journal of Marine Science

Publication date:
2018

Publisher rights:
Copyright © 2017 ICES/CIEM

The re-use license for this item is:
CC BY-NC

The Document Version you have downloaded here is:
Peer reviewed version

The final published version is available direct from the publisher website at:
10.1093/icesjms/fsx184

Link to author version on UHI Research Database

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the UHI Research Database are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights:

1) Users may download and print one copy of any publication from the UHI Research Database for the purpose of private study or research.
2) You may not further distribute the material or use it for any profit-making activity or commercial gain.
3) You may freely distribute the URL identifying the publication in the UHI Research Database.

Take down policy
If you believe that this document breaches copyright please contact us at RO@uhi.ac.uk providing details; we will remove access to the work immediately and investigate your claim.

Download date: 05. Mar. 2020
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, Mark W. Coulson10, Tom F. Cross2, Walter Crozier7, Eileen Dillane2, Jonathan S. Ellis5, Eva García-Vázquez11, Andrew M. Griffiths5, Sigurdur Gudjonsson12, Kjetil Hindar13, Sten Karlsson13, David Knox1, Gonzalo Machado-Schiaffino11, Dorte Meldrup14, Einar Eg Nielsen14, Kristinn Ólafsson15, Craig R. Primmer16, Sergey Prusov17, Lee Stradmeyer1, Juha-Pekka Vähä18, Alexei Je. Veselov18, Lucy M.I. Webster10, Philip McGinnity2A and Eric Verspoor19A

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 BT9 7BL, UK
5Department of Biosciences, Geoffrey Pope Building, University of Exeter, Stocker Road, Exeter EX4 4QD, UK
6Department of Biosciences, Swansea University, Swansea, United Kingdom
7Agri-Food and Biosciences Institute Northern Ireland, Fisheries and Aquatic Ecosystems Branch, Newforge Lane, Belfast BT9 5PX, UK
8GENINDEXE, 6 rue des Sports, 17000 La Rochelle, France
9IBERS, Aberystwyth University, Aberystwyth, United Kingdom
10Rivers and Fisheries Trusts of Scotland (RAFTS), CBC House, 24 Canning Street, Edinburgh, EH3 8EG, UK
11Departament of Functional Biology, Genetics, Universidad de Oviedo, C/Julian Claveria s/n, 33006 Oviedo, Spain
12Marine and Freshwater Research Institute, Skúlagata 4, 101 Reykjavik, Iceland
Running title: Stock identification of European Atlantic salmon
Abstract

Atlantic salmon (Salmo salar L.) populations from different river origins mix in the North Atlantic during the marine life stage. To facilitate marine stock identification, we developed a genetic baseline covering the European component of the species’ range from the Russian River Megra in the north-east, the Icelandic Ellidaar in the west, and the Spanish Ulla in the south, spanning approximately 3700 km North to South and 2700 km East to West. The baseline encompasses data for 14 microsatellites for 26,822 individual fish from 13 countries, 282 rivers and 467 sampling sites. A hierarchy of regional genetic assignment units was defined using a combination of distance based and Bayesian clustering. At the top level subdivision is into Northern, Southern, and Icelandic regions. At the next level of regional subdivision where individual assignments could be accurately there are 18 geographical units and 29 units where accurate mixed stock estimates were possible. The baseline provides the most comprehensive geographical coverage for an Atlantic salmon genetic data-set, and a unique resource in the European marine fisheries context. It 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

Homing of Atlantic salmon to natal rivers, in combination with factors such as founder effects, isolation, selection and genetic drift, and broad scale phylogeographic processes, has resulted significant population structuring at a hierarchy of levels from intra-river to inter-continental (King et al., 2001) and locally adapted populations (Garcia de Leaniz et al., 2007) including variations in marine migratory patterns among populations from different parts of the species range (Webb et al. 2007). However, the full extent of differences in
migratory patterns among populations and how this may be changing in response to shifting environmental conditions remains to be resolved (Jonsson et al., 2016).

