

Annex I

Working Group on Stock Definition

Members: Bravington (Convenor), Acquarone, Alfaro Shigueto, Bachmann, Baker, Bickham, Buchan, Butterworth, Castro, Cerchio, Charrassin, Chilvers, Cipriano, Collins, Donovan, Double, Flores, Fuentes, Garrigue, Givens, Goto, Gregovich, Henriette Payet, Hoelzel, Ilyashenko, Jackson, Jeremie, Kanda, Kaufman, Kitakado, Lyrholm, Miller, Natoli, Nery, Olavarria, Palsbøll, Pampoulie, Park, Pastene, Perez, Perrin, Polacheck, Postma, Punt, Ray, Rosenbaum, Rowntree, Sanino, Skaug, Tala, Tiedemann, Torres-Florez, Urban, Verborgh, Walløe, Waples, Wiig, Yáñez, Young.

1. INTRODUCTORY ITEMS

1.1 Election of chair and appointment of rapporteurs

Bravington welcomed participants and was elected as Chair. Tiedemann and Bravington served as rapporteurs.

1.2 Adoption of Agenda

The Agenda adopted is given as Appendix 1.

1.3 Review of documents

Documents considered were SC/60/SD1, 2, 3, 4, and 5.

2. DNA DATA QUALITY

In recent years, the Scientific Committee has engaged in several in-depth discussions centred on the genetic data that form part of the delineation of stock structure hypotheses, for example in the bowhead whale *Implementation Review*. The Committee's experiences have underlined that a clear understanding of the reliability of each genetic dataset is essential for correct interpretation in terms of stock structure¹, and have re-emphasised the importance of developing suitable quality protocols for genetic data used in providing management advice. There are associated issues in terms of the Data Availability Agreement, and these would be greatly aided by having an initial protocol for the use of genetic data that includes both guidelines and suggestions for minimum standards. At the 59th Annual Meeting, the SDWG developed an initial proposal for such a protocol (JCRM 10 Suppl., Annex I, Appendix 2), which has already made itself felt in the Scientific Committee's discussions during SC 60. This agenda item presents an update of these discussions.

One paper on this specific topic was presented. SC/60/SD1 lays out a template for quality control and quality analysis (QC/QA) during the process of data collection for genetic analysis. The paper outlined the major steps commonly applied to mitochondrial DNA sequences and microsatellite genotypes, including considerations for QC/QA at each step of the data collection process, including: (1) assessment of sample quality prior to genetic analysis, (2) data generation and use of control samples, (3) assessment of sample-specific data quality, (4) assessment of data set quality, and (5) reporting of data.

Drawing both on last year's discussions and on SC/60/SD1, a subgroup under Tiedemann developed an updated set of guidelines on DNA data quality control for genetic studies relevant to IWC management advice. The full text may be found in Annex I Appendix 2, and a brief summary is given in the next three paragraphs. The guidelines address commonly-used DNA marker types (sequences, microsatellites, mitochondrial DNA, Single Nucleotide Polymorphisms (SNPs); possibly nuclear DNA sequencing in the future). The guidelines and considerations on DNA quality 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.

The subgroup identified that the quality of DNA data-based management critically depends on three issues:

- (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 labelling, archiving, and data quality checks); and
- (3) appropriate data analysis and interpretation to provide management advice.

The subgroup recommended that the Scientific Committee should develop guidelines for all three issues in due course. At present, though, the guidelines in Appendix 2 cover only item 2, i.e. the quality of DNA data. As such, the Appendix deals with awareness, minimisation, and control of DNA typing errors². The subgroup was able to provide a general procedural outline regarding how to qualitatively ensure and report DNA data quality, but not (at this stage) to provide quantitative suggestions for benchmarks in quality control. An extensive, and extensible, list of references to published practice is provided in the Appendix.

The subgroup noted that errors can be introduced at various points of a DNA study, and fall into three different primary categories:

- (1) difficulties in reliable 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)

Item 1 calls for marker validation (often addressed in a pilot study), while Items 2 and 3 are addressed by implementing a systematic quality control throughout the entire study. Detailed guidelines on these two separate protocols may be found in the Appendix.

