Validation of mtDNA control-region sequences in GenBank for large baleen whales

H. A. ROSS AND H. SHEARMAN

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

ABSTRACT

The phylogenetic 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 922 sequences from baleen whales published in Genbank prior to 2007. Of these, 42 sequences were identified as belonging to a different species, and 44 to a different subspecies, from that recorded in Genbank. Fourteen blue whale sequences could not be assigned to a subspecies. A species identity could not be assigned unambiguously to seven sequences. A small number of sequences had evidence of poor or unreliable quality, but in each case the species identity as recorded in Genbank was confirmed here. Taxonomic revision is probably the greatest source of disagreement in the identities given by Genbank and DNA Surveillance. To provide better validation of sample origin, all major geographic regions need to be represented for each species in the reference data sets.

INTRODUCTION

Each year large numbers of DNA sequences, mainly of the control-region of the mitochondrial genome (mtDNA) of baleen whales (Mysticeti) are deposited in Genbank (www.ncbi.nlm.nih.gov). These sequences are obtained by scientists throughout the world as part of various studies of the demographics and evolution of these animals. These sequences form the basis of subsequent research which informs the development of conservation policy. Two significant issues arise regarding the reliability of these sequences: the accuracy of the DNA sequencing and the accuracy of the species identification. Errors in DNA sequencing could inflate estimates of genetic distinctiveness while errors in identification erode the reliability of the data. Many of the baleen whales exhibit population structure, with demonstrable genetic differences among regional populations. The accuracy of the source population or location is important for the analysis of such population structure. If sequences have the sampling location recorded incorrectly then errors will arise in the identification of genetic differentiation among populations and the delineation of evolutionary significant units.

The DNA Surveillance (*http://www.dna-surveillance.auckland.ac.nz*) system is a Web-based program in which molecular genetic identification of specimens is achieved by a phylogenetic analysis (Ross et al., 2003). Curated reference sequences have been collated at this site for nearly all cetaceans by Baker and colleagues (2003). These sequences, obtained from expertly identified specimens, represent an authoritative reference collection known as Witness for the Whales (WFTW). Additionally the oceanic region where additional specimens were obtained has been recorded for many sequences. Identification is made by aligning a user-submitted gene sequence of unknown origin against this set of validated reference sequences, computing the evolutionary distances between the unknown and each of the reference sequences, and then building a phylogenetic tree to display the affinity of the unknown sequence with the reference sequences.

The DNA Surveillance system and the reference sequence databases have been used to assess the species identity of cetacean mtDNA sequences in Genbank (Ross and Murugan, 2006). Both sequences labeled as having a cetacean origin, and those with relatively high sequence similarity, were assessed. On the basis of the results from DNA Surveillance/WFTW, all of the 1429 sequences labeled as members of non-cetacean species were confirmed as being non-cetacean. Of the 1628 sequences labeled as cetacean, the species identity of only 25 was disputed by DNA Surveillance/WFTW. Additionally two sequences were unidentifiable, either by DNA Surveillance/WFTW or by BLAST This study confirmed the reliability of the DNA Surveillance system operating in conjunction with the WFTW reference sequence alignments.

The Working Group on DNA recommended in 2007 (Annex N) that work be undertaken to:

1. List the *GenBank* accession number and species identity of each mysticete control region

sequence 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.

3. An evaluation of the types of inconsistencies/errors as agreed by the Committee last year: quality of submitted sequences, accuracy of species identification and 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 reassess all of the mtDNA control region sequences from baleen whales in Genbank.

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 all control-region mtDNA sequences using the query string **'mysticeti[orgn] AND (d-loop OR control)'** and then the resulting hits were sorted by the year of publication. With minor exception this date corresponds to the year of submission, but its relationship to the year of sample collection or sequence determination is unpredictable. A total of 922 sequence records were identified and retrieved (Table 1). These records include not only the DNA sequence but also the species name, and in some cases information about the collection or sampling location.

