Updated guidelines for DNA data quality control and error rate estimation, for genetic studies relevant to IWC management advice

R. Tiedemann, F. Cipriano, P. A. Morin, A. R. Hoelzel, P. Palsbøll, R. Waples, A. Natoli, L. Bachmann, L. Postma, M. Double, C. Pampoulie, H. Skaug, C. S. Baker, Jennifer Jackson

(First presented at IWC-SC60; updated by RT, FC, PM, RW, and JJ April/May 2012)

As genetic data are frequently applied to give advice to the IWC (including, but not limited to, detection of population structure), there is a need to agree on data quality criteria for currently used DNA marker types (mtDNA sequences, microsatellites, Single Nucleotide Polymorphisms (SNPs); nuclear DNA sequencing). The guidelines and considerations on DNA quality provided here represent common practice subject to ongoing discussion and will need future adaptation, as the *state-of-the-art* of DNA analysis in population genetics progresses.

It is also evident that, although compliance with these guidelines is highly desirable - this does not preclude consideration of genetic work failing to fully meet these recommendations. Nonetheless, the issues raised below are intended to assist IWC SC members in judging the reliability of information from genetic studies. In addition, for genetic studies explicitly carried out to provide stock definition and other advice relevant to management to the IWC, adherence to these guidelines is strongly recommended.

The quality of DNA data-based management critically depends on three issues (as summarized in Morin *et al.* 2010):

1) Experimental design (including appropriate sampling scheme with regard to sample size and geographic coverage)
2) Procedural implementation of sample handling and molecular analysis (including labeling, archiving, and data quality checks)
3) Appropriate data analysis and interpretation to provide management advice

Although proper consideration of all three issues above is highly recommended, these guidelines are restricted to explicit coverage of item 2, *i.e.*, the quality of DNA data. As such, this paper mainly deals with awareness, minimization, and control of DNA typing errors. Although assessment of genotyping error rates is more common in published studies now, it is still often neglected. Reported error rates between 0.5% and 1% are typical in many studies (see reviews by Bonin *et al.* 2004, Hoffman and Amos 2005, Pompanon *et al.* 2005) and higher rates are likely in studies involving DNA extracts of poor quantity or quality (Pompanon *et al.* 2005). Conceptually, the reliability of any estimate can be judged based on its proximity to the true value (accuracy) and on its repeatability (precision). Methods for quantifying error rates are provided by Pompanon *et al.* (2005). As true error rates are hard to estimate – requiring special procedures including repeated genotyping or other methods as outlined below - it is evident that most efforts to assess error rates are in fact identifying inconsistencies in data sets. Nevertheless, for simplicity we will use the term 'errors' to include inconsistencies in scoring and recording genotypes. Our objective is to provide a general procedural outline regarding how to qualitatively ensure and report DNA data quality and to provide some

general quantitative suggestions for benchmarks in quality control. Whenever possible, this document shall refer to established published procedures.

Issues related to experimental design and appropriate data analysis and interpretation are summarized in a separate document (Waples *et al.* 2012) that provides guidelines for marker choice, as well as testing the statistical power of molecular analyses to address particular questions.

Errors will be present in virtually every dataset, and thus it is important to be both aware of the most likely types of errors and to incorporate into every study methods for limiting, detecting, and correcting errors. Errors can be introduced at various points of a DNA study (Figure 1); the most important factors that contribute to errors will likely include mislabled samples, data entry errors, etc. – sometimes called "handling errors". In contrast, "systematic errors" are associated with the tendency for particular genetic markers and/or sample types to be susceptible to errors due to their inherent characteristics. In total, errors typically fall into these primary categories:

1) unreliable genotyping due to locus characteristics
2) Insufficient tissue or DNA sample quality
3) Inconsistency of methods, lack of adherence to standards of Good Laboratory Practise (GLP – Seiler 2005) that may exacerbate "handling" errors

Item 1 calls for **marker validation** (often addressed in a pilot study)**,** while Items 2 and 3 are addressed by implementing **systematic quality control and assessment procedures** throughout the entire study.

