Accepted Manuscript

A workflow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon

Luke E. Holman, Daniel Garcia de la serrana, Aubrie Onoufriou, Borghild Hillestad, Ian A. Johnston

PII: S0044-8486(16)31219-4
DOI: doi: 10.1016/j.aquaculture.2017.04.001
Reference: AQUA 632595

To appear in: *aquaculture*

Received date: 16 December 2016
Revised date: 17 March 2017
Accepted date: 1 April 2017

Please cite this article as: Luke E. Holman, Daniel Garcia de la serrana, Aubrie Onoufriou, Borghild Hillestad, Ian A. Johnston , A workflow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Aqua(2017), doi: 10.1016/j.aquaculture.2017.04.001

A workflow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon

Luke E Holman[1,2*], Daniel Garcia de la serrana[1], Aubrie Onoufriou[2], Borghild Hillestad[3], Ian A Johnston[1]

1 - Scottish Oceans Institute, School of Biology, University of St Andrews, St Andrews, KY16 8LB, Scotland, UK

2 - Xelect Ltd, Horizon House, Abbey Walk, St Andrews, KY16 9LB, Scotland, UK

3 - SalmoBreed AS, Sandviksboder 3A, N-5035 Bergen, Norway

Keywords

SNP; parentage assignment; Atlantic salmon; traceability; pedigree; workflow

Abstract

Accurate parentage assignment is key for the development of a successful breeding program, allowing pedigree reconstruction from mixed families and control of inbreeding. In the present study we developed a workflow for the design of an efficient single nucleotide polymorphism (SNP) panel for paternity assignment and validated it in Atlantic salmon (*Salmo salar* L.). A total of 86,468 SNPs were identified from Restriction Site Associated DNA Sequencing (RAD-seq) libraries, and reduced to 1,517 following the application of quality control filters and stringent selection criteria. A subsample of SNPs were chosen for the design of high-throughput SNP assays and a training set of known parents and offspring was then used to achieve further filtering. A panel comprising 94 SNPs balanced across the salmon genome were identified, providing 100% assignment accuracy in known pedigrees. Additionally, the panel was able to assign individuals to one of three farmed salmon populations used in this study with 100% accuracy. We conclude that the workflow described is suitable for the

design of cost effective parentage assignment and traceability tools for aquaculture species

1. Introduction

The provision of genetically improved stock in aquaculture production has greatly increased efficiency and profitability in species with active selective breeding programs. Pedigree information is essential for the implementation of selective breeding strategies to control for inbreeding with the aim of avoiding inbreeding depression and loss of performance (Kincaid, 1983). A pedigree can be maintained by rearing single families in individual tanks until they are large enough to be physically tagged, however this incurs high operational and capital costs (Vandeputte & Haffray, 2014). Molecular methods for pedigree determination, using neutral genetic markers such as microsatellites or single nucleotide polymorphisms (SNPs), allow mixed families to be reared together. Additionally, in some cases it may not be biologically or economically possible to perform individual crosses. Molecular markers allow for the resolution of the contribution of each parent in a group spawning event (Borrell et al., 2011; Morvensen et al., 2013). Vandeputte & Haffray (2014) provide an excellent review of computational methods and technical issues affecting the assignment power of markers including sampling variance and relatedness of parents, Hardy-Weinberg disequilibrium, genotyping errors and null alleles. In general 8-15 polymorphic microsatellite markers provide adequate assignment power for crosses involving a few tens or hundreds of parents. The increasing availability of genomic resources for aquaculture species has led to the adoption of SNPs as the markers of choice for parentage assignment. They offer several advantages over microsatellites including their suitability for high-throughput genotyping, lower genotyping errors and the ability to readily combine and standardize datasets from different laboratories (Yue & Xia, 2014).

The purpose of any SNP panel for pedigree analysis is to achieve 100% assignment accuracy while limiting the number of genotyping assays required in order to reduce cost. Computer simulations indicate around 60-100 SNPs with high (0.3-0.5) minor allele frequency (MAF) are sufficient to give accurate parentage assignment (Anderson & Garza, 2005). Liu *et al.* (2016) showed

increasing accuracy using 36 (92.5%), 48 (99.2%) and 68 (100%) SNPs in rainbow trout (*Oncorhynchus mykiss*) while Weinman et al. (2015) found similar results in cooperatively breeding birds with 10 SNPs giving <20% accuracy and 80 SNPs 100% accuracy. These studies highlight that high information content of markers in the study population, minimal linkage disequilibrium (LD) and marker neutrality are of key importance for the design of good SNP parentage panels. Despite growing consensus on appropriate filters for the selection of SNPs in the development of low-density SNP parentage panels, there has been no formalised effort to establish common workflows. This has little effect on the development of these resources in species with large industry support, but may hinder development in species of lower value or with minimal existing resources. To increase confidence in the development of low density SNP panels, workflows must be established and validated in species with well understood, complex genomes as a worst-case scenario.