Table 1. The number of control-region mtDNA sequences retrieved from Genbank. Sequences using the search string **'mysticeti[orgn] AND (d-loop OR control)'** and then sorted by year of publication.

Year	Number
2006	61
2005	143
2004	300
2003	59
2002	11
2001	159
2000	0
pre-2000	189
Total	922

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 contains 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 repeated using the 'All Cetaceans v3.1' reference alignment. If the query again fell within the toothed whale clade 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 4 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.

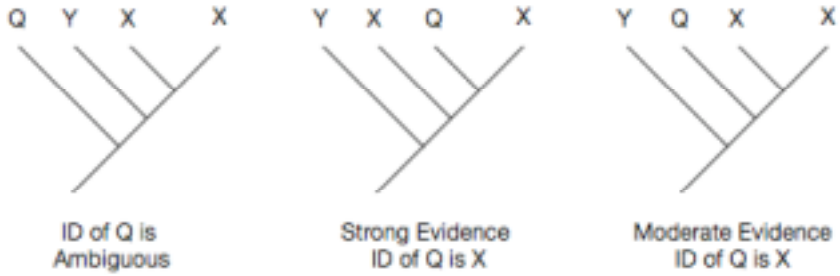

Figure 1. Strength of evidence in making an identification was based on relative position of query sequence to reference sequences.

Five measures of sequence quality were used:

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 2 The number of sequences from each mysticete species published in Genbank each year.

GENBANK ORG NAME	pre-2000	2000	2001	2002	2003	2004	2005	2006	Grand Total
Balaena mysticetus	1		68		4		5		78
Balaenoptera acutorostrata	33				31		5		69
Balaenoptera acutorostrata scammoni							1		1
Balaenoptera bonaerensis							3		3
Balaenoptera borealis	1				2		5		8
Balaenoptera brydei					1		3	53	57
Balaenoptera edeni	13		1		1		2	3	20
Balaenoptera musculus	2				2	13	5	1	23
Balaenoptera omurai					3			3	6
Balaenoptera physalus	56				11	1	16		84
Caperea marginata	1				2		2		5
Eschrichtius robustus	1		60		2	29	3		95
Eubalaena australis			16				2		18
Eubalaena glacialis	3		14			3			20
Eubalaena japonica							2		2
Megaptera novaeangliae	78			11		254	89	1	433
Grand Total	**189**	**0**	**159**	**11**	**59**	**300**	**143**	**61**	**922**

Table 3 The outcomes of comparing the species identification of each sequence, as given in Genbank, with that of 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		6	72					78
Balaenoptera acutorostrata	4			14	28	2	21	69
Balaenoptera acutorostrata scammoni		1						1
Balaenoptera bonaerensis			3					3
Balaenoptera borealis		3	5					8
Balaenoptera brydei		22	35					57
Balaenoptera edeni		1	10			1	8	20
Balaenoptera musculus	14		7	2				23
Balaenoptera omurai		6						6
Balaenoptera physalus	2	71	11					84
Caperea marginata		1	4					5
Eschrichtius robustus			95					95
Eubalaena australis			18					18
Eubalaena glacialis	1	4	5				10	20
Eubalaena japonica			2					2
Megaptera novaeangliae		56	377					433
Grand Total	**21**	**171**	**644**	**16**	**28**	**3**	**39**	**922**

