Salmon tracing: Genotyping to trace back escapees from salmon aquaculture

R.J.W. Blonk

Report number C029/14

IMARES Wageningen UR

(IMARES - Institute for Marine Resources & Ecosystem Studies)

Client: Nofima AS
Mr Matthew Baranski
P.O. Box 6122
NO-9291 Tromsø, Norway

Publication date: 20 February 2014

IMARES WAGENINGEN UR

IMARES is:

- an independent, objective and authoritative institute that provides knowledge necessary for an integrated sustainable protection, exploitation and spatial use of the sea and coastal zones;
- an institute that provides knowledge necessary for an integrated sustainable protection, exploitation and spatial use of the sea and coastal zones;
- a key, proactive player in national and international marine networks (including ICES and EFARO).

P.O. Box 68
1970 AB IJmuiden
Phone: +31 (0)317 48 09 00
Fax: +31 (0)317 48 73 26
E-Mail: imares@wur.nl
www.imares.wur.nl

P.O. Box 77
4400 AB Yerseke
Phone: +31 (0)317 48 09 00
Fax: +31 (0)317 48 73 59
E-Mail: imares@wur.nl
www.imares.wur.nl

P.O. Box 57
1780 AB Den Helder
Phone: +31 (0)317 48 09 00
Fax: +31 (0)223 63 06 87
E-Mail: imares@wur.nl
www.imares.wur.nl

P.O. Box 167
1790 AD Den Burg Texel
Phone: +31 (0)317 48 09 00
Fax: +31 (0)317 48 73 62
E-Mail: imares@wur.nl
www.imares.wur.nl

A_4_3_2-V13.3

Contents

Summary

The overall objective of the project is to assign an escaped salmon back to the farm responsible for the escape with near 100% accuracy. In this report, the potential of a set of genetic markers to assign an escaped salmon was determined for a set of 12 polymorphic microsatellite markers, provided by Nofima, and by using stochastic simulation. Also, the effect of different numbers of sires, and the effect of pooling of multiple sires in crosses was determined.

The effect of the number of sires included was as expected with less sires resulting in lower allocation power. However, for the currently believed common numbers of sires and dams used for production stocks this still resulted in very high allocation power. The effect of pooling of 3 sires in one cross was small, thus leaving room for the salmon breeding companies to put higher selection intensities in their breeding program.

In general, it can be concluded that, based on the genetic data provided, the current set of polymorphic microsatellite markers is enough to trace back most individual salmon back to their farm of origin, assuming that for each farm, the crosses provided are known, and that one cross is only provided to one farm. However, to be 100% accurate, the set of markers needs to be enlarged for ambiguously allocated individuals or a combination of genetic markers and a phenotypic markers such as a tag could be considered.

The Sub delivery provided described in this report is linked to WP3 in the execution of the project “Industry-wide tracing of Norwegian farmed Atlantic salmon”.

1. Introduction

Parental allocation methods using genetic markers and software programs to perform parental allocation have been extensively described in literature (Marshall et al. 1998; Duchesne et al. 2002; Wang 2004). These programs typically use genotypes of individual offspring and putative parents to reconstruct pedigrees. The success of these methods strongly depends on the quality of the genetic markers as determined by potential presence of null alleles, heterozygosity of the markers in the population, number of alleles and allele frequencies. To test the allocation power of a set of markers, most software programs provide options for simulation of populations, based on provided information on the marker quality. However, the currently available programs are not able to simulate large populations, such as true commercial population sizes, and specific breeding structures.

The overall objective of the project is to assign an escaped salmon back to the farm or company responsible for the escape with near 100% accuracy. In this report, the potential of genetic markers to assign an escaped salmon back to its original farm was assessed using basic population genetic data from a part of the commercial salmon breeding population in Norway. To assess the potential of genetic markers in a realistic situation, the power to allocate an individual back to both its parents (and thus to its farm of origin) was determined for a set of 12 polymorphic microsatellite markers in a simulated population nearing the realistic commercial size in Norway and its underlying breeding structures. Also, the effect of different numbers of sires, and the effect of pooling of multiple sires in crosses was determined.

