Applied Conservation Genetics and the Need for Quality Control and Reporting of Genetic Data Used in Fisheries and Wildlife Management

PHILLIP A. MORIN, KAREN K. MARTIEN, FREDERICK I. ARCHER, FRANK CIPRIANO, DEBBIE STEEL, JENNIFER JACKSON, AND BARBARA L. TAYLOR

From the Southwest Fisheries Science Center, National Marine Fisheries Service, La Jolla, CA 92037 (Morin, Martien, Archer, and Taylor); the San Francisco State University, San Francisco, CA 94132 (Cipriano); and the Hatfield Marine Science Center, Oregon State University, Newport, OR 97365 (Steel and Jackson).

Address correspondence to Dr Phillip Morin at the address above, or e-mail: phillip.morin@noaa.gov.

Genetic data are often critical for defining populations for management purposes (e.g., identifying geographic boundaries or diagnostic characters for genetically discrete subunits) but can be called into question by both scientific and legal review. This can result in reversed or delayed implementation of management actions. We discuss methods for data quality control and quality analysis and describe examples of steps applied to 2 of the most common types of genetic data, mitochondrial DNA sequences, and microsatellite genotypes. These steps can serve both as guides to conservation geneticists and as an initial protocol for managers to determine whether genetic data will hold up against legal and scientific challenges. In addition, we suggest types of data and quality measures that should be reported as supplementary materials to published reports. These supplementary data serve to reduce the occurrence of legal and conservation controversies and improve reproducibility over time in population genetics studies where genetic monitoring is likely to play an increasing role.

Key words: conservation, genetics, quality control

Genetics studies are playing an increasingly important role in delimiting units to conserve in an ever-growing number of protected and exploited species. These delimitations in turn affect management decisions with significant impacts on not only the species or populations under consideration but also the human enterprises that depend on or impact them. For example, such decisions have recently resulted in the full or partial closures of some fisheries and constraints on land use or development under implementation of the US Endangered Species Act. Lengthy and costly debates of the merits and interpretation of genetic data to define conservation units have occurred and are likely to become more common (King et al. 2006; Ramey et al. 2006).

The success of management advice based on genetic data critically depends on 3 issues: 1) experimental design (including appropriate sampling scheme with regard to sample size and geographic coverage); 2) procedural implementation of sample handling and laboratory analysis (including labeling, archiving, and data quality checks); and 3) appropriate data analysis and interpretation. It is also critical that the biological question is properly defined, as it will affect the choice and number of markers (loci) used, analytical methods, decision criteria, and interpretation of data, all of which play different roles in the detection of population structure (Schlotterer 2004; Morin and Dizon 2009; Taylor et al. 2009). Here we will focus on the effects of data quality on the interpretation of data under the assumption that the focus of the study has been determined, appropriate markers have been selected, and that those markers include (but are not necessarily limited to) mitochondrial DNA (mtDNA) sequences and nuclear microsatellite genotypes.

Unlike more academic genetic research, conservation genetics for management purposes must be transparent to the users of the research results (often not geneticists) and should be designed to be augmented (perhaps by multiple laboratories in different countries) as more samples and markers become available. These markers should preferably have a low and well-characterized error rate such that analytical results can withstand not only scientific but also sometimes considerable legal scrutiny. Data quality needs to be adequate to support the conclusions reached, but beyond that scientists, managers, lawyers, and policy makers need to be able to evaluate data quality and determine whether there
are technical or analytical issues that have significantly affected the outcome of a study. In addition, effective use of genetic data when sample sizes are expected to increase through time also requires planning to allow tracking of additions, changes and corrections, and careful documentation of laboratory protocols to ensure compatibility with expected future data. Laboratory inventory management systems must be established to ensure access to up-to-date results by all researchers engaged in the analysis and to provide a clear baseline for future genetic monitoring (Schwartz et al. 2007). For these reasons, the methodology used for collecting genetic data and their application in management decisions should be transparent, and methods for assessment of data quality and the results of such assessments should also be clearly described and controlled. These issues are relatively uncommon in typical academic research where outcomes usually do not have large economic or conservation consequences. To apply genetic data in conservation management, a new paradigm of data quality control (QC) and analysis reporting is needed.

