

# Annex I

## Report of the Working Group on Stock Definition

**Members:** Bravington (Convenor), Andersen, Archer, Bickham, Brandon, Cipriano, Donovan, Double, Edwards, Givens, Goto, Gregovich, Hoelzel, Huebinger, Jackson, Jorde, Kitakado, LeDuc, Lyrholm, Martien, Morin, Park, Pastene, Perrin, Prieto, Rosenbaum, Schweder, Skaug, Strand, Taylor, Tiedemann, Waples.

### 1. INTRODUCTORY ITEMS

#### 1.1 Election of Chair and appointment of rapporteurs

Bravington welcomed participants and was elected as Chair. Bravington, Martien, Morin, Strand, and Tiedemann 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/59/SD2, 3, 4, 6, SC/59/FI55, and SC/59/BRG13.

### 2. STATISTICAL AND GENETIC ISSUES RELATING TO STOCK DEFINITION

#### 2.1 DNA data quality

At the start of the meeting, the Chair of the SWG on the AWMP approached the SDWG with a request for advice. Although the general issue of data quality has been on the SDWG for a number of years, discussions surrounding the genetic data used in the analyses providing input to the delineation of stock structure hypotheses for the bowhead whale *Implementation Review* have re-emphasised the importance of developing a suitable protocol for genetic data used in providing management advice. In SC/59/Rep3 the AWMP *agreed that after the Annual Meeting, it will be valuable to develop guidelines for the use of genetic data in Implementations and Implementation Reviews, based inter alia on the valuable experience gained during this review*. The AWMP Chair noted that there are related Data Availability Agreement issues involved and stressed that these would be greatly aided by the development of an initial protocol for the use of genetic data that included both guidelines and suggestions for minimum standards.

The SDWG, which included participants at both the AWMP and BRG discussions, **agreed** that this would be a valuable exercise and formed a subgroup under Tiedemann to begin to address this issue. The WSG was able to make a good start during the 2007 meeting, although it did not have time to finalize its deliberations; for example, there was not enough time to properly review the extensive literature on numerical standards for benchmarks. It discussed the following issues:

- (1) Experimental design (quality control for samples, data, analysis)
- (2) Procedural implementation of data quality checks
- (3) Presentation of data and associated errors
- (4) Assessment of error rates

The WSG further agreed that it: would provide a general procedural outline how to qualitatively ensure DNA data quality; would not provide quantitative suggestions for benchmarks in quality control; and would refer to established published procedures whenever possible. It noted that its efforts were not exhaustive, and encouraged further discussion on the basis of its recommendations. The SDAHWG report is given in Appendix 2.

The SDWG thanked Tiedemann and the WSG for their work, and **agreed** that the suggestions in Appendix 2 would be a very useful contribution to the Committee's work. It was noted that there are some other issues that might usefully be considered, such as (but not limited to) the possible need for higher quality with small datasets that may be more vulnerable to errors, and the use of PHRED scores<sup>1</sup>. The work will continue intersessionally under Tiedemann, leading to a paper to be considered during the 2008 meeting. That paper will contain, *inter alia*, suggestions for numerical benchmarks in relative and absolute terms, and a suggested list of specific analyses expanding on Item V of Appendix 2.

#### 2.1 Scoring errors and mutation rates

SC/59/SD2 presented a new statistical method for estimating genotyping error rates based on mother-fetus pairs. The method was applied to the Norwegian minke whale DNA-register, and it was found that the average error rate for microsatellites was 1.8% per allele for the early period of the register, to 0.4% in the recent period (since 2003).

In discussion, the Group noted the value of the mother-fetus database in providing independent estimates of microsatellite scoring-error rates. It provides a valuable complement to repeat-scoring, in that it addresses some different sources of error (e.g. samples collected in different ways, versus re-scoring the same sample, so eliminating any "sample effect"). Further, the mother-foetus database could potentially provide a direct estimate of mutation rate, which is an important parameter for many population genetics methods but can normally only be inferred indirectly and imprecisely. By repeat-reading those mother-fetus samples that appear inconsistent, the possibility of scoring error can be eliminated, and any remaining differences must be due to mutation. The author of SD2 reported that one instance of apparent mutation was indeed found in the set of animals in SD2. In discussion, it was also noted that the mother-fetus samples could be used to study mutation rates in mtDNA, which will not suffer from scoring errors. In all, the development of the work in SC/59/SD2 was **strongly encouraged**.