2. Materials and Methods

Information of the markers used

Information on the markers was provided by NOFIMA and based on a part of the current commercial breeding population of Atlantic salmon (*Salmo salar*) in Norway (table 1).

Table 1 Locus name, number of alleles and observed heterozygosity in a part of the current commercial breeding population of Atlantic salmon (Salmo salar) in Norway. Assumed to be a representative sample of the total breeding population.

Locus name	Na	Het obs
SAL_CIG_32	22	0.8432
SAL_ICISG_11	16	0.7947
SAL_ICISG_37	15	0.6571
SsaA124-low	11	0.5307
Sssp2216	16	0.8476
SAL_CIG_33	24	0.8287
SAL_ICISG_16	24	0.8771
SAL_ICISG_01	9	0.7591
SAL_ICISG_05	17	0.7731
SAL_CIG_35	29	0.8363
SAL_ICISG_06	10	0.8008
SAL_CIG_37	16	0.8631

Stochastic simulation of populations

To determine allocation success of the selected set of markers (table 1), a stochastic simulation program was written in R (R Development Core Team 2008). In this program, given the marker information provided, random genotypes were generated for a defined number of sires and dams. To define population structure, different mating schemes were defined, including pooling of 1 or 3 males with one female.

Allocation power

To determine the allocation power of a set of markers within a given population size and structure, *matching of parental alleles in pairwise compared crosses of parents* was scored for each locus. For example, consider a 1 locus model where 2 crosses are compared. Here, alleles *a* in cross 1 are represented by a_i for sire *i* and by a_k, for dam *k,* where alleles *a* in cross 2 are represented by a_j for sire *j* and by a_l for dam *l*. Matching of one or more alleles was scored as $I(a_i,a_j) + I(a_k,a_l)$ and as $I(a_i,a_l) + I(a_k,a_j)$. Here, *I* is the identity between alleles, and *I* was set 1 when one or more of the compared alleles are identical, and zero otherwise. Cases with at least one comparison equal to 2 were considered as one match for this particular locus.

Moreover, to obtain the total matching of loci in a pairwise compared cross, matching of parental alleles was summed for all loci and from this the allocation power of the set of markers follows. For example: in a situation with 12 tested loci and where two pairwise compared crosses of parents have 0 matching loci (i.e. 0 matches), these particular crosses will always produce offspring with different genotypes. In contrast, when two pairwise compared crosses of parents have 12 matches (a “full match”), these particular crosses can both produce genotypic identical offspring, at least for these markers.

Obviously, when comparing genotypes of offspring from two crosses in which the crossed parents have a full match on all loci, it will not always be possible to determine the true parents (and the farm of origin when individual crosses are only kept at one farm) with 100% certainty for all individual offspring. In

contrast, for all pairwise compared crosses with at least one non-matching locus, in principle there will be enough information (unique alleles) to unambiguously allocate individuals to the parents and thus the farm of origin. Therefore this method was considered a good measure for the allocation power of the set of markers in this particular application.

Within each simulated population, the number of matches was calculated for all possible pairwise compared crosses. From this, the frequency (i.e. the number of pairwise compared crosses) for all possible numbers of matches (0 until N_{loci}) in the population was determined. To determine the allocation power of a certain set of markers in a given situation, the frequency of full matches was particularly considered.

Scenario's analysed

So assess effects of the number of loci used, number of sires and dams in the population, the population structure and pooling of males in crosses, several scenarios were simulated (see table *2*). The scenarios with 33.000 dams and 660 sires, i.e. mating one sire on 50 dams, were assumed to represent the current situation of commercial salmon production stocks in Norway, whereas the scenarios with 33.000 dams and 330 sires, i.e. mating one sire on 100 dams, and pooling of 3 sires were assumed to represent the potential future situation. In the situation with pooling of sires, the same sire:dam ratios were used while pooling was simulated by combining genotypes of one dam with 3 sires resulting in effectively mating 1 sire to 300 different dams in the future situation. Each scenario was replicated 10 times, and means and standard deviations of the frequencies for all potential number of matches were calculated.