Advancing understanding of population and stock-specific migration, distribution and feeding patterns, and their implications for marine mortality rates, and the impact of climate change, is hampered by a lack of information relating to the marine-phase of the lifecycle (Crozier et al., 2004). This situation makes it extremely difficult to appropriately target actions to mitigate anthropogenic influences on different stock components e.g. the impacts of mixed-stock fisheries and bycatches. Thus a tool that allows the accurate identification of genetically distinct populations and regional entities (MacKenzie et al., 2011) and discrimination of the stock origins of fish in mixed feeding aggregations or during migratory phases would be invaluable in species’ and North Atlantic marine ecosystem management.

DNA profiling methods for identifying the region or river/tributary of origin of salmonids have advanced significantly over recent decades and widely applied to Pacific salmon (Oncorhynchus spp.) stock management (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 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). However, overall, its use has been more limited due to the lack of useful genetic baselines for many parts of the species range.

Genetic baselines are available for the Western side of the Atlantic (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), that facilitate within-region identification of fish originating from Western Atlantic populations at high geographic resolution. In contrast, only partial baselines have been developed for 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) and no high resolution baseline exists for the species’ non-Baltic, eastern Atlantic range. Yet a such a baseline would allow a DNA-based approach to the GSI of marine samples from the Eastern Atlantic and, in conjunction with ecological studies, would help to provide a more detailed understanding of variations in the North Atlantic migration and distribution patterns of different European Atlantic salmon stocks. Such insight could improve understanding of the factors conditioning marine mortality, and facilitate the implementation of more effective management programmes (Crozier et al., 2004).

Genetic stock identification (GSI) has been carried out using various genetic markers with early work successfully using allozymes (Koljonen and McKinnell, 1996) and mitochondrial DNA (Moriya et al., 2007) for salmonid species in some contexts, including for Atlantic salmon. However, higher levels of resolution and more widespread application has been subsequently achieved using microsatellite loci and they became the genetic marker most widely used in studies of Atlantic salmon stock differentiation, with high resolution coverage of many parts of the species’ range (REFs). Even though, more recently, attention has turned to SNPs (Single Nucleotide Polymorphisms) due to their analytical advantages (REFs), the existing large body of microsatellite data available remains a unique and powerful resource that can be exploited for GSI. However, it also has limitations (reviewed in Moran et al., 2006) related to 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 markers across loci and whether the forward or reverse primer is labelled as well as differences in primer sizes. All of these can result in inconsistent allele-size designations across data sets generated by different laboratories. However, evidence from large-scale standardisation projects for salmonid species such as *Oncorhynchus mykiss* (Stephenson et al., 2009) and *O. tshawytscha* (Seeb et al., 2007), as well as Atlantic salmon (e.g. Ellis et al., 2011), indicate these issues can be addressed and comprehensive, large scale integrated genetic baselines constructed (Moran et al., 2006).

Described here is a trans-European GSI baseline for Atlantic salmon (excluding Baltic salmon stocks which do not migrate to the North Atlantic) constructed by linking existing national and international microsatellite screening programmes. 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). Data was integrated for a common set of 14 microsatellite loci for a geographically representative set 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). Baseline samples encompassed rivers responsible for about ~85% of wild-salmon production in the study region, based on rod-catch data derived from numerous sources. 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. A hierarchical assignment unit approach was used and the baseline resolved into genetically distinctive regional assignment units. Assignment power and accuracy to these units were assessed, using both simulations and test samples, the latter constructed by removing fish from the dataset, to establish the
utility of the baseline for regional assignment of marine-phase European origin salmon in
the North Atlantic.

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 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 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 GenAlEx 6 (Peakall and Smouse, 2006).

**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 $D_A$ (Nei et al., 1983) calculated in POPTREE2 (Takezaki et al., 2010) and visualised in MEGA7 (Kumar et al., 2016). The clustering approach was carried out in 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 2682 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**
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 $\geq 0.8$ 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 $\geq 0.8$ to the Level 2 units was $\geq 80\%$ 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 $D_A$ matrixes detailed in Supplementary data S8.