---

<sup>1</sup> See <http://www.phrap.com/phred> for details of PHRED.

### 3. TOSSM

#### 3.1 Progress on the TOSSM simulation-testing framework

SC/59/SD4 is a user's guide to be distributed with the TOSSM. package. It begins with a general overview of the package and definition of some of the terms used. It then describes all of the arguments to the function `run.tossm`, which is the main function in the TOSSM package. Each of the arguments is described in detail.

In discussion, the Group welcomed the progress that has been made in getting the TOSSM control program to work smoothly, and thanked Martien, Gregovich, Bravington, and the developer of Rmetasim (Strand; see Strand, 2002) for their efforts. This year in particular, people outside the "core" TOSSM group have been able to use the control program successfully (e.g. in SC/59/SD6, discussed below).

The Group raised a number of questions about desirable features currently available in, or planned for, the TOSSM control program. These items are listed in the bullet points below. The control program is written in a modular fashion, so that a number of features that are not implemented in the "vanilla" version can be easily implemented, simply by writing an R routine to replace the default (e.g. to change the way genetic samples are collected). This means that any user of the control program can change the way certain things are done without needing to change the R package itself. Some of the items listed below can be implemented by modules, while others will require some change to the control program itself.

- individual breeding histories (present already)
- retention of individual ID in the simulation, to check assignment (addition)
- breeding-ground samples as distinct from harvest-ground samples (addition)
- age- and sex-structure in sampling & harvest (addition)
- sampling linked to catches (addition)
- multi-year sampling (module)
- scoring error (module)
- abundance survey CVs too low (module or trivial change)
- alternative harvest strategies (module)
- "epigenetic inheritance", e.g. learned preference for feeding grounds, in the "ancient history" phase (addition: Rmetasim?)
- ability to run under Linux/Unix as well as Windows (needs checking)

In addition, more documentation, some bug fixes, and more examples (e.g. possible alternatives to the default modules) are planned. An important task will be to ensure that the documentation and nomenclature is made accessible to the population genetics community outside IWC. There are two reasons for this: first, TOSSM has potential value in applications beyond whales; and second, the IWC will continue to derive great benefit from using TOSSM to bring in the expertise of population geneticists unfamiliar with the world of whales.

To assist people interested in developing BSAs, the Group reiterated the utility of a document giving a worked example of how output from a population genetics method might be used to automate a management boundary decision. It may be possible to use some of the results from SC/59/SD3 or SC/59/SD6 (see below). Donovan and Bravington again promised to take the lead in drafting such a document, which will be circulated intersessionally to the TOSSM Steering Committee (Intersessional Working Group R11).

In order to ensure the comparability of performance tests performed by different analysts, it is necessary to specify an initial set of performance trials. During IWC 59, a subgroup on performance trials gave its consideration was given to a number of detailed specifications in the simulations; its report is attached as Appendix 3. The SDWG agreed with the deliberations of the subgroup on performance trials, and the trials will be set up intersessionally.

Many of the Boundary Setting Algorithms (BSAs) to be tested in the TOSSM framework are very computationally intensive. Consequently, it is necessary to limit the number of performance trials. To that end, it was agreed that prior to the commencement of performance-testing of BSAs, all trials specified in this document will be run using the 'single.MU.BSA,' which is a default BSA included in the TOSSM package in which all FIMAs are automatically combined into a single management unit. This will allow the TOSSM community to focus performance testing on only those trials in which failure to detect population structure would result in poor conservation performance.

Champions<sup>2</sup> of specific population genetics methods will want to explore the performance of their methods more widely and will likely conduct performance trials that are designed to test specific aspects of their method. The intent of the trials specified in Appendix 3 is simply to specify a standardized minimum set of trials that will allow comparison of the performance of different methods.

The SDWG noted that the TOSSM framework is beginning to be used well outside its initial scope, which was to provide a *generic* testbed for population genetics in management, away from the obscuring complexities of *specific* stock-structure issues being considered by the Committee. However, the simulation and testing machinery in TOSSM is now close to being powerful enough that it can be adapted to very specific situations. For example, the detailed simulations of bowhead whale demography, genetics, and whaling history in SC/59/BRG17 could almost have been done with the current TOSSM control program, and may indeed be within the scope of the program after its next round of modifications. Further, the programs developed & experience gained so far in TOSSM made the programming of the bowhead simulations faster and easier.

