Under Development: Guidelines for the analysis of population genetic data used in an IWC management context

ROBIN S. WAPLES[1], A. RUS HOELZEL[2], and members of the IWC working group.
[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.

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 include a few worked examples to illustrate the type of content the document might contain. 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.

GUIDELINES

I. Species identification

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. Information related to tests of Hardy-Weinberg (HW) equilibrium. We assume that HW evaluations have been conducted as part of the DNA data quality control step; here, attention would focus on HW deviations that might provide insight into biological processes (such as inbreeding or population mixtures).

B. Information related to tests of linkage disequilibrium (LD). As was the case for HW, the focus would be on signals that might provide insight into biological processes.
C. Measures of genetic diversity, including rarefaction (controlling for sample size in estimating allelic richness)

III. Estimating population size

A. Census size, *N*
    1. DNA mark-recapture
    2. Analysis of close relatives
    3. Identifying recent population bottlenecks
    4. Signals of historical population expansion/contraction
B. Effective population size, N_e
    1. Historical N_e (including coalescent)
    2. Contemporary N_e
        a. Single-sample methods
        b. Two-sample (temporal) methods
    3. Recent vs ancestral N_e using isolation with migration models

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

This is probably the most common type of management problem that utilizes genetic data. This section might be organized in any of several different ways. One way would be to use a framework based on the major stock archetypes identified in TOSSM. The approach below uses as a first cut whether or not individuals can plausibly be grouped into putative populations before analyses are conducted.

A. Putative populations defined *a priori*.

In this case, the analyses are conducted on groups of individuals (samples); individuals are grouped into samples based on collection location, assumptions about geographic configuration of populations, or other *a priori* hypotheses about population membership.

1. Testing panmixia
2. Describing population structure
    a. F_{ST}, genetic distance, and related measures
    b. Ordination
    c. Isolation by distance
    d. Landscape genetics (units = samples)
3. Estimating migration
    a. Methods that assume migration-drift equilibrium and estimate long-term patterns of gene flow (mN_e)
        i. Island model
        ii. Stepping-stone models
        iii. Coalescent methods and directional gene flow
    b. Assignment methods that estimate contemporary migration rate (m)
    c. Isolation with migration models to estimate splitting times and post-division migration rates
4. Estimating divergence time
5. Mixture analysis (e.g., resolving stock composition of samples from feeding grounds or migration pathways)

B. 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.

1. Clustering programs
2. Clustering based on ordination
3. Landscape genetics (units = individuals)

V. Generic 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. Confidence intervals
F. Sampling and experimental design
G. Underlying assumptions and sensitivity to their violation
H. Bayesian vs maximum likelihood vs frequentist methods
I. MCMC issues (burnin, convergence)
J. Integrating genetic and non-genetic data
K. Possible influence of selection

VI. Summary and conclusions

EXAMPLES
The following examples illustrate the type of treatment and level of detail that might be used for individual topics in the guidance document.

IV.A.3.b. Assignment methods that estimate contemporary migration rate (*m*)
Efforts to estimate levels of connectivity from genetic data have traditionally relied on equilibrium models that integrate information over evolutionary time periods (see section IV.A.3.a.). The last decade has seen increasing interest in so-called 'assignment methods' that do not require equilibrium assumptions and instead can estimate contemporary patterns of migration over time frames encompassed by the samples. 'Assignment tests' (Paetkau 1995; Manel et al. 2005) are a type of discriminant function analysis in which the discriminant functions are based on genetic traits that differ in frequency among potential source populations. Multilocus genotypes are used to 'assign' individuals to the most likely source population, guided by learning samples collected from potential sources. If an individual is assigned to a population other than the one it was sampled from, it can be inferred that the individual is a migrant (Waser and Strobeck 1998; Berry 2004). The program *GeneClass* (Piry et al. 2004) includes several different assignment test methods and offers the user various options for attempting to identify first-generation migrants. Other programs attempt to identify second-generation migrants (Wilson and Rannala 2003) or estimate the fraction of genes in each individual that are derived from each population (Pritchard et al. 2000).

