The Scottish Government
Riaghaltas na h-Alba

Scottish Marine and Freshwater Science

Volume 5 Number 5

Identifying a Panel of Single Nucleotide Genetic Markers for River Level Assignments of Salmon in Scotland

marine scotland

Scottish Marine and Freshwater Science Vol 5 No 5

IDENTIFYING A PANEL OF SINGLE NUCLEOTIDE GENETIC MARKERS FOR RIVER LEVEL ASSIGNMENTS OF SALMON IN SCOTLAND

Published by Marine Scotland Science

ISSN: 2043-7722

Marine Scotland is the directorate of the Scottish Government responsible for the integrated management of Scotland's seas. Marine Scotland Science (formerly Fisheries Research Services) provides expert scientific and technical advice on marine and fisheries issues. Scottish Marine and Freshwater Science is a series of reports that publishes results of research and monitoring carried out by Marine Scotland Science. It also publishes the results of marine and freshwater scientific work that has been carried out for Marine Scotland under external commission. These reports are not subject to formal external peer-review.

This report presents the results of marine and freshwater scientific work carried out by Marine Scotland Science.

Marine Scotland Science
Freshwater Laboratory
Faskally
Pitlochry
PH16 5LB

Copies of this report are available from the Marine Scotland website at
www.scotland.gov.uk/marinescotland

IDENTIFYING A PANEL OF SINGLE NUCLEOTIDE GENETIC MARKERS FOR RIVER LEVEL ASSIGNMENTS OF SALMON IN SCOTLAND

Executive Summary

This report examines utility of assigning fish to river in Scotland using Single Nucleotide Polymorphic (SNP) markers. The analysis has shown that, depending on baseline coverage, it is possible using SNP markers to be able to assign fish to both region, and where baseline coverage is sufficient, river with high accuracy in most, but not all, situations investigated. Exclusion techniques have also been examined as to their effectiveness in screening out fish from reporting regions not represented in the baseline.

The procedures developed and outlined here show a multi-stage process can be employed to assign fish to origin. Firstly fish are assigned to a regional reporting unit. Fish from rivers not represented in the baseline within regional reporting units of interest are then excluded from the analysis, before finally fish are assigned to their river of origin. It has been shown that this can be achieved in most cases analysed with high accuracy (i.e. from 90 to 100 %), although there are situations where differentiating between rivers is problematic (Spey/Dee and rivers within the Kyle of Sutherland).

The study has also identified areas where further SNP baseline coverage is required associated with both regional and river level SNPs. Regional coverage of particularly the north and west requires enhancing as do rivers of particular interest on the East coast.

Introduction

The Scottish Government aims to rapidly increase Scotland’s scope for generating Marine Renewable Energy (MRE). This aim will require the development and installation of novel engineered structures in a number of coastal locations around Scotland. It is not yet possible to predict, with certainty, the specifications of the devices that will be deployed. However, there is a need, and in some cases, a legal imperative, to consider the effects these developments may have on migratory fishes. As such, identification of the natal origin of fish and knowledge of the migration routes of specific stocks will be of great benefit.

Together with the installation of MRE devices, there are a number of interceptory fisheries operating along migratory routes of fish as they return through coastal waters to spawn, generally in their home river. Such fisheries have the potential to differentially impact salmon from a number of different populations, rivers or stocks (defined as “an exploited or managed unit”: Royce, 1984), as differences in marine migratory patterns of stocks from different parts of the species’ range are known to occur, though the full extent of differences among stocks remains to be resolved (Webb *et al.*, 2007; Thorstad *et al.*, 2011). Determination of the exploitation rates within such fisheries of fish from different rivers will again greatly aid management, such that specific stocks, which may be at or beyond their conservation limits, or which have specific legal protection (e.g. from Special Areas of Conservation), can be identified. Management of the fishery, taking into account the exploitation rates of the different stocks, could then be undertaken.

The ability to assign fish to natal origin depends on the degree of genetic differences between different stocks within and between regions and rivers and on the resolving power of the genetic baseline utilised. Genetic differences between assignment units can be resolved using a number of different genetic markers. Typically, microsatellite markers have been used for this purpose (e.g. see Griffiths *et al.*, 2010 and references within). A microsatellite baseline, produced as part of the EU funded SALSEA-Merge project, is now available for the entire (non-Baltic) range of Atlantic salmon in the Eastern Atlantic (Gilbey *et al.*, In Prep.). This baseline has allowed assignments to be made of fish to regions within this range and, in some cases, to individual rivers. However, the resolving power of this baseline when assigning fish to individual rivers within Scotland has been found to be weak (Gilbey *et al.*, 2012). A recent mixed stock fishery analysis examined the baseline in relation to its ability to robustly assign fish to individual rivers within Scotland and found that, although the microsatellite baseline could be used to identify Scottish and English fish, it could not be used to reliably identify and assign fish to individual Scottish rivers (Gilbey *et al.*, 2012).

A second genetic baseline is currently being developed within Scotland. This baseline uses Single Nucleotide Polymorphism (SNP) markers to define the genetic characters of the various stocks. SNP markers are very numerous and relatively cheap to screen, meaning a large number can be utilised, potentially resulting in a significant improvement in assignment power (Hess *et al.*, 2011).

The aims of the current report are to determine the levels of resolution that might be achieved in assigning fish to rivers within Scotland using SNP markers.

Methods

The basic outline of the techniques employed is that firstly samples from numerous rivers were screened at a large number of SNP markers. A regional hierarchy of ‘assignment regions’ was then identified in which adjacent rivers shared genetic characteristics, and then a set of SNPs identified that allowed assignment to these regions. Subsequently further sets of SNPs were identified that allowed finer resolution of regional units until, in areas where the river coverage was sufficient and which were of significant interest, SNPs were identified which allowed assignments to the rivers represented in the baseline. Further, techniques were developed which allowed individual fish to be excluded from the analysis if their assignment confidence was not high (i.e. there was no strong evidence that they originated from regions/rivers represented within the baseline). In this way, firstly fish could be assigned to region, and then within some of the regions individuals could be either assigned to river, or excluded from the analysis.

SNP baseline coverage

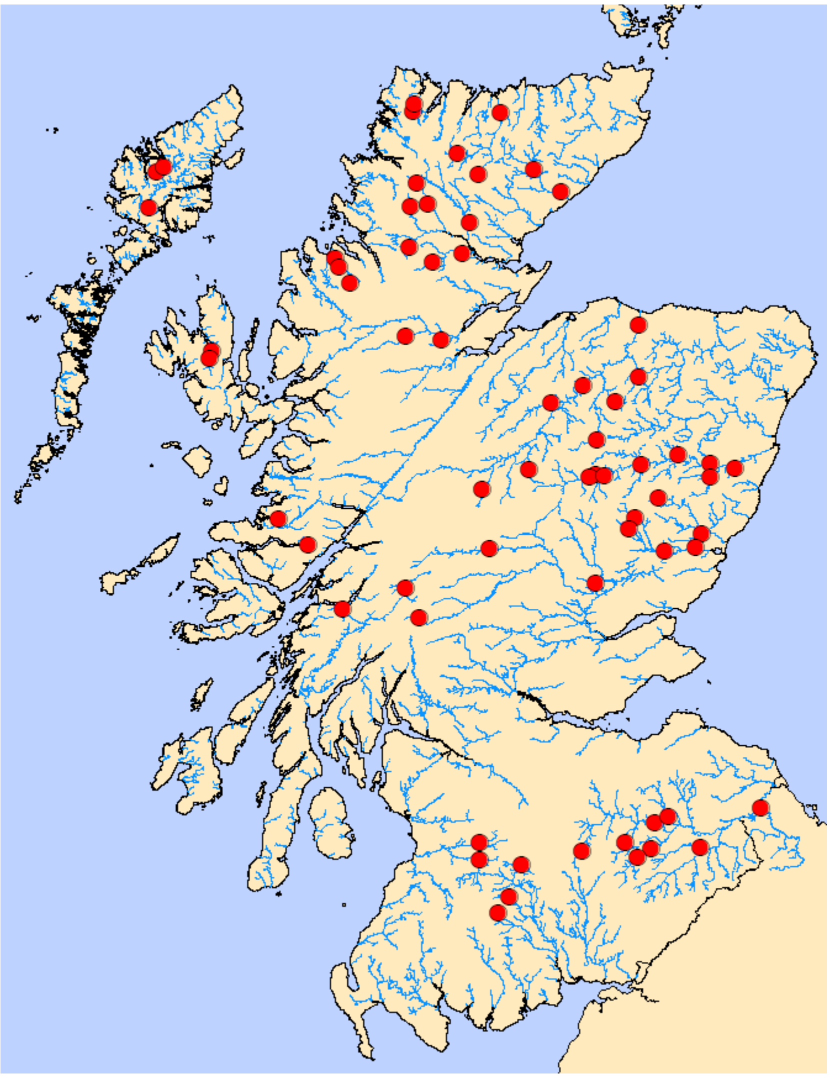


Figure 1 SNP sites included in the analysis.

Figure 1 details the V2 SNP sites available for analysis. The coverage contains 66 sites from 24 rivers containing a total of 1896 fish. The fish were screened using the ‘‘V2-SNP microarray’ on an Illumina Beadstation 500G platform at the Centre for Integrative Genetics (CIGENE) in Norway. This resulted in 5568 SNP genotypes being generated for each fish in the dataset.

Hierarchical analysis

The approach used to identify a useful set of SNP markers for river level analysis from the 5568 available was to follow a hierarchical analysis of the available baseline data, rank the SNPs according to their usefulness in discriminating between reporting groups (using a measure of genetic distance) at each hierarchical level and combine the highest ranked SNPs from each level's analysis into a final panel. This approach was designed to allow the identification of those SNPs that performed best at each level, and to combine these into a single cost-effective set. The hierarchy is outlined in Figure 2.

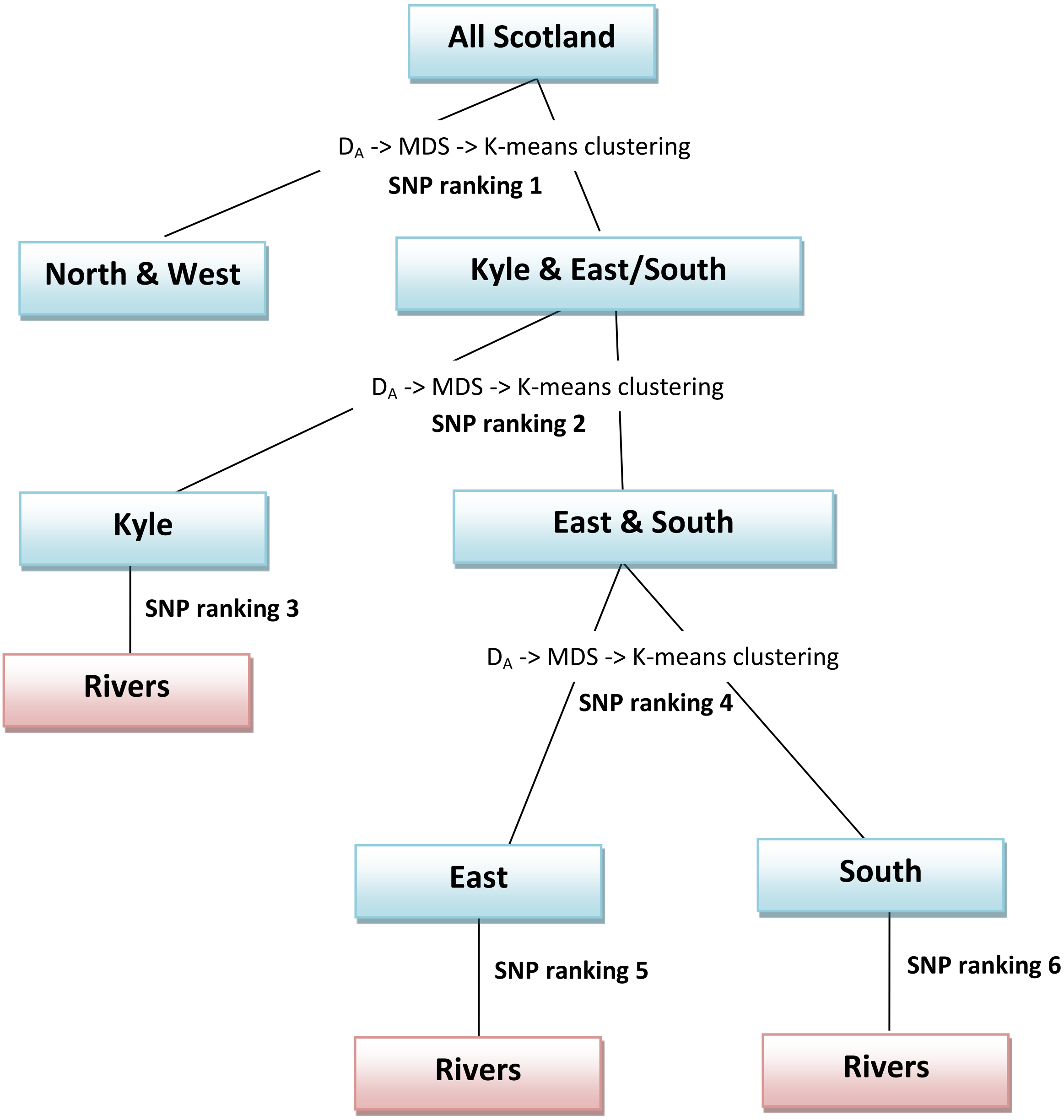


Figure 2 Hierarchical approach to identification of SNPs. Points where SNPs were ranked are noted. For description of assignment units see Fig. 6 and 7 and results section.

