

Genetic data analysis guidelines reworked

ROBIN S. WAPLES¹, A. RUS HOELZEL², OSCAR GAGGIOTTI³, RALPH TIEDEMANN⁴, PER PALSBÖLL⁵, FRANK CIPRIANO⁶, JENNIFER JACKSON⁷, and JOHN W. BICKHAM⁸.

¹*NOAA Fisheries, Northwest Fisheries Science Center, 2725 Montlake Blvd. East, Seattle WA 98112, USA, robin.waples@noaa.gov.* ²*School of Biological and Biomedical Sciences, Durham University, Durham DH1 3LE, UK, a.r.hoelzel@dur.ac.uk.* ³*Laboratoire d'Ecologie Alpine (LECA), Université Joseph Fourier, Grenoble, France, oscar.gaggiotti@ujf-grenoble.fr.* ⁴*Institute of Biochemistry and Biology, University of Potsdam, D-14476 Potsdam, Germany, tiedeman@uni-potsdam.de.* ⁵*Centre for Ecological and Evolutionary Studies, University of Groningen, The Netherlands, p.j.palsboll@rug.nl.* ⁶*Department of Biology, San Francisco State University, San Francisco CA 94132, USA, Frank.cipriano@sfsu.edu.* ⁷*British Antarctic Survey, Cambridge CB3 0ET, UK, Jen.jeck@bas.ac.uk.* ⁸*Department: Forestry and Natural Resources, Purdue University, 715 West State Street, West Lafayette, IN 47907, USA, bickham@purdue.edu.*

ABSTRACT

Recently, an IWC workgroup developed guidelines for quality control of DNA data. Once data have been collected, the next step is to analyze the data and produce results that are useful for addressing practical problems in the management of cetaceans. This is a complex exercise, as numerous analyses are possible and users have a wide range of choices of software programs for implementing the analyses. Here, we provide an outline for a document that would provide guidelines for analysis and interpretation of genetic data in a management context. We encourage comments and suggestions from managers, cetacean biologists, and geneticists to help make the final document as useful as possible.

INTRODUCTION

Recently, guidelines were adopted for quality control of DNA data intended for use within the International Whaling Commission (IWC 2009). Once the data have been collected, the next step is to analyze the data and produce results that are useful for addressing practical problems in the management of cetaceans. This is a complex exercise for two major reasons: 1) a large number of methods can be used to analyze genetic data, and an equally wide range of software programs are available for conducting these analyses; and 2) a key objective is to inform those involved in cetacean management who don't have a background in population genetics. For these reasons, it has been suggested that it would be useful to have a document that provides guidelines for the analysis of population genetic data for use in a management context. Although it is not possible (nor is it desirable) to prescribe specific procedures for all analyses of population genetic data, it can be useful to provide general guidelines for some of the more common types of analyses conducted in a management context. The latter is the objective of this paper. The emphasis will be on a general discussion of issues involved in genetic data analysis rather than detailed comments about specific software programs, but some popular programs will be discussed to make particular points. Because a large number of types of analyses (and software packages to conduct such analyses) are available, to focus on analyses most relevant to a particular study we organize the discussion around some common management problems one might try to address with genetic data. These problems are identified below with roman numerals. We assume that before the analyses considered here begin, the DNA quality control guidelines (IWC 2009) have been consulted and followed to the extent possible, and that any substantial deviations have been documented and explained.

This document is still under construction. During an April 2012 workshop in Cambridge, UK, we completed drafts of almost all of the sections listed in the outline, but editing these is still underway. After the outline of in-depth material, we list some thorny IWC management problems for which genetic data might provide useful insights. This section is also still under construction, but we provide one example to indicate how the material in this document might be used.

OUTLINE of IN-DEPTH MATERIAL

Here we provide an outline of the material that is covered in more depth in an Appendix (which is in preparation).

I. Species identification/delimitation

Issues related to alpha taxonomy come up consistently, esp regarding the boundary between populations and species of small cetaceans (so we expect some overlap with Section IV).

Because a standardized methodology for DNA-based species identification of cetaceans already exists (Baker et al. 2003; Ross et al. 2003), this document will focus on analyses of intraspecific genetic diversity. Information about *DNA Surveillance* and the comprehensive reference database, *Witness for the Whales*, can be found at the following url: <http://www.cebl.auckland.ac.nz:9000/>. Ross and Murugan (2006) present results of a comparison of cetacean DNA sequences in *Witness for the Whales* and *GeneBank*.

II. Analysis of diversity within populations

- A. Measures of genetic diversity, including rarefaction (controlling for sample size in estimating allelic richness)
- B. Information related to tests of Hardy-Weinberg (HW) equilibrium.
- C. Information related to tests of linkage disequilibrium (LD).

III. Estimating population size

- A. Census size, N
 1. DNA mark-recapture
 2. Analysis of close relatives
 3. Identifying recent population bottlenecks
- B. Effective population size, N_e
 1. Historical N_e
 2. Contemporary N_e
- C. N_e/N ratios

IV. Analysis of diversity among populations (aka stock structure)

This is probably the most common type of management problem that utilizes genetic data.