#### 3.2 Testing and development of population genetics methods

Tiedemann reported preliminary results from an application of Principal Component Analysis (PCA) on microsatellite data to detect population structure in harbour porpoises. The Group noted the potential value of Tiedemann's approach as an exploratory tool, especially in cases with a

---

<sup>2</sup> A "champion" is someone who adapts a population genetics algorithm (e.g. STRUCTURE) so that it can be tested in TOSSM, e.g. by adding automated decision rules about whether to manage separately or together based on outputs from the algorithm. See RIWC 58.

*priori* plausible scenarios based on geographic structure; the discussion highlighted the importance of bringing to bear biological insights in suggesting *a priori* plausibility. The Group looked forward to considering a full paper on the approach next year, including application to TOSSM.

SC/59/SD3 presented the results of simulation performance testing of the Bayesian clustering method STRUCTURE. The authors analysed simulated datasets generated for use in the Testing of Spatial Structure Methods (TOSSM) project (JCRM RIWC 2004). They examined three population structure scenarios: (1) a single population with a carrying capacity of 7500; (2) two populations, each with carrying capacities of 3750, exchanging dispersers at the annual rate of  $5 \times 10^{-4}$ ; and (3) two populations with carrying capacities of 3750 and an annual dispersal rate of  $5 \times 10^{-3}$ . The authors found that the probability of correctly inferring the number of populations from the results of STRUCTURE was low. The admixture/correlated-allele-frequencies model performed well for the single population scenario, but consistently underestimated the number of populations for the two scenarios with two populations. The no-admixture/independent-allele-frequencies was biased towards selecting too many populations for scenario 1 (i.e. sometimes splitting), was unbiased when applied to scenario 2 (i.e. always splitting), and was biased towards selecting too few populations for scenario 3 (i.e. sometimes merging). Both models did a poor job of assigning individuals to the correct population, with assignment success only equal to what would be given by assigning individuals at random. Assignment probabilities for the admixture/correlated-allele-frequencies model were close to 0.5, accurately reflecting the degree of uncertainty in the assignments. The no-admixture/independent-allele-frequencies, on the other hand, produced very high assignment probabilities, even when those assignments were wrong.

The appendix to SC/59/SD3 describes a BSA that uses the results of STRUCTURE analyses to define management units. The BSA accepts as arguments the ancestry and allele frequency models to be used by STRUCTURE, the length of the MCMC chain, and the numbers of groups ( $K$ ) STRUCTURE should explore. The BSA determines the number of management units to define by comparing the log-likelihood of the STRUCTURE results for different value of  $K$  and choosing the value that has the highest likelihood. FIMAs are assigned to management units based on the average assignment probability of all individuals in the FIMA. Initial performance testing shows that the BSA performs poorly, as it always resulted in the definition of a single management unit, regardless of the number of populations. To facilitate comparisons between simulation test results from different population genetic methods, the authors suggest that an initial set of performance trials be parameterized to serve as a standard set against which all BSAs would be tested. In order to identify performance trials for which failure to identify population structure will result in conservation risk, they suggest using the TOSSM package to investigate the performance of the CLA under a variety of population structure scenarios assuming that all populations are managed as a single unit.

In discussion, the Group noted that, although the scenarios tested in SC/59/SD3 would be expected *a priori* to be problematic for STRUCTURE because of the high dispersal, reasonable performance might be expected at slightly lower dispersal rates. The author commented that this was borne out in two simulations not included in the paper (for small unequal population sizes, with per capita migration rates  $\sim 10^{-6}$ ), where STRUCTURE had assigned individuals correctly 80% of the time.

It was also noted that the admixture option is known to be difficult for STRUCTURE to deal with; the authors of STRUCTURE have in fact suggested that it may be operationally better to avoid this option even when admixture is *a priori* likely, because the no-admixture model is often more powerful at detecting subtle structure. In an IWC context, the most critical scenarios are those where dispersal rates are likely to be low enough to blur allele frequency differences (so that differentiation is hard to detect) but the proportion dispersing is quite low (so that separate management is desirable). In such cases, there will clearly be correlation between the allele frequencies, but there are unlikely to be many cases of admixture in the genetic samples, so that it would be interesting to try STRUCTURE with correlation ON but admixture OFF.

