Final Report to Pacific Northwest Cooperative Ecosystem Studies Unit
National Park Service

Debbie Steel and C. Scott Baker
Oregon State University
23 December 2011

TASK AGREEMENT NO.:	COOPERATIVE AGREEMENT NO.:	EFFECTIVE DATES:
J8W070900015	H8W07060001	05/01/2010 – 10/01/2011

COOPERATOR: Oregon State University

PRINCIPAL INVESTIGATOR: C. Scott Baker

PROJECT TITLE: Linking genetic and long term sighting histories of individual humpback whales in a collaborative database

PROJECT SUMMARY

This project represents a collaboration between Glacier Bay National Park (GBNP), the University of Alaska, Southeast (UAS), and Oregon State University (OSU) Cetacean Conservation Genetics Laboratory of the Marine Mammal Institute. The objective was to complete primary genetic analysis of the remaining biopsy samples collected in southeastern Alaska and northern British Colombia (SEA/NBC) during the SPLASH program of 2004-2006. SPLASH (Structure of Populations, Levels of Abundance and Status of Humpbacks) is a multi-disciplinary, international collaborative research endeavor aimed at understanding the population structure and abundance of endangered humpback whales in the North Pacific Ocean basin. The final product of this task agreement is an integrated database of photo-identification and genotype records of humpback whales collected during SPLASH.

The completion and addition of these data is essential to maintaining and continuing the legacy of sighting histories individual humpback whales, which began in the waters of Glacier Bay in the early 1970s. Glacier Bay National Park has supported and conducted humpback whale population and behavior research every summer since at least 1981, largely due to concerns about interactions between these whales and vessels that carry visitors into the Park. These data greatly extend our knowledge of the abundance, life history and migratory connections of humpback whales from the Distinct Population Segment thought to inhabit SEA/NBC. In the face of the climate-related ecological changes that marine species in the North Pacific will experience in the coming decades these data will be invaluable to understand how these changes will impact this important Park resource.

The genetic results will also be provided for inclusion in the greater SPLASH database, made available to SPLASH collaborators and the general public in 2011. Understanding the genetic

structure of endangered species populations that inhabit national parks serves the public who has entrusted these exemplary natural areas to the care of the National Park Service.

Methods

The remaining n = 313 tissue samples (Group 2, Table 1) collected during SPLASH from southeastern Alaska (SEA) and northern British Colombia (NBC) were extracted by Southwest Fisheries Science Center (SWFSC) and the DNA sent to Oregon State University. Ten microsatellite loci were amplified for each sample using previously published primers (GT211, GT23, GT575 (Bérubé, et al. 2000), GATA28, GATA417 (Palsbøll, et al. 1997), EV14, EV37, EV96 (Valsecchi and Amos 1996) and rw48, rw4-10 (Waldick, et al. 1999)). Microsatellite loci were amplified individually in 384-well format and co-loaded in two sets (Set 1: EV14, EV37, EV96, GATA417, rw48, Set 2: GATA28, GT211, GT23, GT575, rw4-10) for automated sizing on an ABI 9730xl (Applied Biosystems). Molecular identification of sex and sequencing of the mitochondrial (mt) DNA control region (460 bp) followed methods described in detail by Olavarria *et al.* (Olavarría, et al. 2007). In addition, for confirmation of gel visualization results, one primer of each of the multiplexed sex products was fluorescently labeled and the products coloaded with microsatellite set 2 for sizing on an ABI 3730xl. Genemapper (Applied Biosystems) was used to automatically size and bin alleles, all calls were also visually assessed. Data organization and initial analyses of microsatellite alleles, sex and mtDNA haplotypes were conducted with the program GenAlEx (Peakall and Smouse 2006).

