# GenBank sequence assessment for species assignment - control region sequences published for baleen whales in 2007

H. A. ROSS AND H. SHEARMAN

*School of Biological Sciences*
*University of Auckland*
*Private Bag 92019*
*Auckland, New Zealand*

ABSTRACT

The tree-based methods in DNA Surveillance, in conjunction with the curated reference sequence alignments known as Witness for the Whales, were used to assign species identities to the 499 sequences from baleen whales published in Genbank during 2007. All of the sequences were assigned to the same species as that recorded in Genbank. For the common minke whale, 73 of the 74 sequences were not indentified in Genbank as belonging to one of the subspecies, while they could be assigned unambiguously using the WFTW references. There was uncertainty regarding whether blue whale sequences could be assigned to a subspecies. All of the sequences appeared to be of reliable quality. No geographic information was recorded for nearly all of the sequences.

## INTRODUCTION

Each year large numbers of DNA sequences, mainly of the control region of the mitochondrial genome (mtDNA), derived from baleen whales (Mysticeti) are deposited in Genbank (www.ncbi.nlm.nih.gov). There are several potential sources of error relating to the species identity of these sequences, including misidentification of the animal, mislabelling of the sample, and contamination of the sequencing reaction. One way in which such errors might be identified is to assign each DNA sequence to a species using reference sequences for comparison.

The DNA Surveillance (*http://www.dna-surveillance.auckland.ac.nz*) system allows for the assignment of specimens using tree-based methods (Ross et al., 2003) in conjunction with curated reference sequences (Baker et al., 2003). These sequences, obtained from expertly identified specimens, represent an authoritative reference collection known as Witness for the Whales (WFTW).

Previously this approach was applied to all control region sequences from baleen whales deposited and published in Genbank prior to 2007 (Ross and Shearman, 2008). Very few of these sequences were flagged as being of poor quality. Most disagreements between the identity recorded in Genbank and that assigned by the WFTW database were due to taxonomic revision, or the inability of the reference sequences to differentiate proposed subspecies.

Consequently it was recommended (Scientific Committee, 2008) that this approach be applied to the sequences published during 2007:

1. List the GenBank accession number and species identity of each mysticete control region sequence published in GenBank during 2007 with the species identity as determined using the most recent version of the Witness for the Whale reference sequence alignments (see SC/59/SD5) and the DNA Surveillance software engine;
2. The above list to be supported by phylogenetic trees, one per sequence, showing the placement of the GenBank sequence in relation to the reference sequence; and
3. Evaluation of the types of inconsistencies/errors as agreed by the Committee in 2007: quality of submitted sequences; accuracy of species identification and, where possible, accuracy of geographical location.

This report summarises our attempt to address these issues. Our aim here is to use DNA Surveillance and the most recent version (V4.3) of the WFTW reference alignments to assign all of the mtDNA control region sequences from baleen whales published in Genbank during 2007 to a species, or subspecies wherever possible.

This document summarises the results of this work. Greater detail is provided in the appendices available at www.cebl.auckland.ac.nz/~hros001/cetaceanID.

**METHOD**

Genbank was queried for control region mtDNA sequences in three different ways:

- The query string **`'mysticeti[orgn] AND (d-loop OR control)'`**, with the limit that publication occurred between 1 Jan 2007 and 31 Dec 2007, as used to search the Nucleotide database.
- A randomly chosen humpback whale sequence (DQ768421) was used as a query in a BLAST search of the nr (non-redundant nucleotide) database, with the search limited to **`'cetacea[orgn] AND 2007[pdat]'`**, with a large number (1000) hits to be reported. Otherwise the default search parameters were used.
- A randomly chosen minke whale sequence (EF113863) was also used as a query in a similar BLAST search.

From these three searches we selected records which matched our target: mysticete control region sequences published in 2007. The first search method recovered the largest number of matches, and it included all of the mysticete sequences discovered by BLAST searching. We did not pursue the possibility that some of the non-mysticete sequences in the BLAST results were misidentifications, and might have come from mysticete whales. A total of 499 sequences were downloaded and analyzed.