At each of the 3 regional hierarchical split points a measure of genetic distance (DA, Nei's standard genetic distance; Nei, 1972) was calculated between each site. This measure was then used to visualise the pairwise genetic distance relationships between sites and regions using Multi-Dimensional Scaling (MDS) plots. Groups of similar sites (i.e. regional groupings) were then identified using K-Means Clustering. Finally the SNPs were ranked according to how powerful each one was in its ability to separate the assignment region groups. This was achieved by calculating hierarchical F-statistics (a measure of population differentiation) for each SNP and ranking the SNP loci according to their discriminatory power using the R-package HIERFSTAT (Goudet, 2005).

Within the Kyle, East and South reporting regions identified by this analysis, further SNP rankings were performed but this time focusing on ranking the SNPs according to their power in differentiating between rivers within these regions.

The analysis resulted in 6 sets of ranked SNPs. The final part of the procedure was to combine the top ranked SNPs of each of these ranked sets into a single set of SNPs for assignment of fish to region and/or river. The top 48 SNPs were taken from each ranked set and incorporated into a single panel resulting in a panel of 288 SNPs. Where the same SNP was present in more than one ranked set, the next available one on the ranked set was used.

Data preparation and production of 'Training' and 'Hold-out' test panels

Outlier sites were firstly identified by examination of the MDS plots and then removed from the analysis. Pairwise measures of DA were calculated between each site and summarised in MDS plots. This allowed sites whose genetic character was very different from those in nearby systems to be identified. Such differences could arise due to a number of reasons including those relating to both sampling artefacts and real genetic differences in the sites genetic composition. However, even if the genetic differences are reflecting of the real genetic composition of such groups, the aim of the initial screenings was to identify broad regional genetic signals and as such these 'outliers' were not representative of such signals. The SNP data was then examined for loci where the minor allele frequency was below 10%. These loci were removed as scoring errors when genotyping can result in such frequencies which in turn can give rise to false estimates of allele frequencies (Anderson *et al.*, 2008; Tabangin *et al.*, 2009; Pongpanich *et al.*, 2010) and so negatively influence SNP ranking procedures.

It is very important that an unbiased procedure is used when analysing assignment success rates. Bias can enter the assessment of success rates due to the fact that the same sets of samples are used both to choose a set of SNP loci and to then go on to test the power of such loci. It has been clearly shown that this can lead to what is termed 'ascertainment bias' due to the circularity of the approach (Waples *et al.*, 2008; Anderson, 2010). In order to prevent this, and before any analysis was performed, the dataset was randomly split into two equal parts. At each site in the baseline, an equal number of fish were put into a 'Training

Set' (TS) and a 'Hold-out Set' (HS). Further, in some cases where there were multiple sites within a river, entire sites were randomly removed from the baseline and included only in the HS. The TS was then used to rank and choose the SNP markers. The success rates of the different marker sets were then determined by assigning fish from the HS back to the TS baseline. In this way, ascertainment bias should be greatly reduced. It must be noted, however, that this procedure requires the dataset to be split into 2 and so the baseline is significantly reduced in size and potentially power and so could actually underestimate the power of the assignment accuracy (Waples, 2010).

To try to estimate the power of the assignments without this confounding issue of only using half the fish in the baseline, new baseline and mixture files were assembled. The baseline was the full dataset (i.e. not half of each sites data). The test mixture file comprised an entire site from each river with two or more sites of data. These sites were also those removed at the start of the whole hierarchical analysis and so had never been part of any SNP ranking procedure (i.e. the baseline had that half of the fish that were removed for the original HS put back and the TS was simply now the whole sites form the original TS. As such, they were by definition unaffected by issues of ascertainment bias. Further, as the data from the sites were the entire sites' data, they represented fish from a particular river from locations that were not represented in the baseline. In reality, only a limited number or sites will be present in any baseline so the strictest test of an assignment protocol is how well do fish not from these sites assign to the assignment unit under investigation. The test was therefore very rigorous, and represented the hardest possible scenario of assignment (where a river of origin is represented in the baseline), rather than those often carried out (e.g. self-assignment, or even TS/HS where in both cases the actual site of origin of the test fish is always in the baseline). Sites represented in this mixture were from the rivers: Carron, Conon, Dee, Nith, South Esk, Spey and Tweed (i.e. those from rivers which contained more than two sites).

Exclusion of fish from rivers not in the baseline

The baseline available for assignment of fish to Scottish rivers will never cover every river in Scotland. Many small rivers will not be represented, yet fish from these rivers may be found in any mixed stock group of individuals. Attempts should therefore be made to identify these fish and remove them from the assignment procedure because by definition if they are assigned they will be wrongly assigned.

Exclusion of fish from rivers not in the baseline was examined using the Exclusion method of assignment analysis according to Vasemägi et al. (2001) (see also Ikediashi *et al.*, 2012). When performing assignments, 2 metrics are obtained: the assignment score and the assignment probability. The assignment score shows the proportion of times in the simulated assignment attempts (typically 10,000 attempts) an individual fish assigns to each site in the baseline (and this can be summed for each site within a river to give the value at the river level). The assignment probability will return the probability (typically using Bayesian methods and Monte-Carlo resampling) that an individual belongs to each reference population. This value is calculated for each reference population and so cannot be summed to return a river level result if sites are used in the baseline as reference populations. The exclusion approach as described in Vasemägi et al. (2001) uses the probability values to exclude fish from the assignment analysis if the maximum probability of assigning to any of the reporting groups is less than 0.05 (i.e. no strong evidence of belonging to any of the reporting groups).

A new analysis was performed to examine the efficiency of this technique. New baseline and mixture files were created using the SNP panel identified in the analysis described above. The same sites were removed from the baseline and placed into a mixture file, as had been used above to examine accuracy of assignments. In addition, all the fish from entire rivers were removed from the baseline and added to the mixture file. Assignments were made and both scores and probabilities calculated. Different levels of cut-off for both scores and probabilities were used and accuracy of assignments determined.

The mixture file contained: fish from sites with rivers in the baseline – Dee, Spey, Kyle, Nith, and Tweed; fish where the entire river was not present in the baseline – Ayr, Conon, North Esk, South Esk, Helmsdale, Dionard

Exclusion analysis was performed with the aim of identifying and having the ability to remove from further analysis the fish from rivers not represented in the baseline while at the same time leaving in the analysis assignment of those fish from rivers that were in the baseline.

Regional assignment test

A further analysis was performed to examine the power of the regional level assignments in more detail and as a first limited test of the ability to assign fish to the highest level regional assignment units. Fish were examined from 15 randomly chosen new sites from around the Scottish coastline which had not been screened for the larger SNP panel and so had not been involved in any part of the initial analysis (Figure 3). Six fish were screened at each site using the top 89 SNPs from those identified at the SNP Ranking 1 and SNP Ranking 2 stages as detailed in Figure 2. Assignments were then performed using the methods outlined above and accuracy of assignments determined.

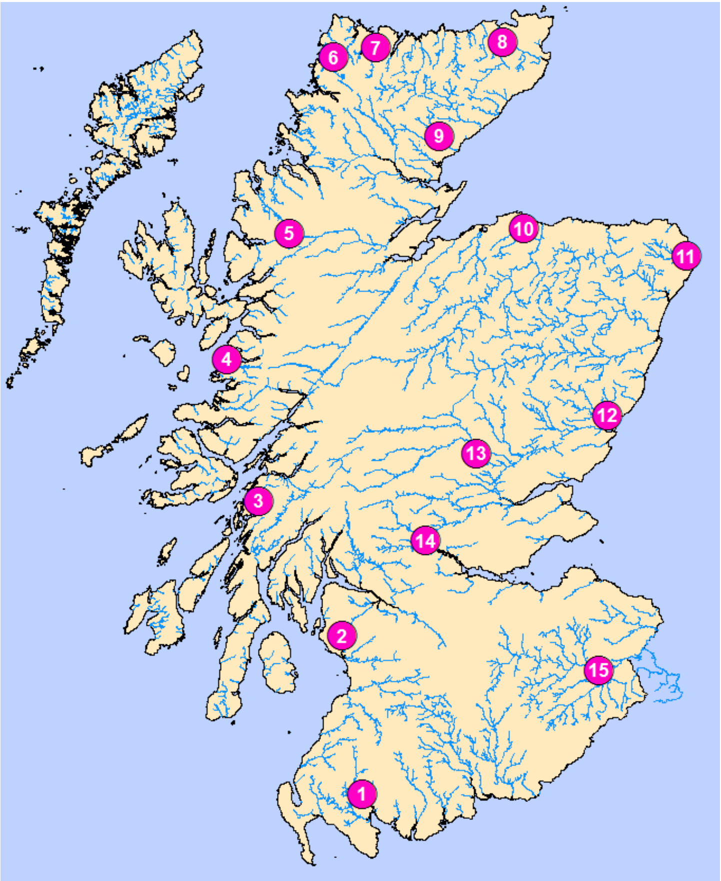


Figure 3 Map showing the sites of the samples used in the regional assignment. 1 River Cree, 2 River Garnock, 3 River Euchar, 4 River Morar, 5 River Ewe, 6 Rhiconich River, 7 River Hope, 8 River Thurso, 9 River Brora, 10 River Lossie, 11 River Ugie, 12 River North Esk, 13 River Tay, 14 River Forth, 15 River Tweed.

Results

Hierarchal splits and SNP ranking

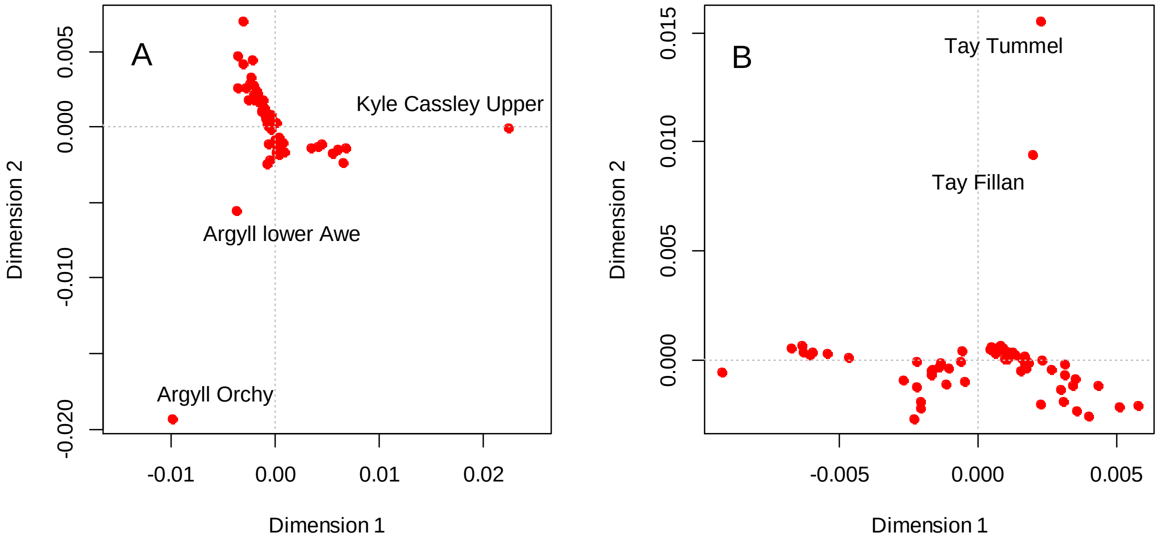


Figure 4 MDS plot of pairwise DA of all sites.

Figure 4A details the pairwise DA relationships between the various sites across the whole of Scotland. Three sites can be seen that fall outwith the central grouping of sites and were removed from the next stage of the analysis. DA was recalculated and the relationships re-plotted in Figure 4B. Two more sites are seen to fall outwith the central grouping and were again removed.

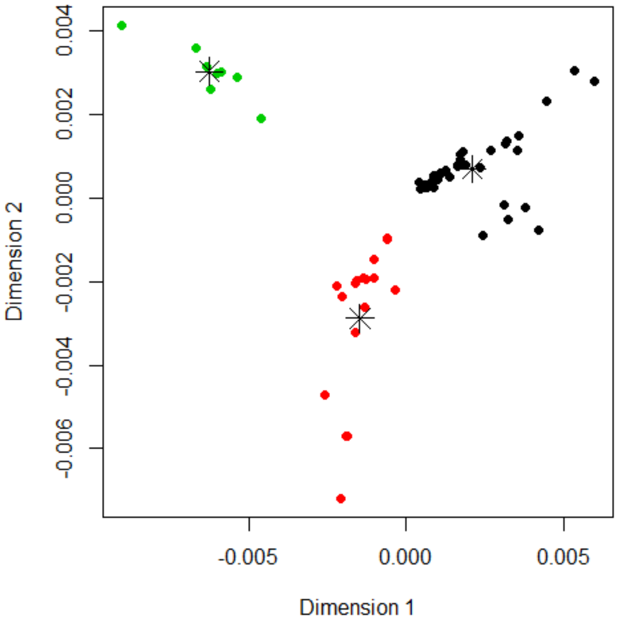


Figure 5 MDS plot of all pairwise DA of all sites after removal of outliers showing how the sites split into 3 groups using K-means clustering. Black are east and south group (ES), green are Kyle of Sutherland group (KY), red are north and west group NW).