SC/59/SD6 reported tests on a BSA based on the program BayesAss. BayesAss estimates the rate of migration per generation between putative populations for which samples are available. The migration rate was used by the BSA to decide whether or not to treat two putative sub-populations into a single management unit. If migration was estimated to be above a pre-specified value (the *critical m*) then a single management unit would be used, otherwise two management units would be used. The BSA was tested in a preliminary fashion on TOSSM datasets. Specifically, Archetypes I (a single panmictic population) and II (two breeding populations joined by migration). Archetype II migration rates of  $5 \times 10^{-6}$ ,  $5 \times 10^{-4}$ ,  $5 \times 10^{-4}$  and  $5 \times 10^{-3}$  were considered. In each case samples were distributed evenly between two FIMAs; for Archetype II, each breeding population was assigned its own FIMA. For each of the five population genetic datasets, the BSA was run once using *critical m* values of 0.0 (always split), 0.1, 0.2 and 1/3 (always merge). It was found that the migration rates estimated by BayesAss were both inaccurate and inconsistent. Consequently the BSA performed poorly, producing a suboptimal description of management units.

In discussion, the Group noted that the results in SC/59/SD6 were consistent with those reported in Faubet *et al.*, 2007, which tested the performance of BayesAss in estimating migration rates and individual assignment using a simulation. With respect to technical issues of MCMC convergence, Waples reported that, following the simulation experience in that paper, some improvements have been proposed to the MCMC algorithm in BayesAss, which may alleviate some of the convergence problems. With respect to overall performance, it was noted that, for any given per capita migration rate, BayesAss and many other population genetic methods will perform worse with increasing population size, because the larger absolute number of migrants per generation will erode the difference between the subpopulations. BayesAss can thus be expected to work well for large populations when there has only been recent contact; a longer period of contact will still leave a signal if the population sizes are small, but this may limit its utility in whaling management.

Many population genetics tools, not just BayesAss, rely on MCMC. Even if better convergence can be obtained in specific cases, computational speed of MCMC-based techniques will likely remain a major stumbling block to simulation-testing. It would be worth exploring whether much faster results could be obtained using something like profile likelihood to identify approximate lower confidence intervals on migration rate. In settings with a large number of nuisance parameters (such as population allele frequency distributions), however, straightforward profile likelihood can be very poorly calibrated, and there may be some value in exploring alternatives such as "confidence distributions" (Schweder and Hjort, 2002) which, while slightly more complex than unmodified profile likelihood, are much less computationally demanding than MCMC.

One of the major unknowns when designing a BSA, is what level of dispersal/migration/etc. would be "enough" to permit single-stock treatment, even if perfect information on that parameter was available and the parameter did not have to be estimated from data. In principle, a BSA designer could simply find out an appropriate "tuning" level for any BSA simply by running enough simulations with different decision thresholds and seeing how the management performance behaved. However, this is a substantial burden of work, and in some sense is also a shared problem that cuts across different BSAs. Some general guidance to BSA developers on sensible criteria would therefore be helpful. The Group agreed that high priority should be given to simulation-testing of some "perfect-information" BSAs under TOSSM. Specifically, this means running trials under the "single.MU.BSA" (Appendix 3) and noting how the conservation performance changes with migration rate. This may suggest a general threshold

migration rate which could be tried in a decision rule by several methods, e.g. if some confidence interval of the estimated migration rate for that method falls below the threshold. It was noted that this may help in future discussions of "unit-to-serve".

#### 4. WORK PLAN AND BUDGET

A number of items relating to further development of the core TOSSM package and the setup of simulations and performance trials are detailed in the report above. These need to be implemented intersessionally. The SDWG noted the progress that has been made since the employment of a full-time TOSSM technical assistant to work with Martien this year (part-funded by IWC). Since funds for this position will run out in March 2008, the SDWG **strongly supported** the funding proposal in Appendix 4, which seeks salary to cover the gap between March 2008 and IWC 60 in May 2008; other sources for funding beyond that date will be sought meanwhile. Continuity in the technical assistant position is crucial to making efficient progress with TOSSM and bringing forward the results into the Committee's work.

The TOSSM Steering Group will continue to function intersessionally, to provide feedback on points of detail in the implementations (see Annex R11 for Terms of Reference and membership).

With respect to testing particular population genetics methods in TOSSM, the SDWG's approach has been to find a "champion" for a method, who will take charge of adapting the method for testing, and actually testing the method on the simulated datasets and in the performance trials. The list of champions was revised at IWC 59 and is given below. Several methods have already had preliminary application to some TOSSM datasets (see RIWC 58 Potsdam rpt and discussion this year) and the SDWG looks forward to seeing next year the results of their application to the performance trials that will be developed intersessionally (see Appendix 3).

