Council

CNL(07)51

ToR (f) - to review and provide recommendations on the application state of the art Genetic Stock Identification methods, with particular emphasis on evaluating the precision for identifying the population of origin of individual Atlantic salmon

ToR (f) - to review and provide recommendations on the application state of the art Genetic Stock Identification methods, with particular emphasis on evaluating the precision for identifying the population of origin of individual Atlantic salmon

Dorte Bekkevold, Tom Cross, Eileen Dillane, Riho Gross, Marja-Liisa Koljonen, and Phil McGinnity

1.1 Introduction

Genetic stock identification (GSI) techniques have been successfully used to resolve a number of, what were presumed until relatively recently intractable, salmon fisheries management questions. For example, in determining the relative proportions of contributing populations in mixed stock fisheries: at the macro scale, separating stocks of North American and European origin at West Greenland (King et al. 2001); at the meso scale, apportioning catches from the Baltic Sea (Koljonen, 2006) and the Irish coastal fisheries (unpublished data) to individual river stocks; and even within river systems (micro scale) such as the Moy and Foyle catchments in Ireland (unpublished data), and the Teno river in Finland, allocating catches to individual river tributaries (e.g. Vähä et al. 2007). GSI, in the context of parental assignment, was also critical in the successful determination of the relative fitness of the progeny of farm escape, wild and hybrid salmon spawning in the wild (McGinnity et al. 2003) and in spatially and temporally determining levels of farm and hatchery introgression in wild populations (Clifford et al., 1998, Nielsen et al. 1999, 2001).

Salmon stocks have declined in both Europe and North America and all evidence points to there having been environmental changes in the ocean phase. One of the key issues in increasing the knowledge of the marine ecology of the species is understanding of the differences which may occur between the sea distribution of different regional stock groups and river/tributary populations. This requires an ability, not only to determine the proportions of contributing populations in a given sample of fish captured in the marine environment, but also to identify the individual fish in that sample to their river or region of origin. Information on individual fish can then be used to map the distribution and migration patterns of different genetic stock groups. This in turn will provide the basis for the development of more informative eco-genetic models linking growth performance, environmental conditions, and the distribution of food organisms.

Genetic individual assignment has many advantages over physical tagging methods such as coded wire tags. The information derived from genetic identification can be done from the fish in the wild, thereby overcoming the experimental error introduced by either handling wild fish when physically tagging them, or using hatchery fish as a surrogate for wild salmon. All individuals captured in the experimental fisheries are of equal value and can be used in subsequent analysis, providing a significant cost advantage over conventional tagging where only those individuals that have tags can be used. With genetic methods there is no loss of tags and no bias due to viability or catchability effects as with external tags. In addition, the time and place of sampling can be chosen more freely and precisely, as preceding tag and release programmes are not required. Moreover, genetic identification is not dependent on fishermen in returning tags or on the detection of internal tags. Furthermore, all samples previously collected from marine surveys, i.e. historic archives of scales/otoliths, etc. can be used, and these data are of value in elucidating temporal trends in migration and distribution patterns.

Genetic methods also have some limitations. The extent of inter-population differentiation will affect the resolution power that can be achieved. Statistically significant differences in allele frequencies often occur, but quantitatively they may be too small to meet the assignment accuracy and precision requirements of the managers. Genetic assignment estimates give probabilistic information about the origin of individuals or populations rather than absolute information, a limitation common to many techniques in fisheries biology, including stock assessment, although satisfactory levels of statistical confidence within regions and individual rivers can generally be achieved for fisheries management requirements. Genetic identification of populations and individuals of Atlantic salmon has been recently reviewed by Koljonen et al. (2007).

1.2 Brief Review of the methods

The genetic analysis of the composition of population mixtures has advanced and diversified in recent years resulting in two types of approach, Mixed Stock Analysis (MSA) and Individual Assignment (IA), in which the goal is either to estimate the proportions of contributing stocks in the catch mixture, or to solve the origin of an individual fish. A commonly used Statistical Program for Analyzing Mixtures (SPAM) is available for the Windows environment (Debevec *et al.,* 2000). SPAM searches for maximum likelihood estimates of population proportions using three numerical algorithms: conjugate gradient (CG), iteratively reweighed least squares (IRLS) and expectation-maximation (EM). However, Bayesian modelling has been shown to provide the most reliable estimates of the relative contributions of different populations in mixed stock fisheries when compared with other methods (Beacham et al. 2006) and also for individual assignment (IA) when compared to GENECLASS (Cornuet et al. 1999, Luikart & England 1999)(see Koljonen et al. 2005). A number of statistical packages are available, which provide a range of methods for assigning individuals to population of origin, and these have been evaluated in the literature (e.g. Manel et al. 2005; Hauser et al 2006). The IA option incorporated into the software BAYES of Pella and Masuda (2001) seems to offer the best levels of correct assignment.