Marker validation

Microsatellites

Microsatellite data quality can be affected by repeat complexity, the number of alleles, the size range of alleles (Morin *et al.* 2009b), tendency of microsatellite PCR products to "stutter" (produce multiple peaks adjacent to the "true" peak, van Oosterhout *et al.* 2004) or be adenylated (also called "plus-A"), and variation in experimental conditions (Davison and Chiba 2003; LaHood *et al.* 2002). To validate a microsatellite locus, the characteristics of the repeat type need to be verified by DNA sequencing in the species to be analyzed. This is particularly important for the plausibility check on allele length during allele calling (see below). A pilot study should then investigate reliability of amplification and identify technical problems. These can include localization of adenylation peaks (extra adenine added to the end of amplicons by the Taq polymerase resulting in a product that is one base larger than predicted) or "stutter" (additional peaks adjacent to the "true" allele often in multiples of the repeat unit and probably related to errors in amplification through repeat regions), null alleles (failure to amplify one allele in heterozygotes), and frequency of allelic dropout (failure to recognize or amplify particular alleles, leading to a false estimation of heterozygote frequency) - see Goossens *et al.* 1998 and Tiedemann *et al.* 2004 for further descriptions). The pilot study should include all relevant sample populations and a sufficient sample size per putative population to permit a statistical test of Hardy Weinberg Equilibrium (HWE) expectations. A consistent deviation from HWE can be an indicator of such technical problems (in particular, if it occurs in particular loci across populations), although HWE departure can also have biological reasons (then, however, it is more likely to occur

within a population across loci). In addition, the genotypic data should be examined for patterns of linkage disequilibrium (LD; non-random associations of alleles at different gene loci).

Like departures from HWE, LD can result from a variety of biological factors as well as artifacts or errors. LD occurs due to genetic drift in all finite populations, and the magnitude of LD can be used to estimate population size. However, many analyses depend on the assumption that different loci are independent. Analysis of LD can identify locus pairs that are consistently out of equilibrium (linked), in which case this should be accounted for in subsequent treatment of the data (e.g., by dropping one of the loci from the analysis if independence is assumed). Both HWE and LD can be examined using a variety of software packages, e.g., GENEPOP (Raymond and Rousset 1995); FSTAT, etc.

It should be noted that using HWE departure for error detection may have an impact on later population genetic analyses and conclusions. For instances, if genetic markers are removed from the data set because they showed significant deviations from the expected HWE genotype frequencies, then later conformation with HWE is likely due to the selection of markers and not related to the underlying population genetic structure. Tests of HWE and LD often involve multiple tests of the same hypothesis. 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. Statistically, this test is designed for interdependent pairwise comparisons (=non-orthogonal study design). Hence, it is correctly applied for LD analysis among all pairs of loci, but not for repeatedly analyzing HWE for different loci. Moreover, this correction is known to be conservative and hence will fail to detect some departures from the null hypothesis. This problem is partially overcome by the sequential Bonferroni-method, but an even better option can be to use the false discovery rate (e.g., Garcia 2003), which adjusts for multiple testing without sacrificing as much power as the Bonferroni correction.

There are established routines to assess marker quality that can be used to decide whether markers should be included or excluded from analysis (e.g. Givens *et al.* 2007). If the marker appears unreliable at this stage, it should not be used. When preliminary analyses identify marker quality to be questionable but not obviously poor, analyses of data with and without that marker can help to determine whether a single marker is causing a particular result.

Mitochondrial DNA (mtDNA)

If using primers not validated in the species to be studied, the mitochondrial origin should be demonstrated. In particular, the possibility of erroneously sequencing nuclear pseudogenes (Numts; Lopez *et al.* 1994; Benssason *et al.* 2001) should be ruled out, as 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. Several methods have been described that can in most cases help to identify Numts (Bensasson *et al.* 2001; Dunshea *et al.* 2008; Lopez *et al.* 1994). After identification of Numts, primers should be re-designed such that they specifically amplify mtDNA (Tiedemann and von Kistowski

1998). Generally, sequences should be compared to GenBank (BLAST) and run through DNA surveillance routines, when available. Note – however – that GenBank itself lacks a stringent control of sequence authenticity, such that additional sequence validation might be necessary.

Single Nucleotide Polymorphisms (SNPs)