1. MIXPROP (Kitakado)
2. BayesAss (Gaggiotti/Edwards)
3. Geneland/Structure (Martien)
4. Sequential hypothesis testing (Punt)
5. Boundary Rank (Martien)
6. LAMARC (Jackson)

In addition, the SDWG recalled from last year's discussion that TOSSM provides a very useful way to simulation-test the performance of commonly-used allele-checking software such as MICROCHECKER. This would be a feasible student project, and members agreed to investigate this intersessionally.

The proposed agenda for IWC 60 is again:

1. Statistical and genetic issues relating to stock definition (including further discussion of DNA data quality)
2. Progress on TOSSM
3. Discussion of possible criteria for unit-to-serve

#### ADOPTION OF REPORT

The report was adopted at 12:40 on 13<sup>th</sup> May 2007.

#### REFERENCES

Faubet, P., R. S. Waples, O. E. Gaggiotti. 2007. Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Molecular Ecology* 16:1149-1166.

Schweder, T. and Hjort, N.L. 2002. Confidence and likelihood. *Scandinavian Journal of Statistics* 29: 309-332.

Strand, A. 2002. METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Mol. Ecol. [Notes]* 2:373-376.

#### 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
6. TOSSM (Testing of Spatial Structure Models)
  - 6.1 Update on progress
  - 6.2 Directions for further work
7. Work plan and budget
8. Adoption of report

## Appendix 2

### REPORT OF THE SUBGROUP ON DNA DATA QUALITY

Members: Tiedemann (convenor), Andersen, Bickham, Donovan, Double, Edwards, Hoelzel, Huebinger, Jackson, LeDuc, Morin, Schweder, Skaug, Strand, Waples

This subgroup was formed in response to a Committee request to discuss and suggest general procedures on how to ensure DNA data quality. The subgroup agreed on the following issues to be discussed:

- 1) Experimental design (Quality control for samples, data, analysis)
- 2) Procedural implementation of data quality checks
- 3) Presentation of data and associated errors
- 4) Assessment of error rates

It was noted that most error rate assessments identify inconsistencies in data sets such that the true error rate is hard to estimate. The AWMP Chair explained the wish of IWC to receive an agreed protocol for each marker type (sequences, microsatellites, SNPS; possibly include nuclear DNA sequencing in the future).

The subgroup agreed that:

- it will provide a general procedural outline how to qualitatively ensure DNA data quality.
- it will not provide quantitative suggestions for benchmarks in quality control.
- it will refer to established published procedures whenever possible.
- it considers this effort not as exhaustive, but instead encourages further discussion on the basis of the recommendations made below.

The subgroup further agreed to discuss error types in the context of three different primary categories:

- 1) consistency of methods
- 2) locus characteristics
- 3) samples characteristics

It was also agreed on that these guidelines do not apply to historical samples; for these, it is possible to refer to literature for well-established methods to handle 'ancient DNA'. It was noted that established DNA protocols from forensic sciences potentially mark the feasible optimum of data quality and can be referred to where appropriate. Generally, however, they cannot be demanded as this level of scrutiny is very expensive and appears unnecessary for standard DNA analyses.

The group developed a flow chart that indicates the points where errors can be introduced (Figure 1). This is supposed to be generic to all genetic marker types and can be modified for each marker type, where appropriate.