Table 4 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>	21						1	22
Balaena mysticetus		6	72					78
Balaenoptera acutorostrata acutorostrata				8	22			30
Balaenoptera acutorostrata scammoni		1		2	2			5
Balaenoptera acutorostrata subsp. (dwarf)					1			1
Balaenoptera acuturostrata scammoni				4	3			7
Balaenoptera bonaerensis			3			2	20	25
Balaenoptera borealis		3	5					8
Balaenoptera edeni (common form)		23	45					68
Balaenoptera edeni (Kochi form)							7	7
Balaenoptera edeni (omurai)		6				1	1	8
Balaenoptera musculus			7					7
Balaenoptera musculus (brevicauda?)				2				2
Balaenoptera physalus		71	11					82
Caperea marginata		1	4					5
Eschrichtius robustus			95					95
Eubalaena australis			18				1	19
Eubalaena glacialis		4	5					9
Eubalaena japonica			2				9	11
Megaptera novaeangliae		56	377					433
Grand Total	**21**	**171**	**644**	**16**	**28**	**3**	**39**	**922**

RESULTS

Identification

A total of 922 control region sequences were downloaded from Genbank. The number of sequences published varied substantially each year (Table 1) and among species (Table 2). Among these sequences, 42 have species identities which disagree with those given by WFTW (Table 3). These inconsistencies are restricted to the minke whale, Bryde's whale and right whale groups. For a further 21 sequences, DNA Surveillance and WFTW gave ambiguous identities. In the following the results for each species are described in detail.

Bowhead whale (*Balaena mysticetus*)

A total of 78 sequences labelled as this species were found in Genbank. There was no uncertainty regarding the identification of any of them. None of these sequences had any information relating to their geographical origin.

Common minke whale (*Balaenoptera acutorostrata*)

Of the 69 sequences in Genbank with this species label, 64 were identified, based on the WFTW reference dataset, as belonging to two species, *B. acutorostrata* and *B. bonaerensis*, with *B. acutorostrata* divided in three subspecies, *acutorostrata*, *scammoni* and a dwarf form (Table 5).

Table 5 The identities given by WFTW to the sequences labelled as *Balaenoptera acutorostrata*.

B. a. acutorostrata		*B. a. scammoni*	*B. acutorostrata* dwarf	*B. bonaerensis*	
AF487467	AF487485	AJ226101	DQ145048	AJ226093	AJ226112
AF487468	AF487486	AJ226103		AJ226094	AJ226113
AF487469	AF487487	AJ226105		AJ226095	AJ226114
AF487470	AF487488	AJ226106		AJ226096	AJ226115
AF487471	AF487489	AJ226107		AJ226097	AJ226116
AF487472	AF487490	AJ226108		AJ226098	AJ226117
AF487474	AF487491	AJ226110		AJ226099	AJ226119
AF487475	AJ554054	AY822111		AJ226100	AJ226121
AF487476	AP006468	AY822112		AJ226102	AJ226122
AF487477	AY230267	DQ145040		AJ226104	M60408
AF487478	AY352278	Y17160		AJ226109	X87774
AF487479	AY352280				
AF487480	NC_005271				
AF487481	X72006				
AF487482	X87773				

For a single sequence (AJ226111) there was no evidence that it was derived from a mysticete when analysed with the All Cetaceans datasets, but when analysed using the Mysticete data set, or when submitted to NCBI's BLAST, there was strong evidence that it belonged to *B. bonaerensis*.

A further four sequences (AF487473, AF487483, AF487484, AY352279) were judged to have an ambiguous identity from their placement on a phylogenetic tree. Each sequence fell as sister to a mixed clade containing two or more species of minke whales. However, in each case the genetic distance was shortest between the test sequence and a *B. a. acutorostrata* reference.

Common minke whale (*Balaenoptera acutorostrata scammoni*)

The single sequence with this label had its identity confirmed by the reference sequences.

Antarctic minke whale (*Balaenoptera bonaerensis*)

Three sequences were labelled as this species. WFTW agreed with strong evidence.

Sei whale (*Balaenoptera borealis*)

Eight sequences were labelled as this species. WFTW agreed with both moderate and strong evidence.

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, all of the 57 sequences labelled as *B. brydei* were identified as belonging to this species, with a mixture of moderate and strong evidence.