A variety of methods exist to examine Single Nucleotide Polymorphisms (SNPs), ***i.e.***, traditional Sanger-Sequencing, Next-Generation-Sequencing (NGS), analysis on standard or custom-made microarrays. Although SNP genotyping assays are in general very accurate and handling errors can be limited by appropriate automated data entry and transfer procedures, handling errors can still occur and systematic errors are sometimes found in a small percentage of "problem" SNP loci that exhibit unusually high error rates (Sheet and Stephens 2008). Two types of errors can occur, ***i.e.***, (1) an erroneous determination of the single nucleotide (genotyping error) or an erroneous comparison of non-orthologous loci. The former error rate depends on the typing technology: In Sanger sequencing, this error can be minimized by sequence quality checks (see below). The different NGS-platforms have their specific quality scores to evaluate sequencing reliability, most prominent the coverage (in the case of SNPs, the number of repeated sequence reads covering the same SNP). The reliability of microarray studies crucially depends on the source of sequence information used to create the microarray (ideally, sequence information for specimens of the target species encompassing the suspected most divergent target populations). SNP locus validation often relies on pilot studies and typically excludes individual SNPs with a high proportion of missing data or showing extreme deviations from Hardy-Weinberg equilibrium (Hosking *et al.* 2004); in studies with a high density of SNP loci an alternative approach using Linkage Disequilibrium analysis can detect problematic loci using correlation of alleles among nearby genetic loci and can reduce genotyping error rates by automatically correcting some genotyping errors (Sheet and Stephens 2008, implemented in the software package fastPHASE).

Systematic quality control and assessment

Assessing sample quality prior to genetic analysis

For many genetic studies, variation in sample quality (e.g., degraded samples from stranded animals, non-invasively collected samples such as faeces and sloughed skin, samples degraded from long-term storage or improper handling, co-purification of inhibitors, potential contaminants etc.) will be a factor. Many publications discuss methods to assure data accuracy for samples known to be of poor quality (e.g., McKelvey and Schwartz 2004; Morin *et al.* 2001; Navidi *et al.* 1992; Paetkau 2003; Taberlet *et al.* 1996) and the need to estimate error rates (Bonin *et al.* 2004; Broquet and Petit 2004; Morin *et al.* 2009a). Analysis of DNA sample quality prior to genetic data generation can ensure, for example, 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 likely to be of poor quality (e.g., non-invasive fecal samples, sloughed skin, poorly preserved and historical "ancient DNA" samples; McKelvey and Schwartz 2004; Morin *et al.* 2001; Morin and McCarthy 2007; Paetkau 2003; Taberlet *et al.* 1996). Indeed, the presence of even a single poor

quality sample in a small population sample can result in false inference of population structure (Morin and LeDuc 2004; Morin *et al.* 2009a).

If problems are detected with particular samples or if quality issues are expected, it is strongly recommend that samples be pre-screened for DNA concentration and quality (*i.e.*, degree of degradation, presence of inhibitors) prior to beginning a study with nuclear markers. Purification of DNA for PCR can co-purify PCR inhibitors (Hoelzel 1998) and this varies for different tissues (e.g., cetacean skin extracts may amplify better at lower concentrations due to these contaminants). When samples are expected to meet a minimum threshold level of DNA (e.g., 20 ng per PCR reaction), 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. When samples are expected to be of low quality or concentration, more sensitive methods such as quantitative PCR (qPCR) 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; Morin *et al.* 2007; Morin and McCarthy 2009a). When DNA concentration is low, potential for contamination is increased. When multiple pieces of sloughed skin are stored in the same vial, the potential for cross-contamination is also higher. When DNA is fragmented it is advisable to target smaller microsatellite or mitochondrial amplicons.

Duplicated analysis approaches for error detection

Duplicate analyses of the same sample has been used to detect genotyping, sample labeling, and handling errors (e.g. Taberlet *et al.* 1996, Bellemain *et al.* 2005, Paetkau 1993, Glover *et al.* SC/63/SD1), but this necessarily involves increased effort and expense (Schwartz *et al.* 2006). Some projects re-analyze a random subset of samples for error checking ("targeted replication", first applied by Palsboll *et al.* 1997), while others reanalyze every sample at every locus ("multi-tube analysis" - sometimes from separate tissue samples, e.g. Glover *et al.* 2011 SC/63/SD1) and then compare results.

Targeted replication involves a subset of the entire dataset and so is efficient in terms of time and cost, but simulations have suggested that this will be less efficient to detect and correct genotyping errors and associated biases when per locus error rates are >5% (Roon *et al.* 2006). The "multi-tube" approach is useful for both error checking and parentage analysis, since true biological offspring should share one allele at each bi-allelic locus with each parent (*e.g.* Skaug *et al.* 2004 SC/53/SD3, Pampoulie *et al.* 2011 SC/63/RMP1, Tiedemann *et al.* 2011). A novel alternative to targeted replication for estimation of error rates is comparison of genotypes from mother/fetus pairs - because mother and offspring necessarily must share at least one allele per locus such data provide an opportunity to estimate error rates (Haaland and Skaug 2007, SC/59/SD2).