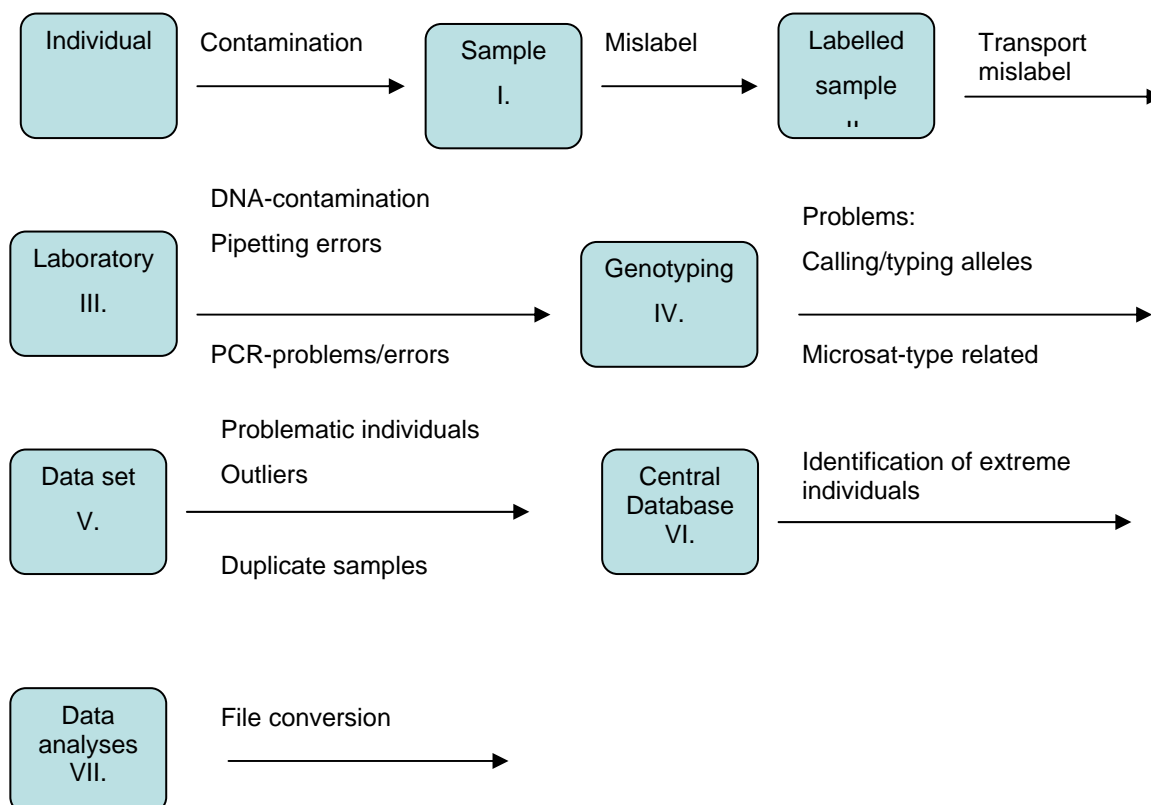


Figure 1. Flow chart of DNA analysis procedures and potential error sources. Roman numbers refer to suggestions for quality control below.

It was noted that two phases of quality control can be distinguished: Marker validation (before use of molecular marker); and Quality Control (while using molecular marker).

For marker validation, the following is demanded:

- Microsatellites: (1) 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 (see below). (2) A pilot study should investigate reliability of amplification and identify technical problems (e.g., frequent allelic dropout). An indicator can be consistent deviation from HWE (although it was noted that HWE departure can also have biological reasons). If the marker appears unreliable at this stage, it should not be used. (3) Assessment of genotype error by blind duplicate analysis of a subset of samples. (4) Check for linkage disequilibrium among loci analyzed. If linkage disequilibrium is detected, use only one of the linked loci.
- mtDNA: If using non-validated primers, the mitochondrial origin must be demonstrated. In particular, the possibility of erroneously sequencing nuclear pseudogenes (Numts) must be ruled out. Generally, sequences should be compared to Genbank (BLAST) and run through DNA surveillance routines, when available. 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.

During the analysis, the following measures are recommended (Roman numbers refer to figure 1).

- I. Sampling: Provide prelabelled (numbered) sample vials prefilled with appropriate storage buffer to the field worker. Provide explicit easy-to-read instructions for contamination-minimizing sampling (see literature for established procedures).
- II. Sample labelling: Provide prelabelled (numbered) sample vials (barcoded, if possible). Provide unambiguous instructions about additional labels. Double-label every vial with waterproof pen, do not use tape for labelling (might fall off later on). Advice to always start with vial with lowest number and strictly following numbers, such that they reflect order of sampling. Vials that had been erroneously left out should not be used later on.
- III. Establish standardized procedure of handling sample upon receipt. In particular, create data base entry with field number and unambiguous lab number. Double check data base entries to minimize copying error. Divide sample and store backup sample in separate fridge.
- IV. Work according to established procedures for Good Laboratory Practice (GLP). Establish standardized routine to avoid mislabelling of tubes in the process of genotyping. For microsatellites: Check allele calling for subset of samples by double-blind genotype calling involving at least two persons.
- V. Check data for consistency and plausibility. For microsatellites, use quality control software (e.g., MICROCHECKER) to check for nullalleles and stutter/short allele dominance effects; 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). If inconsistencies occur, re-type these specimens. If data do not appear plausible, 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 demanded, 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 sample size is low, be particularly thorough on quality checks. Produce reference data set for which consistency **you** hold responsibility, even though data are shared or submitted to a central data base.
- VI. **Central database holds responsibility for combined data set.** In coordinated data acquisition efforts (e.g., as in BCB-bowhead whales), there should be two deadlines: First deadline 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 producing laboratory holds full responsibility.
- VII. Data analysis: Manual file conversion should be avoided (because of copying error). Use automated routines (e.g. MSA) wherever possible.