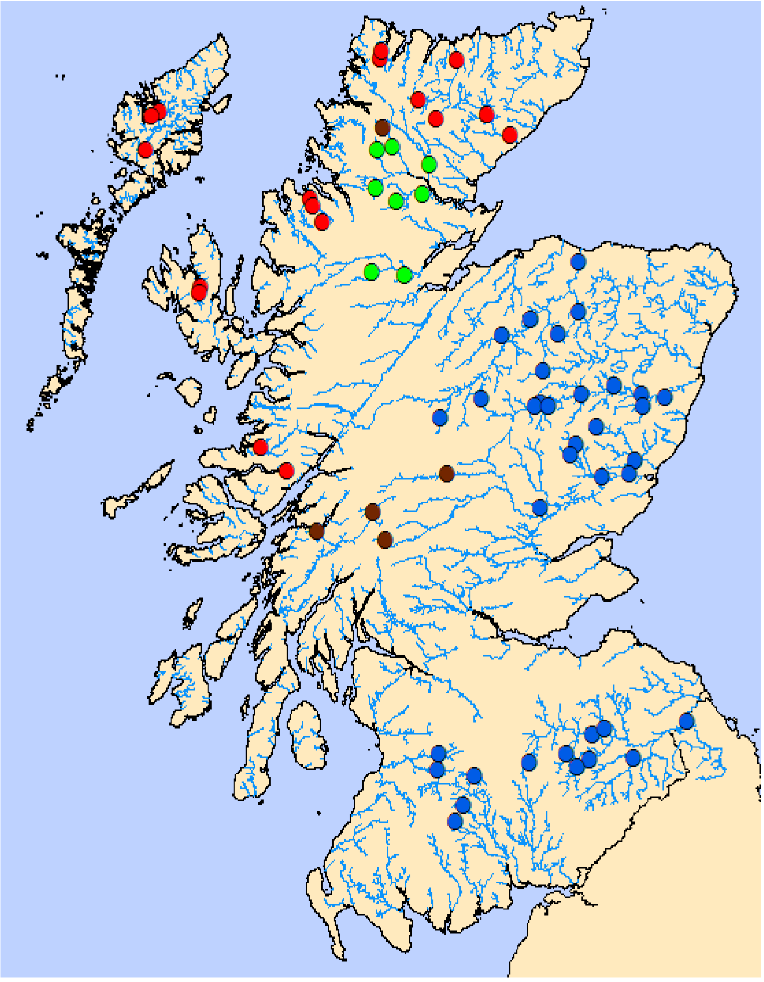


Figure 6 Map showing the top level regional groupings of sites as defined by K-means clustering. Green are Kyle of Sutherland group (KY), blue are east and south group (ES), red are north and west group (NW), brown are the outlier sites.

Figure 5 details the split of sites into the 3 groupings identified using K-means clustering. Fig 5 shows that these 3 clusters represent regional genetically similar site groups that also are related well to geographic location. Due to the particular interest of being able to distinguish fish at the first level which originate from the NW or KY and ES groups the first SNP ranking procedure was carried out using the NW group as one regional assignment unit and the KY and ES groups as the other (SNP Ranking 1 on Figure 2). SNPs were thus identified which had greatest power in splitting up these two units.

As described in Figure 2 the second level analysis was focused on being able to differentiate between fish which originate from the KY and ES groups and so the second level analysis was performed on these data only (SNP Ranking 2 on Figure 2).

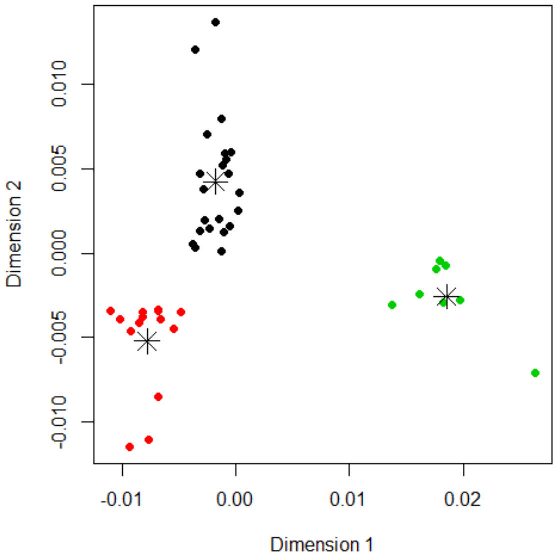


Figure 7 MDS plot of all pairwise DA of KY and ES sites showing how the sites split into 3 groups using K-means clustering. Red are south group (S), green are Kyle of Sutherland group (KY), black are east group (E).

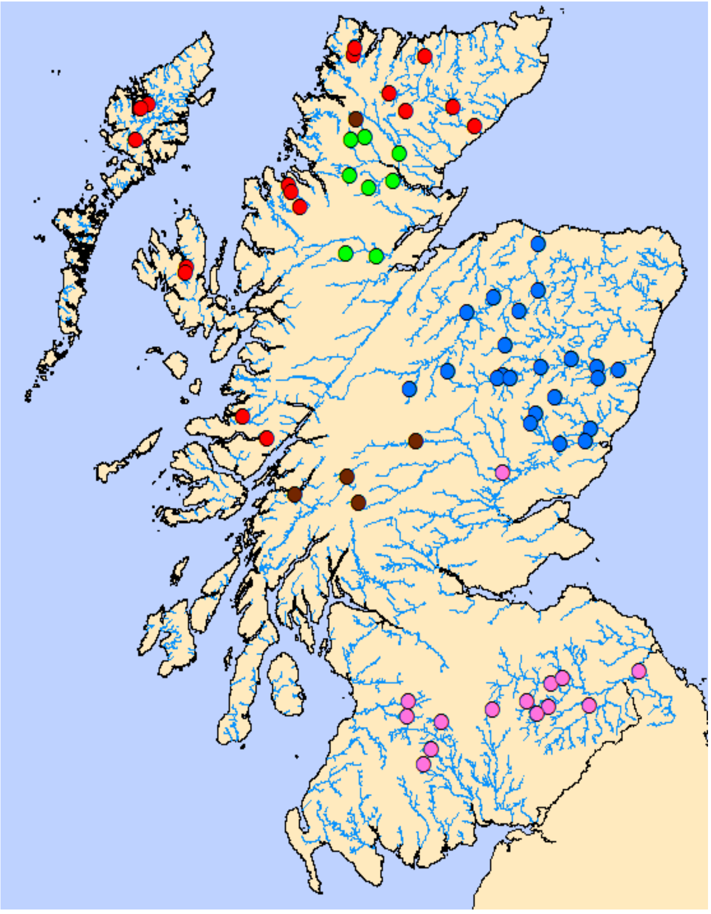


Figure 8 Map showing the regional groupings of sites as defined by K-means clustering. Green are Kyle of Sutherland group (KY), blue are east group (E), pink are south group (S), red are north and west group (NW), brown are the outlier sites.

As before, Figure 7 shows that the KY group are separated from the remaining fish, but in this second stage analysis the previous ES group is seen to split into 2 further groups, one associated with the east coast and one with the south of Scotland on both coasts. Again as is shown in Figure 8 the groupings are geographically consistent apart from a single site on the Tay which seems to fall into the S regional group.

After identification of the E and S split, a further round of SNP ranking and identification of those most powerful in separating these two groups was performed (SNP Ranking 3 on Figure 2). Once this was completed SNP ranking were available which allowed identification of loci which had the most power in separating all regional assignment units identified. The next stage of the process was in areas where there was sufficient river coverage, and which were of particular interest to be able to assign fish to the river level (i.e. east coast and southern regions) further SNP ranking was performed separately on the data from the different regional assignment units, this time focusing on ranking the SNPs according to their ability to differentiate rivers (SNP Ranking 4, 5 and 6 on Figure 2).

The outcome of the procedure described above was 6 sets of ranked SNPs each detailing the most powerful loci for separating the particular regions or rivers upon which their ranking was based. The top 48 SNPs from each ranking set were taken and combined into a single screening panel of 288 SNPs. Where SNPs were duplicated in more than one ranking set (33 loci), the next on the list was add as a substitute.

Assignment accuracy to region

At each stage of the hierarchical ranking stages assignment, power was analysed by assigning fish back to the particular split under investigation using the HS data set. From each ranked list, assignments of the HS to the TS data sets were performed with varying numbers of SNPs (12, 24, 48, 96, 192 and 288 SNPs) to determine assignment accuracy. Correct assignments to reporting group (region or river) were determined using all assignments, and also using an illustrative assignment cut-off score of 80 which achieves a balance between assignment vigour (i.e. only assigning fish with a strong assignment score) and the number of fish assigned (i.e. not being too strict as to only leave very few fish assigned).

Assignments to the three top level assignment units are shown in Figure 9. As expected, the more SNPs that are used the greater the accuracy of the assignments up to an asymptote beyond which little if any extra power is obtained. The full HS set can be seen to be very well assigned with an accuracy of between 90 to 100% correct with 288 SNPs. If focus is made on just the HS data from complete sites that were removed from the data-set before SNP ranking but which had other sites in their river represented in the baseline then the accuracy is seen to be much the same as the full HS set (which of course includes half the data from all sites). Finally, if the data from rivers not represented in the baseline and which were not included in the SNP ranking procedure is examined, a small drop in accuracy is seen with

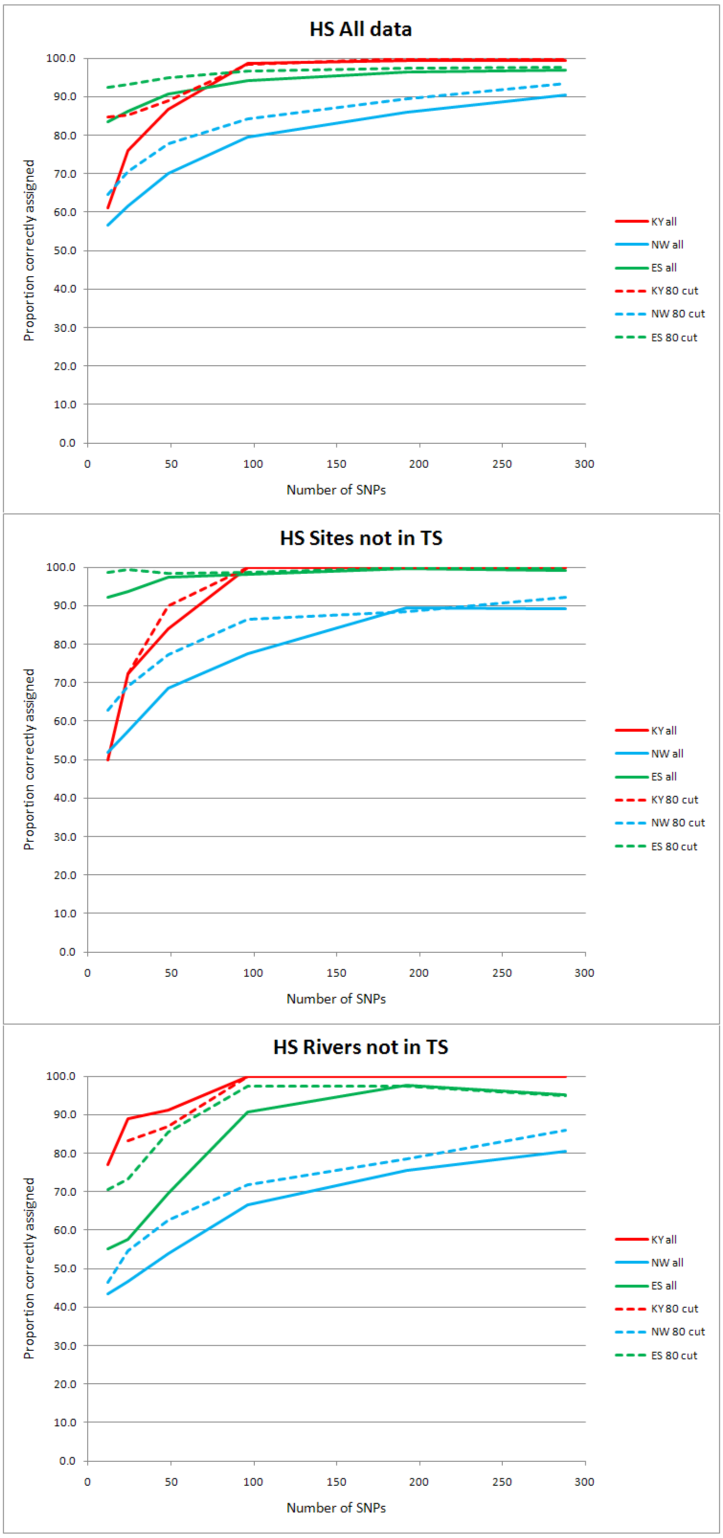


Figure 9 Proportion of fish assigned to a reporting region that are correctly assigned with varying numbers of SNPs to top levels regional assignment units: KY is Kyle, NW is North and West, ES is East and South. Solid lines are all data dashed are assignments using a cut-off value of 80 for the assignment score.

