**Electronic Supplementary Material**

**Male-female relatedness at specific SNP-linkage groups influences cryptic female choice in Chinook salmon (*Oncorhynchus tshawytscha*)**

Cornelia Geßner1, Sheri L. Johnson1,2, Paul Fisher3, Shannon Clarke3, Kim Rutherford1, Jane Symonds4,5, Neil J. Gemmell1

**SM1 – Materials and Methods**

**SNP discovery and SNP-linkage mapping**

The SNP detection was performed by Fisher and Clarke (unpublished). A total of 55 randomly selected Chinook salmon belonging to a commercial breeding population in New Zealand were used for the SNP detection. The *Rsa*I restriction enzyme was utilized to produce reduced representational sequencing (RRS) libraries; the genome of one individual was sequenced using the 454 technology (Roche) and the remaining 54 individuals were sequenced using the SOLiD platform (Applied Biosystems). The sequence data obtained from the 454 was assembled into contigs (Newbler, 454 Life Sciences, Roche Diagnostics) and together with singleton reads, mapped [[1](#_ENREF_1)] against the Zebrafish and Medaka genomes, and Atlantic salmon RefSeqs (ie expressed gene based products). Note at the time this work was carried out the Atlantic salmon genome assembly was not available and nor was the rainbow trout genome, a species closer to the Chinook salmon that Atlantic salmon. SNP distribution across the Zebrafish and Medaka genomes showed that the SNP markers were evenly spread across all chromosomes. The RRS SOLiD data generated from the 54 individuals were mapped to 454 contigs [[2](#_ENREF_2)] to establish a SNP data set. Approximately 800K putative SNPs were discovered that were heavily filtered to generate a set of SNP markers suitable for an Illumina Infinium SNP array. Briefly, variants that were supported by at least 2 SOLiD reads not present at the end of reads and had ~100 bases flanking sequence without repeats and allowance of 1 or more flanking SNPs within 10 bases on one side were considered for further for section (~95K SNPs). Further filtering criteria (no neighboring SNP (within 10 bp), no repeat elements within 80 bp, only 2 alleles and read depth ≥5 and ≤10) reduced SNP set to ~11K. The ~11K SNPs with ~150 bp flanking sequence were sent to the Illumina technical team to asses with the Illumina Assay Design Software to QC the SNPs and further assist in the selection process. SNPs were given a score from 0-1; 63% of the SNPs had a score of ≥0.95. The SNPs with a score ≥ 0.7 (85%) were further assessed by alignment to the Chinook singleton and contig sequences, the Medaka, Atlantic salmon and Zebrafish genomes and filtered for distance of neighboring SNPs. The final SNP selection was based on two thirds unique to the Chinook salmon sequence and a third that mapped to the other species. For ~5000 SNPs unique to Chinook salmon sequence, a QC score of ≥0.95 and minimum distance of 50 to neighboring SNP was required. Less stringent criteria was used to capture ~2000 SNPs from the other species (QC score of ≥0.70 and minimum distance of 20 to neighboring SNP). The locus, position, flanking sequence and alleles for the above selected SNPs was checked back with the original SNP file and the flanking sequences (left and right per SNP) were then mapped back onto the contigs and singletons to give an indication if the SNPs are in unique positions or in repetitive regions. A uniqueness criteria was used (i.e. evalue; this was based on BLAST results where each hit had to have an evalue of 1e-5 or less. If there were multiple hits for a query an "adjusted evalue" of the top hit was calculated. This is the evalue of the top hit, divided by the evalue of the second to top hit. This adjusted evalue had to be 1e-20 or less). Without this majority maps back. It is highly possible that the SNPs that do not have unique flanking regions are SNPs that are in regions of the duplicated genome (and not just repetitive sequence regions). The final SNPs selected to build the Chinook salmon infinium SNP array were:

1. SNPs mapped to Medaka, Zebrafish and Atlantic salmon
* 1907 SNPs (nb this included 32 of the ‘none\_mapped’ SNPs; 1422 from ‘both\_mapped both\_unique’; 230 ‘both\_mapped one\_unique’; 65 from ‘one\_mapped one\_unique’ and 158 from ‘none\_unique’ as indicated as in Table X).
1. SNPs unique to Chinook
	* + 3441 with both flanking regions mapped and both unique (ie both\_mapped both\_unique from Table X)
		+ 652 with both flanking regions mapped but only one unique (ie both\_mapped one\_unique from Table X). The QC score was used to select the 652 from the 919 SNPs

Table 1: Mapping back of the left and right flanking sequences of the selected SNPs (Total 7290) .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   |   | both\_unique | one\_unique | none\_unique |
| none\_mappeda | 178 |   |   |   |
| both\_mapped | 6498 | 4862 | 1149 | 487 |
| one\_mapped | 614 |   | 261 | 353 |
| a a uniqueness criteria was used (i.e. e-score ). None mapped indicates that neither the left or right flanking sequence passed the uniqueness criteria. Without this everything maps back.  |

Additionally, 37 Gbp of Illumina HiSeq 1500 DNA sequence data from the 55 individuals was analysed to support SNP selection.