To provide a guide for improved QC and quality analysis (QC/QA), we outline the major steps that help to assure data quality and transparency and provide an example of the process as applied to 2 of the most common types of genetic data, mtDNA sequences and microsatellite genotypes. Many of these steps can and should also be applied to other types of genetic data used to make management decisions (e.g., single nucleotide polymorphisms [SNPs]). A summary of the steps, with the types of QC that can be implemented to maximize data quality and transparency, is shown in Table 1. Generally, errors and inconsistencies can be introduced at various points of a DNA study and fall into 3 different primary categories: 1) difficulties in reliable genotyping due to locus characteristics; 2) insufficient tissue or DNA sample quality; and 3) inconsistency of methods and lack of adherence to accepted standards, for example, “good laboratory practice” (Seiler 2005). Avoiding errors associated with the first item requires marker validation (often addressed in a pilot study), whereas the second and third items are addressed by implementing systematic QC throughout the entire study.

We structure the following by the steps in the data generation process from receiving samples to making available quality-controlled data ready for analysis. We begin with steps to quantify the quality of the DNA extracted from tissue samples and then separately treat mtDNA sequencing and microsatellite genotyping and the relevant laboratory and data quality checks.

### QC/QA Considerations at each Step of Genetic Analysis

**Step 1: Sample Collection and Assessing Sample Quality Prior to Genetic Analysis**

For all genetics studies, the process starts with sample collection. The effects of sample preservation methods, sample labeling and transport, and data collection in the field can have downstream effects in data quality and interpretation. As with marker selection and validation, we recommend careful study design and a pilot project that will help to identify problems with handling and preservation methods before a large number of samples have been collected (for an overview and examples, see Schwartz and Monfort 2008).

<table>
<thead>
<tr>
<th>Table 1. Application of QC steps in each phase of a conservation genetics study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project stage</strong></td>
</tr>
<tr>
<td>Study design</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Data generation</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Data presentation and archiving</td>
</tr>
</tbody>
</table>

Conclusion: With careful planning, implementation of systematic QC throughout a study, and effective data management, the application of genetic data to conservation management decisions can be made with greater transparency and reliability.
Even with a robust collection and preservation methodology established, variation in sample quality will still be a consideration (e.g., degraded samples from dead animals, non-invasively collected samples such as feces, hair, and sloughed skin, samples degraded from long-term storage or improper handling, etc.). Few studies can claim that sample quality is equal for all samples and that variation in quality has not affected the researchers’ ability to generate accurate data. There are many publications discussing methods to assure data accuracy for samples known to be of poor quality (e.g., Navidi et al. 1992; Taberlet et al. 1996; Morin et al. 2001; Paetkau 2003; McKelvey and Schwartz 2004) and the need to estimate error rates (Bonin et al. 2004; Broquet and Petit 2004; Morin, LeDuc, et al. 2009).

Variation in sample quality can be quantified prior to attempting to generate genetic data, resulting in cost and time savings as well as higher quality data. Although not strictly necessary, analysis of DNA sample quality prior to genetic data generation can ensure that low-quality (and therefore highly error prone) samples are either removed from the study or replicated sufficiently to ensure accuracy. This is particularly important for studies involving sample types that are very likely to be of poor quality (e.g., noninvasive and historical samples; Taberlet et al. 1996; Morin et al. 2001; Paetkau 2003; McKelvey and Schwartz 2004; Gilbert et al. 2005; Morin and McCarthy 2007) but can also be important for any study, as sample quality can vary significantly even when samples initially appear to be of relatively high and uniform quality. Indeed, the presence of even a single poor-quality sample in a population sample can result in false inference of population structure (Morin, LeDuc, et al. 2009), leading to controversial data interpretations and pursuit of potentially unnecessary additional analyses (e.g., Givens et al. 2009).

For these reasons, we strongly recommend that samples be prescreened for at least DNA concentration prior to beginning a study with nuclear markers. When samples are expected to meet a minimum threshold level of DNA (e.g., 20 ng per polymerase chain reaction [PCR]), quantification by absorbance or fluorescence spectrophotometry (e.g., Pico Green) can be rapid and inexpensive, allowing sample concentrations to be normalized to produce consistent results. For samples expected to yield low quality or limited concentration of DNA extracts, more sensitive methods such as quantitative PCR can provide highly accurate data on DNA concentration, and even on relative abundance of DNA at multiple fragment sizes, to optimize sample selection and data replication criteria (Morin et al. 2001, 2007; Morin and McCarthy 2007). Empirical testing of some samples through the entire data generation procedure may be necessary to see how DNA quantity affects data quality, as an absolute relationship cannot be assumed across studies. Information on average sample quality can affect how additional protocols are implemented or suggest that extra precautions be observed, for example, when DNA concentrations are generally low, potential for contamination is increased. If DNA from available samples is highly fragmented (degraded), it is advisable to target smaller microsatellite or smaller mitochondrial amplicons or to use SNPs instead of microsatellites (Morin and McCarthy 2007).