Each downloaded sequence, the query sequence, was analysed individually using DNA Surveillance and the WFTW reference alignments by the following steps:

1. The query sequence was aligned against a reference alignment containing representatives of all of the cetacean families, called 'All Cetaceans v4.3', using a profile alignment method.
2. The evolutionary distances among all of the aligned sequences, reference and submitted, are then calculated using the F84 model of evolution with transition/transversion ratio ($T_s/T_v$) = 2 and empirical nucleotide frequencies (Felsenstein, 1984).
3. A phylogenetic tree is built from the table of evolutionary distances using the Neighbor-Joining (NJ) algorithm (Saitou and Nei, 1987). The tree is rooted with an outgroup comprising the sperm whale (*Physeter macrocephalus*) and the pygmy sperm whale (*Kogia breviceps*). This tree always contained two major clades representing the baleen whales and the toothed whales.
4. If the query sequence was placed within, or sister to, the clade of baleen whales then it was considered to be from a putative baleen whale and the analysis was continued. If the query fell elsewhere on the tree then the analysis was stopped.
5. Then, once the sequence had been identified as from a putative baleen whale, the process was repeated using the reference alignment 'Mysticetes v4.3'. Table 2 lists the mysticete taxa, species and subspecies, recognised by WFTW.
6. The identity of the query sequence was taken from that of the clade in which it fell. If the clade contained two or more species, then the identification was considered ambiguous. If the query was embedded in a single-species clade (e.g., (X, (Q, X)) ), then the identification evidence was considered to be strong. If the query was in a sister position with respect to a single-species clade (e.g., ((Q, (X, X))) ), then it was considered to be only moderate (Figure 1).
7. For each analysis, a table of genetic distances was recorded and the phylogenetic tree was saved.

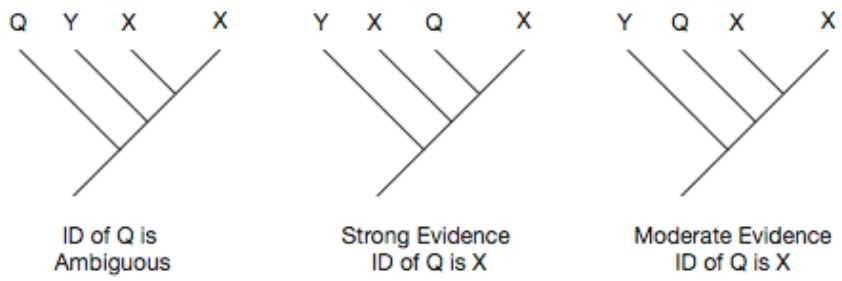


Figure 1. Strength of evidence in making an identification was based on relative position of query sequence to reference sequences.

The quality of each sequence can be inferred from some simple measures. It is expected that sequences containing errors will differ from the references sequences in novel ways. Since we do not have access to the original electropherograms or other evidence, we have used the following measures as indices of sequence quality:

1. Number of nucleotides with IUPAC ambiguous nucleotide codes (R, Y, M, K, S, W, H, B, V, D, N). These represent direct evidence of uncertain or ambiguous sequence.
2. Number of single-position gaps. Sequencing error might result in false deletions of a single nucleotide. When the query sequence is aligned against the reference sequences, these sequencing errors will result in single-base deletions. Some such deletions, of course, are real and so this measure is only indicative of problematic sequences.
3. Number of positions at which the observed state did not occur in the reference sequences. When the query sequence is aligned against the reference sequences, there may be a nucleotide, an inserted nucleotide or a gap in the query sequence which does not occur at that position in any of the reference sequences. Some such differences may be real and so this measure is only indicative of problematic sequences.
4. The number of positions at which the query does not match any of the references can be scaled by the length of the overlapping alignment between the references and the query to give a relative frequency of such mismatches. This measure is only indicative of problematic sequences.
5. Genetic distance to closest reference sequence. In addition to the introduced errors detected by the preceding measures, some component of genetic distance might be due to other forms of sequencing error. This measure is very indirect and will depend on the degree to which the reference sequences reflect naturally occurring genetic variation.

For each sequence, these measures of sequence quality were computed. Then the sequences with the extreme 5% of values were identified. Sequences with multiple measures in these 5% tails might warrant closer examination.

Table 1 The outcomes of comparing the species identification of each sequence, as given in Genbank, with the assignment attained using the Witness for the Whales references in the DNA Surveillance system. The relative strength of evidence, moderate or strong, is discussed in the text.

| GENBANK ORG NAME | Ambiguous | Same Species | | Different Subspecies | | Different Species | | Grand Total |
|---|---|---|---|---|---|---|---|---|
| | | moderate | strong | moderate | strong | moderate | strong | |
| *Balaena mysticetus* | | 9 | 90 | | | | | 99 |
| *Balaenoptera acutorostrata* | | | | 1 | 72 | | | 73 |
| *Balaenoptera acutorostrata subsp dwarf minke whale* | | | 1 | | | | | 1 |
| *Balaenoptera bonaerensis* | | | 115 | | | | | 115 |
| *Balaenoptera brydei* | | 20 | 32 | | | | | 52 |
| *Balaenoptera musculus* | 11 | 6 | 22 | 5 | | | | 44 |
| *Megaptera novaeangliae* | | 64 | 51 | | | | | 115 |
| **Grand Total** | **11** | **99** | **311** | **6** | **72** | **0** | **0** | **499** |