100% simulations

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 rise 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, encompassing the largest analysis of Atlantic salmon population structure in the Eastern Atlantic, for the first time, provides a genetic framework to exploit the power of microsatellite variation to assign Atlantic salmon from this part of the species’ range to smaller scale regional stock groups. As such, the reported genetic baseline provides a powerful resource that can be used to increase understanding of the biology of European
Atlantic salmon stocks in the North Atlantic marine environment, such as the causes of differential mortality among salmon stocks, and help to inform actions to achieve a more efficient management of Atlantic salmon fisheries (Crozier et al., 2004).

Distance-based and Bayesian cluster analyses both reveal hierarchical structuring of river populations of European and Icelandic salmon into regional groups. At the highest level, this structure encompasses large-scale geographical discontinuities between Scandinavia-Russia, Iceland, and the southern region (Britain-Ireland-France-Denmark-Spain) populations, and have been identified in previous analyses of Atlantic salmon population structure. 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 next highest level, two assignment units shared the largest average degree of distinctiveness from other units, the two also being on opposite extremes of the neighbour-joining tree (Fig. 2). The Baltic unit had a mean $D_A$ of 0.236 to other units (Supplementary data S8), a level of differentiation to other European rivers seen in previous studies (e.g. Bourret et al., 2013) and consistent with the restricted migration of Baltic stocks (Karlsson and Karlstrom, 1994) and their long history of geographical isolation (Bourret et al., 2013). A second assignment unit, the English Chalk streams, also shared a similarly high mean $D_A$ of 0.236. Griffiths et al. (2010) also reported these rivers to be highly differentiated from others in the southern part of the European range. However, it is unexpected in the context of the entire European and Icelandic range, that the degree of differentiation matches that
of the Baltic. At the same time, the salmon populations in Iceland segregate into Northern and Western Icelandic units as was also reported by Olafsson et al. (2014).

Initially the Northern Level 2 unit 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. However, their use of 33 microsatellites and a more comprehensive geographical coverage allowed them to define structure at further hierarchical levels within these groups unresolved in the present study using only 14 microsatellites and a more limited population coverage.

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 accords with that reported by Griffiths et al. (2010) based on 12 microsatellites, 11 of which form part of in the panel used in the present study. Their study encompassed fish from 57 rivers across the Southern region but excluded rivers from the East coast of Scotland and Northern Ireland an showed similar geographic patterns of genetic structure (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). However, some differences were seen with some of the units between the two methods used to resolve 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). In contrast, using the distance-based approach the various Scottish and Irish units were clearly
separated (Fig. 2) to which generally good assignments of fish could be made. Nevertheless,
some misassignment was still evident (Supplementary data S7) suggesting a degree of
homology between the units.

Accurate assignments to the initial 32 Level 2 units was not possible at the individual
level but was achieved for mixed stock fishery estimates. Acceptable levels of individual
assignments could be made to some defined units using the initial split but some areas
proved problematic at this scale particularly for Britain and Ireland. This difference 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 a priori. In turn, this will depend on the specific questions
being examined and the tools being utilised.

Overall, the two levels of genetic structure are geographically consistent and in basic
agreement with major regional phylogeographic groups previously reported using a variety
of markers, suggesting the higher level regional structuring is geographically and temporally
robust. In contrast, differentiation between regional units identified at the finer geographic
scales may in part be conditioned 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. Such genetic structuring,
if defined by such contemporary influences, may not have temporal stability and such lower
level units thus will need to be monitored to determine if they are stable. Encouragingly, in
a previous assessment 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 scale regional level units are likely to be temporarily stable.
However, this should not be assumed to always be the case and a program of resampling should be incorporated if the baseline is exploited in the future. Assuming temporal stability, at Level 1 and the final Level 2 regional units, all tests of power suggest high accuracies can be achieved with both individual assignments and mixed stock analysis. Accuracies can be further improved by use of a probability cut-off of 0.8 for individual assignments, which may be useful in some contexts. However, this will reduce the proportion of fish assigned. Thus in application, the best cut-off will depend on the question address and will need to be decided by each individual user. This will also apply to the assignment units used; if reduced accuracies to some of the combined units are acceptable these may also be used in specific circumstances.