The n = 313 genotypes were then combined with n = 338 genotypes (Group1, Table 1) previously analyzed SEA and NBC. Variation in the number of microsatellite loci amplified successfully suggested relatively poor quality DNA for a few samples. Following a quality control (QC) review, samples with less than 8 microsatellite loci were identified as poor quality (PQ) samples. These samples were kept in the dataset but were flagged for greater scrutiny in replicate analysis. Replicates were identified with Cervus (Kalinowski, et al. 2007) using 'relaxed' matching criteria to avoid false exclusion of true matches due to genotyping errors (Waits and Leberg 2000, Waits, et al. 2001, Hoffman and Amos 2005, Morin, et al. 2010). The relaxed conditions required a minimum of 5 matching loci between any pair of samples but allowed for up to 3 initial mismatches. These 'likely matches' were then reviewed for errors at any mismatching loci to confirm identity and, where possible, errors were corrected (e.g., electropherograms were rechecked by eye for allelic dropout, sizing errors or data entry error). Where available, variation in mtDNA control region sequences (i.e. haplotypes) and sex were used to confirm matches. The average probability of identity (PI: the probability that two genotypes could match by chance) for a minimum criterion of 8 matching loci ranged from 7.7×10^{-8} to 8.24×10^{-10} as calculated following (Paetkau, et al. 1995). Given these low values, we assumed that genotypes matching at 8 or more loci represented replicate samples (true recaptures) of the same individual whales and any mismatching loci were likely to represent genotype error (Hoffman and Amos 2005). In these cases errors that could not be corrected following a recheck were changed to 0. All other likely matches samples were rerun.

To track matches, groups of replicates were allocated 'match numbers'. If a match had previously been identified during SPLASH the match number given in that dataset was carried over, if it was a new match, numbers were allocated following on from SPLASH, No. 210 onwards. In addition, once a match had been confirmed groups of replicates were allocated a Genetic ID. This was usually the first sample number in the series of replicates and again if a Genetic ID had already been allocated during SPLASH it was carried over.

As an external check, replicate groups identified by genotype matching were compared to 'resights' identified by SPLASH photo analysis using the SPLASH ID field from a file received from Erin Falcone at Cascadia Research, "SPLASH all samples by LABID Sept08-CRC".

Results

Of the n = 650 genotypes, n = 608 passed QC in the initial run and a further 3 passed QC after reruns; n = 39 (6.0%) were flagged as PQ Samples (Figure 1) (n = 4 from NBC (2.5%) and n = 35 from SEA (7.2%); n = 16 from Group 1 (4.7%) and n = 23 from Group 2 (7.3%)). These are noted by PQS in the match number and Genetic ID columns of the datasheet.

Within the n = 611 genotypes that passed QC, 60 groups of replicates (i.e. at least 2 matching samples) were identified that matched at a minimum of 8 loci during the initial analyses. A further 8 groups of replicates were confirmed following reruns of samples, for a total of 68 replicate groups. These were attributed match numbers and Genetic ID's.