**Step 2: Data Generation**

**Sample Controls**

To estimate error rates once data have been generated, it is necessary to conduct a series of blind control replicates. These controls, which should be specified in the initial study design, can be used to compare genotypes generated throughout the data generation process and will serve the following purposes:

1. Identification of random and systematic errors via random sample replication. A subset of samples (a few percent of the total) scattered throughout the samples and genotyped/sequenced at all loci will help to identify errors that have to do with either sample handling or raw data interpretation.
2. Verification of alleles and normalization of sizes across time, laboratories, and technologies. This requires including 2–3 standard samples or allelic ladders replicated in every genotyping experiment (PCR and electrophoresis) (LaHood et al. 2002; Morin, Manaster, et al. 2009).
3. Targeted replication of samples after the majority of data are generated. This will allow verification of data quality and can also detect sample-handling errors (e.g., reversal of a sample plate). We recommend that this involves some samples from every sample group run together, with >10% replication of the data set (in combination with the controls from (1) and (2)).

**Double-Blind Genotype Scoring**

In addition to controls, genotype scoring can be prone to biases and common error types. Several authors have recommended that at least 10% of microsatellite genotypes (across all loci) be scored (blind to the original scores) by a second experienced genotyper. This serves to identify particular loci that may be difficult to score consistently and detects biases in the way that 1 genotyper interprets raw data. As any experienced genotyper will attest, calling microsatellite genotypes correctly and consistently (as well as checking sequence database calls and interpreting other types of genetic data) requires experience; for potentially contentious data, such as those used in management decisions, it is inappropriate to assign genotyping to anyone without proper training, supervision, and documented error checking procedures.

**Electronic Capture of Raw and Scored Genotypes**

Most genotyping now makes use of fluorescently labeled PCR products detected by automated systems (e.g., capillary electrophoresis) followed by automated or semiautomated allele size scoring. The data that can be captured automatically from the software include sample names, raw allele sizes, fluorescence peak heights, scored genotypes, fluorophore, run date, instrument type, capillary length, polymer, quality score, etc. When data may be compared across time or laboratories, these ancillary data can be
critical for proper allele binning and data quality checking. Electronic capture of the data and direct import into a simple database (e.g., MICROSOFT ACCESS) can reduce errors due to manual transcription of genotypes to spreadsheets and has the added benefit of easily capturing all experimental results, including those that failed to be scored. Failed genotypes can be indicative of both sample and marker quality problems, providing important information for quality analysis. An example of some of the data that can be automatically captured for storage in a database is shown in Table 2. This type of database storage of genotype data simplifies analysis of replicates and calculation of error rates, as well as facilitating more sophisticated queries of the data for QC, data analysis, and data reporting.

**Allele Size Binning**

One of the biggest technical difficulties for microsatellite analysis is variability in allele size estimates. Electrophoretic migration rate can be affected by both size and nucleotide composition of the alleles plus the addition of various fluorescent molecules used for detection. Inferred allele size is, therefore, not always perfectly correlated with the actual size of the amplified alleles; allele sizes can differ by more or less than the size of the microsatellite repeat unit (e.g., a CA repeat can have alleles that differ on average by 1.8–2.2 bp; Amos et al. 2007). In addition, electrophoretic migration rates are influenced by variation in chemical and physical conditions and can cause allele size differences of up to 7 bp across time, technologies, and instruments (LaHood et al. 2002; Davison and Chiba 2003). Several methods have been introduced to facilitate normalization of allele sizes, but all require that standard samples or allelic ladders (LaHood et al. 2002) are included with each run to verify that alleles are correctly assigned to bins (Amos et al. 2007; Morin, Manaster, et al. 2009). To facilitate replicability, specification of the standard samples or allelic ladders used should be part of the data report, along with agreement to distribute the standards to other researchers, as is done by several research consortia and programs (e.g., salmon genotyping laboratories; LaHood et al. 2002). Keeping conditions as similar as possible (fluorescent dye label, primer sequences, PCR profile, etc.) can help minimize allelic size differences. Used in conjunction with size standards, this can reduce errors in binning, particularly between different researchers wanting to compare data.