Atlantic salmon (*Salmo salar* L.) is a major aquaculture species with an annual global production that has grown by 7%/year over the last several decades to reach over 2 million metric tonnes per annum (FAO, 2014). This growth rate has been supported by large-scale breeding programs established in Norway in the early seventies based on family selection for traits of interest such as late maturation, fast growth, disease resistance and flesh quality (Gjedrem et al., 1991). Family selection has recently advanced to individual selection by incorporating biotechnological approaches such as marker-assisted selection (Moen et al., 2015; Gonen et al., 2015) and genomic prediction (Jonas and de Koning, 2015) to further increase the rate of genetic gain per generation. The recent publication of the Atlantic salmon genome (Lien et al., 2016) adds to the already large collection of genomic resources available for the species including high-density linkage maps (Gonen et al., 2014, Lien et al., 2011) and validated 6K (Lien et al., 2011), 132K (Houston et al., 2014) and 151K (Yáñez et al., 2016) SNP genotyping arrays. Several microsatellite panels for parentage assignment have been developed for Atlantic salmon (O'Reilly et al., 1998; Norris et al., 2000). However, to date no publication has described a validated SNP panel for parentage assignment in Atlantic salmon.

In the present study we describe a generalized workflow for the design and validation of an efficient SNP panel for parentage assignment from RAD-seq data, using Atlantic salmon as an example, which should be widely applicable to other aquaculture species. Following the application of quality control filters, mapping to a reference genome and the design of Fluidigm SNPtype SNP genotyping assays (Fluidigm Ltd, San Francisco, USA) the most efficient panel was selected by employing a "training" set of known parents and offspring. The final panel of 94 SNPs distributed across the salmon genome achieved an accuracy assignment of 100%. We additionally determined whether the selected parentage panel retained sufficient information on population structure to be of utility as a cost effective, traceability tool able to discriminate between farmed salmon populations of different origins.

2. Methods

The workflow used to produce an efficient SNP panel consisted of a series of filters summarised in Figure 1.

2.1 SNP Discovery

Fast skeletal muscle samples were obtained from 102 adult Atlantic salmon from three commercial strains: two of Norwegian origin (n=40,41) and one of Scottish origin (n=21). For DNA extraction, 20-40mg of tissue was homogenized in SSTNE buffer (Pardo et al., 2005) containing 0.1% SDS (m/v) and 50µg of proteinase K for 3 hours at 55^{o}C and 15 minutes at 70^{o}C. RNase A was added to the extraction and incubated for 1 hour at 37^{o}C and proteins precipitated by adding 5M NaCl. DNA was recovered from the supernatant and precipitated in isopropanol. The resulting pellet was washed several times with cold ethanol 70-75%. DNA samples were subject to single digestion RAD-seq protocol by Floragenex, Inc (Portland, USA) as described by Baird et al. (2008). Briefly, samples were digested using *Sbf*1 restriction enzyme, individually barcoded with custom Floragenex adapters followed by PCR amplification of the fragments. Pooled libraries were sequenced over two lanes of the Illumina HiSeq 2000 platform (Eurofins Genomics, Ebersberg, Germany). Sequence data was de-multiplexed

and trimmed to a length of 90 base pairs using custom Floragenex scripts and mapped to the Atlantic salmon genome (ICSASG v_1, NCBI WGS Project ID AGKD03) using BOWTIE v.0.12.8 allowing up to three mismatches (Langmeid et al., 2012). SAMTOOLS (Li et al., 2009) and custom Floragenex scripts were used for SNP calling and variants were output as a Variant Call Format (VCF) file.

2.2. SNP Selection: Discovery Population

The discovery SNPs were subjected to quality control filters using VCFtools v.0.1.12b (Danecek et al., 2011) as follows: 15x minimum sequencing depth, Phred scaled genotype quality per sample of 20 and a minimum of 90% of the samples genotyped. These steps minimise the effect of sequencing errors and bioinformatic artefacts and ensure only high confidence variants are retained for downstream analysis.