Table 2 Summary of the circumstances under which each taxonomic group in the WFTW was the assigned identity of a sequence from Genbank.

| WFTW GENUS SPECIES NAME | Ambiguous | Same Species | | Different Subspecies | | Different Species | | Grand Total |
|---|---|---|---|---|---|---|---|---|
| | | moderate | strong | moderate | strong | moderate | strong | |
| <blank> | 11 | | | | | | | 11 |
| *Balaena mysticetus* | | 9 | 90 | | | | | 99 |
| *Balaenoptera acutorostrata acutorostrata* | | | | 1 | 28 | | | 29 |
| *Balaenoptera acutorostrata scammoni* | | 1 | | | 35 | | | 35 |
| *Balaenoptera acutorostrata subsp. (dwarf)* | | | 1 | | 9 | | | 10 |
| *Balaenoptera bonaerensis* | | | 115 | | | | | 115 |
| *Balaenoptera edeni (common form)* | | 20 | 32 | | | | | 52 |
| *Balaenoptera musculus* | | 6 | 22 | | | | | 28 |
| *Balaenoptera musculus (brevicauda?)* | | | | 5 | | | | 5 |
| *Megaptera novaeangliae* | | 64 | 51 | | | | | 115 |
| **Grand Total** | **11** | **99** | **311** | **6** | **72** | | | **499** |

## RESULTS

### Species Assignment

A total of 499 control region sequences were downloaded from Genbank. The number of sequences published varied substantially among species (Table 1). The sequences comprised six nominate species, assignable to a total of nine taxa (species and subspecies) using the WFTW reference sequences. Among these sequences, none have species identities which disagreed with that assigned by WFTW (Table 2). The inconsistencies in assignment were due either to the lack of a subspecies designation in the Genbank record (common minke whale), or to the inability of WFTW to distinguish between subspecies or well-marked forms (blue whale). In the following the results for each species are described in detail.

### Bowhead whale (*Balaena mysticetus*)

Of the 99 sequences labelled in Genbank as belonging to this species, 9 were assigned with moderate evidence, and 90 with strong evidence, to this species using the WFTW references.

### Common minke whale (*Balaenoptera acutorostrata*)

Of the 73 sequences labelled in Genbank as belonging to this species, all were assigned to one of the three subspecies, *acutorostrata*, *scammoni* and a dwarf form using the WFTW references (Table 3). The Genbank record reported the subspecies for only one of these sequences (EU285375). Consequently, all of the remaining sequences were scored as belonging to a different subspecies.

Table 3 The identities given by WFTW to the sequences labelled as *Balaenoptera acutorostrata*.

| *B. a. acutorostrata* | | *B. a. scammoni* | | *B. acutorostrata* dwarf |
|---|---|---|---|---|
| AP006468 | EF113882 | EF113824 | EF113842 | EF113860 |
| EF113859 | EF113883 | EF113825 | EF113843 | EF113862 |
| EF113861 | EF113884 | EF113826 | EF113844 | EF113863 |
| EF113870 | EF113885 | EF113827 | EF113845 | EF113864 |
| EF113871 | EF113886 | EF113828 | EF113846 | EF113865 |
| EF113872 | EF113887 | EF113829 | EF113847 | EF113866 |
| EF113873 | EF113888 | EF113830 | EF113848 | EF113867 |
| EF113874 | EF113889 | EF113831 | EF113849 | EF113868 |
| EF113875 | EF113890 | EF113832 | EF113850 | EF113869 |
| EF113876 | EF113891 | EF113833 | EF113851 | EU285375 |
| EF113877 | EF113892 | EF113834 | EF113852 | |
| EF113878 | EF113893 | EF113835 | EF113853 | |
| EF113879 | EF113894 | EF113836 | EF113854 | |
| EF113880 | EF113895 | EF113837 | EF113855 | |
| EF113881 | | EF113838 | EF113856 | |
| | | EF113839 | EF113857 | |
| | | EF113840 | EF113858 | |
| | | EF113841 | | |

There was strong evidence that every sequence came from the subspecies listed, with a single exception where the evidence was only moderate. The Genbank record reported the location where the specimen was collected for only one sequence (EU285375).