To establish a linkage map, four full-sib families (607 fish) were genotyped with the Illumina Chinook salmon 6K SNP array. Following manufacture, 88% passed manufacture genotyping QC. Of the ~5.3K SNPs, ~1.3K SNPs were scored as diploid or tetraploid, however, the remaining SNPs were either all-heterozygous or “messy” to score. The all-heterozygous SNPs are indicative of fixed homozygous variation within duplicated regions of the genome ie AA in one, compared to BB in duplication. Unlike the true tetraploid SNPs exhibiting variation within homologs. To establish a linkage map only diploid SNPs were considered and filtered on HWE, Mendelian inheritance, MAF and PIC scores across-family. A total of 698 diploid SNPs that exhibited correct parentage assignment were analyzed for linkage mapping utilizing CRIMAP (chrompic and build options as well as flips). Genetic maps, with 598 diploid SNPs were built where male- and female-specific maps were produced for 30 main linkage groups (and 7 very small male specific groups), which were assumed to correspond to ~30 Chinook salmon chromosomes. The map coverage was good overall; the combined female linkage groups were slightly longer than the published Atlantic salmon map, and the male maps were shorter.

**SNP selection for this study**

Salmonid fish are tetraploid but undergoing rediploidization [[3](#_ENREF_3)]. 698 SNPs that were identified as unambiguously diploid with a high call rate (Fisher and Clarke, unpublished) were selected for further analysis. Of these 698 SNPs, 588 SNPs could be assigned to 30 LGs (Fisher and Clarke, unpublished). We then excluded another 16 SNPs that were either monomorphic or had a call rate below 90% in our population. Genome-wide relatedness estimates between mates are based on the remaining 682 SNPs. Five of the 16 excluded SNPs were part of the 588 SNPs used for linkage mapping, leaving 583 SNPs across all LGs. LG-specific relatedness between mates was calculated for a total of 29 LGs with at least 8 SNPs (Table S1). On average, LGs contained of 19 SNPs. All individuals had a call rate above 95% and thus, no individual was excluded from further analysis. To verify the reliability of our genotyping results, a subset of 29 individuals was genotyped twice and genotypes of all 29 individuals were identical between genotyping runs.

Table S1: A total of 588 SNPs could be assigned to 37 SNP-linkage groups in Chinook salmon (*Oncorhynchus tshawytscha*) (Fisher and Clarke, unpublished). This table shows the 583 SNPs that passed the quality filtering, i.e. were not monomorphic in our population and had call rates above 90%, and thus, were used in further analyses. Relatedness of mates was computed for linkage groups (n = 29) holding a minimum of 8 SNPs (average 19 SNPs). Linkage groups not used for further analysis are coloured in grey.



**Statistical analysis**

All MCMCglmm models were run with the settings of 1,300,000 iterations, a thinning interval of 1,000 and 300,000 burn-in. The model fit was assessed using trace and posterior density plots. We ensured the non-independence between samples in the chain (autocorrelation) was below 0.1, indicating a good mixing. There were neither within-seasonal effects on sperm velocity (pMCMC = 0.17 – 0.47) and fertilization success (pMCMC = 0.5 – 0.9 across models), nor was there a seasonal effect (pMCMC = 0.34 – 0.98) on fertilization success when the date of experiment and the year were modeled as fixed effects. To avoid overparameterization, we did not include these parameters into our final model. For all models, the R2 value was calculated according to Nakagawa and Schielzeth (2013), note that when random-slope models were used, we report R2 values from corresponding random-intercept models. The lower and upper bound of the 95% credible interval (CI) are reported as an indication of the uncertainty around our estimates of regression coefficients and variance components. For all models, we used both inverse gamma priors and parameter-expanded priors for random effects [[5](#_ENREF_5)]. We report results from the parameter-expanded prior because the results were qualitatively identical and quantitatively similar between corresponding models using the two priors and also because the gamma inverse prior can incorrectly influence random effect estimates, especially when sample sizes are small to moderate.

All fertilization trials were performed twice and the two replicates (first trial denoted ‘replicate 1’, second trial ‘replicate 2’) were analyzed separately. We additionally randomly assigned the data to replicate 1 and 2 using the dplyr package [6] in R to show that our results were independent of the trial assignment to the replicate ID (pMCMCLG12 = 0.054 and 0.028; pMCMCLG23 = 0.034 and 0.056 for replicates 1 and 2, respectively).

References

[1] Li, H. & Durbin, R. 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760.

[2] Chen, Y., Souaiaia, T. & Chen, T. 2009 PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics* **25**, 2514-2521.

[3] Allendorf, F.W. & Thorgaard, G.H. 1984 Tetraploidy and the evolution of salmonid fishes. In *Evolutionary genetics of fishes* (pp. 1-53, Springer.

[4] Nakagawa, S. & Schielzeth, H. 2013 A general and simple method for obtaining R2 from generalized linear mixed‐effects models. *Methods in Ecology and Evolution* **4**, 133-142.

[5] Hadfield, J. 2015 MCMCglmm course notes.

[6] Wickham, H. & Francois, R. 2015 dplyr: A grammar of data manipulation. *R package version 0.4* **1**, 2.