**Step 3: Assessing Sample-Specific Data Quality**

Several methods can be used to investigate data quality for individual samples (Table 3) ranging from simple calculations, like the numbers of missing and of homozygous genotypes per individual, to more complicated analysis of the effects of individual genotypes on deviations from Hardy–Weinberg equilibrium (HWE) (Morin, LeDuc, et al. 2009). Indeed, if genomic DNA quantity and/or quality is assayed, the correlation of DNA concentration (or other qualitative or quantitative quality measures) with error rates can be used to predict which samples might require additional replication or...
error checking to ensure correct genotypes and low error rates (Morin et al. 2001; Morin and McCarthy 2007).

With or without such DNA characterization, individual samples with unusual characteristics warrant extra scrutiny to verify genotypes and sequences, as these samples are both more likely to contain errors and more likely to bias analytical results. For microsatellite data, a simple analysis of the number or percentage of homozygous and missing genotypes per individual can rapidly identify individuals likely to have experienced high levels of “allelic dropout” (failure to amplify 1 of the alleles in a heterozygote). Plotting the values indicates which samples are outliers from the general population or at the high end of the distribution (Figure 1), so that genotypes can be replicated to correct seemingly homozygous genotypes that are due to allelic dropout. If exclusion does not significantly impact sample size in the population strata or introduce bias (e.g., poor-quality samples disproportionately present in 1 stratum or loss occurs of alleles that significantly alters allele frequencies in 1 stratum), it might be better to simply exclude the poor-quality samples from further analysis rather than spend the time and money to replicate analyses.

Mitochondrial sequence data also need to be carefully scrutinized for sample-specific errors. Sequence quality can be quantified either with Phred scores or the roughly equivalent Applied Biosystems Inc. (ABI) “QC scores” (Ewing and Green 1998; Ewing et al. 1998). These scores represent the probability of the miscall of an individual base, as reported on a log scale (e.g., 30 is a 1 in 1000 error rate and 40 is a 1 in 10 000 error rate). As average QC scores can be misleading, we also suggest establishing criteria for threshold values that take into account variation in quality, for example, 90% of all bases must have a quality score >30 and 100% must be >20. As a general practice, sequences falling below the chosen threshold should be visually inspected and repeated if necessary, though there are instances where even apparently high-quality sequences can contain systematic or random errors. If a sample is found to have a unique haplotype sequence (i.e., a haplotype not found in any other sample), then the electropherogram should be checked to ensure that all nucleotides were correctly called. We recommend that these samples also be sequenced again, preferably from a new PCR product, to confirm the unique sequence. This is especially true if the haplotype differs from other haplotypes by only a single nucleotide. Similarly, any nucleotide substitution that is unique to a single haplotype should be confirmed by resequencing. To reduce costs, some laboratories routinely sequence only 1 DNA strand but use similar protocols to determine which samples should additionally be sequenced in the “reverse” direction and/or resequenced. Nuclear sequences should be treated in a similar way, recognizing that with direct nuclear sequencing the presence of heterozygous sites will strongly reduce local base call scores and thus reduce the overall quality score. In such cases, the

### Table 3. Sample-dependent error checking methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Application</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA quantification</td>
<td>Identifies samples likely to produce poor-quality data (allelic dropout,</td>
<td>Morin et al. (2001, 2007); Morin and McCarthy</td>
</tr>
<tr>
<td></td>
<td>spurious alleles, and short-allele dominance)</td>
<td>(2007); Morin, LeDuc, et al. (2009)</td>
</tr>
<tr>
<td>Excess homozygosity</td>
<td>Identifies samples with unusual (outlier) levels of homozygosity that could</td>
<td>Taberlet et al. (1996); Miller et al. (2002);</td>
</tr>
<tr>
<td></td>
<td>be due to allelic dropout.</td>
<td>Johnson and Haydon (2007); Morin, LeDuc, et al.</td>
</tr>
<tr>
<td>Genetic identity or similarity</td>
<td>Identifies known and unknown sample duplicates and types of genotyping errors</td>
<td>McKelvey and Schwartz (2004)</td>
</tr>
<tr>
<td></td>
<td>found in duplicates that do not match perfectly.</td>
<td></td>
</tr>
<tr>
<td>Effect on HWE (Jackknife analysis)</td>
<td>Jackknife analysis of microsatellite data for the effect of individual</td>
<td>Morin, LeDuc, et al. (2009)</td>
</tr>
<tr>
<td></td>
<td>samples on significant deviations from HWE; identifies rare homozygous</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genotypes and influential samples.</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 1](http://jhered.oxfordjournals.org)  
Proportion of homozygosity for 138 bottlenose dolphin samples genotyped for 11 microsatellites. The proportion is the number of homozygous genotypes divided by the number of completed genotypes. Samples that fall outside the distribution or at the high end (e.g., ≥0.5 in this example) may be more likely to have allelic dropout or other issues.
### Table 4. Reporting of genetic and QC data