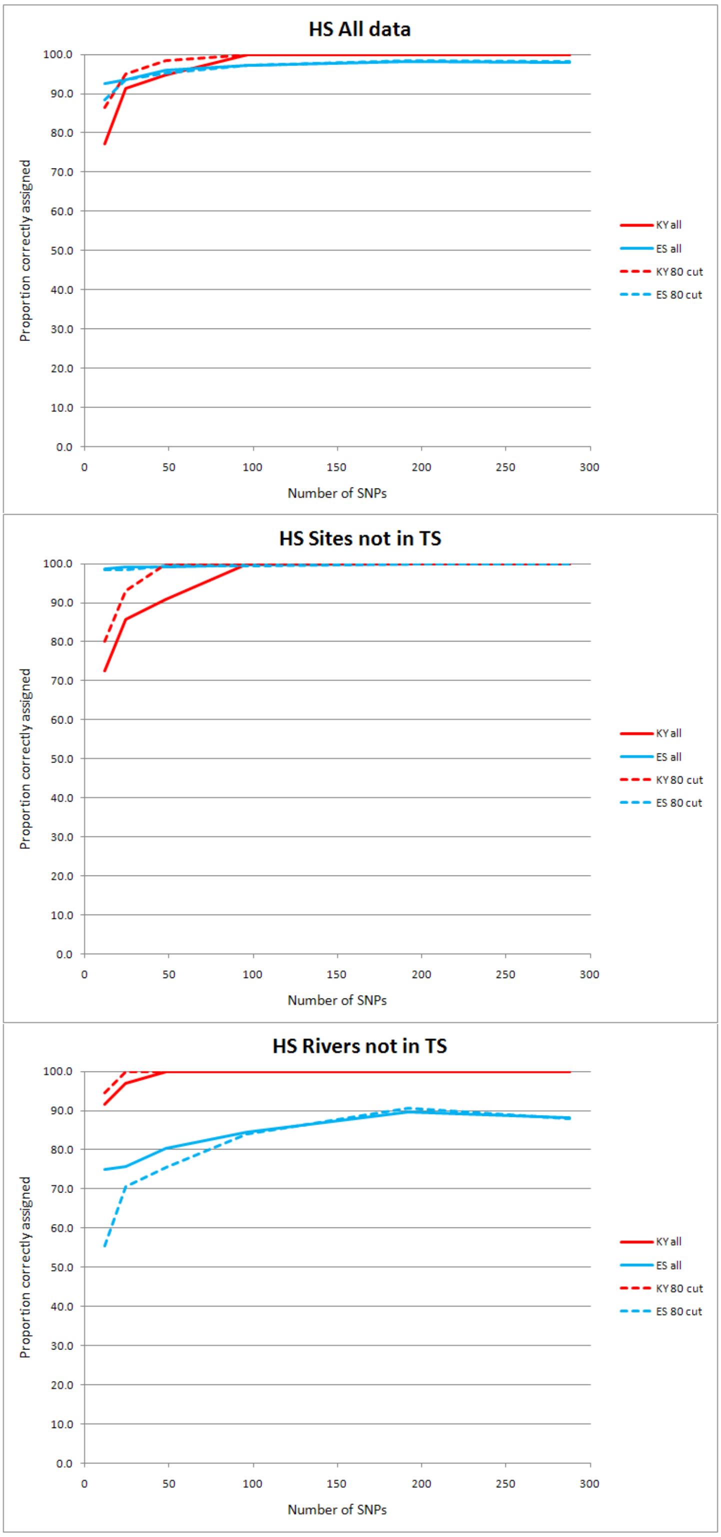


Figure 10 Proportion of fish assigned to a reporting region that are correctly assigned with varying numbers of SNPs to the first East and South coast regional split level: KY is Kyle, ES is East and South. Solid lines are all data dashed are assignments using a cut-off value of 80 for the assignment score.

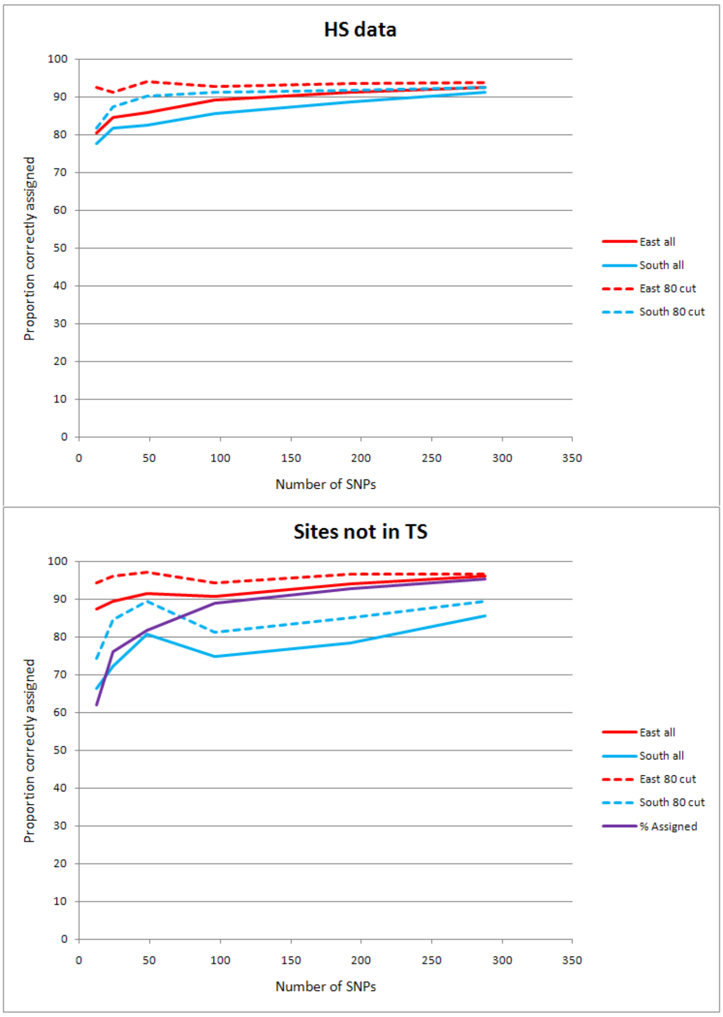


Figure 11 Proportion of fish assigned to a reporting region that are correctly assigned with varying numbers of SNPs to the East and south coast regional split level. Solid lines are all data dashed are assignments using a cut-off value of 80 for the assignment score. Note. Whole rivers were not removed for analysis due to the small number in the southern group.

the NW assignment unit, however assignments of the other 2 units are still very good (i.e. >90 % accuracy with 48 SNPs or more) . This finding can be explained by the fact that there were some small number of incorrect assignments of fish from both the KY and ES units into the NW unit but no or very low levels of incorrect assignments from the NW to either of these units. There was also no or very little incorrect assignments between the KY and ES units.

Figure 10 details the assignment success to the second level regional groupings, Kyle and East/South. Again very good assignment success is seen whether all the HS data are considered, sites which have been removed from the analysis at ranking or even whole rivers which had been removed at the SNP ranking stage. The lowest level regional split of the ES region also shows very good assignment accuracy success as shown in Figure 11. Although here the low number of rivers in the dataset meant that whole rivers were not removed from the analysis to form the HS set, the levels of assignment success with the full

HS dataset and the site level analysis are very similar to the higher level success and so there is no reason to suppose the river level assignments would not follow the same pattern.

Combined SNP panel

The hierarchical analysis was performed at 6 points in the analysis and a ranked set of SNPs identified at each point:

Stages	Level
Rank_1	All Scotland
Rank_2	Kyle & East/South
Rank_3	Kyle river level
Rank_4	East & South
Rank_5	East river level
Rank_6	South river level

The top 48 SNPs were selected from each stage giving 288 SNPs. In this panel were 33 SNPs that were duplicated. A further 33 SNPs were selected from the ranking performed at the East river level which examined river level assignments on the East coast and was thus of most interest. The correlations of the 6 different SNP sets at all SNPs are shown in Table 1, followed by the ranking relationships between the 6 sets in Figure 12.

Table 1 Correlations between SNP rankings of the 6 different SNP sets identified during the hierarchical analysis below the diagonal, probabilities associated with the correlations above diagonal.

	Rank_1	Rank_2	Rank_3	Rank_4	Rank_5	Rank_6
Rank_1		0.000	0.798	0.731	0.545	0.021
Rank_2	0.355		0.000	0.000	0.000	0.513
Rank_3	-0.004	-0.1213		0.329	0.789	0.738
Rank_4	-0.005	0.1991	0.016		0.000	0.000
Rank_5	-0.010	-0.0736	-0.004	-0.218		0.024
Rank_6	-0.037	-0.0104	0.005	-0.384	-0.036	

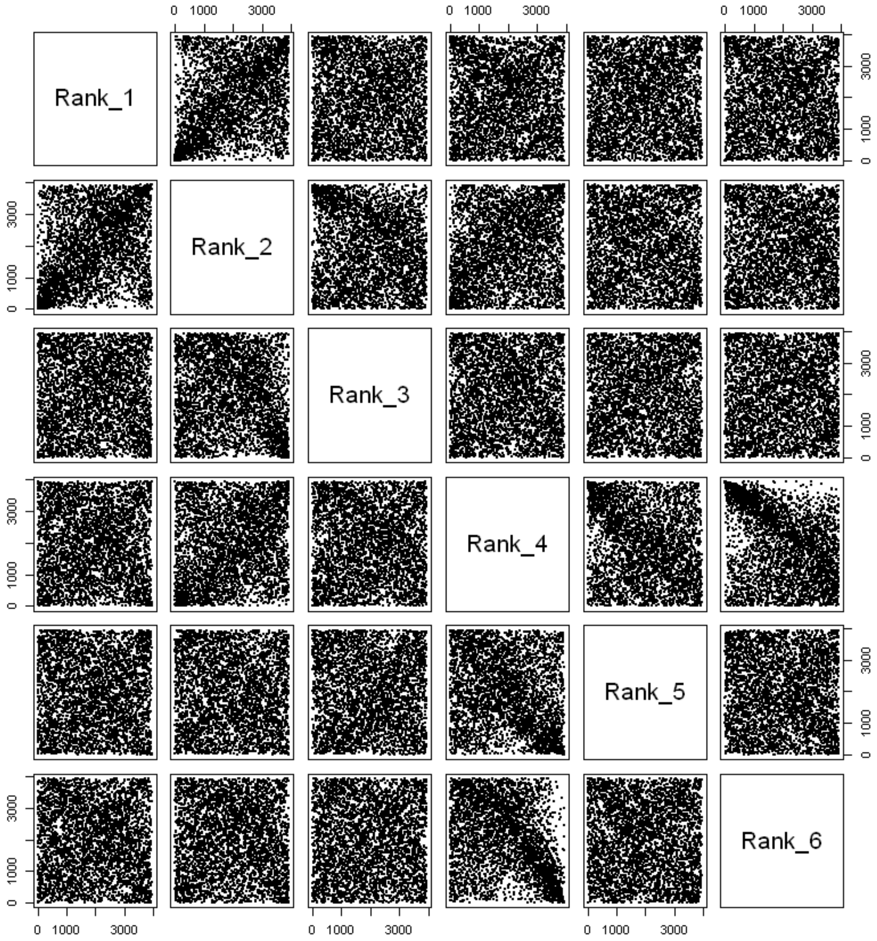


Figure 12 Relationship between SNP rankings of the 6 different SNP sets identified during the hierarchical analysis.

Determination of accuracy of hierarchical panel

Assignment to region has been examined at the different hierarchical scales identified. It has already been shown that it is possible to identify and exclude fish from the NW region from subsequent river level analysis. Table 2 details the accuracy of assignments using the 288 SNP panel at the river level in fish from outwith the NW region using the new baseline and the test assignment files which were assembled (i.e. the baseline was the full baseline and the assignment mixture file comprised an entire site from each river with two or more sites of data, the Carron, Conon, Dee, Nith, South Esk, Spey and Tweed).

As expected from the regional initial analysis, Table 2 shows that indeed there are very few miss-assignments of fish from the Eastern regions to the NW baseline, and not a single miss-assignment of fish from the Kyle to the ES regions or vice-versa

Table 2 Assignments of fish using the 288 SNP panel examined at the regional level. KY is Kyle, NW is North and West, ES is East and South. Assignments using all data are shown and those using the illustrative assignment score cut-off score of 80.