¹ Genetic data are also used for other management purposes besides stock structure inference, such as abundance estimation and species ID. Discussions in SDWG have, understandably, focussed on stock structure applications, but most of the fundamental data quality issues also apply more generally.

² True error rates in genetic studies are hard to estimate, and it is evident that most efforts to assess error rates are in fact identifying inconsistencies in data sets. Nevertheless, for simplicity, the term 'error' here implies inconsistencies both in scoring and in recording genotypes.

The SDWG thanked Tiedemann, Morin, the other authors of SC/60/SD1, and the other working group members for their efforts, and endorsed the conclusions of Annex I Appendix 2. The material in the Appendix should be seen as a 'living document' that will be updated by the SC to keep up with the rapid progress in genetic techniques. Although accordance to the guidelines in Appendix 2 is highly desirable, this does not preclude consideration of genetic work failing to fully meet these standards (though all studies should endeavour at least to report on whether the guidelines in Appendix 2 have been met). The guidelines, if met, should assist IWC SC members in judging the respective reliability of information from genetic studies. In addition, for studies explicitly carried out to give stock definition advice to the IWC, adherence to the guidelines is strongly recommended.

The SDWG recalled its past intention to provide actual numerical guidelines for measure of data quality. There was insufficient time at this year's meeting to address this aspect of guidelines, but the DNA Quality subgroup agreed to consider it next year, along with reviewing and updating its qualitative guidelines. To fuel the discussion of quantitative aspects/thresholds of adequacy associated with data quality, Tiedemann will lead the group intersessionally in conducting a literature review of the range of error rates and other quality measures associated with studies published in peer reviewed journals. This shall be presented as a regular paper to SC 61.

Aside from the specific data quality issues in Appendix 2, a number of areas were identified where further guidelines might usefully be developed next year and beyond. In particular, it should be possible to provide guidelines for some of the more common types of statistical analyses of genetic data that are employed in IWC management contexts. There are two aspects to this: comments on general statistical usage (e.g. multiple comparison tests in the context of stock structure), plus more general summaries on the appropriate domains of application of different stock structure tools such as STRUCTURE, BayesAss, etc. The SDWG recalled its efforts at the latter during the 2nd TOSSM workshop (IWC, 2007) and decided that an update would be appropriate. An Intersessional Email group convened by Waples was established to progress both aspects, with a view to producing a paper for consideration at SC 61.

3. TESTING OF SPATIAL STRUCTURE METHODS

3.1 Progress on the TOSSM simulation-testing framework

Paper SC/60/SD2 describes the intersessional developments to the TOSSM package, including implementation of fully-spatial sampling and harvest (using polygons of arbitrary shape), a plotting function, and an alternative catch-limit algorithm. A simple example of package use for a very simple BSA is included, with a complete description of the parameters used in simulations.

In discussion, the SDWG thanked the Gregovich and Martien for their efforts on TOSSM intersessionally, as summarised in SD 2, and welcomed the news that a version of the paper is in preparation for submission to peer-reviewed literature. Discussion of future plans is summarised below under Work plan.

Paper SC/60/SD3 tests the CLA under a range of population structures, to better understand the range of situations under which the CLA fails to protect substocks. Parameters varied across trials were the relative carrying capacities (K) of the populations, the dispersal rate among them, maximum sustainable yield rate ($MSYR_{1+}$), and the precision in simulated abundance estimates. All of these parameters had strong effects on population persistence under the CLA. Trials with a low $MSYR_{1+}$ (1%) generally ended with the abundance of population 1 below 0.54K, regardless of the dispersal rate or relative carrying capacities of the two populations. The same was true of trials in which the carrying capacity of population 1 represented only 20% or less of the total landscape carrying capacity, even when dispersal between populations was high ($5 \times 10^{-3}/yr$) and $MSYR_{1+}$ was 4%. These results highlight the value of spatially diffuse harvest that avoids potential overharvest of unrecognised stocks. The results also underline the need for powerful genetic methods as a tool in management, as populations connected even by dispersal rates as high as 5×10^{-3} are shown here to be vulnerable to overharvest if not managed separately.