For example, up to 95% of 700 salmon caught in a recreational fishery in the tidal part of the Moy fishery in Ireland assigned to the Moy river catchment (with high levels of confidence). In this test fishing nearly 100% fish could be assumed to originate from the River Moy. Table 1 gives a comparison of this result with that from some other packages and methods.

Table 1

	cBAYES	cBAYES	SPAM	GENECLASS
method	IA	MSA	MSA	IA
Moy	95.3	89.8	77.5	43.6
Proximate catchments	2.7	5.6	10.8	11.6
Regional catchments	1.4	2.5	4.6	8.2
Outside region	0.6	2.1	7.1	36.6

In the Moy example, as well as providing the highest levels of correct assignment within the regions expected, the cBAYES assignment method gave high degrees of confidence for these assignments with 61% of the fish assigned with greater that 95% probability, and 86% with greater than 75% probability and the remainder assigned with 50% probability. These results were achieved using just 10 microsatellite loci (we recommend 15-20 for such studies which will substantially improve levels of confidence in the assignment). The probabilities may, however, be lower, when more stocks are contributing into the mixture.

When the identification of the stock of origin of the individual fish has been unsuccessful, in the sense that the probabilities for each of the stock of origin are low or too even (for example 0.3, 0.3, 0.3), individuals can be assigned to originate from groups of genetically similar stocks or regional grouping with higher levels of confidence (higher probabilities).

1.3 Sampling

Currently the most reliable individual assignments are achieved in combination with mixed-stock-analysis, in which information obtained from the genetic composition of the mixture can be utilized in addition to the multilocus genotype information of the particular individual, to determine the river of origin (Koljonen et al. 2005). Variation in the estimates may thus be derived from the mixture sample, the baseline sample, or both. However, the bias of the proportion estimates is mainly due to the baseline data, and is at its greatest when genetically similar stocks differ markedly in abundance. The variance resulting from mixture sampling depends on the size and stock composition of the mixture sample. To achieve high levels of precision, the number of fish per stock sampled within the mixture is important. If the number of contributing stocks is high, a large mixture sample is needed for reliable estimates. In the mixed-stock analysis of Atlantic salmon catches in the Baltic Sea, the 95% probability interval (confidence interval) for the stock group estimates was about ±10%, when mixture sample size of 300 fish, 8 microsatellite loci, and a baseline with 32 river stocks were used (Koljonen 2006).

1.4 Issues related to data quality

A number of issues in relation to assuring data quality must be addressed when planning and carrying out MSA and/or IA for Atlantic salmon. These include issues related to

1) assumptions for baseline samples,
2) genotyping errors and
3) choice of genetic marker system to be applied.

1.4.1 Assumptions for baseline samples

Several factors in baseline sampling may affect the amount of variation and bias in the estimates:

Hardy-Weinberg equilibrium: All currently available statistical procedures for MSA and IA assume that the genotypes used for individual baseline population information conform to Hardy-Weinberg equilibrium proportions. Including information for samples not exhibiting Hardy-Weinberg proportions may therefore pose a problem to the outcome of the statistical analyses. Reasons for failure to conform include the presence of cryptic population sub-structure, sib-group sampling and genotyping errors (for the latter, see section 2 below). The effects of such non-representative sampling on IA performance are likely to vary among specific analysis aims, and precautions can be taken by performing detailed examinations of baseline data prior to performing IA analyses, and by not causing H-W deviations by unjustified pooling of baseline data (e.g. by statistically testing for presence of sub-structure and/or sib-groups).

Temporal stability of allele frequencies: If allele frequencies vary within baseline populations over time, such changes will affect the performance of MSA and IA procedures, reducing the statistical power for correct assignment of individuals to specific rivers, and should be taken into account as a potential source of error in GSI. In theory, any effects of changes caused by genetic drift can be compensated for to a marked extent by collecting baseline data over several years (Waples, 1990). The importance of repeated sampling depends on the life history of the species concerned and on the degree of overlapping in the year-classes. Atlantic salmon have overlapping generations, and partly for that reason the temporal variation of allele frequencies in large natural stocks may be of little significance to IA performance, whereas in small natural populations or in hatchery stocks, genetic drift can cause pronounced changes. Regular validation and updating of population samples used to define baselines need to be planned as part of the estimation routine to encompass such variation.