A. Testing for heterogeneity

1. Putative populations defined *a priori*
2. No *a priori* basis (or a questionable basis) for grouping individuals into populations.
In this case, the analyses are conducted on individuals rather than groups of individuals.
 - a. Standard clustering programs
 - b. Clustering based on ordination
 - c. Landscape genetics (units = individuals), including temporal structure
 - d. Kinship

B. Describing Population structure

1. Estimating degree of divergence
 - a. F_{ST} , genetic distance, and related measures
 - b. Ordination
 - c. Isolation by distance/Landscape genetics (units = samples)
2. Estimating migration
 - a. Methods that assume migration-drift equilibrium and estimate long-term patterns of gene flow (mN_e)
 - b. Isolation with migration models to estimate splitting times and post-division migration rates
 - c. Assignment methods that estimate contemporary migration rate (m)
 - d. Kinship/individual
3. Mixture analysis (e.g., resolving stock composition of samples from feeding grounds or migration pathways)

V. Generic/cross-cutting issues

Some issues will apply to many of the above analyses. Examples include:

- A. Choice of markers (mtDNA, microsats, SNPs)
- B. Ascertainment bias
- C. Multiple testing
- D. Mutation rates
- E. Sampling and experimental design, including confidence intervals, underlying assumptions and sensitivity to their violation
- F. Bayesian vs maximum likelihood vs frequentist methods
- G. MCMC issues (burnin, convergence)
- H. Integrating genetic and non-genetic data
- I. Possible influence of selection
- J. Interpreting negative results
- K. Validation/transparency

MANAGEMENT NIGHTMARES

Plan for this section: for each scenario, provide a short summary of underlying biology and the key management questions, and then discuss the major issues related to genetic data analysis, drawing on material discussed in the in-depth sections. Within the sections themselves, we could also make specific comments about one or more of these management scenarios.

North Pacific common minke whales

Management context: The breeding areas for common minke whales in the western North Pacific have not been identified but are presumed to exist to the south of Japan. Animals are observed and sometimes taken as they migrate along both coasts of Japan, as well as in oceanic waters farther to the east. The Okhotsk Sea is a primary feeding ground. Convincing evidence (including differences in morphological and genetic traits and conception date) exists for the occurrence of at least two stocks, which have been termed O (more oceanic) and J (more coastal); however, rigorous characterization of stocks is difficult because of the lack of samples from breeding areas. Existing analyses are anchored by samples from areas thought to contain mostly pure J individuals (SA6) and mostly pure O individuals (SA9 and perhaps parts of SA8 and SA7). Some analyses also suggest additional heterogeneity might exist within either the J-like animals, the O-like animals, or both. In one view, this heterogeneity simply represents different mixture fractions of the same O and J stocks; in another view, this heterogeneity indicates the presence of 2-3 additional stocks.

Application of genetic data analysis guidelines: This section needs to be developed, but could include discussion of the following:

- relatively large sample sizes are available, which creates the possibility of high power to detect heterogeneity that might not be biologically meaningful
- lack of samples from breeding grounds raise questions about applicability of many standard methods that require *a priori* grouping of individuals into putative populations or stocks
- some issues of data standardization involving two laboratories (Japan and Korea)
- how to interpret results of HW and LD tests
- usefulness of and caveats associated with clustering methods like STRUCTURE
- application of PCA
- methods that can be used to try to 'cleanse' datasets to construct samples of 'pure' stocks, and caveats associated with this approach
- interpretation of temporal patterns
- combination of genetic and non-genetic data

The following nightmarish scenarios are awaiting input to make them come to life:

B-C-B bowhead whales

N. Atlantic common minke whales

Southern hemisphere humpback whales

N. Pacific gray whales

SUMMARY AND CONCLUSIONS

Under construction.

Members of the IWC Workgroup on Guidelines for Population Genetic Data Analysis include Scott Baker, Mark Bravington, Greg Donovan, Mike Double, Rus Hoelzel, Jennifer Jackson, Phil Morin, Ada Natoli, Per Palsböll, and Robin Waples.

REFERENCES

- Baker CS, Dalebout ML, Lavery S, and Ross HA. 2003. www.DNA-surveillance: applied molecular taxonomy for species conservation and discovery. *Trends Ecol. Evol.* 18:271-272.
- IWC. 2009. Report of the Scientific Committee. Annex I. Report of the Working Group on Stock Definition. Appendix 2. Guidelines for DNA data quality control for genetic studies relevant to IWC management advice. *J. Cetacean Res. Manage. (Suppl.)* 11: 252-256.
- Ross HA, Lento GM, Dalebout ML, Goode M, Ewing G, McLaren P, Rodrigo AG, Lavery S, and Baker CS. 2003. DNA surveillance: web-based molecular identification of whales, dolphins and porpoises. *J. Heredity* 94, 111-114.
- Ross HA, and Murugan S. 2006. Using phylogenetic analyses and reference datasets to validate the species identities of cetacean sequences in GenBank. *Mol. Phylog. Evol.* 40:866-871.