In discussion, the SDWG welcomed the results in SD3, which address a basic question long pondered by the Committee: when is separate-stock management actually required for avoidance of local depletion? Of course, the results are very specific to the scenario tested, but the paper shows that people outside "core TOSSM" could now re-run the analyses with different stock and harvest structures, if desired. One technical problem was noted with the TOSSM implementation of the CLA, in cases where the CV on abundance estimates is very low. This, and other details of appropriate CLA-related defaults, will be discussed and incorporated into TOSSM intersessionally.

With the addition to TOSSM of finer-scale spatial structure on the sampling/harvest grounds, it was also noted that results of some stock structure methods will be sensitive to the precise way that spatial distribution on the sampling/harvest grounds is simulated. For example: do animals from a given breeding population go to a spot in the sampling grounds that is chosen freshly and independently each year? and do newly-migrated animals (from one breeding population to another) take up the same spatial distribution on the sampling/harvest grounds as unmigrated animals?³ Much of this is embodied in principle in the general description of the TOSSM Archetype that is being used, but the details may require further specification, and also need to be described in all papers which use a particular scenario.

As a general comment on all papers reporting results from TOSSM tests (including SD3, 4, and 5 this year), it was noted that presentation could be assisted by including some standard background material on TOSSM itself and on the parameters of the particular simulated datasets in question. This material would summarise not just IWC-specific parameters, but also parameters more familiar to the population genetics community, such as number and "power" of loci, migration rates expressed in terms of m/N_e per generation, and average F_{ST} in cases where population structure really does exist. All this could be facilitated by (i) developing some boilerplate paragraphs about TOSSM in general, and about specific archetypes and catch-control rules, that could simply be downloaded from the TOSSM website (<http://swfsc.noaa.gov/tossm.aspx>) and incorporated into TOSSM-related papers for IWC and beyond, and (ii) perhaps adding some simple modules to the TOSSM program itself, to automatically calculate not just performance statistics *sensu* IWC but also other "metadata" statistics such as F_{ST} . The TOSSM Steering Committee (Q18) will develop these ideas further intersessionally.

3.2 Development and testing of statistical methods

SC/60/SD4 tests the comparative performance of three genetic analytical methods– Wombling, the Monmonier algorithm, and the Waples/Gaggiotti algorithm. Performance of these methods was evaluated with respect to how well they detect population genetic structure and use this information to

³ The IWC parlance for these questions is "mixing". In the particular case of the scenarios tested in papers SD3, 4, and 5 this year, the probability distribution of each animal's location on the sampling/feeding grounds each year depends only on the animal's breeding population, not on where it was last year or whether it has migrated away from its natal breeding population. This differs in the different TOSSM archetypes.

construct appropriate management units for use in the IWC's Revised Management Procedure (RMP). Trials of each genetic method were performed in a simulated management setting across a range of population structure scenarios. The scenarios varied with respect to the number and sizes of populations and the annual rate of dispersal among them. The methods generally detected populations and managed them appropriately when the annual dispersal rate was low (5×10^{-6} /yr). At intermediate dispersal rates (5×10^{-5} and 5×10^{-4}), there was a large difference in the performance of the methods, with the Monmonier algorithm and Waples methods performing very well and Wombling performing poorly. None of the methods was able to detect population structure when the annual dispersal rate was 5×10^{-3} . Consequently, populations were frequently over-harvested in these trials. Nonetheless, our results indicate that the Monmonier algorithm and the Waples/Gaggiotti algorithm may prove to be useful tools for defining management units for use with the RMP. Further testing to fully characterise the performance of these methods is required before final conclusions can be drawn.

The SDWG noted that the results of SD4 concerning the Monmonier and W/G methods were more encouraging than for any stock structure method previously tested under TOSSM. It was agreed that an up-to-date summary of results should be developed and maintained before SC 61. The highest dispersal rate tested, 5×10^{-3} animals per year, is equivalent to several tens of migrants per generation given the population size (7500 adults), and thus the sub-populations are nowhere near separate in *genetic* terms; hence, when the genetic methods fail to detect structure at the highest dispersal rates, they are in fact making the "correct" decision in terms of their original design purpose. Nevertheless, as shown in both SD4 and SD3, this dispersal rate is low enough to be important for IWC management, and in this range the genetic methods are nevertheless not successful in *management* terms, at least so far. Although the methods will inevitably do better with larger sample sizes and numbers of loci, these parameters are already fairly generous in the SD4 simulations (600 animals and 30 microsatellite loci), so the practical scope for improvement in that direction may be limited.