Surviving SNPs were further filtered for specific properties with the aim of retaining only highly informative, neutral SNPs with minimal LD. Firstly, we omitted SNPs with more than two alleles, as tri-allelic markers are not suitable for probe-based genotyping, using VCFtools. Remaining SNPs were further filtered using SNPRelate v.1.2.0 (Zheng et al., 2012) implemented in R version 3.2.2 (R Core Team, 2014). SNPs deviating from Hardy-Weinberg equilibrium (HWE) within the three discovery populations at the $p<0.10$ level were discarded. The $p<0.10$ level was used to avoid discarding potentially useful SNPs at the $p<0.05$ threshold due to variation introduced during sampling. SNPs with a minor allele frequency, below 0.15 across all populations, were also removed. Finally, pairwise LD between SNPs was calculated and for any pair with a correlation coefficient of $R > 0.46$ one SNP of the pair was randomly excluded from further analysis.

As the reference used for the original mapping of the RAD-tags was not linked to chromosomes, the information was added by aligning a region 1kb up and downstream of each SNP to the most recent assembly of the Salmon genome with chromosome ID (ICSASG v_2, NCBI WGS Project ID AGKD04). Variant sites were annotated onto contigs, and any regions where any two SNPs were within 30bp were excluded from assay design to avoid SNPs in primer binding sites.

Finally, 50bp up and downstream of the SNP were extracted from contigs and used for Fluidigm SNPtype genotyping assay design.

2.3. SNP Genotyping

SNP genotyping using the Fludigm SNPtype assays proceeded as follows: 20-40mg of tissue was subjected to crude lysis using Proteinase K (Qiagen, Hilden, Germany) and Chelex 100 (Sigma-Aldrich, St Louis, USA). PCR template consisted of a 1:100 lysis dilution in distilled water. SNP genotyping was carried out using the Fluidigm EP1 platform. A pre-amplification step was performed using a combination of a Locus Specific Primer (LSP) and a Specific Target Amplification primer (STA). The pre-amplified product was diluted 1:100 in distilled water and subject to a second round of PCR amplification using the LSP and a set of fluorescently labelled Allele Specific Primers (ASP). SNPs were called using pre-defined algorithms implemented in the Fludigim SNP Genotyping Analysis software. The clustering setting *automatic confidence threshold* was set at 85 and genotypes were manually confirmed for each SNP.

2.4. SNP Selection: Training Population

A second set of Atlantic salmon sample crosses obtained from the SalmoBreed AS breeding program (Bergen, Norway) was used to experimentally filter the SNPs. A total of 10 families, each family consisted of dam, sire and eight offspring were used. Two sets of two families were selected to share a sire to test the ability for the panel to distinguish highly related individuals. Parental adipose fins and entire fry were preserved in 70% (v/v) ethanol and stored at room temperature. DNA extraction and SNP genotyping proceeded as described above.

PLINK v1.07 (Purcell et al., 2007) was used to check for Mendelian errors in known crosses. The surviving SNPs were ranked according to their ability to correctly assign offspring to known parent groups using COLONY v2.0.6.2 (see Parentage Assignment section). Briefly, subpanels of 45 SNPs were randomly sampled without replacement from the post-filtered SNPs for a total of 1000 replicate parentage assignment runs. The SNPs were ranked according to the ratio of successful (100%) or unsuccessful (<100%) parentage assignment runs

from the 1000 replicates. Finally, SNPs were discarded from the ranked list sequentially until the parentage assignment was 100% correct.

2.5. Parentage Assignment

Parentage assignment using the results of the SNP genotyping was performed using COLONY v.2.0.6.2 (Jones and Wang, 2009). See supplementary information for individual run parameters. Parentage assignment may include a small number of closely related candidate parents, or a much larger set of distantly related candidate parents depending on the research question. In order to model both these scenarios, two different sets of candidate parents were assigned to the offspring. One set included only the eight sires and nine dams from the experimental crosses for a total of 17 possible parents. The second set included the experimental cross parents and also the 102 samples described in the *SNP Discovery* section used as potential parents. The second set of parents featured the discovery samples twice, as both dams and sires for a total of 110 potential sires and 111 potential dams for a total of 221 possible parents. In order to determine the effect of a diminishing number of markers on parentage assignment, an R script was written to bootstrap the random selection without replacement of SNPs to produce COLONY runs with varying numbers of SNPs. For each quantity of SNPs 100 repeats were performed. GNU parallel was used to parallelize computation of serial COLONY runs (Tange, 2011).