In scenario A and C, 17 alleles per marker were used and random heterozygosity was programmed. In scenario B and D, the number of alleles and the heterozygosity was based on a genotyped part of the current commercial breeding population of Atlantic salmon in Norway (data provided by Nofima). In these analyses effects of null alleles and erroneous genotyping are neglected.

Table 2. Simulated scenarios to assess allocation power of genetic markers to assign an escaped salmon back to its original farm.

Scenario	Dams	Sires	Pooling sires	Loci	Na	Het obs
A	1000	50,100	1	3	17	Random
	1000	50,100	1	6	17	Random
	1000	50,100	1	9	17	Random
	1000	50,100	1	12	17	Random
B	1000	50,100,200,500	1	12	Sample	Sample
C	33000	660	1	3,6,9,12	17	Random
D	33000	660	1,3	12	Sample	Sample
	33000	330	1,3	12	Sample	Sample

Pooling sires = the number of sires mated with one female. Na = number of alleles; Het obs = heterozygosities used; Sample = information based on genotyped part of the current commercial breeding population of Atlantic salmon in Norway; For scenarios A and C, the number of alleles was set to 17 (the average number of alleles in the real sampled population;

3. Results

Distribution of frequencies for all possible numbers of matches in populations of 1000 dams, 50 sires and 12 markers (scenario A) are shown in figure *1*. In this situation, 17 alleles per marker were used and random heterozygosity was programmed. In this situation, most of the compared pairs in the simulated population show 4 to 5 matches, whereas almost no pairs had either 0, 10, 11 or 12 matches.

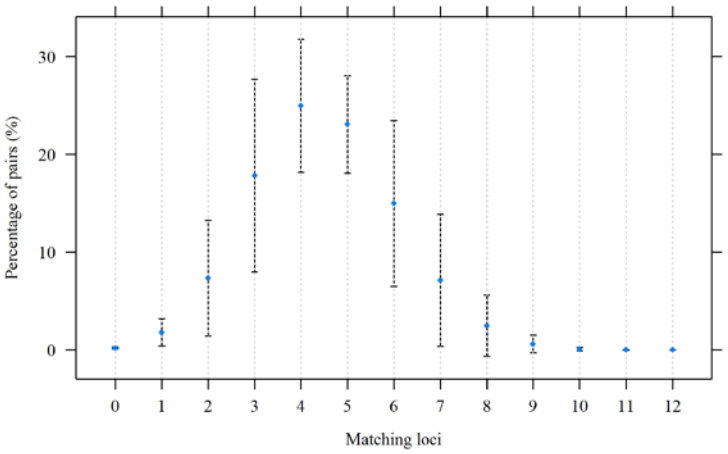


Figure 1. Distribution of frequencies (percentage of all possible pairwise compared crosses) for all possible numbers of matches (matching loci), in a population of 1000 dams, 50 sires and 12 loci, all with 17 alleles and random heterozygosity.(Scenario A)

The effect of the number of used loci on the percentage of full matches in a population of 1000 dams with 50 or 100 sires is shown in figure *2*. In this scenario, 17 alleles and random heterozygosity were programmed (scenario A). From figure *2* it is clear that, although standard deviations are high, most full matches occur with 3 to 6 loci, indicating that both for cases with 50 and 100 sires, such sets of markers are less suitable to accurately allocate fish to farms, at least in the current population size and structure.