The tests in SD4 did not attempt to tune the various parameters in the underlying stock structure methods (e.g. p -value threshold used to aggregate units), so there may be scope for increasing the upper limit of dispersal rate that the methods are able to detect reliably. However, doing this will of course increase the chance of false positives (i.e. decision to use separate management when there truly is panmixia), which may be important for management.

The SDWG recalled its previous discussions of Performance Measures (JCRM Vol 6 Suppl., p476-478) which include false-positive rate and a variety of other measures. As TOSSM moves towards investigation of scenarios closer to actual management applications, it will be increasingly important for authors to include a full set of performance measures, and the TOSSM program itself must of course be set up to make it easy for developers to extract & report the appropriate statistics.

In comparing the methods tested in SD4, it was noted that the aggregative approaches such as Waples/Gaggiotti (and Boundary Rank, discussed in numerous previous SC reports) are potentially sensitive to the size of initial units, and in particular to whether those initial units are each truly drawn from a single population. This restriction could prove limiting in settings where mixing occurs on the sampling grounds.

In paper SD5, a new boundary setting algorithm (BSA) based on genetically based close-kin information was presented. In essence, the method compares the within- and between degree of relatedness for two sampling sites, to test whether individuals from the two sites come from the same breeding population. The performance of this two-site BSA was evaluated on TOSSM data for different levels of dispersal and sample size. A clustering algorithm was also presented aimed at the situation with more than two sampling sites, but it was noted that this work had a preliminary status. The criteria for determining relatedness was targeted at parent-offspring relationships, and future work will investigate other types of relatedness.

In discussion of SD5, the SDWG remarked upon the impressive performance and potential of an approach that is still at an early stage of development. Unlike any of the population genetic methods, there is no requirement that the breeding populations be genetically distinct, so that neither high rates of dispersal nor historically recent separation are problematic; a corollary, though, is that the method does not provide information about gene flow. A number of issues were identified that could require attention in different scenarios: these included the potential sensitivity to persistent family structure, the need to relax the strict-exclusion criterion (called L_1 in the paper) if scoring error is a risk and large numbers of loci are involved, the potential utility of including higher-order relatedness in the model, and the requirement of the method for a large sample of animals when population size is large (so that enough closely-related pairs are actually found). Palsbøll reported that a similar approach has just been published in *Ecology*, to investigate movement rates in a marbled murrelet population with source-sink dynamics. The SDWG encouraged further work on close-kin methods for stock structure.

In order to share out the substantial workload of testing population methods within TOSSM—which is something that the methods are rarely designed to do from scratch—TOSSM has invited developers to act as Champions for particular methods that appear to be worth testing. At the 59th AM, Jackson had agreed to act as Champion for the LaMarc program, and this year the SDWG heard a summary of intersessional progress thereupon. LaMarc has several powerful features, but its principal appeal compared to other stock structure methods tested in TOSSM lies in its ability to estimate migration rates. However, because Lamarc requires a pre-specification of population structure (rather than deciding itself whether and where such structure is present), it is necessary to link it with a "hypothesis generator front-end", for example a sequential hypothesis test, in order to incorporate it into TOSSM. This is far from trivial, and in addition Lamarc is very slow to run. With core TOSSM assistance, it was possible intersessionally to run about 60 simulated datasets under Archetype II (2 stocks, limited migration, no mixing on sampling grounds) through Lamarc, but further checking is required and the process may prove too unwieldy.

In terms of other stock structure methods that might be investigated through TOSSM, it was noted that a faster version of the program IM is now available (IMA), and that work on a three-population version is in progress.