The assignment tests carried out indicate that the described baseline can be exploited to help investigate patterns of ocean utilisation and associated differences in marine mortality operating at the regional stock level. However, important quantitative variation linked to how individual population components use the ocean, 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, likely to exist in some contexts, will require assignments at the individual river level, for which the current baseline appears to have limited usefulness. However, even if river level identification is problematic, identification of region of origin may allow finer scale analysis using higher resolution region specific baselines.

Resolution of intra-regional population contributions in mixed oceanic samples would be facilitated by further increases in the coverage and resolution of the baseline. 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.,
Ideally, future work will likely increase baseline coverage towards the estimated 2000 rivers in the North-East Atlantic Commission area. However, this will involve diminishing returns given that the rivers currently in the baseline represent an estimated ~85% of the non-Baltic European adult salmon production. However, 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 started 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).

**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


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></th>
<th></th>
<th>Post-QC</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rivers</td>
<td>Sites</td>
<td>Fish</td>
<td></td>
<td>Rivers</td>
<td>Sites</td>
<td>Fish</td>
<td></td>
</tr>
<tr>
<td>Denmark¹</td>
<td>3</td>
<td>6</td>
<td>253</td>
<td></td>
<td>2</td>
<td>4</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>England²,³</td>
<td>24</td>
<td>38</td>
<td>1652</td>
<td></td>
<td>23</td>
<td>35</td>
<td>1498</td>
<td></td>
</tr>
<tr>
<td>Finland⁴</td>
<td>2</td>
<td>5</td>
<td>395</td>
<td></td>
<td>2</td>
<td>5</td>
<td>393</td>
<td></td>
</tr>
<tr>
<td>France²,³,⁵,⁶</td>
<td>13</td>
<td>16</td>
<td>759</td>
<td></td>
<td>9</td>
<td>9</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>Iceland⁷</td>
<td>17</td>
<td>25</td>
<td>2352</td>
<td></td>
<td>16</td>
<td>22</td>
<td>1986</td>
<td></td>
</tr>
<tr>
<td>Ireland⁸</td>
<td>29</td>
<td>45</td>
<td>2345</td>
<td></td>
<td>29</td>
<td>40</td>
<td>2053</td>
<td></td>
</tr>
<tr>
<td>Northern Ireland⁹</td>
<td>9</td>
<td>20</td>
<td>1469</td>
<td></td>
<td>7</td>
<td>18</td>
<td>1302</td>
<td></td>
</tr>
<tr>
<td>Norway⁴,ⁱ⁰,¹¹</td>
<td>90</td>
<td>109</td>
<td>7749</td>
<td></td>
<td>81</td>
<td>99</td>
<td>7008</td>
<td></td>
</tr>
<tr>
<td>Russia⁴,¹⁰,¹²</td>
<td>33</td>
<td>36</td>
<td>2506</td>
<td></td>
<td>30</td>
<td>33</td>
<td>2350</td>
<td></td>
</tr>
<tr>
<td>Scotland²</td>
<td>87</td>
<td>230</td>
<td>11625</td>
<td></td>
<td>69</td>
<td>185</td>
<td>8884</td>
<td></td>
</tr>
<tr>
<td>Spain⁶</td>
<td>7</td>
<td>7</td>
<td>342</td>
<td></td>
<td>4</td>
<td>4</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>Sweden¹,⁴</td>
<td>4</td>
<td>4</td>
<td>180</td>
<td></td>
<td>4</td>
<td>4</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>Wales²</td>
<td>7</td>
<td>10</td>
<td>375</td>
<td></td>
<td>6</td>
<td>9</td>
<td>347</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>325</td>
<td>551</td>
<td>32002</td>
<td></td>
<td>282</td>
<td>467</td>
<td>26822</td>
<td></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.
Table 2. Individual assignment accuracy using fish removed from the reference baseline. Initial assignment units in first column defined by distance and Structure based analysis. Remaining assignment represent amalgamations of units where assignment accuracy is <80%. Assignment accuracy calculated using only fish with individual assignment probabilities ≥ 0.8. Values in bold represent accuracy of at least 80% to assignment units. Sample sizes represent baseline/mixture size.