<u>Comparison to SPLASH photo-id matches:</u>

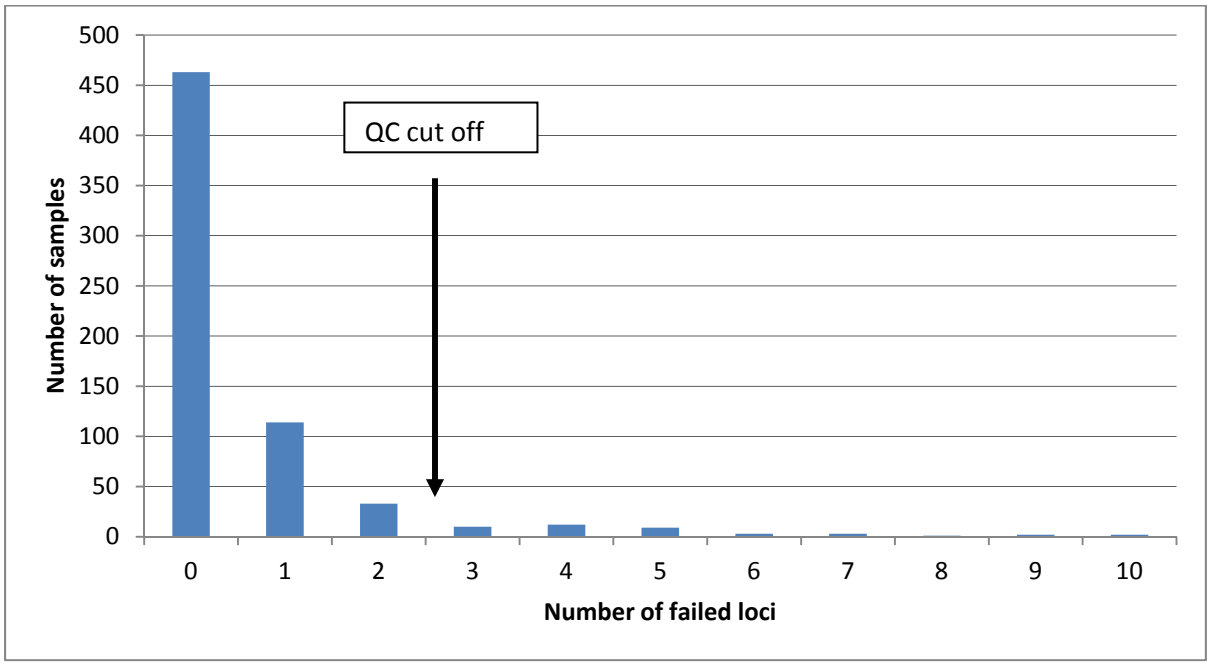
Of the 68 replicate groups identified through genotype matching, 47 also had photo-ID information for at least 2 samples within the group based on the Sept08 SPLASH database. For 38 of the 47 replicate groups, genotype and photo-ID matching agreed, 7 were duplicate records where photo-ID information disagreed and 2 were triplicate records within which 2 samples had photo-ID information that agreed and 1 sample where photo-ID information disagreed.

A further 9 pairs of replicates were identified by photo-id matching but were not supported by the associated genotype matching. Five of these involved PQ genotypes, which were not considered sufficient for individual identification. The remaining four were good quality genotypeswhich mismatched at a minimum of 5 loci and cannot be considered to be genotype replicates (i.e., the photo-ID records report a match but the associated samples do not match). One of these pairs mismatched at all 9 of the overlapping loci and at dlp, another pair mismatched at 6 of the 8 overlapping loci. In the remaining two cases however the genotypes share at least one allele at each locus and as such could be first order kin (i.e. parent-offspring). Following communication with Jan Straley (27 Dec 2011) and review of both photo-ID and genotypes some of the disagreements were resolved. These will be corrected in the final integration of the databases. Others were considered likely to be uncorrectable allocation errors in which samples were misattributed in the field or mislabeled at some point in handling.

Table 1: Summary of SEA and NBC samples by year for each group analyzed. Group 1, analyzed as part of SPLASH; Group 2, analyzed for NPS under this Task Agreement.

Region - Year	Group 1	Group 2	Total
NBC-2004	68	38	106
NBC-2005	55	0	55
NBC total	**123**	**38**	**161**
SEA-2004	214	130	344
SEA-2005	0	145	145
SEA total	**214**	**275**	**489**
Total	**337**	**313**	**650**

Figure 1: Number of samples that failed for a given number of loci. Samples that failed for more than 2 loci were considered to be poor quality.