<table>
<thead>
<tr>
<th></th>
<th>Data to present</th>
<th>Method of determination</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Microsatellites</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype data</td>
<td>Uncorrected raw allele sizes</td>
<td>Output from electropherogram analysis program (GENEMAPPER, GENESCAN, GENOTYPER, etc.)</td>
<td>Optional</td>
</tr>
<tr>
<td></td>
<td>Scored genotypes and normalized allele sizes (if relevant)</td>
<td>Genotypes scored based on allelic ladders or control samples (made available for others to use)</td>
<td>Critical</td>
</tr>
<tr>
<td>Summary statistics</td>
<td>Number of alleles, allele size range, repeat size, and heterozygosities</td>
<td>MICROCHECKER, GENEPOL, and other programs for summary statistics</td>
<td>Critical</td>
</tr>
<tr>
<td></td>
<td>Deviation from HWE for all loci, by locus</td>
<td>Fisher's Exact test for deviation from HWE (e.g., GENEPOL and ARLEQUIN)</td>
<td>Critical</td>
</tr>
<tr>
<td>Error rates</td>
<td>Effects of individual samples on HWE</td>
<td>Jackknife analysis</td>
<td>Desirable</td>
</tr>
<tr>
<td></td>
<td>Per-marker and whole data set per-allele error and PCR failure rates</td>
<td>Systematic blind replication of ~10% of samples and consistent use of controls.</td>
<td>Critical</td>
</tr>
<tr>
<td>Check for unintentional duplicates</td>
<td>Find matched and nearly matched composite genotypes, re-genotype mismatched loci to verify</td>
<td>Individual identity or genetic mark-recapture programs (DROPOUT and MS TOOLS)</td>
<td>Desirable</td>
</tr>
<tr>
<td>Exclusion from analysis</td>
<td>Justification for exclusion of markers and samples from all or some analyses</td>
<td>Based on HWE, jackknife, and other sample or marker summaries or characteristics</td>
<td>Critical</td>
</tr>
<tr>
<td>QA/QC</td>
<td>Description of all QC and quality analysis methods used and their results (if not included above, e.g., in estimating error rates)</td>
<td></td>
<td>Desirable</td>
</tr>
<tr>
<td><strong>MtDNA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haplotype sequences</td>
<td>Submitted to public sequence databases (e.g., GenBank) with haplotype designations</td>
<td>Comparison to reference haplotype set or generate new list (DNASP and phylogenetic software)</td>
<td>Critical</td>
</tr>
<tr>
<td>Summary statistics</td>
<td>Haplotypic diversity and number of haplotypes</td>
<td>Output from analysis programs (DNASP and ARLEQUIN)</td>
<td>Critical</td>
</tr>
<tr>
<td>Error rates</td>
<td>Whole data set, based on replication and double checking of novel haplotypes differing by 1 nucleotide</td>
<td>DNA alignment and analysis programs (SEQUENCHER, SEQSCAPE, DNASP, MEGA, etc.)</td>
<td>Critical</td>
</tr>
<tr>
<td></td>
<td>Description of sites that are difficult to call, inconsistent between sequence directions, or consistently have double peaks</td>
<td>Visual inspection and comparison of sequences by experienced sequencer</td>
<td>Desirable</td>
</tr>
<tr>
<td>QA/QC</td>
<td>Description of all QC and quality analysis methods used and their results (if not included above, e.g., in estimating error rates)</td>
<td></td>
<td>Desirable</td>
</tr>
</tbody>
</table>

* Listed programs are suggestions, but not an exhaustive list.

* Critical, must be included in all studies; Desirable, should be included where possible; Optional, useful in some circumstances, but does not impact interpretation of results.
quality of sequence on either side of putative heterozygous sites can be used to infer reliability of the variable site.