### Antarctic minke whale (*Balaenoptera bonaerensis*)

All of the 115 sequences labelled as this species in Genbank were assigned to this species with strong evidence using WFTW.

**Bryde's whale (*Balaenoptera brydei*)**

In the WFTW references, this species is indicated by the common name 'Bryde's (common)' associated with the species name *Balaenoptera edeni* (common form). Given this synonymy, of the 52 sequences labelled as *B. brydei* in Genbank, 20 were assigned with moderate evidence, and 32 with strong evidence, to this species using the WFTW references.

**Blue whale (*Balaenoptera musculus*)**

All of the 44 sequences labelled in Genbank as belonging to this species were unambiguously assigned to this species using the WFTW references. Twenty-eight sequences were assigned to the nominate subspecies. Five sequences (EU093927, EU093928, EU093929, EU093949, EU093951) were assigned to the pygmy blue whale labelled in WFTW as *Balaenoptera musculus (brevicauda?)* but only with moderate evidence to distinguish it from the nominate subspecies. The assignment of the remaining 11 sequences was ambiguous to the extent that it was not possible to determine whether they should be assigned to the nominate blue whale species or to the pygmy blue whale (Table 4). In every case the test sequence fell in a sister position relative to the clade containing both blue whales.

Table 4 The sequences labelled as *Balaenoptera musculus* which could not be assigned to either the nominate form or the pygmy form.

| | |
|---|---|
| EU093926 | EU093945 |
| EU093931 | EU093947 |
| EU093936 | EU093950 |
| EU093939 | EU093960 |
| EU093943 | EU093962 |
| EU093944 | |

**Humpback whale (*Megaptera novaeangliae*)**

Of the 115 sequences labelled in Genbank as belonging to this species, 64 were assigned with moderate evidence, and 51 with strong evidence, to this species using the WFTW references.

**Location**

Where possible, information on the geographical origin of each sequence was extracted from the Genbank records. Then, that location was compared with the source location of the reference sequence with the shortest genetic distance. Table 5 shows the results of the comparison.

Table 5 Agreement between the location information given in the Genbank record and that associated with the most similar reference sequence in WFTW.

| GENBANK ORG NAME | Disagree | Agree | no info | Grand Total |
|---|---|---|---|---|
| *Balaena mysticetus* | 4* | 95* | | 99 |
| *Balaenoptera acutorostrata* | | 2 | 72 | 74 |
| *Balaenoptera bonaerensis* | | | 115 | 115 |
| *Balaenoptera brydei* | | | 52 | 52 |
| *Balaenoptera musculus* | | | 44 | 44 |
| *Megaptera novaeangliae* | | | 115 | 115 |
| **Grand Total** | **4** | **97** | **398** | **499** |

* location inferred from publication title

**Bowhead whale (*Balaena mysticetus*)**

The geographical origin of these sequences was not directly recorded in Genbank but is indicated by the annotation:

| | |
|---|---|
| AUTHORS | Borge,T., Bachmann,L., Bjornstad,G. and Wiig,O. |
| TITLE | Genetic variation in Holocene bowhead whales from Svalbard |
| JOURNAL | Mol. Ecol. 16 (11), 2223-2235 (2007) |

Four of the sequences were most similar to WFTW reference sequences from the North Pacific and the remainder to reference sequences from the Atlantic Ocean.

**Common minke whale (*Balaenoptera acutorostrata*)**

Two sequences had their origins recorded, and they agreed with that for the most similar WFTW reference sequence. No location was recorded for the other sequences.

**Antarctic minke whale (*Balaenoptera bonaerensis*)**

No location was recorded in the Genbank records of these sequences.

**Bryde's whale (*Balaenoptera brydei*)**

No location was recorded in the Genbank records of these sequences.

**Blue whale (*Balaenoptera musculus*)**

No location was recorded in the Genbank records of these sequences.

**Humpback whale (*Megaptera novaeangliae*)**

No location was recorded in the Genbank records of these sequences. The location was hinted in the title of the overarching publication "Population structure of South Pacific humpback whales and the origin of the eastern Polynesian breeding grounds". Of these sequences, 48 were most similar to references from the North Atlantic and 67 were most similar to references from the North Pacific. However, the Genbank records do not report the sampling regime in this study. Consequently we cannot judge whether the suggested eastern Polynesian location is correct or not.

**Sequence Quality**

No IUPAC ambiguity codes were reported in any of the sequences indicating that by this measure all of the sequences were of high quality.

Only a small number of single-site gaps were observed when the downloaded sequences were aligned with the references. In some cases these gaps were due to insertions in one of the reference sequences. Give that no more than two such gaps were found in any alignment, this measure indicates that the sequences were of high quality.