Detection of multiple samples with exactly the same multi-locus genotype can indicate duplicate samples and should be investigated (e.g. Glover *et al.* 2011 SC/63/SD1). Uncorrected genotyping errors will result in a spike in the number of pairwise comparisons that match at "all but one" locus ("1MM-pairs") and error correction should not be considered complete until all such pairs are confirmed through carefully documented data replication at the locus or loci in question (Waits and Paetkau 2005).

It has been suggested that researchers should seek to decrease error rates to less than 1 percent through duplicate analysis error checking and corrections (Waits and Paetkau 2005), but a more pragmatic approach might instead use sensitivity testing to determine the error rate that is tolerable given the number of available samples, analytical objective, and appropriate effect size. If errors occur stochastically (i.e., not systematically), the larger the total number of independent loci analyzed the less of an affect a particular error rate will have on the ability to distinguish "signal" from "noise" . Some additional time and expense should be included in project budgets for error checking via replicated sample analysis, but whether this is targeted replication or complete duplication of all samples and all loci will depend on the overall sample size and other details of the particular sample set and research question.

Blind testing

A somewhat controversial variant of targeted duplicate analysis involves repeated testing of a subset of samples where the analysts are ignorant of the specimen identifiers until testing is complete, often called "blind testing" (the same objective can be accomplished by sending samples to an independent laboratory for reanalysis, but this will require careful inter-laboratory calibrations). This approach can improve attention to detail and decrease complacency, but can also erode morale and the attitude of research staff who may feel they are being subjected to unnecessary surveillance (Waits and Paetkau 2005).

Automated screening analysis for error detection

A third approach for error detection is to use computer algorithms to detect samples containing genotyping errors, and unlike the quantification and multi-tube approach, this may additionally allow detection of errors created when scoring gels or transcribing data (Schwartz *et al.* 2006). For example, the DCH test ("difference in capture history", implemented in the software package DROPOUT, Schwartz *et al.* 2006), is applicable for multi-locus genotyping such as microsatellite and SNP loci. The approach estimates the number of genetic markers required to eliminate "shadow effects" (lack of discrimination of individuals because of low variability or sampling too few markers, Evett and Weir 1998, Mills *et al.* 2000, Waits *et al.* 2001), then adds additional loci one at a time to determine if there is an increase in number of individuals generated by adding additional loci; rotating the loci so that every locus is added late in the process allows an efficient screen to determine which loci are erroneously adding new individuals (McKelvey and Schwartz 2004). Screening datasets with various computer algorithms (Paetkau 2003, McKelvey and Schwartz 2004, McKelvey and Schwartz 2005, Schwartz *et al.* 2006), allows for both error removal and statistical evidence that the data have few or no errors (Schwartz *et al.* 2006).

Correcting errors

Whenever potential errors are detected, it is necessary to re-run samples that might contain errors to clarify the source of discrepancies and correct them (Ewen *et al.* 2000, Miller *et al.* 2002, Valiere 2002, van Oosterhout *et al.* 2004, McKelvey and Schwartz 2005). Re-running samples should be considered a necessary part of the data analysis and error checking process, and also included in project budgets.

Ensuring consistent data generation

In order to limit the number of errors and provide clear and transparent documentation of experimental procedures, the following measures are recommended (Roman numbers as in Figure 1)

I. Sampling: Preferentially provide prelabelled (numbered) sample vials prefilled with appropriate storage buffer to the field worker. Provide explicit easy-to-read instructions for contamination-minimizing sampling. It is essential that each sample is uniquely identified. Methods for ensuring that samples are uniquely identified can include: Providing prelabeled (numbered) sample vials (barcoded, if possible), providing a pre-numbered data sheet against which sample numbers are checked off as vials are filled, etc. Double-label every vial with waterproof pen, do not use tape for labeling (might fall of later on). It is advisable to start with the vial with lowest number and strictly following numbers, such that they reflect order of sampling.