Assignment methods have some advantages for estimating migration: they don't require one to assume migration-drift equilibrium, as do most standard models; they can potentially provide very detailed information about connectivity (both magnitude and direction); and they provide information about contemporary dispersal, which might be of interest for a variety of reasons. However, assignment methods also have some substantial limitations for studying dispersal. First, these methods provide information about movement of individuals but not reproductive success of the migrants; therefore, they do not provide any direct information about gene flow. Second, assignment methods provide information about dispersal only for the time frames encompassed by the sampling. Because dispersal is a stochastic process, samples taken from only one or a few years might not provide a representative picture of migration. This can be contrasted with equilibrium models, which can provide an estimate of long-term patterns of gene flow from samples taken at a single point in time. For any given application, these two factors might or might not represent serious limitations, depending on the nature and objectives of the research program.

A third factor—statistical power—is potentially a more general limitation on use of assignment methods to study contemporary dispersal. Power to detect migrants with genetic methods depends on two things: the amount of data one has (samples of individuals, gene loci, and alleles), and the magnitude of genetic differences among populations. The researcher has control over the former but not the latter, and therein lies a conundrum: power is highest when genetic differences among populations are large, but in that case migrants will be rare and difficult to detect without a very ambitious sampling program; conversely, if migration is high enough to provide reasonable prospects for finding migrants, the resulting levels of gene flow should erode most differences among populations, making it difficult to genetically distinguish migrants from residents.

Two examples illustrate the inherent difficulty related to power. Paetkau et al. (2004) used computer simulations to evaluate power to detect first generation migrants. They found that even with fairly large amounts of data (50 individuals sampled per population; 20 microsatellite-like gene loci), power to detect true migrants was <50% when gene flow rates were high enough ($mN \geq 5$) to keep F_{ST} values below about 0.05. These conditions would apply to a substantial fraction of potential applications for cetaceans.

Second, the power issue sets up an inherent tradeoff between Type I (incorrectly labeling a resident as a migrant) and Type II (failing to detect a true migrant) error rates, either of which can seriously bias estimates of migration. Consider this hypothetical example: a group of populations with $N = 100$ individuals each are connected by 1% migration per generation ($m = 0.01$). This leads to $mN = 1$ (a low level of gene flow) and relatively large genetic differences among populations. Optimistically, assume that these large differences lead to ~100% power to detect migrants using assignment methods (as found by Paetkau et al. 2004 for data-rich scenarios). So, a large sample would on average contain 1% true and correctly-identified migrants. But if the standard tolerance for Type I error is used ($\alpha = 0.05$), then 5% of the sample would also be incorrectly identified as migrants. In this case, even with perfect statistical power, the estimate of migration rate (0.01 + 0.05 = 0.06) would be six times the true level. The only solution to this problem is to adopt a very low α level, but doing so is likely to compromise power unless genetic differentiation is very strong.

The conundrum regarding power does not necessarily represent an insurmountable problem for using assignment methods to study contemporary dispersal—for example, Berry et al. (2004) reported reasonably good agreement between genetic and mark-recapture estimates of dispersal in a series of populations of the grand skink, *Oligosoma grande*, for which F_{ST} values ranged between 0.04 and 0.11. However, the issues discussed above do indicate that careful attention to experimental design is essential, as is a realistic assessment of prospects of producing useful information. Three general strategies can help improve performance. First, in theory at least, it is possible to achieve high power to identify migrants among populations with very low levels of genetic differentiation, provided that arbitrarily large numbers of loci and alleles can be scored. At present the ability to do this with non-model species is limited, but that situation could change in the future. Second, adopting a very low tolerance for Type I errors (e.g., $\alpha \leq 0.01$) can help reduce some of the most serious sources of potential bias, but this will likely compromise power unless genetic differences are moderately large and/or very large amounts of data are available. Third, the major challenges to these methods arise from uncertainty in identifying individual migrants. Using an analogue to Genetic Stock Identification (resolution of mixed-stock fisheries using genetic data—Shaklee et al. 1999; Manel et al. 2005), if focus is shifted from identifying individual migrants to estimating an overall migration *rate*, then uncertainty about origins of individuals might not preclude precise and accurate estimates of migration. This would require developing, or at least refining, some new analytical techniques. One software program, *BayesAss* (Wilson and Rannala 2003) does actually attempt to estimate migration rate, but its performance with weakly differentiated populations has not been encouraging (Faubet et al. 2007).