## Appendix 3

### REPORT OF THE WORKING GROUP TO SPECIFY INITIAL PERFORMANCE TRIALS FOR TOSSM

**Members:** Martien (convenor), Strand, Gregovich, Edwards, Punt, Hoelzel, Tiedemann.

There are a number of parameters that must be specified in order to conduct simulation performance testing using the run.tossm function of the TOSSM package. In order to ensure the comparability of performance tests performed by different analysts, it is necessary to specify an initial set of performance trials. This small subgroup was tasked with developing this set.

Many of the Boundary Setting Algorithms (BSAs) to be tested in the TOSSM framework are very computationally intensive. Consequently, it is necessary to limit the number of performance trials. To that end, the subgroup agreed that prior to the commencement of performance-testing of BSAs, all trials specified in this document will be run using the 'single.MU.BSA,' which is a default BSA included in the TOSSM package in which all FIMAs are automatically combined into a single management unit. This will allow us to focus our performance testing on only those trials in which failure to detect population structure would result in poor conservation performance.

#### Defining FIMAs

The subgroup agreed that the number of FIMAs in performance trials should exceed the number of breeding populations so that it is possible to make overprotection errors (defining too many management units) as well as under-protection errors (defining too few management units). It is analytically convenient for the number of FIMAs to be an even multiple of the number of populations. The TOSSM datasets include scenarios containing one, two, or three populations. Consequently, we agreed that all performance tests will be conducted with six FIMAs.

For Archetypes 1, 2, and 5, there is no mixing of breeding populations on FIMAs. For scenarios of these archetypes where abundance of all populations are equal, the FIMAs will be split evenly between the breeding populations (Table 1). In some Archetype 2 scenarios, there are 2 populations of unequal abundance. For these scenarios, all individuals from the smallest population will go to FIMA 1, while the remaining 5 FIMAs will contain individuals from population 2. This will result in lower abundance on FIMA 1 than on FIMAs 2 through 6. The subgroup recognized that this unequal abundance undesirable, as it could provide 'clues' to a BSA regarding the number of populations in the simulated dataset. The subgroup felt that in the initial performance trials this was acceptable, while acknowledging that at later stages of performance testing it will be necessary to design trials with more FIMAs so that unequal abundance scenarios cannot be identified based on abundance estimates on the FIMAs.

Table 1. Mixing matrices specifying the distribution of breeding populations (BPs) across FIMAs for performance trials of Archetypes 2 and 5.

	FIMA1	FIMA2	FIMA3	FIMA4	FIMA5	FIMA6
2-population even abundance scenarios						
BP1	0.333	0.333	0.333	0	0	0
BP2	0	0	0	0.333	0.333	0.333
2-population uneven abundance scenarios						
BP1	1.0	0	0	0	0	0
BP2	0	0.2	0.2	0.2	0.2	0.2
3-population scenarios						
BP1	0.5	0.5	0	0	0	0
BP2	0	0	0.5	0.5	0	0
BP3	0	0	0	0	0.5	0.5

In Archetype 4, breeding populations mix on feeding grounds. At the first TOSSM workshop it was agreed that we would examine two variants of this Archetype: 1) complete overlap of feeding grounds and 2) 50% overlap of feeding grounds. For the former variant, both breeding populations will be distributed evenly across all six FIMAs. For the partial overlap variant, the subgroup agreed that the middle two FIMAs would be evenly split between the two breeding populations, and that the total abundance of all FIMAs will be equal. That specification results in the mixing matrix shown in Table 2. This matrix does actually reflects 33% mixing rather than the 50% specified at the TOSSM workshop (one-third of the FIMAs are mixed, and one-third of each population is in a mixed FIMA). The subgroup acknowledged this departure from the original specifications, but felt that 33% overlap was preferable as it is more likely to result in undesirable conservation outcomes if a BSA fails to detect population structure.

Table 2. Mixing matrices specifying the distribution of breeding populations (BPs) across FIMAs for performance trials of Archetypes 2 and 5.