3.3 The state of the TOSSM project: summary

Progress with TOSSM has leapt forward this year, in three respects. First, the simulations in SD3 have, for one particular stock structure scenario, identified *quantitatively* the threshold of dispersal rate at which local depletion becomes a real risk under single-stock RMP management. Second, the results in SD4 are considerably more encouraging than those presented for other stock structure methods in previous years, in that the methods in SD4 appear capable of successfully identifying the presence of population structure at dispersal rates rather higher than normally considered by population geneticists. Taken in conjunction, the two papers identify the "danger zone" between those dispersal rates that are so high that single-stock management is satisfactory, and those low enough for stock structure methods to reliably detect the need for separate-stock management. This danger zone is about one order of magnitude in SD3 and SD4, and could probably be reduced somewhat by adjusting the methods. While the qualitative notion of such a danger zone is very familiar, this is the first time that it has been quantified within TOSSM—and it appears to be smaller than previous years' results

might have suggested. The third leap is the presentation of a stock structure method based on completely different principles (SD5), which actually seems capable of bridging the danger zone. Again, it must be emphasised that all these quantitative results are provisional and specific to the precise scenarios and methods tested, but the point is that they now can be obtained other specific scenarios, too.

With these results in mind—and not forgetting that they still need to be explored in more general population structure scenarios than the very simple two-stock no-mixing Archetype II considered to date—it is timely to turn some attention onto the long-awaited Phase II of TOSSM, in which more specifically realistic—and likely more challenging—scenarios are developed and tested. Donovan and Punt agreed to review the Scientific Committee's current areas of enquiry, and to liaise with themselves and the other members of the TOSSM Steering Committee to prioritise which scenarios to implement.

Aside from the high-priority issue of implementing these more specific scenarios, a number of key issues were proposed during SC 60 for development within the near future of TOSSM, as discussed below. Some of these require core TOSSM work, whereas others might be addressed by individual methods developers. Because of the absence of the keystone TOSSM member (Martien) this year, it was not possible to finalise all plans in detail or completely prioritise intersessional developments during SC 60. However, the TOSSM Steering Committee will discuss and refine these ideas intersessionally. There are four general areas: Presentation; Archetypes requiring retuning of genetic parameters; Scenario development not requiring genetic retuning; adjustments to the Genetics; development of Methods.

Presentation

Provide standard text for TOSSM generally, for specific population archetypes, and for specific harvesting/sampling/genetic scenarios. Add routines to the TOSSM program that will automatically report population genetic summary statistics such as allelic diversity. Revisit previous discussions of Performance Statistics, and implement code in TOSSM to automatically report appropriate statistics.

Archetypes

Development of new breeding-biology archetypes is very time-consuming, because of the need to tightly define the population dynamics and to retune the genetic models (e.g. for mutation rates) to produce plausible parameter ranges (e.g. for allelic diversity). Of the original set of 5 proposed for TOSSM, only Archetype III (spatial autocorrelation/clinal structure) has not yet been implemented.

Priority will be given to specific archetypes identified by the Steering Committee as of most relevance to current SC priorities.

Scenario development not requiring genetic retuning

Reconsider the implementation of dispersal rates in multi-stock models, to ensure consistency with unequal equilibrium population sizes.

Minor changes to the RMP-CLA formulation in TOSSM.

The improved spatial flexibility in new TOSSM allows the possibility to specify a much wider range of mixing, sampling, and harvest schemes, without the major workload required to develop entire new archetypes. A range of plausible default scenarios should be available within TOSSM. Beyond that, the flexible nature of TOSSM means that this should be within the ambit of individual users, provided the code is sufficient examples are provided. Ease-of-use enhancements to allow users to vary the population dynamics would also enhance uptake, though there may be implications for genetic retuning. Priority will be given to specific archetypes identified by the Steering Committee as of most relevance to current SC priorities.

Genetic models

Highest priority is to define and introduce realistic models for scoring error.

3.4 Overall plans for progress

It is particularly encouraging to see that a number of different developers (some new to TOSSM) have been able to get stock structure methods running inside TOSSM. It was also noted that TOSSM is being used as a test bed for non-IWC genetic simulations; apparently it is much easier to implement genetic population dynamics appropriate to animals than the RMETASIM package (originally for plants) which underlies TOSSM. From the outset, one of the goals of TOSSM was to reach outside the IWC community into the wider world of population genetics, where similar management questions are often of great importance; it seems that TOSSM is starting to achieve this. This impressive level of uptake is undoubtedly due to the major improvements intersessionally in developing and documenting the TOSSM code. These improvements have been made possible by the employment of a dedicated programmer (Gregovich), whose position at SWFSC has over the last 2 years been part-funded by the US Govt. and part-funded (50%) by the Scientific Committee. A further request for 50% support during 2008/9 was received this year, in order to continue the development tasks identified above (Appendix 3). The subcommittee strongly recommended that the IWC provide support for this venture. By the 2009 meeting, all the major developments to the TOSSM code should have reached stability, so that further modifications can be made by developers outside "core TOSSM". It is therefore not anticipated that IWC funding support will be required after the 2009 SC meeting.