<table>
<thead>
<tr>
<th>Assignment unit</th>
<th>Sample size</th>
<th>Assigned %</th>
<th>Correct %</th>
<th>Assignment unit</th>
<th>Assigned %</th>
<th>Correct %</th>
<th>Assignment unit</th>
<th>Assigned %</th>
<th>Correct %</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Sea</td>
<td>758/86</td>
<td>68.6</td>
<td><strong>90.2</strong></td>
<td>White Sea</td>
<td>70.9</td>
<td><strong>90.3</strong></td>
<td>White Sea</td>
<td>72.1</td>
<td><strong>90.3</strong></td>
</tr>
<tr>
<td>Kola</td>
<td>1561/160</td>
<td>50</td>
<td><strong>82.1</strong></td>
<td>Kola</td>
<td>51.9</td>
<td><strong>82.1</strong></td>
<td>Kola</td>
<td>53.1</td>
<td><strong>82.1</strong></td>
</tr>
<tr>
<td>Kola (Tuloma Basin)</td>
<td>287/39</td>
<td>61.5</td>
<td><strong>100</strong></td>
<td>Kola (Tuloma Basin)</td>
<td>66.7</td>
<td><strong>96</strong></td>
<td>Kola (Tuloma Basin)</td>
<td>66.7</td>
<td><strong>96</strong></td>
</tr>
<tr>
<td>Finnmark</td>
<td>1109/107</td>
<td>54.2</td>
<td><strong>84.7</strong></td>
<td>Finnmark</td>
<td>59.3</td>
<td><strong>82.9</strong></td>
<td>Finnmark</td>
<td>59.3</td>
<td><strong>82.9</strong></td>
</tr>
<tr>
<td>Tana</td>
<td>271/28</td>
<td>42.9</td>
<td>10</td>
<td>Mid挪威</td>
<td>68.3</td>
<td><strong>84.4</strong></td>
<td>Mid &amp; SW挪威</td>
<td>69.2</td>
<td><strong>84.4</strong></td>
</tr>
<tr>
<td>Mid Norway</td>
<td>3195/369</td>
<td>54.5</td>
<td><strong>84.1</strong></td>
<td>South Norway</td>
<td>45.8</td>
<td><strong>82.4</strong></td>
<td>South Norway</td>
<td>47.0</td>
<td><strong>82.4</strong></td>
</tr>
<tr>
<td>South West Norway</td>
<td>816/95</td>
<td>42.1</td>
<td>73.8</td>
<td>Enningdalselva</td>
<td>87.5</td>
<td><strong>100</strong></td>
<td>Enningdalselva</td>
<td>87.5</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>South Norway</td>
<td>693/83</td>
<td>32.5</td>
<td><strong>81.25</strong></td>
<td>Sweden</td>
<td>41.7</td>
<td><strong>100</strong></td>
<td>Sweden</td>
<td>41.7</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>Finmark</td>
<td>1109/107</td>
<td>54.2</td>
<td><strong>84.7</strong></td>
<td>Baltic</td>
<td>80.0</td>
<td><strong>100</strong></td>
<td>Baltic</td>
<td>80.0</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>Denmark</td>
<td>176/13</td>
<td>61.5</td>
<td>100</td>
<td>Denmark</td>
<td>76.9</td>
<td>100</td>
<td>Denmark</td>
<td>76.9</td>
<td>100</td>
</tr>
<tr>
<td>Central Scotland/North England</td>
<td>1711/200</td>
<td>32</td>
<td>73.5</td>
<td>Scotland/North East England</td>
<td>66.3</td>
<td><strong>80.4</strong></td>
<td>Scotland/North East England</td>
<td>76.4</td>
<td><strong>87.2</strong></td>
</tr>
<tr>
<td>North East Scotland</td>
<td>2183/233</td>
<td>42.5</td>
<td>56.5</td>
<td>Leven</td>
<td>82.9</td>
<td><strong>96.9</strong></td>
<td>Leven</td>
<td>85.4</td>
<td><strong>96.9</strong></td>
</tr>
<tr>
<td>Kyle/Ness</td>
<td>814/99</td>
<td>42.4</td>
<td>78.7</td>
<td>English Chalk</td>
<td>88.9</td>
<td><strong>100</strong></td>
<td>English Chalk</td>
<td>100</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>North &amp; West Scotland</td>
<td>2005/255</td>
<td>35.7</td>
<td>72</td>
<td>North France</td>
<td>78.9</td>
<td>80.0</td>
<td>North France</td>
<td>68.0</td>
<td><strong>91.7</strong></td>
</tr>
<tr>
<td>Water of Luce</td>
<td>225/20</td>
<td>30</td>
<td>40</td>
<td>South France/Spain</td>
<td>72.5</td>
<td><strong>100</strong></td>
<td>South France/Spain</td>
<td>100</td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>West Central Scotland</td>
<td>242/28</td>
<td>46.4</td>
<td><strong>83.3</strong></td>
<td>North Ireland</td>
<td>59.6</td>
<td><strong>85.9</strong></td>
<td>North Ireland</td>
<td>64.0</td>
<td><strong>87.0</strong></td>
</tr>
<tr>
<td>Irish Sea</td>
<td>1992/214</td>
<td>39.7</td>
<td>77.3</td>
<td>South West Ireland</td>
<td>52.2</td>
<td>77.4</td>
<td>South West Ireland</td>
<td>93.3</td>
<td><strong>98.7</strong></td>
</tr>
<tr>
<td>Leven</td>
<td>324/41</td>
<td>75.6</td>
<td>100</td>
<td>Ireland</td>
<td>66.7</td>
<td><strong>93.9</strong></td>
<td>Ireland</td>
<td>66.7</td>
<td><strong>93.9</strong></td>
</tr>
<tr>
<td>English Chalk</td>
<td>134/9</td>
<td>88.9</td>
<td>100</td>
<td>North Iceland</td>
<td>95.5</td>
<td><strong>96.3</strong></td>
<td>North Iceland</td>
<td>95.5</td>
<td><strong>96.3</strong></td>
</tr>
<tr>
<td>North France</td>
<td>283/35</td>
<td>45.7</td>
<td>79.8</td>
<td>West Iceland</td>
<td>92.1</td>
<td><strong>98.7</strong></td>
<td>West Iceland</td>
<td>93.3</td>
<td><strong>98.7</strong></td>
</tr>
</tbody>
</table>
Figure 1. Map of sampling region. Points represent sample sites and/or river mouths. Full site information is contained in Supplementary data S1 and an expanded map with all rivers identified is in Supplementary data S2. Regions noted are all those referred to in the text. The Level one assignment units (see text) are delineated by the dashed line and the initial Level 2 units by coloured points.