Although it is not practical to detect and correct every error, some errors have potentially greater impact on analysis than others. One example of this is the presence of erroneous homozygous genotypes at rare alleles. Presence of a single rare homozygous genotype in a stratum has been shown to cause significant deviations from HWE, resulting in false inference of population structure (Morin, LeDuc, et al. 2009). Jackknife analysis of genotypic data (repeated analysis with the removal of 1 sample at a time) can reveal which samples have the greatest effect on HWE, so that they can be rechecked to verify the genotypes (Morin, LeDuc, et al. 2009). The effect of genotyping errors on individual identification (e.g., genetic mark–recapture) and paternity studies are generally higher than on allele frequency–based studies, so in some cases different levels of error will be considered acceptable in a study depending on the applications and goals.

Finally, sample replication can include both intentional and unintentional replicates, for example, samples included multiple times in the data generation process specifically to estimate error rates and to control for variability across sample runs and between laboratories or samples unknowingly replicated because individuals were sampled multiple times or because of sample-handling errors. These replications can be identified by using software that identifies genotypes that match at more loci than is typical for unrelated individuals (e.g., DROPOUT; McKelvey and Schwartz 2004). When no genotyping errors are present, the samples are easy to identify because of perfect genotype matches at all loci. Because genotyping errors introduce uncertainty, however, software that identifies near matches can help to identify potential replicates to be rechecked. It is important to note that even with low error rates, the chances that replicated samples will not match at all loci increases with the number of loci. For example, an error rate of 1% per allele, which is typical of microsatellite studies of wildlife (Bonin et al. 2004; Hoffman and Amos 2005; Hoffman et al. 2006; Morin, LeDuc, et al. 2009), would result in an 80% probability that the same sample individual genotyped twice for 20 microsatellites would not match at all 40 alleles.

Table 5. Sperm whale microsatellite genotyping QC/QA summary for genotyping project with 8 microsatellite markers and 320 samples, including remote tissue biopsies from live animals, skin biopsies from stranded animals, and sloughed skin collected from water after an animal has dived.

<table>
<thead>
<tr>
<th>QC/QA step</th>
<th>Results</th>
<th>Sample No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checked allele binning for all experiments using control samples</td>
<td>Could not bin alleles for loci EV37 and EV30; excluded loci from further analysis</td>
<td>320</td>
</tr>
<tr>
<td>Checked for genotype mismatches between replicated samples</td>
<td>Sixteen Out of 1849 replicated genotypes (0.9%) did not match; most were resolved by looking at the raw data, and remaining were re-genotyped to verify the genotype</td>
<td>320</td>
</tr>
<tr>
<td>Checked for samples with ≤50% completed genotypes</td>
<td>Twenty-eight samples had 3 or fewer completed genotypes. All had been attempted multiple times from ≥2 extractions, so they were excluded from further analysis.</td>
<td>292</td>
</tr>
<tr>
<td>Calculated % homozygosity for all samples</td>
<td>Six samples had &gt;50% homozygosity across 6 loci; re-genotyped homozygous loci; and 5 samples were excluded due to high failure rate and evidence of allelic dropout at several loci</td>
<td>287</td>
</tr>
<tr>
<td>Used MICROCHECKER to analyze all remaining data</td>
<td>Two loci have potential null alleles, but effect is limited; all loci retained for analysis.</td>
<td>287</td>
</tr>
<tr>
<td>Jackknife analysis of deviations from HWE</td>
<td>Six samples caused 2 markers to deviate from HWE because of homozygous rare alleles (odds ratio &gt; 2; Morin, LeDuc, et al. 2009). Four were re-genotyped; 2 had been previously replicated. After re-genotyping, no samples had odds ratios &gt; 2.</td>
<td>287</td>
</tr>
<tr>
<td>Checked for duplicate samples using program “DROPOUT” to find multilocus genotype matches and near matches among samples</td>
<td>Identified 27 perfect matches across 6 loci plus 10 potential matches (1 or 2 differences, usually where 1 genotype was homozygous for an allele for which the other individual was heterozygous). After re-genotyping homozygotes, all were confirmed to be perfect matches. (mtDNA haplotypes also verified to match.) One sample of each matched set was retained, and the rest were excluded from further analysis.</td>
<td>250</td>
</tr>
<tr>
<td>Final data set released for population structure analysis</td>
<td>Six microsatellite loci completed for 250 samples, quality checked for common genotyping problems, with &gt;97% completion of genotypes for all loci, and an estimated error rate of 0.9%. Duplicate and poorest quality samples removed.</td>
<td>250</td>
</tr>
</tbody>
</table>

*Most duplicated samples were from animals biopsied from the same group, so represent accidental double sampling of the same individual. At least 1 duplication represents genetic resampling identification of the same individual at different locations in the migratory route.
Step 4: Assessing Data Set Quality