Baseline sample size: The precision and accuracy of the estimates can be improved by increasing the baseline sample size for each baseline stock from the commonly used about 50 to about 100 fish, and ensuring that it is representative of at least two cohorts (50 from each).

Gaps in the baseline: Standard individual assignment procedures (like BAYES or GENECLASS) does not allow for the assignment of individuals to unknown source populations (i.e. populations not present in the baseline). This shortcoming may be overcome by the program HWLER (Pella and Masuda 2006), which allows unknown individuals to be assigned to a hypothetical baseline sample or samples. However, the statistical properties and levels of attainable resolution of such approaches are likely to differ from, for example, the standard method proposed by Pella and Masuda (2001).

Introgression: In relation to allele frequency stability, an important issue for Atlantic salmon is concerned with genetic effects of farmed salmon escapes and deliberate stocking from hatcheries and subsequent introgression in wild populations. In such cases MSA and IA performance may be negatively affected, as baselines generated using information for non-introgressed populations may at some point no longer adequately reflect the genetic composition of contemporary catches. Correspondingly, baseline information may have been collected at a point of time where one or more of the baseline populations were affected by genetic input from reared salmon that was lost over subsequent generations (e.g. due to selection against reared salmon genotypes under natural conditions). The impact of such introgression dynamics on MSA and IA can be assessed through simulation studies and needs to be routinely monitored for running MSA and IA programmes, by frequent re-sampling of populations/spawning rivers potentially affected by escapes or stocking in order to update baseline allele frequency information. Moreover, baseline samples need to include information for farmed salmon strains/populations. Especially for small farm brood-stocks, allele frequencies may change rapidly over time, and such variation also needs to be incorporated into sampling strategies. Simulation analyses can be used to assess the effects of introgression on MSA and IA performance.

1.4.2 Genotyping errors

Microsatellite genotyping errors Microsatellite genotyping has been shown to be error-prone (reviewed by DeWoody et al. 2006) and even modest error rates can bias estimates of population allele and genotype frequencies and thus, cause artefact deviations from Hardy-Weinberg equilibrium, which is a fundamental assumption for baseline samples of many MSA and IA methods. Compounded over multiple loci, even a small per-locus genotyping error rate can result in relatively large probabilities of a multilocus genotype containing at least one error (Creel et al. 2003; Bonin et al. 2004; Hoffman & Amos 2005), although error rates are rarely equal across loci, and dropping a single locus may provide a disproportionate decrease in error rate. Checks can be built into the system to address these (Taberlet et al. 2004). See Appendix 1for Quality Control measures.

1.4.3 Marker selection

Number of loci: The required number of loci depends on the level of differentiation among stocks and is case-specific. However, there obviously is a level, where there is no diagnostic advantage in having additional markers. At the moment 15 microsatellites should be adequate for most MSA and IA analyses. In specific situations, however, the optimum number of loci has to be determined.

New markers: To date, most MSA and IA analyses in Atlantic salmon have been carried out employing genetic information from microsatellite DNA markers, which has proven to perform well in terms of statistical properties for assignment and technical reproducibility. Nonetheless, other approaches also exist, such as analysis of Major Histocompatibility gene Complex variation and single nucleotide polymorphisms (SNPs), which are being routinely employed for GSI approaches in other species, including salmonids (e.g. Beacham et al. 2004; Smith et al. 2005). In comparison with microsatellite markers, SNP screening is expected to be less affected by DNA quality and inter-laboratory variation. However, as SNPs are commonly bi-allelic, GSI analyses normally require screening of a larger number of loci (commonly > than 2-3-fold) compared with microsatellites. It is envisaged that in the future SNPs are going to be the population markers of choice across fish species and taxa, but whether Atlantic salmon GSI approaches would benefit from including SNP marker application remains to be examined.

Further research is needed to identify DNA markers associated with protein variation and other genetic variation defining regional groupings of populations, which can be used to achieve regional assignment in a practical cost effective way in support of marine ecological studies. There are now available a number of classes of DNA markers which could be applied for this purpose. For example, existing work shows point mutations in mtDNA with highly restricted regional distributions that could be informative for some regional groups but further work is needed to confirm their diagnostic potential and to identify a suite of markers to comprehensively cover the European range of salmon. Additionally work to date shows regionally restricted distributions of microsatellite alleles and varying levels of regional differentiations among different microsatellite loci. Approximately 1700 microsatellite loci have been identified in Atlantic salmon and those optimal for use in regional discrimination remain to be identified. Furthermore, as can be inferred from genetic protein studies of loci such as MEP-2*, there is considerable potential for identifying single nucleotide polymorphisms (SNPs) (McMeel et al. 2001) with the capacity to contribute regional assignment (Rengmark et al., 2006).