Finally, the conundrum regarding the inverse relationship between the level of migration and genetic differentiation largely disappears if the system one is analyzing involves populations that historically have been strongly isolated (and hence are well differentiated genetically) but which are currently exchanging sizeable numbers of migrants. This non-equilibrium situation cannot persist for long unless the migrants have little or no reproductive success, but in the interim could provide a large number of migrants to sample *and* high power to distinguish them from residents. This scenario, in fact, is one that the Wilson and Rannala (2003) program was designed to study. It is not clear how often this scenario might occur with cetaceans.

V.C. Multiple testing

Some analyses routinely involve multiple tests of the same hypothesis (e.g., tests of HWE and LD, or pairwise tests of heterogeneity between populations). In these applications, it is common practice to use a correction for multiple testing, such as the Bonferroni correction, in which the critical *P* value is inversely proportional to the number of tests. Two points should be kept in mind when using this type of correction for multiple tests.

1) The Bonferroni correction is widely known to be conservative and hence will fail to detect some actual departures from the null hypothesis.
2) If the correction is performed, then the expectation is (with probability 1-*P*) that the number of adjusted significant tests will be zero. Therefore, even a single adjusted significant test cannot easily be attributed to chance and requires an explanation.

If a multiple testing correction is performed, a better option might be the false discovery rate (FDR; the fraction of tests in which the null hypothesis is falsely rejected; Benjamini and Hochberg 1995), which adjusts for multiple testing without sacrificing as much power as the Bonferroni correction. In addition, it is recommended that results are also presented for unadjusted tests, as the distribution of unadjusted *P* values provides valuable information about agreement with the underlying null hypothesis.

V.D. Mutation rates

The parameter $\theta = 4N_e\mu$ plays a key role in both theoretical and applied population genetics. θ is a composite parameter, proportional to the product of effective population size (N_e) and mutation rate (μ). Although this fact adds complexity to some analyses, it can be used to advantage by a simple rearrangement of the above equation:

$$N_e = \theta/(4\mu). \quad (1)$$

This means that if θ can be estimated from population genetic data (as is routinely done with microsatellites or sequence data for mitochondrial or nuclear DNA), then insights into N_e can be obtained if one can also estimate mutation rate. The effective size that is estimated in this way is a long-term, or historic, N_e that depends (among other things) on the assumption that measured levels of genetic diversity reflect an equilibrium between mutation and genetic drift (see Section III.B.1. above). This general approach has a variety of practical applications, such as estimating historical effective population size; estimating divergence times between populations or species; and estimating population demographic patterns over time (see Beaumont and Rannala 2004, Nielsen and Beaumont 2009). Several factors, however, contribute to uncertainty and limit the practical usefulness of these approaches. First, only four kinds of DNA bases occur (termed A, T, C, G for short), and DNA sequences are typically compared by counting the fraction of sites at which they have different bases. Once a mutation has occurred at a particular site (e.g., from A to T), a subsequent mutation at that site will still result in only a single difference compared to the reference sequence (if the mutation is from T to G or C) or will negate the original change (if the mutation is a back mutation from T back to A). This saturation effect is of particular relevance for estimates of mutation rate at mtDNA, which are typically obtained by the 'phylogenetic method' that involves comparing sequences from different species. In addition to making duplicate mutations more likely, this introduces potential sources of error in developing calibration points for divergence times—typically derived from the fossil record, which is relatively poor for cetaceans.

Second, mutation rates can vary considerably among species and among regions of the genome within species. For many years, a '2% rule' was used for mtDNA, based on calculations using the phylogenetic method, suggesting that for vertebrates the average rate of base substitution was about 2% per million years (Wilson et al. 1985). However, rates vary among regions of the mtDNA molecule, and for the widely used control region, the estimates were considerably higher (12% to 38% per million years in humans; see review in Henn et al. 2009). Furthermore, the mtDNA control region itself is heterogeneous for mutation rate, with the central, very conserved, region being flanked by two 'hypervariable' regions (HVR1 and HVR2).

Finally, recent estimates of mutation rate over shorter time frames for intraspecific comparisons often differ greatly from those based on the phylogenetic method. An extensive analysis (Howell et al. 2003) provided an estimate for the human HVR1 of 95% per million years (0.95 changes / site/ million years). Similar approaches applied to other species including *C. elegans* (Denver et al. 2000) and Adélie penguins, *Pygoscelis adeliae* (Millar et al. 2008) have also produced estimates that are 1-2 orders of magnitude higher than suggested by the 2% rule. Henn et al. (2009) suggested that for humans the high mutation rates decay after about 15,000 years, but for penguins the elevated rate seemed to extend back further in time (Millar et al. 2008).