All data	Assigned origin				
Origin	KY	NW	ES	Correct	% Correct
KY	58	6	0	58	90.6
NW	0	0	0	0	N/A
ES	0	5	229	229	97.9
% correct	100	N/A	100		

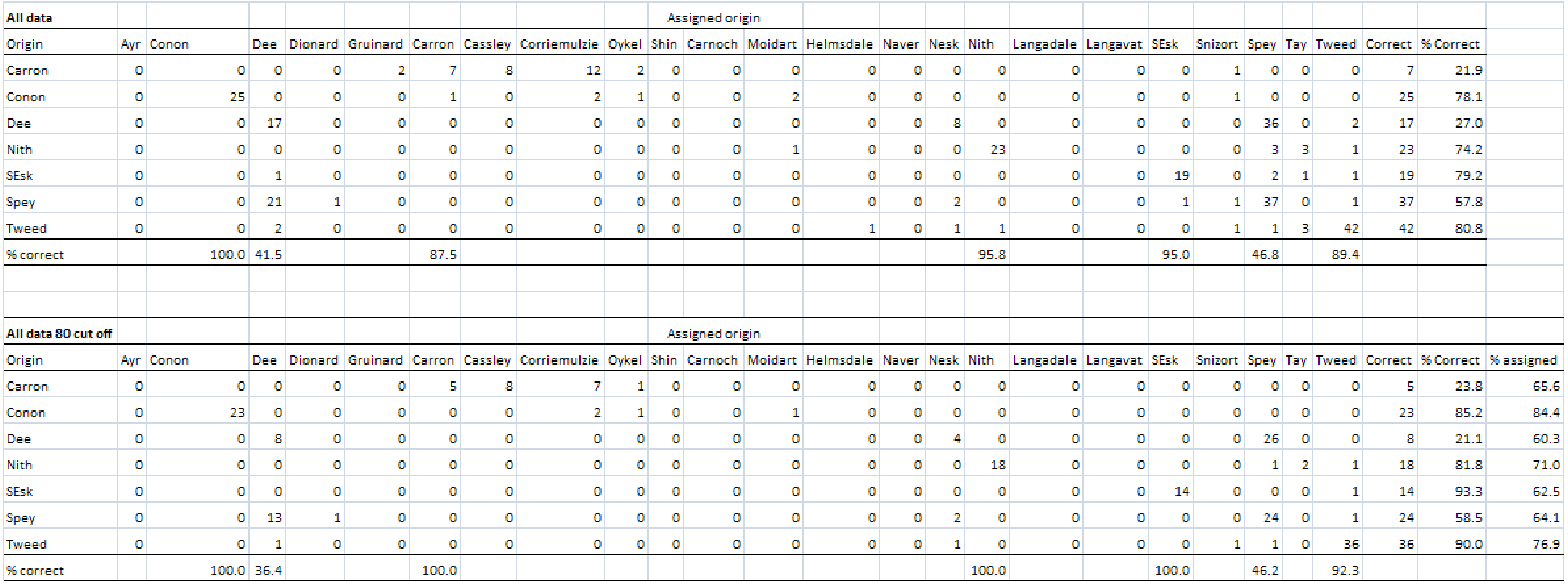
All data 80 cut off	Assigned origin					
Origin	KY	NW	ES	Correct	% Correct	% assigned
KY	56	1	0	56	98.2	89.1
NW	0	0	0	0	N/A	N/A
ES	0	4	224	224	98.2	97.4
% correct	100	N/A	100			

Table 3 shows that the Carron fish are seen to have a relatively low assignment success to river, but that the miss-assignments from the Carron are mainly to other rivers in the Kyle region. As all these rivers flow into the same estuary and significant mixing is thought to occur between them, the Kyle rivers were subsequently grouped and successful assignments of the Carron fish examined at this new level as shown in Table 4. As expected, the assignments of these fish are now significantly improved when assigning to this new regional grouping.

Tables 3 and 4 also show that there are relatively low assignment successes in both the Spey and Dee systems. Again these two rivers have been combined and assignment success to the new assignment unit is detailed in Table 5.

Overall, the river level assignments to the combined groups are seen to be very strong (i.e. ≥ 92.3% accuracy in fish from an assignment group assigning back to that group, Table 5), and considering that the fish being assigned are from sites not represented in the baseline (i.e. they are not the TS/HS split data), this suggests that assignments to the river level are likely to be robust where there is sufficient river level SNP coverage to characterise individual rivers well. However, it can be seen that the Dee and Spey are hard to separate in the preceding analysis. It may also be the case that as other rivers are characterised with SNP markers similar patterns might be seen between 2 or more rivers, but at present the extent that this pattern may be manifest is hard to predict.

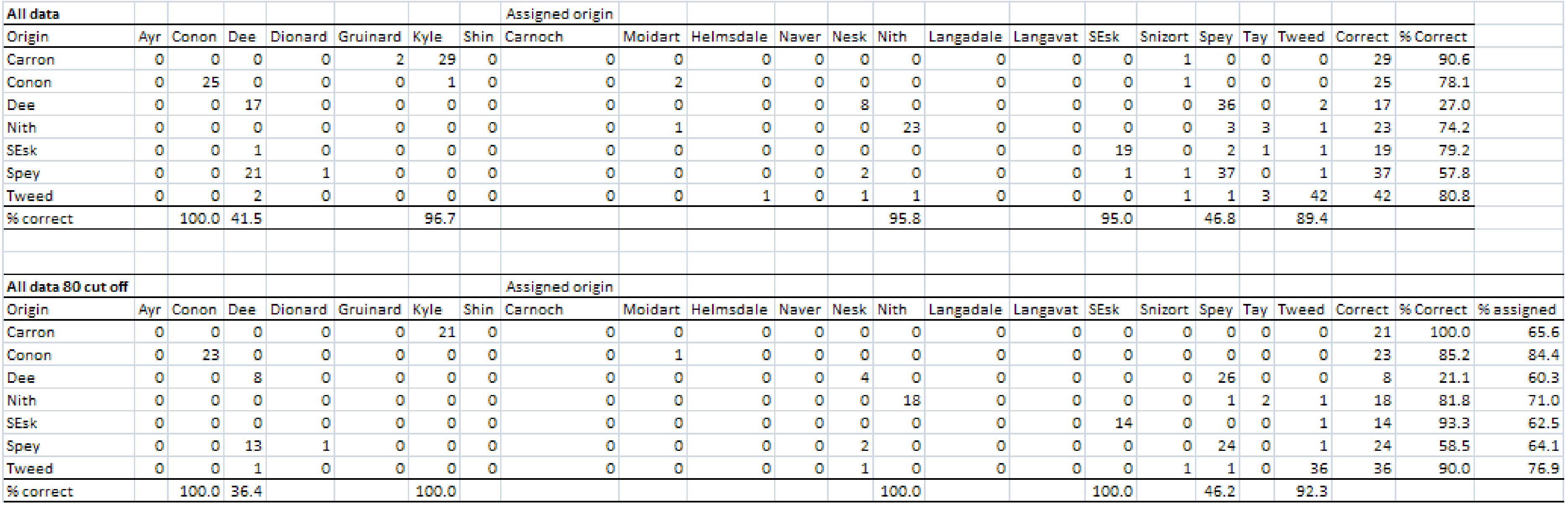
.



All data											Assigned origin															
Origin	Ayr	Conon	Dee	Dionard	Gruinard	Carron	Cassley	Corriemulzie	Oykel	Shin	Carnoch	Moidart	Helmsdale	Naver	Nesk	Nith	Langadale	Langavat	SEsk	Snizort	Spey	Tay	Tweed	Correct	% Correct	
Carron	0	0	0	0	2	7	8	12	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	7	21.9	
Conon	0	25	0	0	0	1	0	2	1	0	0	2	0	0	0	0	0	0	0	1	0	0	0	25	78.1	
Dee	0	0	17	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	36	0	2	17	27.0	
Nith	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	23	0	0	0	0	3	3	1	23	74.2	
SEsk	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	2	1	1	19	79.2	
Spey	0	0	21	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	37	0	1	37	57.8	
Tweed	0	0	2	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	1	3	42	42	80.8	
% correct		100.0	41.5			87.5										95.8			95.0		46.8		89.4			

All data 80 cut off											Assigned origin															
Origin	Ayr	Conon	Dee	Dionard	Gruinard	Carron	Cassley	Corriemulzie	Oykel	Shin	Carnoch	Moidart	Helmsdale	Naver	Nesk	Nith	Langadale	Langavat	SEsk	Snizort	Spey	Tay	Tweed	Correct	% Correct	% assigned
Carron	0	0	0	0	0	5	8	7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	23.8	65.6
Conon	0	23	0	0	0	0	0	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	23	85.2	84.4
Dee	0	0	8	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	26	0	0	8	21.1	60.3
Nith	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	1	2	1	18	81.8	71.0
SEsk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	1	14	93.3	62.5
Spey	0	0	13	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	24	0	1	24	58.5	64.1
Tweed	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	36	36	90.0	76.9
% correct		100.0	36.4			100.0										100.0			100.0		46.2		92.3			

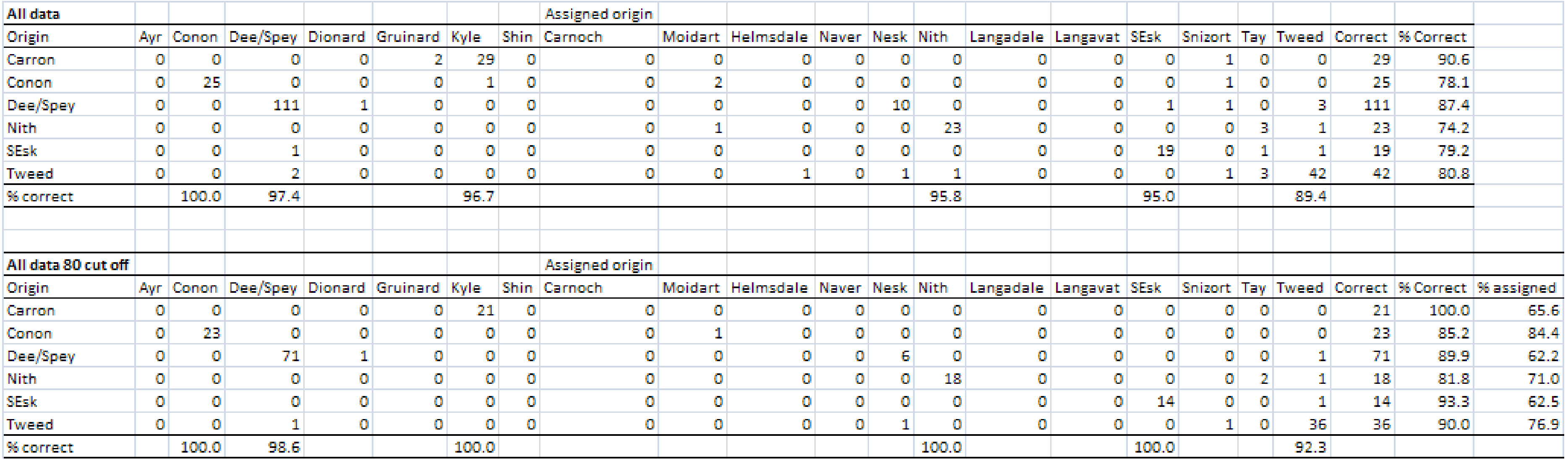
Table 3 Assignment success using the 288 SNP panel to river level.



All data								Assigned origin														
Origin	Ayr	Conon	Dee	Dionard	Gruinard	Kyle	Shin	Carnoch	Moidart	Helmsdale	Naver	Nesk	Nith	Langadale	Langavat	SEsk	Snizort	Spey	Tay	Tweed	Correct	% Correct
Carron	0	0	0	0	2	29	0	0	0	0	0	0	0	0	0	0	1	0	0	0	29	90.6
Conon	0	25	0	0	0	1	0	0	2	0	0	0	0	0	0	0	1	0	0	0	25	78.1
Dee	0	0	17	0	0	0	0	0	0	0	0	8	0	0	0	0	0	36	0	2	17	27.0
Nith	0	0	0	0	0	0	0	0	1	0	0	0	23	0	0	0	0	3	3	1	23	74.2
SEsk	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	19	0	2	1	1	19	79.2
Spey	0	0	21	1	0	0	0	0	0	0	0	2	0	0	0	1	1	37	0	1	37	57.8
Tweed	0	0	2	0	0	0	0	0	0	1	0	1	1	0	0	0	1	1	3	42	42	80.8
% correct		100.0	41.5			96.7							95.8			95.0		46.8		89.4		

All data 80 cut off								Assigned origin															
Origin	Ayr	Conon	Dee	Dionard	Gruinard	Kyle	Shin	Carnoch	Moidart	Helmsdale	Naver	Nesk	Nith	Langadale	Langavat	SEsk	Snizort	Spey	Tay	Tweed	Correct	% Correct	% assigned
Carron	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	100.0	65.6
Conon	0	23	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	23	85.2	84.4
Dee	0	0	8	0	0	0	0	0	0	0	0	4	0	0	0	0	0	26	0	0	8	21.1	60.3
Nith	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	1	2	1	18	81.8	71.0
SEsk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	1	14	93.3	62.5
Spey	0	0	13	1	0	0	0	0	0	0	0	2	0	0	0	0	0	24	0	1	24	58.5	64.1
Tweed	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	36	36	90.0	76.9
% correct		100.0	36.4			100.0							100.0			100.0		46.2		92.3			

Table 4 Assignment success using the 288 SNP panel to river level with the individual Kyle rivers combined.