MtDNA sequence data are rarely evaluated for error rates or types, as it is generally assumed that automated technologies have advanced to the point where sequencing is routine and generally of high quality. Using Phred scores or ABI QC scores is a relatively effective way to confirm this assumed sequence quality. Nevertheless, there can be systematic errors in mtDNA sequencing that should be checked and reported (i.e., if any nucleotide positions are consistently difficult to call or show a tendency to exhibit double peaks, this should be noted in supplementary materials to alert other researchers working on the same species). Replication of a portion of the samples is important to identify random and systematic sequencing errors. In addition, a potential error that is often ignored and difficult to detect is the presence of nuclear copies of mitochondrial sequences or NuMts (Lopez et al. 1994; Bensasson et al. 2001). NuMts are pervasive in some species (e.g., Tursiops sp.; Dunshea et al. 2008) and can easily be mistaken for actual mitochondrial haplotypes, potentially leading to false inference of population structure or other analysis errors. NuMts have been reported in >60 mammalian species (Bensasson et al. 2001) and are sometimes amplified instead of, or together with, true mtDNA, especially from poor-quality DNA such as from hair and feces or when using degenerate primers or primers designed from other species (Sorenson and Fleischer 1996; Bensasson et al. 2001; Dunshea et al. 2008). There are several ways to detect NuMts, including the presence of highly divergent sequences (Pereira and Baker 2004), detection of unexplained phylogenetic placements (Bensasson et al. 2001), and nucleotide substitutions in phylogenetically conserved positions such as those important to secondary structure of tRNAs or coding of amino acids (Dunshea et al. 2008). Some recent NuMt insertions may not be recognized using these methods and can result in believable and analytically robust phylogenies that are nevertheless incorrect. Some methods for avoiding amplification of NuMts include amplification of long PCR products from representative haplotypes to verify the sequence, extracting DNA from tissue that is rich in mtDNA (e.g., muscle), and cloning and resequencing of products that appear to have ambiguous nucleotides or ghost bands in the sequence (Bensasson et al. 2001). After identification of NuMts in a target species (or population), primers should be redesigned such that they specifically amplify mtDNA (Tiedemann and von Kistowski 1998). Generally, sequences should be compared with those in GenBank (e.g., using BASIC LOCAL ALIGNMENT SEARCH TOOL) and/or run through DNA sequence comparison routines (e.g., Baker et al. 2007) when they are available. Note, however, that GenBank itself lacks a stringent control of sequence authenticity, so additional sequence validation might be necessary (Ross and Murugan 2006).

For microsatellite data, the use of a set of replicated standard samples results in an estimate of the overall, averaged per-allele error rate for the whole data set under the assumption that the standard samples reflect the quality of the overall data set. In reality, however, standard samples rarely reflect average quality. This is because these standards are often chosen because they represent samples that have yielded high-quantity and/or high-quality DNA that would not be exhausted during the data generation period and which will yield data that can be used for interexperiment normalization and validation as well as estimates of error rates due to experimental factors (sample handling, variation in experimental conditions, and systematic errors) and genetic marker characteristics (e.g., short-allele dominance, allelic stutter, and PCR product adenylation; reviewed in van Oosterhout et al. 2004). Unintentional replicates and samples chosen at random to be replicated may give a more accurate assessment of average per-allele error rate and data quality. Sample-specific error rates for low-quality samples may remain an issue, however, and need to be assessed in order to minimize the effects of variation in sample quality (Morin, LeDuc, et al. 2009).

Finally, genetic marker characteristics, especially for microsatellites, can be highly variable. Microsatellite data quality can be affected by repeat frequency, the number of alleles, the size range of alleles, tendency of microsatellite PCR products to “stutter” (van Oosterhout et al. 2004) or be adenylated (also called “plus-A”), and variation in experimental conditions (LaHood et al. 2002; Davison and Chiba 2003). Many papers have been published describing these issues and suggesting methods to deal with them, but every data set differs, and it is up to individual researchers to decide which markers can be genotyped reliably and accurately. However, there are some analytical methods to assess marker quality that can be used to decide whether markers should ultimately be included or excluded from analysis (McKelvey and Schwartz 2004; van Oosterhout et al. 2004; Givens et al. 2009). Marker exclusion has been the subject of intense debate for some data sets used for making management recommendations, and ultimately, the decision to exclude a marker must be justified by objective description of marker issues (such as identification of null alleles, high failure rates, inconsistent genotypes, binning problems, etc.). Significant deviation from Hardy–Weinberg expectations within a population can be indicative of marker problems but should not in itself be considered reason to exclude a marker from analysis. For this reason, it is imperative that researchers routinely assess and report marker QC issues thoroughly enough to justify exclusion of such data from analysis. Reasons for marker exclusion include (but are not limited to) evidence of systematic errors or unusually high error rates, high-frequency null alleles, and high frequency of PCR failure or failure to score genotypes (relative to other markers). Sequencing of selected microsatellite alleles at each locus is encouraged and can provide insight into complexity of size variation especially for heterologous primers. When marker quality is questionable but not obviously poor, analysis of data with and without a marker can help to determine whether a single marker is causing a particular result.