2.6. Population Genetics

To test the ability for a small subset of SNPs to retain population substructure contained within a larger genetic dataset, the filtered panels were subject to a principal component analysis (PCA) using SNPRelate v.1.2.0. Following PCA analysis the data was subject to a leave-one-out population assignment implemented in GENECLASS2 (Piry et al., 2004) to test the ability for the markers to assign to population level. The test removes the population data from each sample sequentially and uses the remaining sample population and genetic data to assign the removed sample. In GENECLASS2 assignments the Rannala and Mountain criterion was used with an assignment threshold of 0.01 (Rannala and Mountain, 1997).

3. Results and Discussion

3.1. Sequencing Results

Although there are a large number of SNPs published for Atlantic salmon, we included an initial sequence discovery step in the workflow to make it applicable to a wide range of other aquaculture species for which there might be limited genomic resources. RAD-seq protocols provide a reduced representation of the genome and are acknowledged as a cost effective method of SNP discovery (Baird et al., 2008). In our case sequencing produced a raw total of 452.9 million reads, with a mean of 4.4 million reads per individual. Following mapping, an average 56.7% of the reads were unambiguously mapped to the reference genome. Sequence data is publically available through the EMBL-EBI Short Read Archive (SRA) under the study accession number PRJEB17687.

3.2. SNP Discovery and workflow for designing paternity assignment panel

A total of 86,468 SNP variants were initially obtained from the RAD-seq data. Following first round of quality filtering a total of 17,283 high confidence variants were retained. SNPs were further filtered for property characteristics yielding a total of 1,517 suitable SNPs. Fluidigm SNPtype Assays were ordered in three batches of 96, 45 and 40 assays. In all cases SNPs were randomly selected while balancing selections across chromosomes.

Following initial SNP genotyping trials for assay validation, only 54 (56.3%) of the assays used from the first batch produced clear genotyping clusters in the test population. For the subsequent two batches the assays were subjected to *in silico* validation before synthesis of the assays, resulting in a higher success rate. In each case primer sequence information was subject to a BLAST (v2.2.30+) search against the Atlantic salmon genome. Assays were synthesised only if primers had a single clear match to the salmon genome. Success rates were higher for the second and third batches, 33 (73.3%) and 27 (67.5%) respectively. Overall from the 181 Fludigim SNPtype assays trialled a total of 111 (61.3%) confidently distinguished between genotypes, 23 (12.7%) exhibited poor

clustering or unclear genotypes and 47 (26.0%) exhibited no clustering or poor success rates.

The success rate found here is lower compared to other published parentage panels based on Fludigim SNPtype assays such as Liu et al 2016 (79.1%), and much lower than the >99% success achieved using iPLEX gold assays in a MassARRAY system (Agena Biosciences, Hamburg, Germany) (Weinman et al., 2014). The cost of assays is a significant proportion of the total cost in the development of SNP genotyping panel, and success rate should be taken into account during budgeting. Current assay conversion success rates indicate that between 130-170 SNPs, with suitable properties, will be sufficient to develop a panel of 96 genotyping assays depending on the SNP genotyping platform used.

Additionally, researchers should be wary of organisms with complex genome architecture, such as those with recent whole genome duplications (WGD) in their evolutionary history. Salmonids underwent a WGD around 88 Mya (Macqueen & Johnston, 2014) and around 50% of the duplicated genes are retained in extant species, complicating bioinformatic analyses and primer design. This may have contributed to the lower success rates of assays between salmonid and non-salmonid species. *In silico* primer validation and a high-quality reference sequence are recommended for good assay success rate.

Out of the 111 assays that could be used for genotyping, a total of 16 exhibited more than 1 Mendelian error in the 10 families sampled. Analysis for Mendelian error at the filtering stage also revealed a single dam with 31 incompatible genotypes against her offspring. The dam and her offspring were subject to further genotyping using an additional 10 validated microsatellite markers and Mendelian errors were found in seven out of 10 markers (see supplementary information for methods). It was concluded that data from this entire family was unreliable and so it was removed from further analyses. Unreliable data can result from the unintentional crossing of brood stock animals in aquaculture facilities (Morvensen et al., 2013), or from a mislabelled sample. The use of existing validated markers to increase confidence in samples of known pedigree is recommended in cases with large numbers of Mendelian errors.

The remaining panel of 95 SNPs were randomly subsampled into 1,000 panels of 45 SNPs. The subsampled panels had a mean of 99.0% parents correctly assigned with a maximum of 100% and a minimum of 94.1%. Following the omission of the lowest ranking SNP parentage assignment was 100% correct, the final panel consisted of 94 assays (supplementary table 1).

3.3. Parentage Assignment Success

Analysis of subsets containing 17 or 221 parents with the panel of 94 SNPs markers gave 100% successful parentage assignment for 68 offspring. The effect of number of SNPs on parentage assignment was trialed using a total of 3,400 COLONY runs. Our results show that using less than 25 SNPs reduces the accuracy below 90% while using over 60 SNPs increases the accuracy up to 100% in >85% of simulations (Figure 2).