All data								Assigned origin													
Origin	Ayr	Conon	Dee/Spey	Dionard	Gruinard	Kyle	Shin	Carnoch	Moidart	Helmsdale	Naver	Nesk	Nith	Langadale	Langavat	SEsk	Snizort	Tay	Tweed	Correct	% Correct
Carron	0	0	0	0	2	29	0	0	0	0	0	0	0	0	0	0	1	0	0	29	90.6
Conon	0	25	0	0	0	1	0	0	2	0	0	0	0	0	0	0	1	0	0	25	78.1
Dee/Spey	0	0	111	1	0	0	0	0	0	0	0	10	0	0	0	1	1	0	3	111	87.4
Nith	0	0	0	0	0	0	0	0	1	0	0	0	23	0	0	0	0	3	1	23	74.2
SEsk	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	19	0	1	1	19	79.2
Tweed	0	0	2	0	0	0	0	0	0	1	0	1	1	0	0	0	1	3	42	42	80.8
% correct		100.0	97.4			96.7							95.8			95.0			89.4		

All data 80 cut off								Assigned origin														
Origin	Ayr	Conon	Dee/Spey	Dionard	Gruinard	Kyle	Shin	Carnoch	Moidart	Helmsdale	Naver	Nesk	Nith	Langadale	Langavat	SEsk	Snizort	Tay	Tweed	Correct	% Correct	% assigned
Carron	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	21	100.0	65.6
Conon	0	23	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	23	85.2	84.4
Dee/Spey	0	0	71	1	0	0	0	0	0	0	0	6	0	0	0	0	0	0	1	71	89.9	62.2
Nith	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	2	1	18	81.8	71.0
SEsk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	1	14	93.3	62.5
Tweed	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	36	36	90.0	76.9
% correct		100.0	98.6			100.0							100.0			100.0			92.3			

Table 5 Assignment success using the 288 SNP panel to river level with the individual Kyle rivers combined and also the Spey and Dee combined.

Dee/ Spey separation

A new analysis was performed on just the Dee and Spey data to examine in more detail the possibility of separating Dee and Spey fish. Firstly, relationships between sites in the two rivers was examined using a Neighbour-joining tree (Saitou and Nei, 1987) and MDS plots using D_A as before. Outlier sites were removed as before and the data were split in half at each site creating a set of individuals to use to rank the SNPs (Training Set, TS) and a set of individuals to test the accuracy of the assignments (Holdout Set, HS). Further, a random site from each river was also completely removed from the TS and added to the HS set. As before, pairwise F_{ST} was used to rank loci based on their discriminatory power between rivers. Assignments were then performed using different numbers of loci from this ranked list and summing assignment success to the river level.

Three outlier sites are seen when all data is analysed as can be clearly seen on both the tree in Figure 13 and the MDS plot in Figure 14A (Dee Sheeoch, Dee Water Dye, Spey Avon Lyon). When these are removed the two rivers show separation on the Y-axis of the MDS plot in Figure 14B apart from a single Dee site which falls into the Spey grouping on this plot.

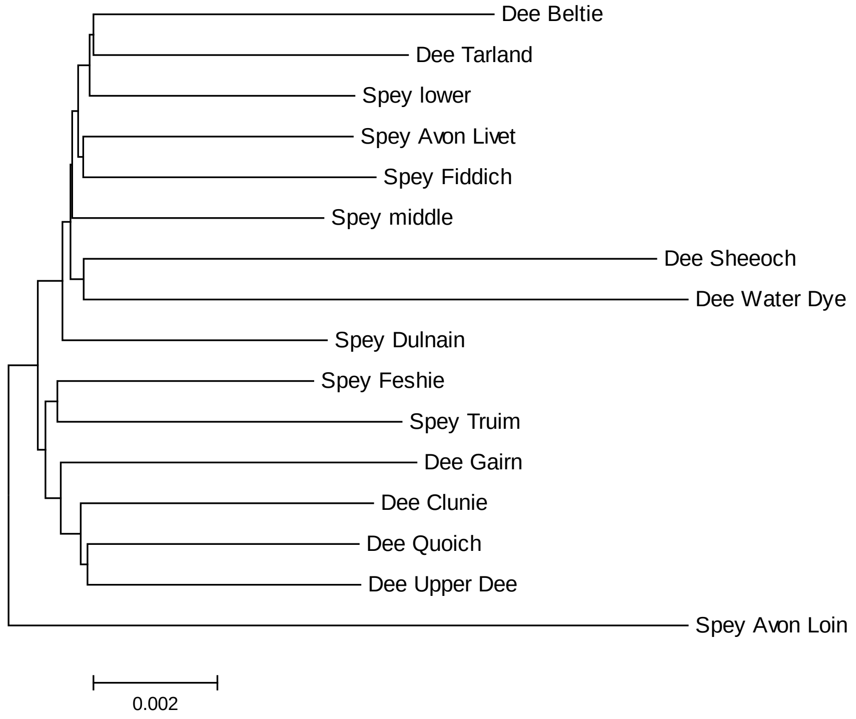


Figure 13 Neighbour-joining tree showing relationships between sites on the Spey and Dee.

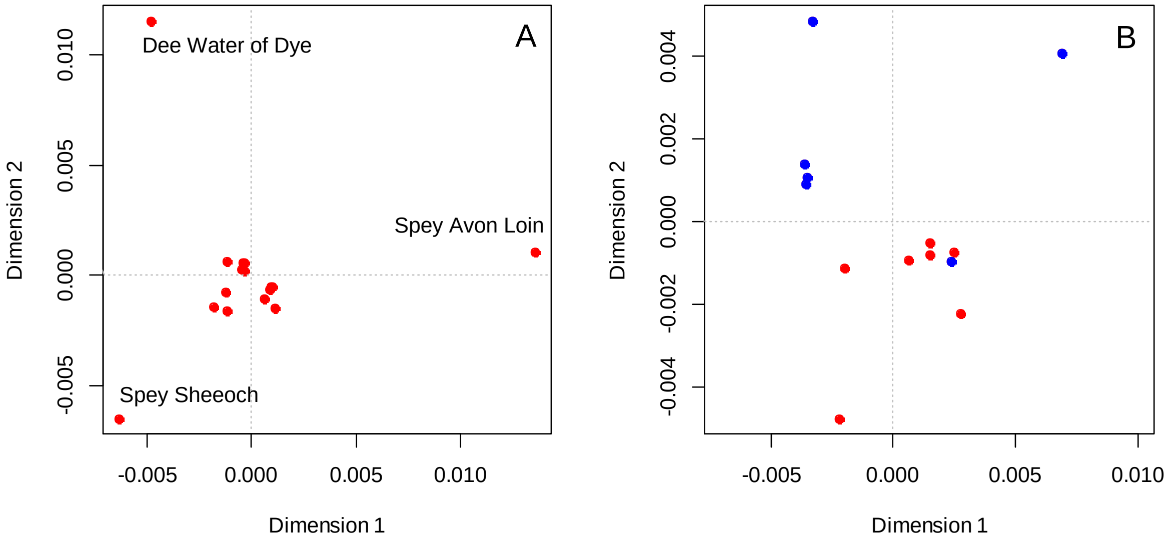


Figure 14 MDS plots of sites pairwise DA between sites on the Spey and Dee. A has all sites and shows the three outliers sites which were removed from the analysis before pairwise DA was recalculated. B shows the relationship between sites with the three outliers removed with Dee sites in blue and Spey in red.

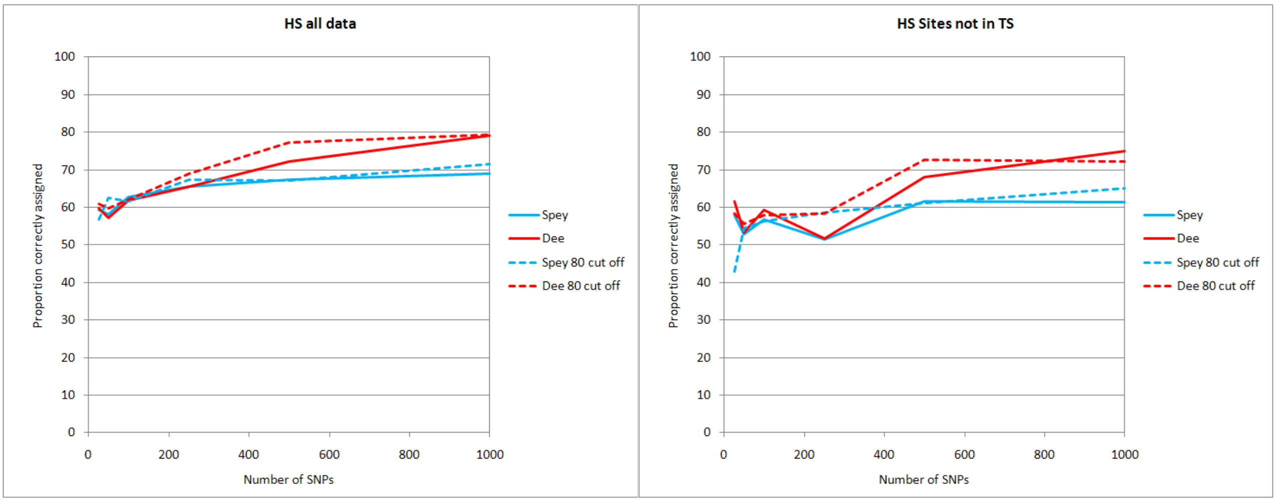


Figure 15 Proportion of fish assigned to a river that are correctly assigned with varying numbers of SNPs. Solid lines are all data dashed are assignments using a cut-off value of 80 for the assignment score.

As before, assignment success of the HS dataset to the TS baseline is shown in Figure 15. Accuracy is seen to be increased compared to the previous analysis (see Table 4). However, the proportion of fish assigning back to a river that actually came from that river is still relatively low (~55 – 80 % of all data and just ~50 – 75 % of sites not in the baseline). By pure chance alone it would be expected that there would be 50% right (two rivers to assign to) and the values are better than this, but not by much until a very large number of SNPs are utilised.

Exclusion of fish from rivers not in the baseline

Exclusion analysis was performed with the aim of removing the fish from rivers not represented in the baseline while at the same time leaving in the assignment of those fish from rivers that were in the baseline. As was hoped it can be seen that the maximum assignment score of fish from rivers not in the baseline is significantly lower than for those fish where their river of origin was present in the baseline (Figure 16). If a cut-off of 0.05 is used a large proportion of fish from rivers not in the baseline can be removed from the analysis without losing a significant proportion of the other fish.

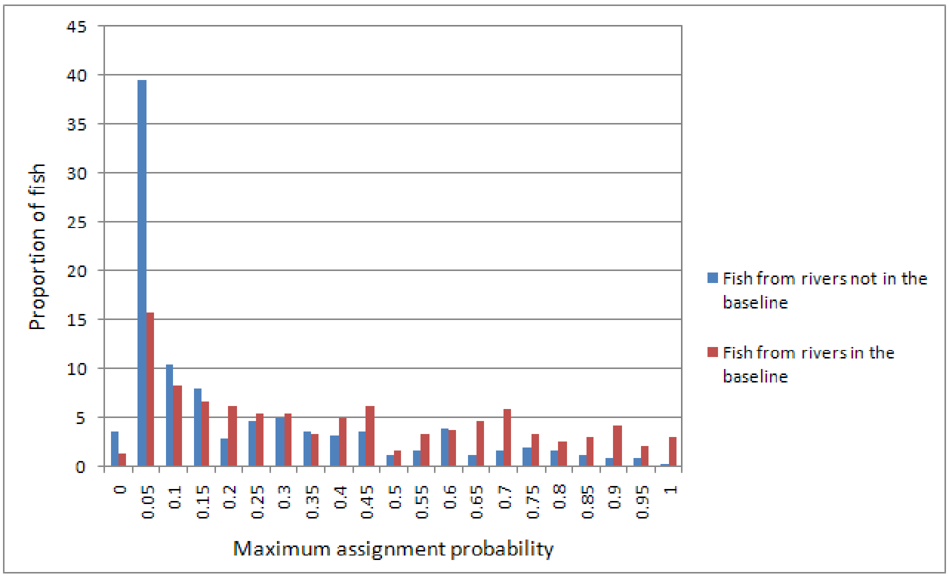


Figure 16 Maximum assignment probabilities of all fish in the analysis.

Variation in the cut-off level of the assignment probability together with variation in the assignment score value cut-off was then examined to determine the optimum level of cut-off for both these metrics where the maximum numbers of fish from rivers not in the baseline are removed while leaving in the analysis those fish from rivers that are represented. As the river level assignments are particularly focused on identifying and assigning fish from the East coast (i.e. where river level coverage is greatest along with productivity and the presence of Special Areas of Conservation for salmon) the first step in this assignment analysis was to remove any fish that were assigned to the North and West region.

Figure 17 shows the proportion of fish remaining to be assigned after removal of those assigned to the NW region. In this situation, the assignment probability is not yet being used, just removal of all fish assigned to the North and West, which in the mixture file consists of fish from the Helmsdale and Dionard. It can be seen that no matter what the assignment cut-off used, a large proportion of these fish can be removed as they assign to the NW region even though they are from rivers not in the baseline. The regional assignments thus reflect those seen to region in the preceding analysis and are thus very robust.