II. Sample handling: Establish standardized procedure for receipt of samples at the analytical laboratory. In particular, create database entry with field number and unambiguous lab number. Double check data entries to minimize transcription errors. It is advisable to have a backup whenever possible, so samples can be divided and sub-samples kept in separate storage locations (*i.e.*, when samples are shared between laboratories or before shipping samples from a remote location)

III. Laboratory Practice: Work according to established procedures for Good Laboratory Practice (GLP, e.g., Seiler 2005). Establish standardized routine to avoid mislabeling of tubes in the process of genotyping. Electrophoretic migration can be affected by both size and nucleotide composition of the alleles, as well as the addition of fluorescent molecules for visualization, although this is less of a problem when using modern capillary analyzers. 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 1.8-2.2bp; Amos *et al.* 2007). In addition, electrophoresis is itself variable and can cause allelic size differences of up to 7bp across time, technologies, and instruments (Davison and Chiba 2003; LaHood *et al.* 2002). Several methods have been introduced to facilitate normalization of alleles, but all require that controls are run to verify that alleles are correctly sized (Amos *et al.* 2007, Morin et al. 2009b). It is advisable to maintain all original data for reanalysis, and periodically check consistency of allele calling ("binning") for a subset of samples by double-blind genotype calling involving at least two persons. It is good practice, when inconsistencies are found or when starting to use new microsatellite primers (especially on a different species), to compare allele calling to absolute length information obtained by sequencing (part of marker validation, see above).

IV. Check data for consistency and plausibility. For microsatellites, use quality control software (e.g., MICROCHECKER, van Oosterhaut *et al.* 2004 and DROPOUT, McKelvey and Schwartz 2005) to check for null alleles and stutter/short allele dominance effects. Be aware that (1) such analysis packages do not necessarily find all potential errors and (2) non-rejection of the null hypotheses about non-existence of these effects can also originate from lack of statistical power; check HWE and, if heterozygote deficiency occurs, inspect data for rare allele homozygotes; check for plausibility of allele calls (referring to known repeat characteristics, see above; e.g.,

a tetranucleotide microsat should be expected to typically yield alleles differing by multiples of 4). Individual samples with unusual characteristics warrant extra scrutiny to verify genotypes, as these samples are both more likely to contain errors and more likely to bias analytical results. A simple analysis of the number or percentage of homozygous genotypes per individual can rapidly identify individuals likely to have experienced high levels of allelic dropout. Plotting the values indicates which samples are outliers from the general population, so that genotypes can be replicated to correct seemingly homozygous genotypes that are due to "allelic dropout" (failure to amplify one of the alleles in a heterozygote, usually the larger fragment). Another approach is to use Jackknife analysis of HWE, where samples are sequentially removed to detect samples and genotypes that have a strong effect on HWE (Morin et al. 2009a)

A similar approach can be used to evaluate the distribution of missing datapoints across individuals and markers. If data do not appear plausible after retyping, repeat entire typing starting with new DNA extraction from back-up sample, eventually sequence microsatellite in this specimen. For mtDNA sequences: sequence both strands (not required, but highly recommended), check quality of sequence with regard to ambiguous (mixed) bases, uneven spacing between bases; check sequence in BLAST for authenticity; check polymorphisms for plausibility (e.g., identify sequences which might show far more than expected polymorphisms and/or a bias towards a single nucleotide in several polymorphisms); if sequence is considered not plausible, re-sequence. If inconsistencies occur, re-sequence these specimens. From the entirety of unambiguously genotyped specimens, produce reference data set for which consistency **the laboratory/researcher of origin** holds primary responsibility, even though data are shared or submitted to central data bases. If microsatellite data from different laboratories are to be jointly analyzed, type a set of reference samples in both labs in order to synchronize allele calling (binning).

V. **Central databases hold responsibility for combined data sets**. In coordinated data acquisition efforts (e.g., as in BCB-bowhead whales), there should be a stringent time schedule for quality checks on composite data sets, implemented by two types of deadlines. The first deadline is for data submission. After that, a predefined period of quality control starts in which (1) the individual laboratory can still correct the submission and (2) the central database also performs plausibility checks on data consistency (along the lines mentioned under V.). If inconsistencies occur, they will be communicated to the laboratory of origin. If no consensus can be reached, this ambiguity will be reported in all occasions where the data are used. After the quality control period, data **must not** be changed, except for very specific reasons for which the laboratory of origin holds full responsibility.

VI. Data analysis: Manual file conversion should be avoided (because of copying error). Use automated routines for file conversion whenever possible. The program CREATE (Coombs *et al.* 2008) provides efficient conversion for a wide variety of analysis programs. In addition, consistent formatting of data through automated formatting from a database (as opposed to copy and paste in spreadsheets) can greatly reduce the introduction of errors due to data frame shifts (application of wrong samples to genotypes) or other data handling errors.