1.5 Recommendations

As an overriding recommendation we are convinced that in most circumstances IA can give valuable information for Atlantic salmon management and specifically identify the population of origin of individual Atlantic salmon with relatively high probabilities.

1) We recommend that genetic stock identification methods be applied to addressing salmon biology and fisheries management questions, e.g. contribution of individual rivers to fisheries, identification of migration and distribution patterns, introgressions between hatchery and wild fish, temporal changes in population structure. This should be reported within the appropriate probabilistic framework.

2) Presently, according to performed comparison tests, the Bayesian approach of Pella & Masuda (2001) appears to provide the most accurate results with regard to individual assignment, and we therefore recommend its use. In specific cases other methods can be useful.

3) Ideally all contributing stocks should be included in the baseline, and discrete entities within rivers should be recognised, as well as those among rivers (i.e. the sampled units should be in Hardy-Weinberg equilibrium). Efforts must be made to make baseline samples representative of each population, which may subsequently be sampled at sea. Sampling programs should be prioritised by targeting the most productive rivers, rivers where conservation limits are not being reached, and should also include hatchery stocks (if not identifiable by physical markers).

4) Baseline sample sizes of 100, representative of at least two cohorts (50 from each), from each population are recommended. Sample size should be consistent across baselines.

5) It is recommended that baseline populations be re-sampled every 5-10 years. However, where introgressions from cultured fish are suspected, or where population sizes are quite small, it is recommended that this should be done on a more frequent basis.

6) The sufficient size of a mixture sample is dependent on the required precision level and the number of stocks occurring in the mixture, and can be determined by simulations prior to sampling plan.

7) If microsatellites are the marker of choice, recent reviews suggest, that a minimum of 15 polymorphic loci should be used for individual assignments. These should be investigated to ensure that null alleles, linkage, allelic dropouts and stutter bands do not cause problems e.g. microsatellite loci having a tetranucleotide repeat motif are preferred to loci having a dinucleotide repeat motif as they are easier to score and typically display no stuttering patterns.

8) To guarantee consistency of genotyping results, calibration of allele sizes among participating labs is warranted by exchange of reference samples, and baseline samples should be preferably genotyped in a single lab.

9) Accuracy and precision (depending on level required) of the assessment estimates must be examined by simulation studies, and using true mixtures of fish of known origin, not included in the baseline data set, before applying to true catch data. Simulations should be used for defining the needed catch mixture sample sizes, in order to achieve the required confidence. This information can be incorporated into the fisheries sampling protocols.

10) The highest resolution to be aimed for should be assignment to river of origin (corresponding H-W unit). In cases where the probability levels of the assignments for individual rivers remain low, pooling of probabilities over individual rivers can be employed to achieve a regional assignment for individuals. In designing programmes and models for handling assignment data, managers should be aware of the probabilistic nature of the data and the required level of probabilities and statistical confidence needed for utilising the data for combined studies, (e.g. for some ecological or fisheries studies, it may be sufficient to combine information from high and low assignment values to make a composite sample).

11) Current methods utilise microsatellite technology, however the development and application of novel microsatellite markers, as well as nuclear and mtDNA SNPs etc. should be investigated with regard to their diagnostic usefulness as river and regional specific identifiers. Also, some consideration should be given to the integration of genetic and biological data for assignment.

12) Development of statistical methods for utilising probabilistic individual assignment results is recommended.

1.6 References

Beacham, T.D., M. Lapointe, J.R. Candy, B. McIntosh, C. MacConnachie, A. Tabata, K. Kaukinen, L. Deng, K.M. Miller, and R.E. Withler. 2004. Stock identification of Fraser River sockeye salmon (Oncorhynchus nerka) using microsatellites and major histocompatibility complex variation. *Trans. Am. Fish. Soc*. 133: 1106-1126.

Beacham, T.D., Candy, J.R., Jonsen, K.L., Supernault, J., Wetklo, M., Deng, L., Miller, K.M. & Withler, R.E. (2006) Estimation of stock composition and individual identification of Chinook salmon across the Pacific rim by use of microsatellite variation. *Transacations of the American Fisheries Society* 135, 861-888.

Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetic studies. *Molecular Ecology*, 13, 3261–3273.

Clifford, S.L., McGinnity, P. & Ferguson, A. (1998) Genetic changes in Atlantic salmon (*Salmo salar*) populations of northwest Irish rivers resulting from escapes of adult farm salmon. *Canadian Journal of Fisheries and Aquatic Sciences* **55**, 358-363.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., and Solignac, M. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics **153**: 1989-2000.