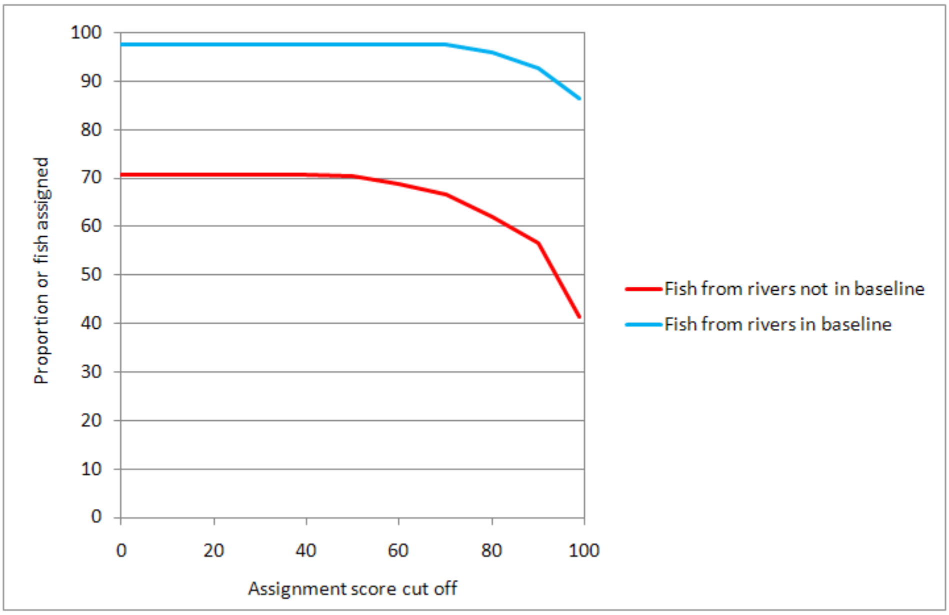


Figure 17 Proportion of fish assigned after removal of fish assigned to the NW region.

The next stage of the exclusion analysis was to utilise the exclusion probability cut-off levels to try to remove the remainder of fish from rivers not represented in the baseline. Different cu-off levels can be used which results in a trade-off between accuracy and numbers of fish remaining in the analysis to be assigned.

As expected by the distribution of assignment probabilities between the two groups of fish, using a cut off probability allows a large proportion of fish from rivers not in the baseline to be removed (Figure 18). Using an exclusion score of ≤ 0.05, together with an assignment score cut off of 90, is seen to remove ~70 % of fish from rivers not in the baseline while at the same time leaving in ~75 % of fish from rivers represented in the baseline. If a stricter exclusion score of ≤ 0.1 is used together with an assignment score cut off of 90 this is seen to remove ~75 % of fish from rivers not in the baseline while at the same time leaving in ~70 % of fish from rivers represented in the baseline. The lack of significantly greater power to remove more fish from rivers not in the baseline reflects the differential exclusion scores of the two groups of fish as seen in Figure 16. Using a score of 0.05 will remove a much greater proportion of fish from rivers not in the baseline than those in it. However, moving to 0.01, the differential between these two groups is much less and as such the increase in ability to screen out fish from rivers not in the baseline is not increased by a significant amount.

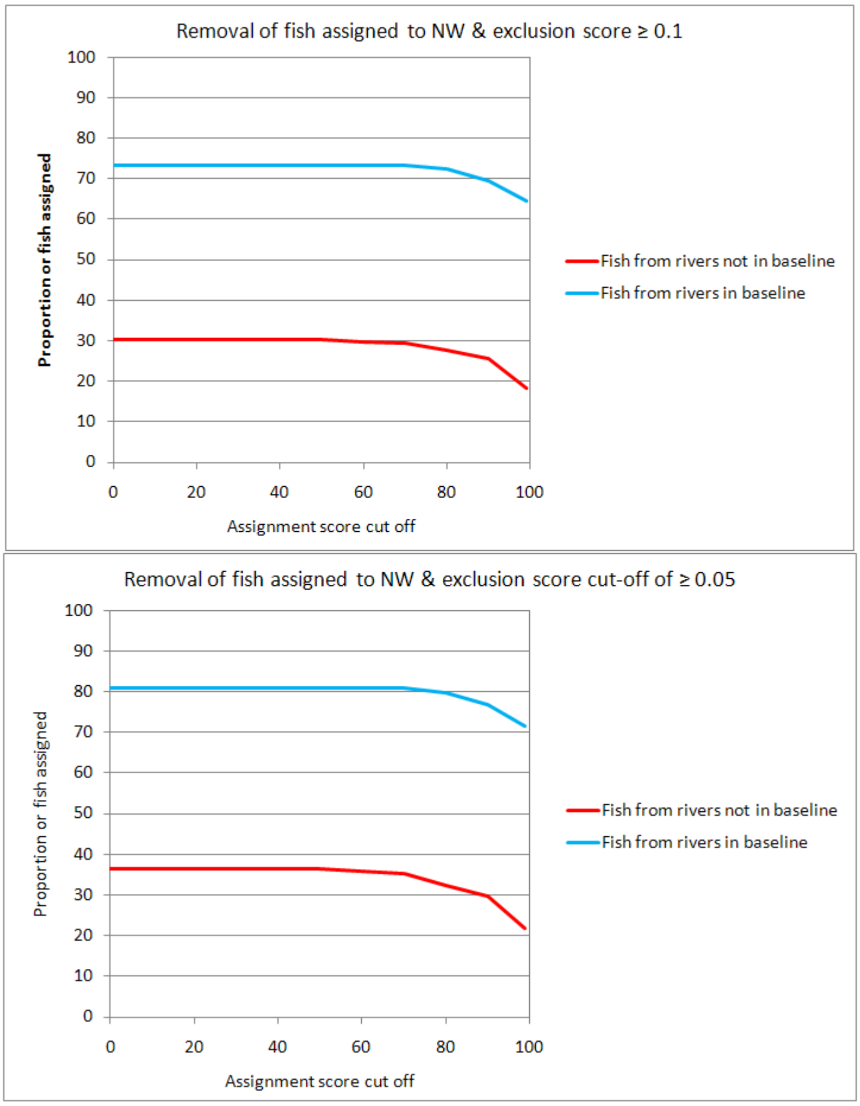


Figure 18 Proportion of fish assigned after removal of fish assigned to the NW region with varying cut-off scores (x-axis) and also utilising the exclusion probabilities of 0.05 and 0.01.

It should also be remembered that in reality the assignment accuracy figures in a real situation might be expected to be better than those reported above. The full baseline will include representation from rivers responsible for the vast majority of production on the East coast. In a realistic mixed stock analysis focusing on this region, the exclusion procedure will be focused on removing fish from outwith the East coast assignment unit. It has been shown that this can be done with accuracy; leaving fish to be assigned within the East coast assignment unit which again has been shown can be (apart from the separation of the Spey/Dee) performed with good accuracy (at least in the rivers analysed so far).

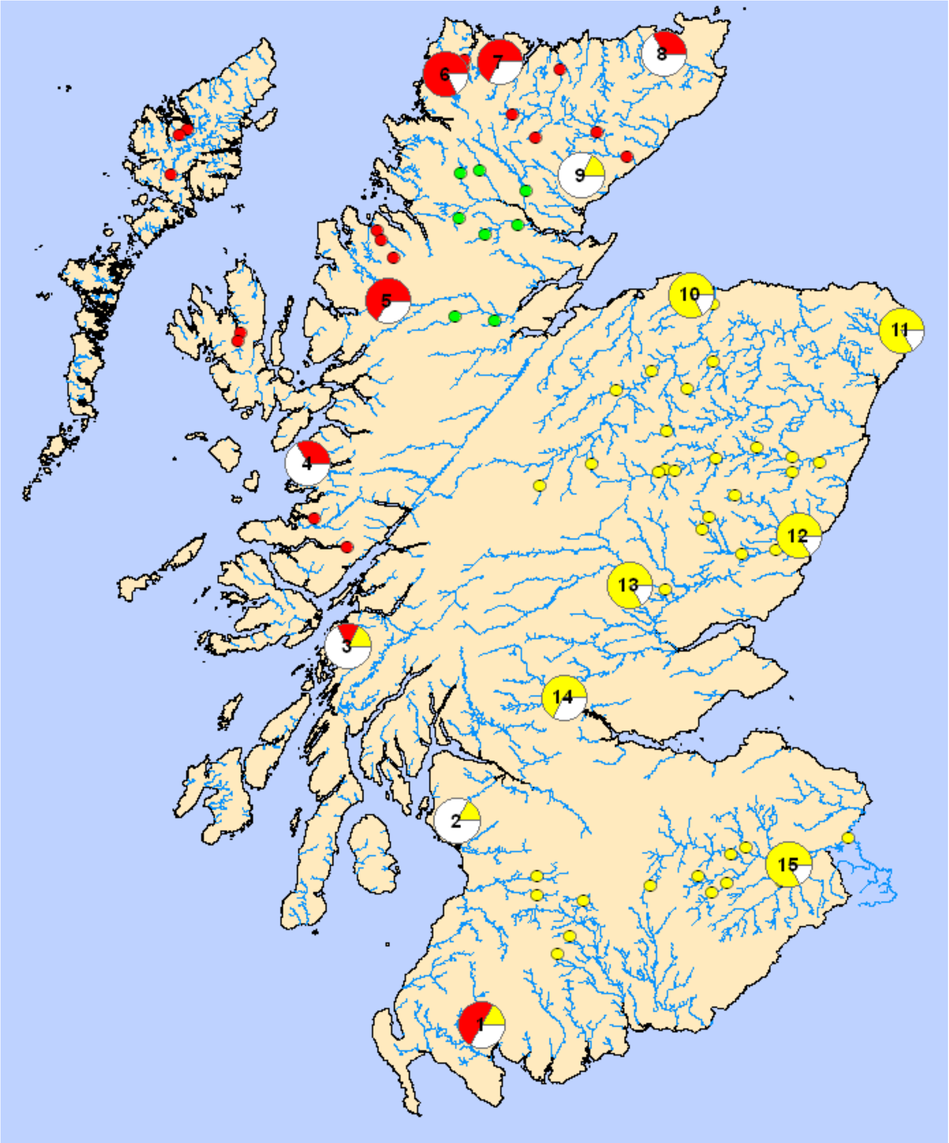


Figure 19 Regional assignment summary of test fish to second level regional assignment units assignments using an assignment confidence cut-off score of 90 and an exclusion probability cut-off of 0.5. Baseline sites are shown as small circles with their colour representing their assignment group: Green are Kyle of Sutherland group (KY), yellow are east and south group (ES), red are north and west group (NW), white are unassigned fish. Pie-charts show the assignments of the test fish. As there were 6 test fish at each site pie-chart segments can be seen to relate to individual fish. 1 River Cree, 2 River Garnock, 3 River Euchar, 4 River Morar, 5 River Ewe, 6 Rhiconich River, 7 River Hope, 8 River Thurso, 9 River Brora, 10 River Lossie, 11 River Ugie, 12 River North Esk, 13 River Tay, 14 River Forth, 15 River Tweed.

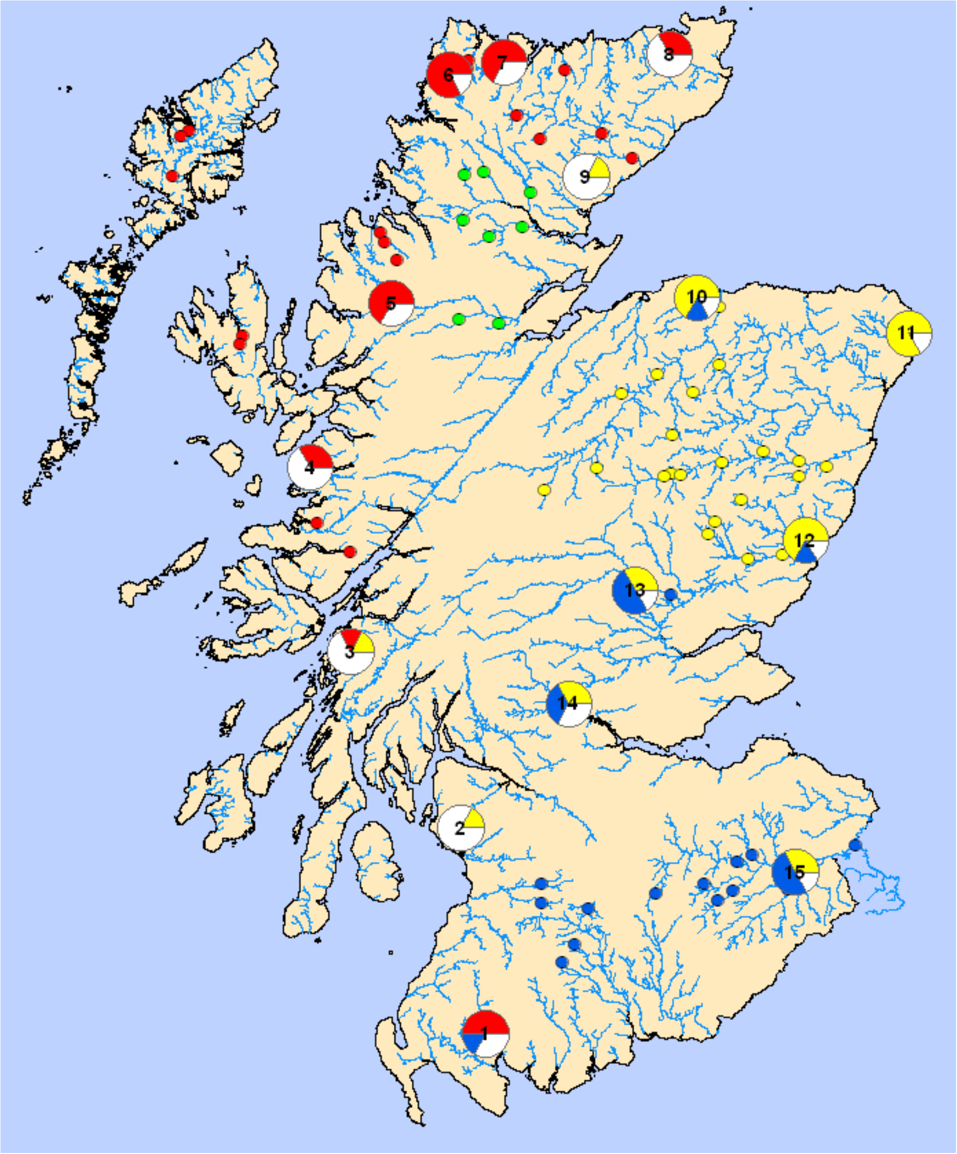


Figure 20 Regional assignment summary of test fish to third level regional assignment units assignments using an assignment confidence cut-off score of 90 and an exclusion probability cut-off of 0.5. Baseline sites are shown as small circles with their colour representing their assignment group: Green are Kyle of Sutherland group (KY), yellow are east group (ES), red are north and west group (NW), blue are south (S) group white are unassigned fish. Pie-charts show the assignments of the test fish. As there were 6 test fish at each site pie-chart segments can be seen to relate to individual fish. 1 River Cree, 2 River Garnock, 3 River Euchar, 4 River Morar, 5 River Ewe, 6 Rhiconich River, 7 River Hope, 8 River Thurso, 9 River Brora, 10 River Lossie, 11 River Ugie, 12 River North Esk, 13 River Tay, 14 River Forth, 15 River Tweed.

