

May 23, 2012

Methods used to generate NFH Genetic Profiles

The following document describes methods used to analyze genetic data for the National Fish Hatchery broodstock profiles. Questions regarding these methods may be addressed to Christian Smith (Christian_smith@fws.gov).

Genotype Quality Assessment

Genotypes generated at Abernathy Fish Technology Center were scored by two independent readers (double-scoring). Further, 10% of all individuals were extracted and genotyped a second time as part of AFTC's Quality Assessment / Quality Control Standard Operating Procedure.

In order to avoid using data from potentially degraded or contaminated samples, any individuals with multi-locus genotypes less than 75% complete (i.e. for which less than 75% of loci under consideration had allelic states assigned) were deleted. Finally, data sets were screened for individuals with identical genotypes (which typically represent either contamination or individual fish sampled multiple times).

Statistical Analyses

Testing for genotypic ratios that departed from Hardy-Weinberg Equilibrium (HWE) was conducted using Fisher's exact tests in GENEPOP version 4.0 (Rousset 2008). The log likelihood ratio statistic (G test) was used to test for genotypic disequilibrium (referred to as Linkage Disequilibrium (LD) in the reports) between each pair of loci in each collection, and to

23 test for allele frequency differences between collections. Settings for Markov chains were:
24 dememorization number = 10,000, number of batches = 200, and iterations per batch = 5000. For
25 all tests $\alpha=0.05$. In order to make results for HWE and LD comparable among profiles
26 employing different sets of collections, **no corrections for numbers of simultaneous tests were**
27 **made.**

28
29 Observed and expected heterozygosity, allelic richness (AR; number of alleles observed per
30 individual, corrected for unequal sample sizes) and F_{IS} (Weir & Cockerham 1984) were
31 calculated for each collection using FSTAT (Goudet 2000). Allelic richness scores for any
32 collection may vary greatly with the total number of alleles at each locus. In order to summarize
33 AR scores per collection we thus ranked all n populations by the sums of AR ranks over all loci.
34 The collection with the largest number of high AR scores (relative to scores in the other
35 collections) will thus be ranked “1”, and the collection with the largest number of low AR scores
36 will be ranked n . Effective population size for each collection based on LD was estimated
37 following Waples (2006) with confidence intervals determined based on jackknifing across loci.
38 Microsatellite alleles with frequencies of 5% or greater were included for estimates based on
39 microsatellites and SNP alleles with frequencies of 1% or greater were included for estimates
40 based on SNPs. These calculations were performed using LDNe (Waples & Do 2008).

41
42 Two-dimensional representations of genetic data were generated in order to allow visual
43 representation of the data. This was done using one of two methods, as identified in each report.
44 Either 1) genetic chord distances (Cavalli-Sforza & Edwards 1967) were calculated using
45 PHYLIP (Felsenstein 1993), and a principal component analysis was performed on the distances

46 using GENELEX (Peakall & Smouse 2006); or 2) correspondence analysis was performed on
47 allele frequencies for all loci using the program GENETIX (Belkhir *et al.* 2004). GENETIX was
48 also used for calculation of F_{ST} (θ ; Weir & Cockerham 1984). Significance of pairwise F_{ST}
49 values was evaluated by comparison to a null distribution based on 10^4 replicate datasets in
50 which individuals were permuted among collections (α was set to 0.05).

51
52 LITERATURE CITED
53

- 54 Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) GENETIX 4.05, logiciel sous
55 Windows TM pour la génétique des populations. *Laboratoire Génome, Populations,*
56 *Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France).*
- 57 Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation
58 procedures. *American Journal of Human Genetics* 19, 233-257.
- 59 Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) Version 3.5c. Department of
60 Genetics, University of Washington, Seattle. Distributed by the author.
- 61 Goudet J (2000) FSTAT, a program to estimate and test gene diversities and fixation indices
62 (version 2.9.3). Available from <http://www.unil.ch/izea/software/fstat.html>. Updated
63 from Goudet (1995).
- 64 Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic
65 software for teaching and research. *Molecular Ecology Notes* 6, 288-295.
- 66 Rousset F (2008) GENEPOP '007: a complete re-implementation of the GENEPOP software for
67 Windows and Linux. *Molecular Ecology Resources* 8, 103-106.
- 68 Waples R (2006) A bias correction for estimates of effective population size based on linkage
69 disequilibrium at unlinked gene loci. *Conservation Genetics* 7, 167-184.
- 70 Waples RS, Do C (2008) LDNE: a program for estimating effective population size from data on
71 linkage disequilibrium. *Molecular Ecology Resources* 8, 753-756.
- 72 Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure.
73 *Evolution* 38, 1358-1370.
- 74