4. Work plan and agenda for SC/61

There will be three general intersessional tasks, aside from work by individual developers and champions of stock structure methods, steering committees, and

DNA data quality (item 2): Tiedemann to lead preparation of review paper on numerical guidelines for discussion at SC 61.

Development of guidelines for stock structure analysis (item 2): Intersessional working group convened by Waples to develop paper for discussion at SC 61; terms of reference in Annex Q.

Development of the TOSSM code, documentation, and core set of simulations (items 3.3 and 3.4): code maintenance and development of new simulated datasets.

The proposed agenda for IWC 61 is similar to this year's:

1. Statistical & genetic issues relating to stock definition (including further discussion of DNA data quality, and guidelines for appropriate analysis);
2. Progress on TOSSM; and
3. Discussion of possible criteria for unit-to-convert.

Appendix 1

AGENDA

1. Convenor's opening remarks
2. Election of Chair and appointment of rapporteurs
3. Adoption of Agenda
4. Review of documents
5. Statistical and genetic issues relating to stock definition
 - 5.1 DNA Data Quality
 - 5.2 Statistical methods
6. TOSSM (Testing of Spatial Structure Models)
 - 6.1 Update on progress
 - 6.2 Directions for further work
7. Overall work plan and adoption of report

Appendix 2

GUIDELINES FOR DNA DATA QUALITY CONTROL FOR GENETIC STUDIES RELEVANT TO IWC MANAGEMENT ADVICE

Members: 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

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 (sequences, microsatellites, Single Nucleotide Polymorphisms (SNPs); possibly nuclear DNA sequencing in the future). 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 accordance to these guidelines is highly desirable - this does not preclude consideration of genetic work failing to fully meet these standards. Nonetheless, the issues raised below are intended to assist IWC SC members in judging the respective reliability of information from genetic studies. In addition, for studies explicitly carried out to give stock definition advice to the IWC, adherence to these guidelines is strongly recommended.

It was identified that the quality of DNA data-based management critically depends on three issues:

- (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 labelling, archiving, and data quality checks)
- (3) Appropriate data analysis and interpretation to provide management advice

Although consideration of guidelines for all items is 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, minimisation, and control of DNA typing errors. As true error rates are hard to estimate, it is evident that most efforts to assess error rates are in fact identifying inconsistencies in data sets. Nevertheless, for simplicity in what follows 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, but – at this step – not to provide quantitative suggestions for benchmarks in quality control. Whenever possible, this document shall refer to established published procedures.

Generally, errors can be introduced at various points of a DNA study (Figure 1) and fall into 3 different primary categories:

- (1) Difficulties in reliable 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)

Item 1 calls for marker validation (often addressed in a pilot study), while Items 2 and 3 are addressed by implementing a systematic quality control 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, 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 & 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 analysed. 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 (e.g., localisation of adenylation peak, null alleles, frequent allelic dropout; Goossens *et al.* 1998, Tiedemann *et al.* 2004). 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, although HWE departure can also have biological reasons. 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 artefacts 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 analyses 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 & Rousset 1995); FSTAT, etc. It should be noted that deliberately 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. In addition, 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. This correction is known to be conservative and hence will fail to detect some departures from the null hypothesis. If a multiple testing correction is performed, a better option is to use the false discovery rate (e.g., Garcia 2003), 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.

There are established routines to assess marker quality that can be used to decide whether markers should ultimately be included or excluded from analysis (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, analysis 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 & 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)

For sequence analysis of SNPs, sequence quality checks outlined below generally apply. Other SNP technologies (SNaPshot, quantitative PCR) are not covered here, as they are so far not very common in IWC-related studies.