From the form of Equation 1, it is easy to see that errors in estimating the mutation rate directly translate into the same proportional errors in estimating N_e. For example, if mutation rate is underestimated by a factor of two, N_e will be overestimated by the same amount. This fact, together with the wide range of published estimates of mutation rates, has helped spawn much of the controversy that has surrounded some attempts to estimate historical N_e based on existing levels of genetic diversity. For example, Roman and Palumbi (2003) calculated that there must have been many more whales in pre-whaling oceans than had previously been thought, based on an estimate of mutation rate in the cetacean mtDNA control region derived from the phylogenetic method (e.g., Hoelzel et al. 1991)—about 2% per million years. However, the relevant timeframe suggests that the much higher rate estimates derived from intra-specific genealogies might be more appropriate (e.g. Henn et al. 2009, Millar et al. 2008). If those higher rates were instead applied, the cetacean population size estimates would fall in line with what had been previously interpreted from historical catch data.

In summary, current levels of genetic diversity and other patterns in the DNA of contemporary populations contain information about historic size and demographic processes. However, deciphering this information is tricky and depends heavily on obtaining a reliable estimate of mutation rate. It is not enough to have an estimate of mutation rate for the focal species; it is also important to have estimates for the regions of the genome that produced the genetic data being analyzed, and to apply the correct mutation rate to the relevant time frame – higher rates for more

recent events (Ho et al. 2005). Because of the saturation effect, estimation of mutation rates might best be confined to the time period during which each mutation occurs at a unique site and the cumulative number of substitutions increases linearly with time.

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.

Beaumont MA, and Rannala B. 2004. The Bayesian revolution in genetics. *Nature Reviews Genetics* 5: 251-261.

Benjamini Y, and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Berry O, Tocher MD, and Sarre SD. 2004. Can assignment tests measure dispersal? *Mol. Ecol.* 13, 551–561.

Denver DR, Morris, K, Lynch M, Vassilieva LL and Thomas K. 2000. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* 289:2342-2344.

Faubet F, Waples RS, and Gaggiotti OE. 2007. Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Mol. Ecol.* 16:1149–1166.

Henn BM, Gignoux CR, Feldman MW and Mountain JL. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Mol. Biol. Evol*. 26:217-230.

Ho SYW, Phillips MJ, Cooper A, and Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* 22: 1561–1568.

Hoelzel AR, Hancock JM and Dover GA. 1991. Evolution of the cetacean mitochondrial D-loop region. *Mol. Biol. Evol.* 8:475-493.

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM and Herrnstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am. J. Hum. Genet.* 72:659-670.

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.

Manel S, Gaggiotti O, and Waples RS. 2005. Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. Evol.* 20:136-142.

Millar CD, Dodd A, Anderson J, Gibb GC, Ritchie PA, Baroni C, Woodhams MD, Hendy MD, and Lambert DM. 2008. Mutation and evolutionary rates in Adélie Penguins from the Antarctic. *PLoS Genetics* 4: e1000209.

Narum SR. 2006. Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conserv. Gen.* 7:783–787.

Nielsen R, and Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol. Ecol.* 18, 1034–1047.

Pritchard JK, Stephens P, and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Paetkau D, Calvert W, Stirling I, and Strobeck C. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol*. 4:347–354.

Paetkau D, Slade R, Burden M, Estoup A. 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Mol. Ecol.* 13:55–65.

Piry S, Alapetite A, Cornuet J-M, Paetkau D, Baudouin L, and Estoup A. 2004. *GENECLASS2*: a software for genetic assignment and firstgeneration migrant detection. *J. Heredity* 95:536–539.

Roman J, and Palumbi SR. 2003. Whales before whaling in the North Atlantic. *Science* 301:508-510.

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.

Shaklee JB, Beacham TD, Seeb L, and White BA. 1999. Managing fisheries using genetic data: case studies from four species of Pacific Salmon. *Fish. Res.* 43, 45–78.

Waser PM, and Strobeck C. 1998. Genetic signatures of interpopulation dispersal. *Trends Ecol. Evol*. 13:43–44.

Wilson AC et al. 1985. Mitochondrial-DNA and 2 perspectives on evolutionary genetics. *Biol. J. Linn Soc*. 26:375-400.
Wilson GA, and Rannala B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163:1177–1191.