	FIMA1	FIMA2	FIMA3	FIMA4	FIMA5	FIMA6
BP1	0.33	0.33	0.167	0.167	0	0
BP2	0	0	0.167	0.167	0.33	0.33

#### Simulation schedule

The performance trials will consist of a single pre-RMP year during which all historical catches are removed. The RMP phase will last 100 years, with abundance estimates collected and quotas calculated every 5 years. The BSA will be called and management units defined at the beginning of the RMP phase only. Though it will eventually be necessary to conduct trials in which the BSA is re-run at regular intervals during the RMP

phase, the computational burden of the BSAs necessitates postpone such 'full feedback' testing until the performance and utility of different performance trials and BSAs is better understood.

The subgroup agreed that there would be no post-RMP phase in the performance trials.

#### **Sample size**

At the first TOSSM workshop, it was agreed that sample sizes of 50 and 100 samples per population would be used. The subgroup found two problems with this specification. First, it would result in the number of samples varying as a function of the number of populations. Second, the larger of these two sample sizes (100 for one-population scenarios, 200 for two-population scenarios, 300 for three-population scenarios) is not large enough to encompass the range of sample sizes that are likely to be available for large whale species.

We agreed that rather than specifying the number of samples per population, the sample sizes should be equal across FIMAs (for the initial set of performance trials). This will result in total sample size being independent of population structure scenario. The subgroup agreed to sample sizes of 20 samples and 100 samples per FIMA (total sample sizes of 120 and 600).

It was noted that for some BSAs, computational burden is roughly proportional to sample size. In such cases, it may not be feasible to examine a total sample of 600 individuals due to computational constraints.

#### **Initial depletion and tuning of the CLA**

Initial depletion is specified indirectly in the run.tossm through the specification of historical catches. The subgroup agreed that the number of historical catches would be chosen so that the 'coastal' population in a two-population scenario would have an initial depletion of 30%, while the other population remains at carrying capacity. The historical catch will be the same for all scenarios, with the result that initial depletion will vary depending on the number of populations. This choice was made so that trial specification would be independent of the number of populations.

The life history matrices used in generating the TOSSM datasets result in an MSYR of approximately 4%. Consequently, Punt stated that it is unlikely that any performance trials would produce conservation risk if we use the default tuning of the CLA. Run.tossm includes the argument mult.CLA, which is a multiplier that can be used to increase (or decrease) the quotas calculated by the CLA. We agreed to set the value of mult.CLA so that the median final depletion for a performance trial with a single-population scenario is 72% of carrying capacity. This is the equilibrium value achieved in the default CLA when MSYR equals 1%.

#### **Other issues**

The TOSSM datasets included genotypic data for 30 microsatellite loci and a 500bp mitochondrial haplotype sequence. However, the number of microsatellite loci and length of mitochondrial sequence available for analysis varies quite widely between case studies. In order to bracket the range of values likely to be available, the initial performance tests of methods that use microsatellite data will examine the performance of methods using 10 loci and 30 loci. Similarly, methods that rely on mitochondrial data should be tested using sequence lengths of 250bp and 500bp.

Finally, the coefficient of variation of the abundance estimates generated by run.tossm need to be tuned so that it is comparable to values typically observed in real datasets. The subgroup did not discuss this issue in detail, but a value of 0.3 to 0.4 was mentioned as being reasonable.

Method champions will want to explore the performance of their methods more widely and will likely conduct performance trials that are designed to test specific aspects of their method. The intent of the trials specified herein is simply to specify a standardized minimum set of trials that will allow comparison of the performance of different methods.

## **Appendix 4**

### **BUDGET PROPOSAL FOR TOSSM DEVELOPMENT**

Testing of Spatial Structure Methods (TOSSM) is a continuing project aimed at evaluating the performance of different genetic methods of identifying population structure in a management context. At SC58, SD requested and received GBP 16K to pay half the salary of a fulltime technician to assist with the completion of phase I of TOSSM. An additional GBP 16K was contributed by the US government, allowing a fulltime hire to assist Dr Martien for one year. That position was filled in early March 2007 and the money will run out in March 2008.

The TOSSM project is starting to make major contributions to the work of the Committee. There are a number of modifications and extensions that are required (see Annex I), and a need to improve documentation to make it easier for non-IWC population geneticists to get involved.

In order to extend the position through the Annual Meeting in Chile in June, 2008, Dr Martien requests an additional three months of fulltime funding (\$18,000). Meanwhile there will be opportunities to seek further funding from other sources to continue the work after May 2008.

#### **Timetable**

March 2008 – May 2008

#### **Researchers' names**

Karen Martien / Dave Gregovich

#### **Estimated total cost**

GBP 9K (salary)