Creel S, Spong G, Sands J. L. et al. (2003) Population size estimation in yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, 12, 2003–2009

Debevec, E.M., Gates, R.B., Masuda, M., Pella, J., Reynolds, J. & Seeb L.W. (2000) SPAM (Version 3.2): Statistics program for analyzing mixtures. *Journal of Heredity* 91, 509-510.

Dewoody J, Nason JD, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. Molecular Ecology Notes, 6 , 951–957.

Hauser, L., Seamons, T.R., Dauer, M., Naish, K.A. & Quinn, T.P. (2006) An empirical verification of population assignment methods by marking and parentage data: hatchery and wild steelhead (*Oncorhynchus mykiss*) in Forks Creek, Washington, USA. *Molecular Ecology* 15, 3157-3173.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology, 14, 599–612.

King, T.L., Kalinowski, S.T., Schill, W.B., Spidle, A.P., Lubinski, B.A. (2001) Population structure of Atlantic salmon (*Salmo salar* L.): a range wide perspective from microsatellite DNA variation. *Molecular Ecology*, **10**, 807-821.

Koljonen, M-L. (2006). Annual changes in the proportions of wild and hatchery Atlantic salmon (*Salmo salar*) caught in the Baltic Sea. *ICES Journal of Marine Science* 63: 1274-1285.

Koljonen, M.-L., Pella, J. J., Masuda, M. (2005). Classical individual assignments versus mixture modeling to estimate stock proportions in Atlantic salmon (*Salmo salar*) catches from DNA microsatellite data. *Canadian Journal of Fisheries and Aquatic Sciences* 62, 2143-2158.

Koljonen, M-L. King, T. and Nielsen, E. (2007). Genetic Identification of populations and individuals. In: *The Atlantic salmon: Genetics, Management and Conservation* (eds. E.Verspoor, J. Nielsen and L. Stradmeyer). Blackwell Publishing. In press.

Luikart, G., and England, P.R. 1999. Statistical analysis of microsatellite DNA data. TREE **14**: 253-256.

Manel, S., Gaggiotti, O.E. & Waples, R.S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution* 20, 136-142.

McMeel, O.M., Hoey, E.M., Ferguson, A. (2001) Partial nucleotide sequences, and routine typing of polymerase chain reaction-restriction fragment length polymorphism, of the brown trout (*Salmo trutta*) lactate dehydrogenase, LDH-C1*90 and *100 alleles. *Molecular Ecology*, 10, 29-34.

McGinnity, P., Prodöhl, P., Ferguson, A., Hynes, R., Ó Maoiléidigh, N., Baker, N., Cotter, D., O'Hea, B., Cooke, D., Rogan, G., Taggart, J. & Cross, T. (2003) Fitness reduction and potential extinction of wild populations of Atlantic salmon *Salmo salar* as a result of interactions with escaped farm salmon. *Proceedings of the Royal Society of London Series B* **270**, 2443-2450. DOI 10.1098/rspb.2003.2520.

Nielsen, E. E., Hansen, M. M. & Loeschke, V. (1999). Genetic variation in time and space: microsatellite analyses of extinct and extant populations of Atlantic salmon. *Evolution* **53**, 261-268.

Nielsen, E. E., Hansen, M. M. & Bach, L. A. (2001). Looking for a needle in a haystack: Discovery of indigenous Atlantic salmon (*Salmo salar* L.) in stocked populations. *Conservation Genetics* **2**, 219-232.

Pella, J. & Masuda, M. (2001) Bayesian methods for analysis of stock mixtures from genetic characters. *Fisheries Bulletin* 99, 151-167.

Pella, J. & Masuda, M. (2006) The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences* 63, 576-596.

Rengmark, A.H., Slettan, A., Skaala, O., Lie, O., Frode, L. (2006) Genetic variability in wild and farmed Atlantic salmon (*Salmo salar*) strains estimated by SNP and microsatellites. *Aquaculture*, 253, 229-237.

Smith, C.T., Templin, W.D., Seeb, J.E., Seeb, U.W. (2005). Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of US and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* 25 (3): 944-953.

Taberlet P, Griffin S, Goossens B et al. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research, 24, 3189–1394.

Vähä, J-P, Erkinaro, J., Niemela, E., Primmer, C.R. (2007) Life-history and habitiat features influence within-river genetic structure of Atlantic salmon. *Molecular Ecology*, in press.

Waples, R.S. (1990) Temporal changes of allele frequency in Pacific salmon: implications for mixed-stock fishery analysis. *Canadian Journal of Fisheries and Aquatic Sciences* 47, 968-976.