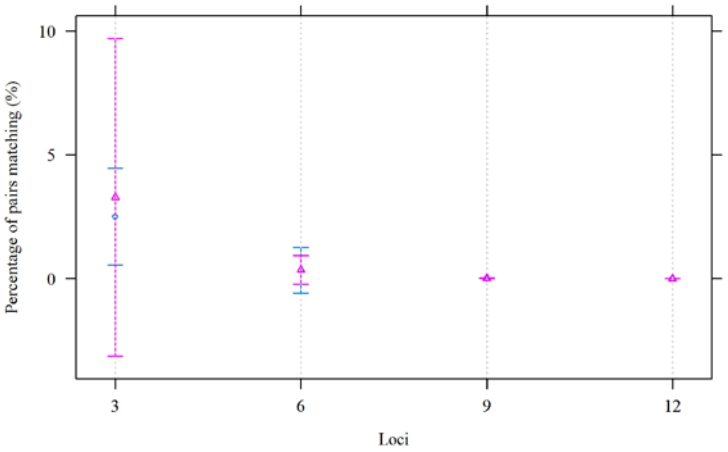


Figure 2. The relation between percentage of full matches and the number of used loci in a population of 1000 dams with 50 or 100 sires, 17 alleles and random heterozygosity (scenario A).

The effect of the number of sires on the percentage of full matches in a population of 1000 dams genotyped for 12 markers with sampled numbers of alleles and heterozygosity (scenario B) is shown in figure 3. From this figure it can be seen that at this population size, relatively small numbers of sires imply higher frequency of full matches, thus corresponding to lower allocation power. This is probably caused by the fact that the same number of dams, using more sires implies production of more families and thus more divergent genotypes, resulting is a smaller chance of matching loci.

When comparing results from scenario A and B, at 50 and 100 sires, the effect of using a realistic marker set can be determined as the marker set as used in scenario B is based on information of a genotyped part of the current commercial breeding population of Atlantic salmon, whereas A is not. When using 50 sires, the percentages of full matching were 0.001502 (sd=0.002095) for a situation with equal numbers of alleles and random heterozygosity (scenario A) whereas for a more realistic situation (scenario B), a much lower value 0.000701 (sd = 0.000380) was found. This implies that the realistic markers set is more informative. However, when using 100 sires, results were much less different with percentages of 0.000350 (sd = 0.000192) for scenario A and percentages of 0.000561 (sd = 0.000363) for B. Very likely the effect of the number of sires and families plays a role in the latter.

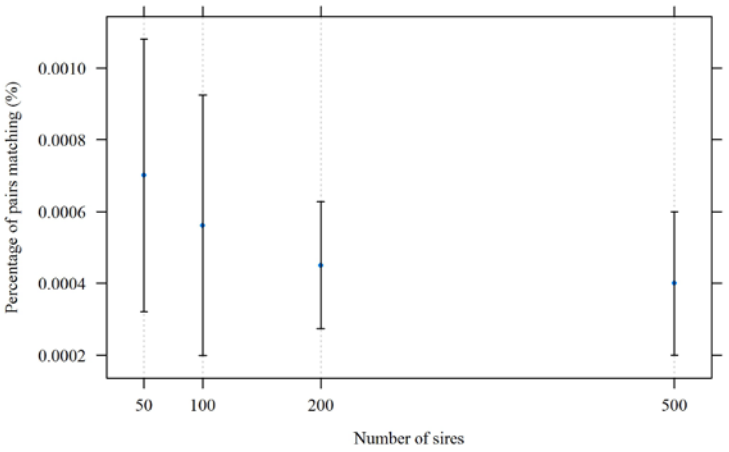


Figure 3. Relation between the percentage of full matches and the number of sires, in a population of 1000 dams, with 12 markers and sampled numbers of alleles and heterozygosity (scenario B).

The effect of the number of used loci on the percentage of full matches in a population of 33,000 dams with 660 sires, i.e. mating one sire on 50 dams, is shown in figure 4. In this scenario, 17 alleles and random heterozygosity were programmed (scenario C). From figure 4 it can be seen that most full matches occur with 3 loci. This shows that with large realistic population sizes, really more than 6 markers are required to accurately allocate escaped fish to parents and farms. The percentages of full matches where 0.008641 (sd = 0.018309) for 9 loci and 0.000143 (sd = 0.000256) for 12 loci.