The success of variable numbers of SNPs agrees broadly with published theoretical (Anderson et al., 2005) and empirical (Liu et al., 2016; Weinman et al., 2014) findings. Comparisons between success rates and sizes of panels are frequently provided, but in most cases only a single (Liu et al., 2016) or two contrasting panels (Kaiser et al., 2016; Weinman et al., 2014) are evaluated for each subset of SNPs. Our results indicate that there is significant variation in suboptimal panels, and that studies presenting SNP panels for parentage should consider randomly subsampling many SNP subsets to report panel success rate as a distribution. Furthermore, before using published SNP panels commercially some further validation with the specific populations in the breeding program is highly desirable due to possible variations in MAF, null and private alleles.

3.4. Population Identification

Traceability is of growing importance in aquaculture. The escape of individuals derived from selective breeding schemes into the wild has become increasing frequent and has a significant, negative effect on wild stocks (Hindar et al., 1991). Identifying the origin stock of potential escapees is important for regulatory agencies, and for producers to evaluate good and bad practice in preventing escape of fish raised in sea pens.

Using a reduced panel of 94 SNPs we observed that much of the population level information was conserved (Figure 3). The reduced panel carried a population assignment success of 100% using the GENECLASS2 algorithm.

This result highlights that neutral markers, optimized for parentage assignment, are able to provide sufficient information for population assignment. Molecular population assignment of escapees in Norwegian Atlantic Salmon is well reviewed in Glover et al. (2010) and a panel of 70 highly discriminatory SNPs and 29 putatively neutral SNPs for population assignment are available (Glover et al., 2013). However, one of the difficulties in assignment of escapees to their brood stock parents is the cost and effort of genotyping parents. The use of our workflow has resulted in a dual purpose, highly cost effective SNP panel, for pedigree reconstruction in selective breeding programs while also offering accurate population assignment for product traceability or the tracking of escapees from net pens.

4. Conclusions

Overall, we present an efficient and robust workflow to develop low density SNP panels for parentage assignment and a multi-purpose validated SNP panel for Atlantic salmon. The utility of high-throughput SNP panels for parentage assignment will undoubtedly become increasingly important as demand on worldwide aquaculture production puts pressure on the development of selective breeding programs for farmed species. The workflow and pipeline described for SNP panel design in Atlantic salmon should have wide applicability to other aquaculture species.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 654008.

References

Anderson, E.C., Garza, J.C., 2005. The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference. Genetics 172, 2567–2582. doi:10.1534/genetics.105.048074

Baird, N. a., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z. a., Selker, E.U., Cresko, W. a., Johnson, E. a., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, 1–7. doi:10.1371/journal.pone.0003376

Borrell, Y.J., Gallego, V., García-Fernández, C., Mazzeo, I., Pérez, L., Asturiano, J.F., Carleos, C.E., Vázquez, E., Sánchez, J.A., Blanco, G., 2011. Assessment of parental contributions to fast- and slow-growing progenies in the sea bream Sparus aurata L. using a new multiplex PCR. Aquaculture 314, 58–65. doi:10.1016/j.aquaculture.2011.01.028

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi:10.1093/bioinformatics/btr330.

FAO, 2014. FAO Yearbook. Fishery and Aquaculture Statistics. FAO yearbook.Fishery and aquaculture statistics.

Gjedrem, T., Gjøen, H.M., Gjerde, B., 1991. Genetic origin of Norwegian farmed Atlantic salmon. Aquaculture 98, 41–50. doi:10.1016/0044-8486(91)90369-I

Glover, K., 2010. Forensic identification of fish farm escapees: the Norwegian experience. Aquac. Environ. Interact. 1, 1–10. doi:10.3354/aei00002

Glover, K., Pertoldi, C., Besnier, F., Wennevik, V., Kent, M., Skaala, Ø., 2013. Atlantic salmon populations invaded by farmed escapees: quantifying genetic

introgression with a Bayesian approach and SNPs. BMC Genet. 14, 74. doi:10.1186/1471-2156-14-74

Gonen, S., Baranski, M., Thorland, I., Norris, A., Grove, H., Arnesen, P., Bakke, H., Lien, S., Bishop, S.C., Houston, R.D., 2015. Mapping and validation of a major QTL affecting resistance to pancreas disease (salmonid alphavirus) in Atlantic salmon (*Salmo salar*). Heredity (Edinb). 115, 405–414. doi:10.1038/hdy.2015.37