Step 5: Data Reporting, Archiving, and Access

For management decisions based on genetic data, transparency of the data quality assessment and methods of analysis
are critical, as the survival of species and populations is potentially at stake and economic and personal (e.g., private property) impacts can be substantial. Scientific and legal challenges can be costly and delay implementation of management decisions. For these and other reasons, we advocate standards for genetic data reporting, summarized in Table 4, that go beyond what is typical for academic research reporting, particularly the recommendation that raw data need to be made available for replication of the analysis and quality evaluation by independent researchers. As an example of how samples and genetic markers have been checked and QA/QC steps documented, we have provided a summary of data QA/QC for an ongoing sperm whale genetics study (Table 5).

The anonymous peer-review process required for publication in an international journal is one of the best assurances of data quality and appropriate analyses. Given increasing restrictions on the amount of methodological detail allowed in journal articles, we support the principles of the Joint Data Archiving Policy drafted by Michael Whitlock (Editor-in-Chief of The American Naturalist), which recommends that data in the form of electronic supplementary information should be provided with sufficient details that, together with the contents of the paper, allows the primary results to be confirmed or validated (Greenberg 2009). Where results of genetic analyses are reported in advance of publication, as is frequently the case in pressing management issues, we recommend that data holders develop a “data availability protocol” to facilitate verification and transparency. Such a protocol could specify limits to use, similar to requirements of anonymous peer review, with oversight by a neutral third party, such as a professional society or the national or international agency responsible for the final management policy or decisions. This protects the scientific investment of those generating the data while creating necessary transparency for management. For example, the Scientific Committee of the International Whaling Commission developed a data availability protocol intended to allow access to primary data for analysis by multiple parties while ensuring that data owners preserve primary right of publication (Scientific Committee of the International Whaling Commission 2004). Submission of data sets to international data repositories, such as GenBank, Global Biodiversity Information Facility, DRYAD, or Ocean Biogeographic Information System should also be encouraged (Costello 2009; Kaye et al. 2009).

**Conclusions**

In all cases, we recommend a thorough presentation of the QC methods and results as discussed above that will provide scientists evaluating the results with the appropriate context to judge data quality and that can be used by managers to determine whether to request additional analyses to support policy decisions. Further, we would encourage journals that publish papers related to conservation and management decisions to establish guidelines for reporting QC data. This would have the effect of setting standards and formats for data reporting. One simple example is the reporting of genotyping error rates; relatively few papers publish error rates and those that do use a variety of methods to calculate the rates. A policy stating that the per-allele error rate should be used and how that should be calculated would be very useful for evaluating data quality among studies.

For population genetics studies where genetic monitoring is likely to play an increasing role, repeating and extending earlier work to assess changes in abundance or diversity will be much easier with these recommended measures and help to provide a more accurate baseline for comparison. These improvements in documenting QC and archiving of primary data, including genotypes, sequences, and sample records are critical to enhancing the credibility of conservation genetics and for reducing the potential conflict over management decisions based on this scientific advice.

**Funding**

NOAA Fisheries and by their respective institutions.

**Acknowledgments**

We thank Bill Perrin, Kim Parsons, Scott Baker, Michael Schwartz, and an anonymous reviewer for helpful comments on the manuscript and helpful discussion from the Scientific Committee of the International Whaling Commission.

**References**


genetics of bowhead whales (*Balaena mysticetus*): regional differentiation but no structure within the western Arctic. Arctic.


Taylor BL, Dizon AE. 1999. First policy then science: why a management unit based solely on genetic criteria can’t work. Mol Ecol. 8:S11–S16.


Accepted October 29, 2009

Received March 30, 2009; Revised October 11, 2009; Accepted October 29, 2009

Corresponding Editor: C. Scott Baker

10