While validation is an important aspect of any protocol, routine error checking is equally important, since sample type and quality will vary between studies (Waits and Paetkau 2005), and may even vary over time and space within studies (Lucchini *et al.* 2002). In addition to the procedural guidelines below, error rates should be systematically estimated, including incorporation of replicated "blank" controls that can be used to compare genotypes generated throughout the data generation process. These controls serve several purposes:

1. Random sample replication to identify random and systematic errors. 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 both sample handling and raw data interpretation.
2. Control samples (2-3) replicated in every genotyping experiment (PCR and electrophoresis) serve to verify alleles and normalize sizes across time, laboratories and technologies.
3. Targeted replication of samples after the majority of data are generated will allow verification of data quality and can also detect sample handling errors (e.g., reversal of a sample plate). This should involve some samples from every sample group run together, and result in ≥10% replication of the data set.

Although it is not practical to detect and correct every error by the measures suggested above, 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 Hardy-Weinberg equilibrium, resulting in false inference of population structure (Morin et al. 2009a). Jackknife analysis of genotypic data (repeated analysis with the removal of one sample at a time) can reveal which samples have the greatest effect on HWE, so that they can be re-checked to verify the genotypes (Morin *et al.* 2009a).

Based on error rates that are commonly reported in the literature, it is recommended that analysts routinely take steps to decrease the overall error rate to around 1% for microsatellite data used in population studies and less than 1% for studies using SNPs (Bonin *et al.* 2004; Broquet and Petit 2004; Morin *et al.* 2009a). For parentage and genetic mark-recapture studies error rates should be even lower to reduce the number of false positives (Bonin *et al.* 2004, Hoffman and Amos 2005, Waits *et al.* 2001). In all cases, researchers should report the genotype error rates detected in the course of these quality checks (ideally both locus-specific and overall error rates) along with publication of the genetic analysis. If higher rates are estimated in particular studies of importance to IWC management decisions, the burden of proof should be on the investigators to demonstrate that the results, interpretations and conclusions they reach are still appropriate given the level of uncertainty involved.

References

Amos W, Hoffman JI, Frodsham A, *et al.* (2007) Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes* **7**, 10-14.

Bellemain E, Swenson J, Tallmon D, Brunberg S, Taberlet P (2005) Estimating population size of elusive animals using DNA from hunter-collected feces: comparing four methods for brown bears. *Conservation Biology* **19**, 150–161.

Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution* **16**, 314-321.

Bonin A, Bellemain E, Bronken Eidesen P, Brochmann C, Taberlet P. (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**, 3261-3273.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology* **13**, 3601-3608.

Coombs J.A., Letcher B.H., Nislow K.H. (2008) CREATE: A software to create input files from diploid genotypic data for 52 genetic software programs. Molecular Ecology Resources 8(3):578-580.

Davison A, Chiba S (2003) Laboratory temperature variation is a previously unrecognized source of genotyping error during capillary electrophoresis. *Molecular Ecology Notes* **3**, 321-323.

Dunshea G, Barros NB, Wells RS*, et al.* (2008) Pseudogenes and DNA-based diet analyses: a cautionary tale from a relatively well sampled predator-prey system. *Bull Entomol Res*, 1-10.

Evett, IW, Weir BS (1998) Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer, Sunderland, Massachusetts, USA.

Ewen .R, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW, Foote. SJ (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics* 67, 727-736.

Garcia LV (2003) Controlling the false discovery rate in ecological research. *Trends in Ecology and Evolution* 18, 553-554.

Givens GH, Huebinger RM, Bickham JW, George JC, Suydam R (2007) Patterns of genetic differentiation in Bowhead whales (*Balaena mysticetus*) from the western Arctic. IWC SC/59/BRG14, presented to the IWC Scientific Committee, June 2007.

Glover KA, Haug T, Øien N, Walløe L, Lindblom L, Seliussen BB, Skaug HJ (2011) The Norwegian minke whale DNA register: a fully operational database monitoring commercial harvest and trade of whale products. IWC/SC/SD1, presented to the IWC Scientific Committee, June 2011.

Goossens B, Waits LP, Taberlet P (1998) Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. *Molecular Ecology* **7** , 1237–1241.

Haaland O, Skaug HJ (2007) Preliminary estimates of error rates in the Norwegian minke whale DNA register based on mother-fetus pairs. IWC SC/59/SD2, presented to the IWC Scientific Committee, June 2007.

Hoelzel AR (1998) Molecular genetic analysis of populations: a practical approach. Oxford University Press, Oxford.

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

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, *et al.* (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet* 12: 395–399.