Table 6 Assignment proportions of fish in the regional test panel to second and third level assignment regions as depicted in Figures 19 and 20. Assignment units are: Kyle of Sutherland group (KY), East and South group (ES), North and West group (NW), East group, (E) and South group (S). Non-Ass refers to non-assigned fish after the cut-off has been applied (see text).

Site code	River	Second level regions				Third level regions				
		ES	KY	NW	Non-Ass	E	S	KY	NW	Non-Ass
1	River Cree	83.3	0.0	0.0	16.7	33.3	50.0	0.0	0.0	16.7
2	River Garnock	16.7	0.0	16.7	66.7	16.7	0.0	0.0	16.7	66.7
3	River Euchar	83.3	0.0	0.0	16.7	83.3	0.0	0.0	0.0	16.7
4	River Morar	66.7	0.0	0.0	33.3	33.3	33.3	0.0	0.0	33.3
5	River Ewe	16.7	0.0	0.0	83.3	16.7	0.0	0.0	0.0	83.3
6	Rhiconich River	83.3	0.0	0.0	16.7	66.7	16.7	0.0	0.0	16.7
7	River Hope	83.3	0.0	0.0	16.7	33.3	50.0	0.0	0.0	16.7
8	River Thurso	16.7	0.0	0.0	83.3	16.7	0.0	0.0	0.0	83.3
9	River Brora	16.7	0.0	50.0	33.3	0.0	16.7	0.0	50.0	33.3
10	River Lossie	0.0	0.0	33.3	66.7	0.0	0.0	0.0	33.3	66.7
11	River Ugie	0.0	0.0	83.3	16.7	0.0	0.0	0.0	83.3	16.7
12	River North Esk	0.0	0.0	33.3	66.7	0.0	0.0	0.0	33.3	66.7
13	River Tay	83.3	0.0	0.0	16.7	66.7	16.7	0.0	0.0	16.7
14	River Forth	0.0	0.0	66.7	33.3	0.0	0.0	0.0	66.7	33.3
15	River Tweed	0.0	0.0	66.7	33.3	0.0	0.0	0.0	66.7	33.3

Regional assignment test

Assignments of the fish used in the regional assignment test using an assignment confidence cut-off score of 90 and an exclusion probability cut-off of 0.5 are detailed in Figures 19 and 20 and Table 6. Figure 19 shows all top assignments for all fish to the three regional groups at the second regional level examined. In general the assignments are very accurate with fish from the ES unit mostly being correctly identified as coming from the region and fish from the NW from that region. There are 4 miss-assigned fish in the NW/KY regions which are assigning to the ES group.

Examination of the second regional level of assignment in Figure 20 shows interesting patterns of assignment. Firstly, NW fish are being assigned to the NW with none being miss-assigned to this group from the E region. Within what was previously the ES region though, there is some mixing of assignments to the E and S units in many of the sites. There is also a number of fish in the west coast of the S unit (from the river Cree) which do not assign to the S unit but rather the NW unit. In general, the S unit is poorly characterised in the baseline available here. South of the Esks there is just a single site on the Tay and the nothing else between the Tay and the Tweed. Further, within the S unit the Tweed is well characterised, but the western coast rivers which have been characterised as being within this unit have only sites from the far upper sections of their catchments. When a site from the lower catchment has been included on the Cree, it can be seen that the majority of fish from their assign to the NW rather than the S group. These observations suggest that the S group is not well characterised. The upper sites on the west coast rivers that have been included in the S group may not actually be representative of the majority of the main river stocks on the west coast in the S region, which may actually group with the rest of the NW assignment unit. With the baseline used to define the region groups at this level the upper sites may have grouped with the Tweed as there was little else around the southern part of the west coast for them to group with. Better characterisation of rivers from the Clyde and south to the Scottish boarder would help resolve this picture and it might be expected that the sites in on the west coast of what has been characterised here as the S unit might resolve into a well-defined South West grouping.

Similarly, on the southern part of the east coast grouping, the surprising fact that the single site on the Tay grouped with the Tweed could again be reflective of poor and unrepresentative coverage of fish from this part of the region within this assignment unit. More samples are needed from the Tay, Earn and Forth catchments to resolve this picture better. The patterns of assignment of fish from the test panel within this region might then be expected to be better explained.

It should also be remembered that these results come from an analysis using just 89 SNP markers, and when the full set of 288 is employed it would be expected that the accuracy and confidence of the assignments will likely increase.

Discussion

The analysis presented here has shown that, depending on baseline coverage, it is possible, to be able to assign fish to both region, and where baseline coverage is sufficient, river with high accuracy in most situations. The problem of being able to screen out fish from reporting regions not represented in the baseline has also been successfully addressed using hierarchical reporting regions and exclusion techniques.

The identification of a set of SNP markers able to discriminate between fish from the different regional assignment units even if they are from rivers not represented in the baselines of these assignment units means that it would be expected that new rivers can be added to the baseline in these regions without the necessity of rescreening using the full V2 panel and having to choose new regional SNP markers each time. Regional baseline coverage can now be enhanced using just the regional set of markers already identified which will achieve a significant time and cost saving. It should be remembered, however, that the full V2 panel will be needed if river level assignments are required rather than regional level.

Accuracy of assignment to the regional units was very good, and as expected the assignments become more uncertain at more detailed levels of the hierarchy. However, even when examined at the river level (taking into consideration the requirement of combining some rivers), assignments were found to be very good in most cases. Assignments to river, unlike those to region, will still however, require screening of new rivers using the V2 SNP panel when these new rivers are added to the baseline.

Differentiation between the Spey and the Dee was difficult, even with large numbers of markers. In general when looking at the genetic character of populations it is the smaller populations that tend to be most differentiated. This is a reflection of the various mechanisms responsible for influencing the genetic character of a population. Principal among these are included genetic drift, founder effects and adaptive evolution. All of these influences happen with greater speed and/or have potentially greater effects in smaller populations. The Spey and Dee represent some of the largest meta-populations of salmon in Scotland and it may be the case that genetically they are still very closely related using the 'neutral' SNP markers employed here. It will be interesting to look at the Deveron which lies between them to see if it is also similar to these two larger systems or if it has differentiated through the processes outlined above. It will also be of interest to examine in more detail the Tay as another large geographically close system to determine its levels of differentiation. Further work could be performed on identifying a set of markers which differentiate between the Spey and Dee but this would perhaps require techniques such as RAD (Restriction site Associated DNA) Sequencing using next generation DNA sequencing. Using this approach, DNA from groups of individuals from the different rivers would be pooled and large number of SNPs identified that are associated with differences between the groups (for examples in salmonids see Hohenlohe *et al.*, 2011; Houston *et al.*, 2012).

The results of the analysis presented here have developed techniques that allow:

- Regional assignment of fish to different hierarchical levels
- Screening out of fish from rivers not present in the baseline within regions
- Assignment of fish to most rivers represented in the baseline

Future developments of the procedures will aim to:

- Increase baseline coverage using the regional SNP markers identified here
- Increase river level coverage using the V2 SNP panel
- Identify a set of SNP markers for within region regional level assignments to river
- Further investigate techniques of examining situations where rivers cannot be differentiated

Acknowledgments

MSS would like to acknowledge the contribution of a number of individuals and organisations in the production of this report. These include John Gilbey (MSS), Lee Stradmeyer (MSS), Eef Cauwelier (MSS), Anja Armstrong (MSS), the Rivers and Fisheries Trusts of Scotland, and Mark Coulson (RAFTS). MSS would also like to thank individual river trusts for making available some of the baseline SNP data used in this analysis.

References

Anderson, C. A., Pettersson, F. H., Barrett, J. C., Zhuang, J. J., Ragoussis, J., Cardon, L. R. & Morris, A. P. (2008). Evaluating the Effects of Imputation on the Power, Coverage, and Cost Efficiency of Genome-wide SNP Platforms. *The American Journal of Human Genetics* **83**, 112-119.

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources.* **10**, 701-710.

Gilbey, J., Coughlan, J., Wennevik, V., Prodohl, P., McGinnity, P., Cauwelier, E., Cherbonnel, C., Coulson, M. W., Cross, T., Crozier, W., Dillane, E., Ellis, J. S., Ensing, D., Garcia-Vazquez, E., Griffiths, A. M., Gudjonsson, S., Hindar, K., Karlsson, S., Knox, D., Machado-Schiaffino, G., Meldrup, D., Nielsen, E. E., Olafsson, K., Primmer, C. R., Prusov, S., Stradmeyer, L., Stevens, J. R., Vaha, J. P., Webster, L. M. I. & Verspoor, E. (In Prep.). Genetic stock identification of European Atlantic salmon (*Salmo salar* L.) utilising a comprehensive microsatellite based genetic baseline.

Gilbey, J., Stradmeyer, L., Cauwelier, E., Middlemas, S., Shelly, J. & Rippon, P. (2012). Genetic Investigation of the North East English Drift Net Fisheries. Marine Scotland Science.

Goudet, J. (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184-186.

Griffiths, A. M., Machado-Schiaffino, G., Dillane, E., Coughlan, J., Horreo, J. L., Bowkett, A. E., Minting, P., Toms, S., Roche, W., Gargan, P., McGinnity, P., Cross, T., Bright, D., Garcia-Vazquez, E. & Stevens, J. R. (2010). Genetic stock identification of Atlantic salmon (*Salmo salar*) populations in the southern part of the European range. *BMC Genetics* **11:31**.

Hess, J. E., Matala, A. P. & Narum, S. R. (2011). Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia river Basin. *Molecular Ecology Resources* **11**, 137-149.

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W. & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**, 117-122.

Houston, R., Davey, J., Bishop, S., Lowe, N., Mota-Velasco, J., Hamilton, A., Guy, D., Tinch, A., Thomson, M., Blaxter, M., Gharbi, K., Bron, J. & Taggart, J. (2012). Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics* **13**, 244.

Ikediashi, C., Billington, S. & Stevens, J. R. (2012). The origins of Atlantic salmon *(Salmo salar* L.) recolonizing the River Mersey in northwest England. *Ecology and Evolution*, 11.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283.

Pongpanich, M., Sullivan, P. F. & Tzeng, J.-Y. (2010). A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics* **26**, 1731-1737.

Royce, W. F. (1984). *Intoduction to the practice of fishery science.* New York: Academic Press.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406-425.

Tabangin, M., Woo, J. & Martin, L. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings* **3**, S41.

Thorstad, E. B., Whoriskey, F., Rikardsen, A. H. & Aarestrup, K. (2011). Aquatic Nomads: The Life and Migrations of the Atlantic Salmon. In *Atlantic Salmon Ecology* (Aas, O., Einum, S., Klemetsen, A. & Skurdal, J., eds.), pp. 1-32: Wiley-Blackwell.

Vasemägi, A., Gross, R., Paaver, T., Kangur, M., Nilsson, J. & Eriksson, L. O. (2001). Identification of the origin of an Atlantic salmon (*Salmo salar* L.) population in a recently recolonized river in the Baltic Sea. *Molecular Ecology* **10**, 2877-2882.

Waples, R. S. (2010). High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology* **19**, 2599-2601.

Waples, R. S., Kalinowski, S. T. & Anderson, E. C. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* **65**, 1475-1486.

Webb, J., Verspoor, E., Aubin-Horth, N., Romakkaniemi, A. & Amiro, P. (2007). The Atlantic Salmon. In *The Atlantic Salmon: Genetics, Conservation and Management* (Verspoor, E., Stradmeyer, L. & Nielsen, J. L., eds.), pp. 17-45. Oxford, U.K.: Blackwell Publishing.

The Scottish Government
Riaghaltas na h-Alba

Marine Scotland Science
Freshwater Laboratory
Faskally
Pitlochry
PH16 5LB

Copies of this report are available from the Marine Scotland website at www.scotland.gov.uk/marinescotland

ISSN: 2043-7722 ISBN: 978-1-78412-465-6 (web only)

The Scottish Government
St Andrew’s House
Edinburgh
EH1 3DG

Produced for the Scottish Government by APS Group Scotland
DPPAS26004 (05/14)

Published by the Scottish Government, May 2014