Gonen, S., Lowe, N.R., Cezard, T., Gharbi, K., Bishop, S.C., Houston, R.D., 2014. Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics 15, 166. doi:10.1186/1471-2164-15-166

Hindar, K., Ryman, N., Utter, F., 1991. Genetic Effects of Cultured Fish on Natural Fish Populations. Can. J. Fish. Aquat. Sci. 48, 945–957. doi:10.1139/f91-111

Houston, R.D., Taggart, J.B., Cézard, T., Bekaert, M., Lowe, N.R., Downing, A., Talbot, R., Bishop, S.C., Archibald, A.L., Bron, J.E., Penman, D.J., Davassi, A., Brew, F., Tinch, A.E., Gharbi, K., Hamilton, A., 2014. Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). BMC Genomics 15, 90. doi:10.1186/1471-2164-15-90

Jonas, E., Koning, D.-J. de, 2015. Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. Front. Genet. 6, 1–8. doi:10.3389/fgene.2015.00049

Jones, O.R., Wang, J., 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol. Ecol. Resour. 10, 551–555. doi:10.1111/j.1755-0998.2009.02787.x

Kaiser, S.A., Taylor, S.A., Chen, N., Sillett, T.S., Bondra, E.R., Webster, M.S., 2016. A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. Mol. Ecol. Resour. n/a-n/a. doi:10.1111/1755-0998.12589

Kincaid, H.L., 1983. Inbreeding in fish populations used for aquaculture. Aquaculture 33, 215–227. doi:10.1016/0044-8486(83)90402-7

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. doi:10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Lien, S., Gidskehaug, L., Moen, T., Hayes, B.J., Berg, P.R., Davidson, W.S., Omholt, S.W., Kent, M.P., 2011. A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics 12, 615. doi:10.1186/1471-2164-12-615

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Matthew, P., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., Schalburg, K. Von, Rondeau, E.B., Genova, A. Di, Samy, J.K.A., Vik, J.O., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205. doi:10.1038/nature17164

Liu, S., Palti, Y., Gao, G., Rexroad, C.E., 2016. Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture 452, 178–182. doi:10.1016/j.aquaculture.2015.11.001.

Macqueen, D. J., Johnston, I. A., 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc Biol Sci 281, 20132881.

Moen, T., Torgersen, J., Santi, N., Davidson, W.S., Baranski, M., Odegard, J., Kjoglum, S., Velle, B., Kent, M., Lubieniecki, K.P., Isdal, E., Lien, S., 2015. Epithelial Cadherin Determines Resistance to Infectious Pancreatic Necrosis Virus in Atlantic Salmon. Genetics 200, 1313–1326. doi:10.1534/genetics.115.175406

Morvezen, R., Cornette, F., Charrier, G., Guinand, B., Lapègue, S., Boudry, P., Laroche, J., 2013. Multiplex PCR sets of novel microsatellite loci for the great

scallop Pecten maximus and their application in parentage assignment. Aquat. Living Resour. 26, 207–213. doi:10.1051/alr/2013052

Norris, a T., Bradley, D.G., Cunningham, E.P., 2000. Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture 182, 73–83. doi:10.1016/S0044-8486(99)00247-1

O'Reilly, P.T., Herbinger, C., Wright, J.M., 1998. Analysis of Parentage Determination in Atlantic Salmon (*Salmo Salar*). Anim. Genet. 29, 363–370.

Pardo, B. E., Machordom, A., Foresti, F., Porto-Foresti, F., Azevedo, M. F. C., Bañon, R., Sánchez, L., Paulino, M., 2005. Phylogenetic analysis of flatfish (Order Pleuronectiformes) base don mitocondrial 16s rDNA sequences. Sci Mar 69, 531-543.

Piry, S., Alapetite, A., Cornuet, J.M., Paetkau, D., Baudouin, L., Estoup, A., 2004. GENECLASS2: A software for genetic assignment and first-generation migrant detection. J. Hered. 95, 536–539. doi:10.1093/jhered/esh074

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

R Development Core Team, R., 2014. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org

Rannala, B., Mountain, J.L., 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. U. S. A. 94, 9197–9201. doi:10.1073/pnas.94.17.9197

Tange, O., 2011. Gnu parallel-the command-line power tool. USENIX Mag. 36, 42–47.