LaHood ES, Moran P, Olsen J, Grant WS, Park LK (2002) Microsatellite allele ladders in two species of Pacific salmon: preparation and field-test results. *Molecular Ecology Notes* **2**, 187-190.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol* **39**, 174-190.

Lucchini V, Fabbri E, Marucco F, Ricci S, Boitani L, Randi E (2002) Noninvasive molecular tracking of colonizing wolf (Canis lupus) packs in the western Italian Alps. *Molecular Ecology* **11**, 857–868.

McKelvey KS, Schwartz MK (2004) Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. *Journal of Wildlife Management* **68**, 439-448.

McKelvey KS, Schwartz MK (2005) DROPOUT: a program to problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. *Molecular Ecology Notes* **5**, 716-718.

Miller C., Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics* **160**, 357-366.

Mills LS, Citta JJ, Lair KP, Schwartz MK, Talmon DA (2000) Estimating animal abundance using non-nvasive DNA sampling: promise and pitfalls. *Ecological Applications* **10**, 283-294.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative PCR analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes*). *Molecular Ecology* **10**, 1835-1844.

Morin PA, LeDuc RG (2004) Analysis of bowhead DNA quantity and microsatellite characteristics: implications for potential biases in population structure analysis. SC/56/BRG34. presented to the IWC Scientific Committee, June 2004.

Morin P.A., LeDuc R.G., Archer F.I., Martien K.K., Huebinger R., Bickham J.W. and Taylor B.L. 2009a. Significant deviations from Hardy-Weinberg equilibrium caused by low levels of microsatellite genotyping errors. Mol Ecol Res 9:498-504.

Morin P.A., Manaster C., Mesnick S.L. and Holland R. 2009b. Normalization and binning of historical and multi-source microsatellite data: overcoming the problems of allele size shift with ALLELOGRAM. Mol Ecol Res 9:1451-1455.

Morin PA, Martien KA, Archer FI, Cipriano F, Steel D, Jackson J, Taylor BL (2010) Applied conservation genetics and the need for quality control and reporting of genetic data used in fisheries and wildlife management. *Journal of Heredity* 101,1-10. doi: 10.1093/jhered/esp107

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes* **7**, 937-946.

Navidi W, Arnheim N, Waterman MS (1992) A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. *American Journal of Human Genetics* **50**, 347-359.

Paetkau D (2003) An empirical exploration of data quality in DNA-based population inventories. *Molecular Ecology* **12**, 1375-1387.

Palsboll, PJ Allen J, Berube M, Clapham PJ, Feddersen TP, Hammond PS, Hudson RR, Jorgensen H, Katona S, Larsen AH, Larsen F, Lien J, Mattila DK, Sigurjonsson J, Sears R, Smith T, Sponer R, Stevick P, Oien N (1997) Genetic tagging of humpback whales. *Nature* 388:767-769.

Pompanon, F., Bonin A, Bellemain E, Taberlet P. (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* 6, 847-856.

Pompoulie, C. Elvarsson B, Olafsdottir G, Hauksdottir S, Skirnisdottir S, Olafsson K, Magnusdottir S, Chosson V, Halldorsson SD, Olafsdottir D, Gunnlaugsson T, Danielsdottir AK, Vikingsson GA (2011) Mother-fetus pair analyzes in the Atlantic fin whale *Balaenoptera physalus*: an insight into the migration pattern of the species? IWC SC/63/RMP1 presented to the IWC Scientific Committee, June 2011.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Roon DA, Waits L P, Kendall K C. (2006) A simulation test of the effectiveness of several methods for error-checking non-invasive genetic data. *Animal Conservation*, 203-215.

Schwartz MK, Cushman SA, McKelvey KS, Hayden J, Engjker C (2006) Detecting genotyping errors and describing American black bear movement in northern Idaho. *Ursus* **17**, 138-148.

Seiler JP (2005) Good Laboratory Practice: the Why and the How, 2nd edition. Springer, Berlin, Heidelberg, New York.

Sheet P, Stephens M (2008) Linkage disequilibrium-based quality control for large-scale genetic studies. PLoS Genet 4(8): e1000147. doi:10.1371/journal.pgen.1000147

Skaug, H.J., Øien, N. (2004) Genetic tagging of males in north Atlantic minke whales through comparison of mother and fetus DNA-profiles. IWC SC/56/SD3 presented to the IWC Scientific Committee, June 2004.

Taberlet P, Griffin S, Goossens B, *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.