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 & 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 & Petit 2004; Morin *et al.* 2007). 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 & Schwartz 2004; Morin *et al.* 2001; Morin & 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 & LeDuc 2004; Morin *et al.* 2007).

Where problems are detected with particular samples or where quality issues are expected, it is strongly recommend that samples are 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., 20ng per PCR reaction), quantification by absorbance or fluorescence spectrophotometry (e.g., Pico Green) can be rapid and inexpensive, allowing sample concentrations to be normalised 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 optimise sample selection and data replication criteria (Morin *et al.* 2001; Morin *et al.* 2007a; Morin & McCarthy 2007). When DNA concentration is low, potential for contamination is increased. When multiple pieces of sloughed skin are stored in the same vial, the chances for cross-contamination is also more likely. When DNA is fragmented it is advisable to target smaller microsatellite or smaller mitochondrial amplicons.

Ensuring consistent data generation

During the analysis, 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-minimising sampling. It is essential that each sample is uniquely identified. Methods for insuring that samples are uniquely identified can include: Providing prelabelled (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 labelling (might fall off 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 standardised procedure for receipt of samples at the analytical laboratory. In particular, create data base entry with field number and unambiguous lab number. Double check data entries to minimise 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 GLP (e.g., Seiler 2005). Establish standardised routine to avoid mislabelling 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 visualisation although this is less of a problem when using modern capillary analysers. 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 & Chiba 2003; LaHood *et al.* 2002). Several methods have been introduced to facilitate normalisation of alleles, but all require that controls are run to verify that alleles are correctly sized (Amos *et al.* 2007). 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 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 Oosterhout *et al.* 2004) to check for null alleles and stutter/short allele dominance effects. Be aware of 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 microsatellite 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). A similar approach can be used to evaluate the distribution of missing data points 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-type. If inconsistencies occur, re-type 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 analysed, type a set of reference samples in both labs in order to synchronise 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.

In addition to these guidelines, error rates should be systematically estimated. Incorporate replicated blind 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 normalise 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 $\geq 10\%$ 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.* 2007). 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.* 2007; Morin and McCarthy 2007).

REFERENCES

- Amos W, Hoffman JI, Frodsham A, *et al.* (2007) Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes* 7, 10-14.
- Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution* 16, 314-321.
- Bonin A, Bellemain E, Bronken Eidesen P, *et al.* (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.
- 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.
- Dunshie 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.
- 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. SC/59/BRG14, International Whaling Commission, Anchorage, AK.

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.

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

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.

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.

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, International Whaling Commission, Sorrento.

Morin PA, LeDuc RG, Archer FI, Martien KK, Taylor BL, Huebinger R, Bickham JW (2007) Estimated genotype error rates from bowhead whale microsatellite data. SC/59/BRG15, International Whaling Commission, Anchorage, AK.

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.

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

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

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.

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.