Key to Excel database

Genetic information is listed in the attached spreadsheet "SEANBC SPLASH genotypes for NPS". The following are the column headers and definitions;

- Genetic ID: Individual identification number given to each unique genotype. Replicate samples of the same individual will have the same Genetic ID. PQS = Poor Quality Sample, samples with less than 8 microsatellite loci.
- Match number: Number given to each group of replicate samples
- Sample Name: LabID code for each sample.
- Group: Identifying which analysis each sample belonged to. Group 1 samples were processed as part of SPLASH, Group 2 samples were processed as part of NPS contract.
- Pop: Region sample was collected from, NBC or SEA
- Year: Year of sample collection
- GATA417: two columns, one for each allele of microsatellite locus GATA417
- GATA417 QC: QC call for each pair of GATA417 alleles. Based on GQ score given within Genemapper combined with interpretation of researcher. Excellent = Locus had a GQ score higher than 0.75 and was deemed good by eye; Good = Locus had a GQ score lower than 0.75 but was deemed good by eye; Fair = Locus was deemed questionable by eye and in most cases had a GQ score lower than 0.75.
- Ev37: two columns, one for each allele of microsatellite locus Ev37
- Ev37 QC: See GATA417 QC for definition.
- Ev96: two columns, one for each allele of microsatellite locus Ev96
- Ev96 QC: See GATA417 QC for definition.
- rw4-10: two columns, one for each allele of microsatellite locus rw4-10
- rw4-10 QC: See GATA417 QC for definition.
- GT211: two columns, one for each allele of microsatellite locus GT211
- GT211 QC: See GATA417 QC for definition.
- Ev14: two columns, one for each allele of microsatellite locus Ev14
- Ev14 QC: See GATA417 QC for definition.
- rw48: two columns, one for each allele of microsatellite locus rw48
- rw48 QC: See GATA417 QC for definition.
- GATA28: two columns, one for each allele of microsatellite locus GATA28
- GATA28 QC: See GATA417 QC for definition.
- GT23: two columns, one for each allele of microsatellite locus GT23
- GT23 QC: See GATA417 QC for definition.
- GT575: two columns, one for each allele of microsatellite locus GT575
- GT575 QC: See GATA417 QC for definition.
- Sex: Sex of the sample as determined by genetic analysis
- Dlphap: mtDNA haplotype defined from 450bp of the mtDNA control region sequence
- Field ID: sample ID given in the field – obtained from SWFSC
- SPLASH ID: Identification number for an individual whale based on fluke identification photograph, given by Cascadia Research as part of SPLASH analysis.

References;

Bérubé, M., H. Jørgensen, R. McEwing and P. J. Palsbøll (2000). Polymorphic di-nucleotide microsatellite loci isolated from the humpback whale, *Megaptera novaeangliae*. *Molecular Ecology*, 9; 2181-2183.

Hoffman, J. I. and W. Amos (2005). Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, 14; 599–612.

Kalinowski, S. T., M. L. Taper and T. C. Marshall (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, 16; 1099-1106.

Olavarría, C., C. S. Baker, C. Garrigue, M. Poole, N. Hauser, S. Caballero, L. Flórez-González, M. Brasseur, J. Bannister, J. Capella, P. J. Clapham, R. Dodemont, M. Donoghue, C. Jenner, M. N. Jenner, D. Moro, M. Oremus, D. A. Paton and K. Russell (2007). Population structure of humpback whales throughout the South Pacific and the origins of the eastern Polynesian breeding grounds. *Marine Ecology - Progress Series*, 330; 257-268.

Paetkau, D., W. Calvert, I. Stirling and C. Strobeck (1995). Microsatellie analysis of population structure in Canadian polar bears. *Molecular Ecology*, 4; 347-354.

Palsbøll, P. J., M. Bérubé, A. H. Larsen and H. Jørgensen (1997). Primers for the amplification of tri- and tetramer microsatellite loci in baleen whales. *Molecular Ecology*, 6; 893-895.

Peakall, R. and P. E. Smouse (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, 6; 288-229.

Valsecchi, E. and W. Amos (1996). Microsatellite markers for the study of cetacean populations. *Molecular Ecology*, 5; 151-156.

Waldick, R. C., M. W. Brown and B. N. White (1999). Characterization and isolation of microsatellite loci from the endangered North Atlantic right whale. *Molecular Ecology*, 8; 1763-1765.