Tiedemann R, Paulus KB, Scheer M, von Kistowski KG, Skírnisson K, Bloch D, Dam M (2004) Mitochondrial DNA and microsatellite variation in the Eider duck (*Somateria mollissima*) indicate stepwise postglacial colonization of Europe and limited current long-distance dispersal. *Molecular Ecology* **13**, 1481-1494.

Tiedemann R, von Kistowski KG (1998) Novel primers for the mitochondrial control region and its homologous nuclear pseudogene in the Eider duck *Somateria mollissima*. *Animal Genetics* **29**, 468.

Tiedemann R, Paulus KB, Havenstein K, Thorstensen S, Petersen A, Lyngs P, Milinkovitch MC (2011) Alien eggs in duck nests: brood parasitism or a help from Grandma? *Molecular Ecology* 20: 3237-3250.

Valiere N (2002) Gimlet: a computer program for analysing genetic individual identification data. *Molecular Ecology Notes* **2**, 377-379.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology* **10**, 249-256.

Waits, L.P. and D. Paetkau. (2005) Noninvasive genetic sampling tools for wildlife biologists: A review of appications and recommendations for accurate data collection. *The Journal of Wildlife Management* 69:1419-1433.

Waples RS, Hoelzel AR, Tiedemann R, Palsböll P, Gaggiotti O, Jackson JA, Bickham JW. Guidelines for the analysis of population genetic data used in an IWC management context. IWC SC/64/SDX presented to the IWC Scientific Committee, July 2012.

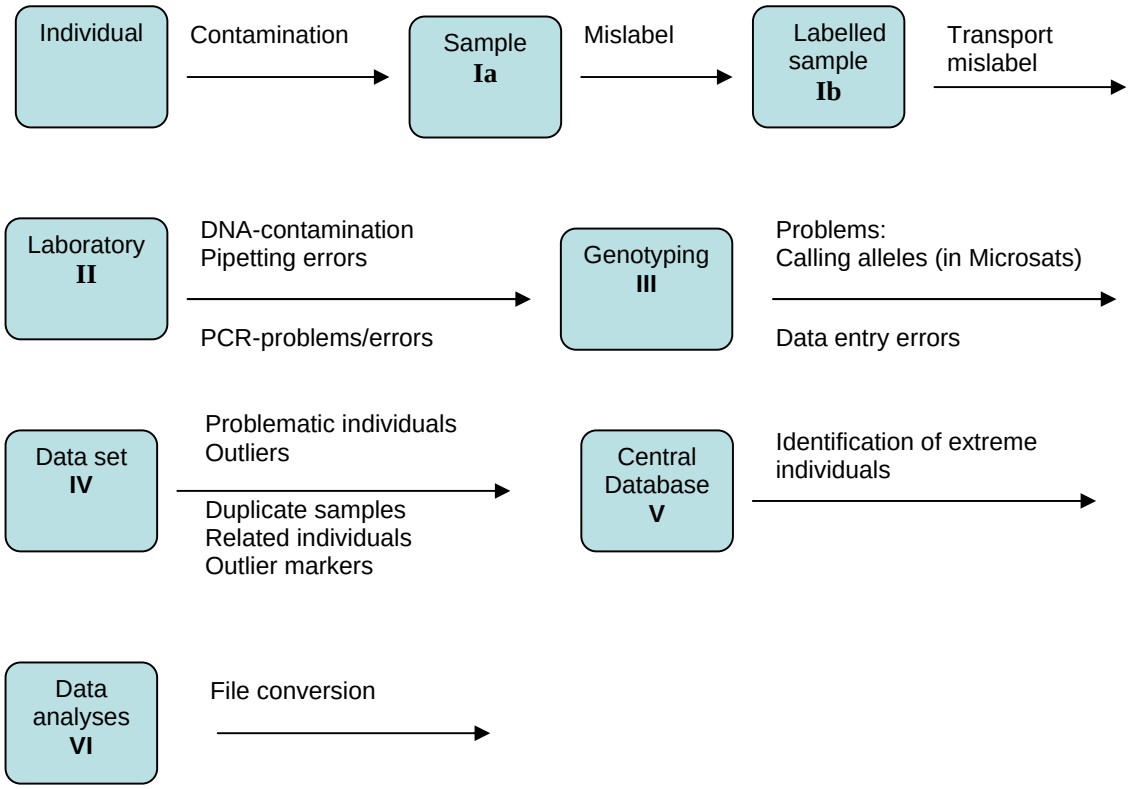

Figure 1. Flow chart of DNA analysis procedures and potential error sources. Roman numbers refer to suggestions for quality control below.