Nineteen sequences had quality scores in the extreme 5% of the distribution for two different quality measures (Table 6). Four species were represented in this group. Although these sequences had relatively extreme quality measures, none of them were, in absolute terms, very different from the other sequences. Many of these, and other sequences, had genetic distances of about 0.05 from the closest reference sequence. In the alignments with the reference sequences, about 2% of the positions in these sequences did not match the references. Perhaps the most noteworthy sequences are DQ231170 and EU093939 which mismatched the references at 11 and 12 sites respectively. These mismatches include both SNPs and indels.

Table 6 Sequences with two extreme measures of sequence quality.

| Accession | Species Identity in Genbank Record |
|---|---|
| DQ231170 | *Balaenoptera brydei* |
| DQ768315 | *Megaptera novaeangliae* |
| DQ768329 | *Megaptera novaeangliae* |
| DQ768355 | *Megaptera novaeangliae* |
| DQ768395 | *Megaptera novaeangliae* |
| DQ768406 | *Megaptera novaeangliae* |
| DQ768409 | *Megaptera novaeangliae* |
| DQ768418 | *Megaptera novaeangliae* |
| EF068036 | *Balaenoptera brydei* |
| EF068046 | *Balaenoptera brydei* |
| EF068053 | *Balaenoptera brydei* |
| EF068058 | *Balaenoptera brydei* |
| EF113748 | *Balaenoptera bonaerensis* |
| EF113790 | *Balaenoptera bonaerensis* |
| EF113797 | *Balaenoptera bonaerensis* |
| EF113809 | *Balaenoptera bonaerensis* |
| EU093936 | Balaenoptera musculus |
| EU093939 | Balaenoptera musculus |
| EU093960 | Balaenoptera musculus |

## DISCUSSION

The accuracy with which one may assign specimens to species using genetic information is dependent upon both the degree of genetic differentiation of the species and the collation of a set of reference sequences from authoritatively identified specimens (Ross et al., 2008). Here the reference sequences in the Witness for the Whale dataset (Baker et al., 2003) were used in conjunction with a tree-based method to assign each of the 499 mysticete control region sequences, published in 2007, to a species or subspecies.

All of the sequences were assigned to the same species as that recorded in Genbank. Disagreements in assignments arose for two species. The common minke whale (*Balaenoptera acutorostrata*) sequences were not assigned to any of the now recognised subspecies, with one exception. In every case such an assignment could have been made unambiguously. For the blue whale (*Balaenoptera musculus*), the main issue appears to be whether the pygmy form can be distinguished using this gene region. Five sequences were assigned to the pygmy form, when they were not labelled as such and a further 11 could not be assigned to either the pygmy or the nominate forms.

Geographic information relating to the sampling location was reported for only two of the 499 sequences. If such information has been published, it has not for the most part been recorded in Genbank.

Overall the quality of the sequences appears to be high. Five different indicators of sequence quality were used, and in almost every case there was little or no indication of sequencing error. The sequences are derived from a small number of studies and laboratories which appear to have high standards of quality control. Also, there is no suggestion that sequences were derived from ancient or museum specimens, in which cases we might expect reduced sequence quality.

## APPENDIX

Archives (http://www.cebl.auckland.ac.nz/~hros001/cetaceanID) of the results, comprising tables of genetic distance and phylogenetic trees for each sequence analysed, and a summary spreadsheet are available.

## REFERENCES

Baker, C. S., M. L. Dalebout, S. Lavery, and H. A. Ross. 2003. www.DNA-surveillance: applied molecular taxonomy for species conservation and discovery. TRENDS in Ecology and Evolution 18:271-272.

Felsenstein, J. 1984. Distance methods for inferring phylogenies: A justification. Evolution 38:16-24.

Ross, H. A., G. M. Lento, M. L. Dalebout, M. Goode, G. Ewing, P. McLaren, A. G. Rodrigo, S. Lavery, and C. S. Baker. 2003. DNA Surveillance: Web-based molecular identification of whales, dolphins, and porpoises. Journal of Heredity 94:111-114.

Ross, H. A., S. Murugan, and W. L. S. Li. 2008. Testing the reliability of genetic methods of species identification via simulation. Systematic Biology 57:216-230.

Ross, H. A., and H. Shearman. 2008. Validation of mtDNA control-region sequences in GenBank for large baleen whales. International Whaling Commission, SC/60/SD6.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

Scientific Committee, I. W. C. 2008. Appendix N Report of the Working Group on DNA.