Figure 2. Neighbour-joining (Saitou and Nei, 1987) phylogenetic tree of sample sites based on $D_A$ (Nei et al., 1983). Major clusters are coloured and named according to the regions detailed on Figure 1. Expanded tree with all sites identified is detailed in Supplementary data S4.

Figure 3. Hierarchical STRUCTURE based clustering analysis of sites. Each cluster analysis is described using three components. Firstly the results of the STRUCTURE analysis are shown with vertical bars representing individual sites and colours relating to cluster membership of that site. A plot of the $\Delta K$ values (Earl and vonHoldt, 2012) associated with the analysis is also shown defining the $K$ identified in that cluster analyses. Finally a map is shown detailing the geographic location of the clusters identified. Cluster names in italics refer to clusters for which further hierarchical analysis was performed. Cluster names in regular text refer to final cluster assignment groups.

Figure 4. Proportion estimates from independent 100% simulation studies of the genetic baseline at Level 1 and all stages of the iterative formation of the Level 2 assignment unit levels. Points represent mean estimates with bars showing 95% confidence intervals.
Figure 5. A) Mixed stock fishery estimates using fish removed from the baseline and used as fishery mixtures. B) Mixed stock fishery estimates using simulated equal proportions of fish from each assignment unit in the mixture. C) Mixed stock fishery estimates using entire rivers removed from the baseline and used as fishery mixtures. Dark bars represent actual proportions in the mixture files and grey bars ONCOR estimates. Bars represent mean estimates with 95% confidence intervals around these estimates. NOTE change of Y-axis scale for the Level 1 and 2 assignment levels.
Fig. 1.
Fig. 3.