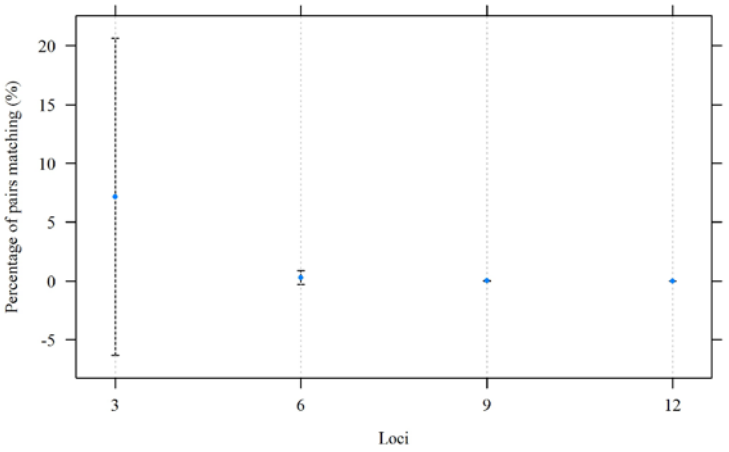


Figure 4. The relation between percentage of full matches and the number of used loci in a population of 33,000 dams with 660 sires, 17 alleles and random heterozygosity (scenario C).

The effect of number of sires in a population of 33,000 dams, and pooling of males per cross (scenario D), can be seen in figure 5. This is the most realistic scenario. It can be seen that in general, using more sires in the population, increases the allocation power of the set of genetic markers. However, in both cases, percentage of full matches is very low, but dropping from 0.00028 (sd = 0.000026) with the use of 330 sires to 0.000211 (sd = 0.000022) for 660 sires. The effect of pooling of males, i.e. mixing sperm of 3 males and then crossing with 50 or 100 females is very limited, mainly showing its effect in populations with 660 males.

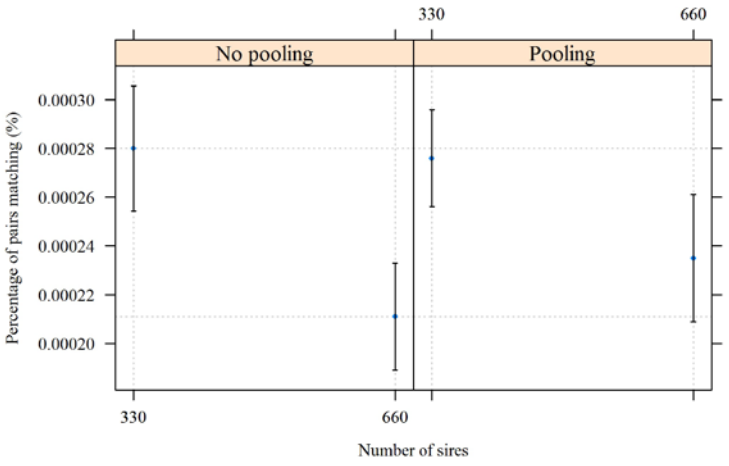


Figure 5. The relation between percentage of full matches and the number of used sires (330 vs 660) in a population of 33,000 dams, with 12 loci and sampled numbers of alleles and heterozygosity (scenario D).

4. Conclusions

The effect of number of loci included in the set or the number of sires in the population, as well as total population size was as expected with less loci, less sires or smaller populations (i.e. less families) resulting in lower allocation power. However, for common numbers of sires and dams used for production stocks in Norway the proposed set of 12 highly polymorphic markers this still resulted in very high allocation power (percentage of parental pairs with matching offspring genotypes between 0.000211 and 0.00028%). The effect of pooling of 3 sires in one cross was small, thus suggesting that there is still room for the salmon breeding companies to put higher selection intensities on their selected stock.

