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

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

Two-dimensional representations of genetic data were generated in order to allow visual representation of the data. This was done using one of two methods, as identified in each report. Either 1) genetic chord distances (Cavalli-Sforza & Edwards 1967) were calculated using PHYLIP (Felsenstein 1993), and a principal component analysis was performed on the distances
using GENELEX (Peakall & Smouse 2006); or 2) correspondence analysis was performed on allele frequencies for all loci using the program GENETIX (Belkhir et al. 2004). GENETIX was also used for calculation of $F_{ST}$ (0; Weir & Cockerham 1984). Significance of pairwise $F_{ST}$ values was evaluated by comparison to a null distribution based on $10^4$ replicate datasets in which individuals were permuted among collections ($\alpha$ was set to 0.05).

LITERATURE CITED


Rousset F (2008) GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources 8, 103-106.