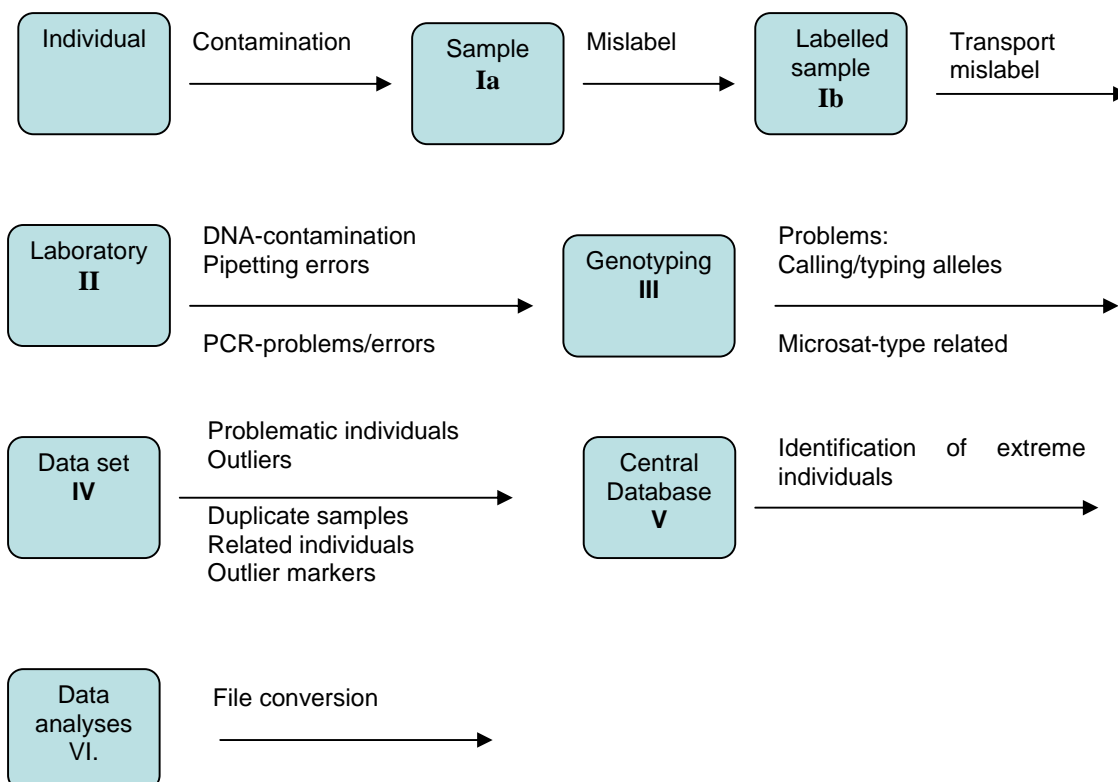


Fig. 1. Flow chart of DNA analysis procedures and potential error sources. Roman numbers refer to suggestions for quality control below.

Appendix 3

REQUEST FOR IWC FUNDING SUPPORT FOR TOSSM DEVELOPMENT

Principle Investigator: Dr. Karen Martien, SWFSC
Amount Requested: \$34,000 (~GBP 17,000)
Period: IWC 60 until IWC 61

Justification: Testing of Spatial Structure Methods (TOSSM) is an ongoing project aimed at evaluating the performance of different genetic methods of identifying population structure in a management context. Prior to SC60, no population genetic method tested in TOSSM had shown much ability to detect the need for separate management when the dispersal rate between subpopulations is high⁴. However, this year several methods have shown promise, even at fairly high dispersal rates, in certain scenarios. It is therefore now worthwhile to expand the range of scenarios tested, both towards general archetypes of population structure and towards more specific scenarios related to issues of immediate concern to the Committee.

Progress on the TOSSM project has been greatly accelerated in the past year and a half because we have been able to hire a fulltime modelling technician to work on the project. The modeller's current contract ends in early July, 2008. We have received \$34,000 from the US Government to extend the position another six months. We request an equal amount from the IWC to fund the position through to SC61, to address the issues identified below which will be overseen by the TOSSM Steering Committee. We do not anticipate extending the position beyond SC61.

Archetypes of whale breeding and feeding biology

Of the original 5 archetypes planned for TOSSM, two have been taken through to a full specification of Initial Performance Trials - an "entry-level" test suite that any candidate population genetics can be applied to. Two more archetypes have been developed at a more basic level. One original archetype stills needs to be developed - number III, clinal structure/isolation by distance, which may be widespread in cetacean populations. In addition to these generic archetypes, there may be a need to develop further more specific archetypes, for example arising from the Committee's *Implementation Simulation Trials* for North Atlantic fin whales. Developing archetypes is time-consuming, because of the need to tightly define the population dynamics and to retune the genetic models (e.g. for mutation rates) to produce plausible parameter ranges (e.g. for allelic diversity).

Modifications to TOSSM code

A number of enhancements are needed to the basic TOSSM code, in four main areas:

- (1) to enhance the implementation of the RMP CLA and of the population dynamics particularly with respect to dispersal rates;
- (2) to introduce realistic models for genetic scoring error;
- (3) to make it easier for users of TOSSM to modify the scenarios (for example, to implement scenarios with sex- and/or age- and/or spatially-biased sampling);
- (4) to facilitate summarising inputs and outputs in terms of both population genetics and management performance.

⁴ "High" means: high in population-genetic terms, not in management terms.