In general, it can be concluded that, based on the genetic data provided, the current set of polymorphic microsatellite markers is able to accurately trace back most individual salmon back to their farm of origin, assuming that for each farm the crosses provided are known, and that one cross is only provided to one farm. However, with a likely scenario where 33,000 dams and 660 sires without pools are used to produce the production stocks in Norway, a percentage of full matches of crosses of 0.000211% can be expected. In practice this means that with 33,000 dams in the population, ((33,000^2-33,000)/2 ≈) 544 $*10^6$ pairwise comparisons can be made, implying approximately 1150 pairs with full matches that produce some offspring with identical genotypes. Thus, in this case one can expect between 1,150 and 2,230 families in a population (i.e. 3-7% of the families produced, see table 3) that include at least some offspring with similar genotypes.

Table 3. Numbers of families (crosses) producing at least some offspring with full matching genotypes with other families, in different scenarios.

Dams	Sires	Pooling	Full matches		Pairwise comparisons	Families with full match			
			mean%	sd		min	max	min%	max%
33000	660	1	0.000211	0.000022	544,483,500	1,151	2,302	3.5%	7.0%
33000	330	1	0.000280	0.000026	544,483,500	1,523	3,046	4.6%	9.2%
33000	660	3	0.000235	0.000026	1,633,450,500	3,835	7,669	11.6%	23.2%
33000	330	3	0.000276	0.000020	1,633,450,500	4,506	9,011	13.7%	27.3%

Pooling = number of sires mated with 1 female.

Should these offspring escape from a farm, it would be difficult to unambiguously allocate them back to the farm of origin. It is therefore suggested that to be 100% accurate, either the set of markers should be enlarged, at least for the animals that ambiguous allocation to parents and farms, and/or a combination of genetic markers and a phenotypic markers such as a tag should be considered.

Recommendations:

1) Full genotypes of all used parents, and mating schemes are required to determine the true power of the set of markers.
2) A larger set of genetic markers is needed to more accurately trace back individuals to their parents and farm of origin.
3) A comparison should be made with wild salmon genotypes for the current markers to determine whether the current set of markers can distinguish wild from farmed salmon.

5. Quality Assurance

IMARES utilises an ISO 9001:2008 certified quality management system (certificate number: 124296-2012-AQ-NLD-RvA). This certificate is valid until 15 December 2015. The organisation has been certified since 27 February 2001. The certification was issued by DNV Certification B.V. Furthermore, the chemical laboratory of the Fish Division has NEN-EN-ISO/IEC 17025:2005 accreditation for test laboratories with number L097. This accreditation is valid until 1th of April 2017 and was first issued on 27 March 1997. Accreditation was granted by the Council for Accreditation.

References

Duchesne, P., Godbout, M. H. and Bernatchez, L., 2002. PAPA (Package for the Analysis of Parental Allocation) : A computer program for simulated and real parental allocation. Molecular Ecology Notes 2, 191-194.

Marshall, T. C., Slate, J., Kruuk, L. E. B. and Pemberton, J. M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7, 639-655.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org.

Wang, J., 2004. Sibship Reconstruction From Genetic Data With Typing Errors. Genetics 166, 1963-1979.

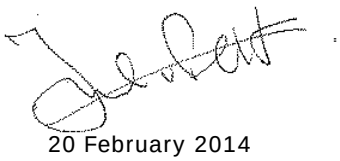
Justification

Report number : C029/13
Project Number : 4304106201

The scientific quality of this report has been peer reviewed by the a colleague scientist and the head of the department of IMARES.

Approved: dr. H. van Pelt-Heerschap
Researcher genetics

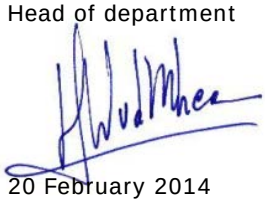
Signature:



Date: 20 February 2014

Approved: Ir. H. van der Mheen
Head of department

Signature:



Date: 20 February 2014