Vandeputte, M., Haffray, P., 2014. Parentage assignment with genomic markers: a major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. Front. Genet. 5, 1–8. doi:10.3389/fgene.2014.00432

Weinman, L.R., Solomon, J.W., Rubenstein, D.R., 2015. A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. Mol. Ecol. Resour. 15, 502–511. doi:10.1111/1755-0998.12330

Yáñez, J.M., Naswa, S., López, M.E., Bassini, L., Correa, K., Gilbey, J., Bernatchez, L., Norris, A., Neira, R., Lhorente, J.P., Schnable, P.S., Newman, S., Mileham, A., Deeb, N., Di Genova, A., Maass, A., 2016. Genome-wide single nucleotide polymorphism (SNP) discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. Mol. Ecol. Resour. n/a-n/a. doi:10.1111/1755-0998.12503

Yue, G.H., Xia, J.H., 2014. Practical Considerations of Molecular Parentage Analysis in Fish. J. World Aquac. Soc. 45, 89–103. doi:10.1111/jwas.12107

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328. doi:10.1093/bioinformatics/bts606

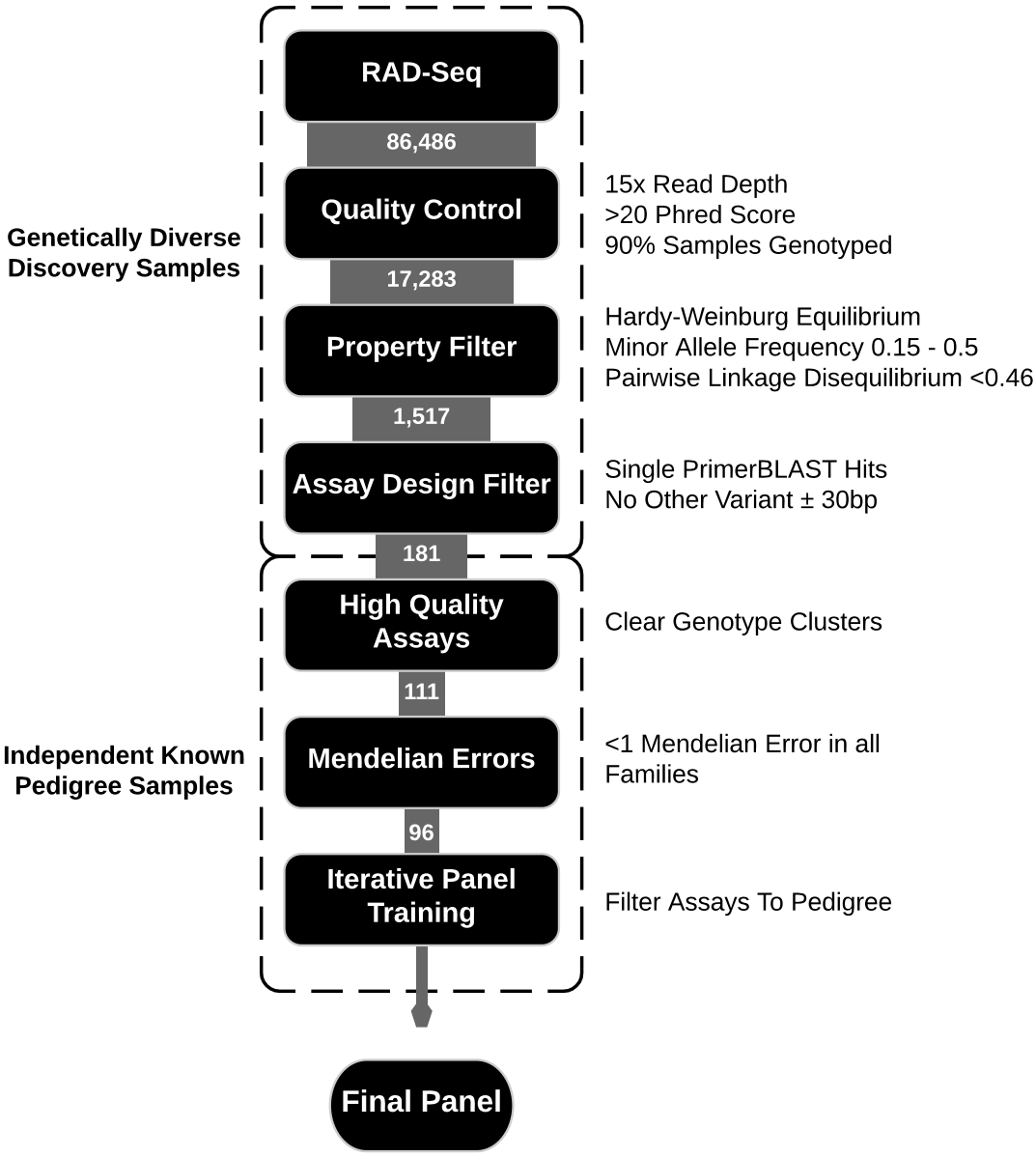

Figure 1- Flowchart detailing workflow for selection of SNP panel for parentage. Each box details a step with the surviving number of SNPs per step detailed in-between boxes. Details of filtering are given adjacent to each box.

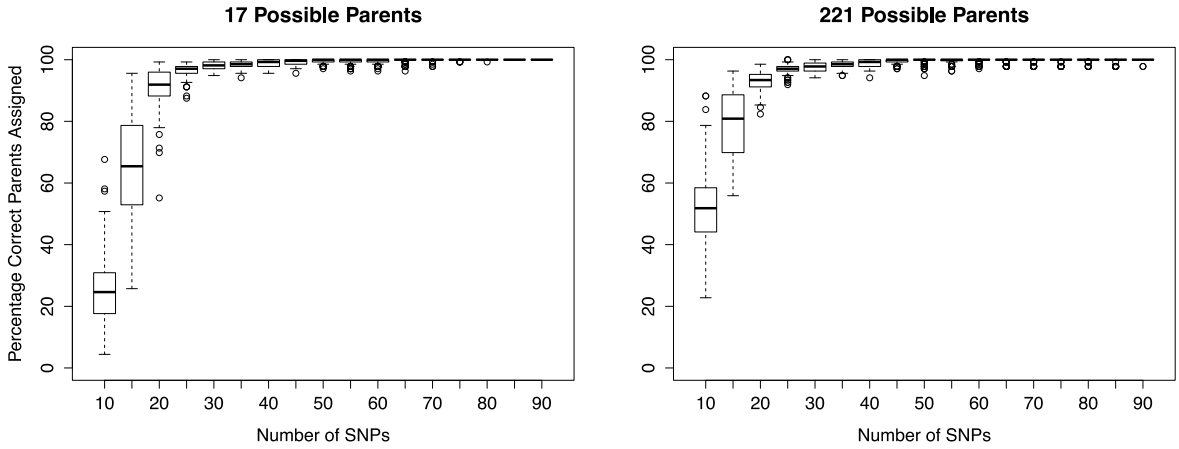

Figure 2 – Boxplot showing effect of number of SNPs on percentage correct parentage assignment for 100 models at each number of SNPs. Results are shown for 17 and 221 possible parents.

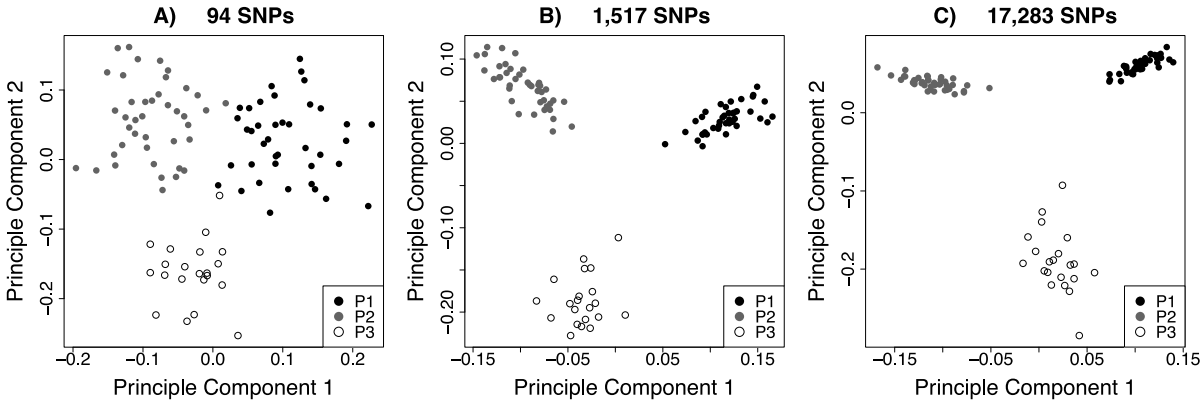

Figure 3 – Plots showing effect of a diminishing number of SNP markers on the first two principal components from a principal component analysis. Populations are as follows – Norwegian origin: P1(Black), P2(Grey). Scottish origin: P3(White).

Highlights

-We present an efficient workflow for designing low density SNP panels for use in aquaculture species.

-The workflow was used to produce a novel panel of 94 validated SNP genotyping assays that provides 100% accurate parentage assignment in Atlantic salmon.

-Additionally, the panel provides 100% population assignment within tested aquaculture populations of Atlantic salmon.