Bryde's whale (*Balaenoptera edeni*)

The 20 sequences with this species label in Genbank appear to belong to three recognized species or forms. (Table 6) These are all labelled in WFTW as *B. edeni* but each has a different common name. Those identified as belonging to the Bryde's form are scored as correct identifications, while those identified as belonging to the other forms are scored as belonging to different species.

Table 6 The identities given by WFTW to the sequences labelled as *Balaeoptera edeni*.

Bryde's	Kochi	omurai
AF146381	AB116099	AF146389
AF146382	AB201258	AF398372
AF146383	AF146378	
AF146384	AF146379	
AF146385	AF146380	
AF146386	EF057443	
AF146387	NC_007938	
AF146388		
AY822091		
AY822092		
X72196		

Blue whale (*Balaenoptera musculus*)

Of the 23 sequences labelled in Genbank as belonging to this species, only 7 were unambiguously identified as such using the WFTW references. Two sequences (AY235201, AY822087) were identified as belonging to the pygmy blue whale labelled in WFTW as *Balaenoptera musculus (brevicauda?)*. The identity of the remaining sequences was ambiguous to the extent that it was not possible to determine whether they belonged to the nominate blue whale species or to the pygmy blue whale (Table 7). In every case the test sequence fell in a sister position relative to the clade containing both blue whales.

Table 7 The sequences labelled as *Balaenoptera musculus* which could not be identified as either the nominate form or the pygmy form.

AY390265	AY390272
AY390266	AY390273
AY390267	AY390274
AY390268	AY390275
AY390269	AY390276
AY390270	AY390277
AY390271	DQ145044

Omura's whale (*Balaenoptera omurai*)

In the WFTW references, this species is indicated by the common name 'Bryde's (omurai)' associated with the species name *Balaenoptera edeni* (omurai). Given this synonymy, all six sequences were given the same identity with moderate evidence.

Fin whale (*Balaenoptera physalus*)

Of the 84 sequences from this species in Genbank, the identity of all but two was confirmed using the WFTW reference dataset. Those two sequences (Y19111 and Y19112) were considered ambiguous because, in the first pass using the All Cetaceans data set, they fell within the clade of mysticete whales, but when assessed using the Mysticetes data set, they fell in an ambiguous position near the root. Both of these sequences are relatively short and include the tRNA-Pro region. When they were submitted to BLAST, we found that the top matched to other sequences from *B. physalus*, sourced from other institutions. The species identification appears to be correct, but it cannot be confirmed using WFTW.

Pygmy right whale (*Caperea marginata*)

The identities of all five sequences from this species were confirmed using WFTW.

Gray whale (*Eschrichtius robustus*)

The identities of all 95 sequences from this species were confirmed using WFTW.

Southern right whale (*Eubalaena australis*)

The identities of all 18 sequences from this species were confirmed using WFTW.

North Atlantic right whale (*Eubalaena glacialis*)

WFTW recognises all three *Eubalaena* species: *E. australis*, *E. glacialis* and *E. japonica* (Table 8). Of the 20 sequences from Genbank labeled as *E. glacialis*, nine have that identity confirmed by WFTW. A further nine are identified as belonging to *E. japonica* and one to *E. australis*.

Table 8 The identities given by WFTW to the sequences labelled as *Eubalaena glacialis*.

E. australis	*E. glacialis*	*E. japonica*
X72199	AF395039	AF275349
	AF395040	AF275350
	AF395041	AF275351
	AF395042	AF275352
	AF395043	AF275353
	AY395733	AF275354
	AY395734	AF275355
	U96647	AF275356
	U96648	AF275357

One remaining sequence (AY821863) has an ambiguous identity because it fell in a sister position to both *E. australis* and *E. japonica*. The sequence is relatively short (158bp) being from 16th century archaeological material. BLAST confirms the genus, but the top hits are to the sequences listed above and labelled as *E. glacialis* in Genbank. Given that the source material is European, and the lack of significant penetration of the Pacific Ocean by Europeans in the 16th century, the identification of this sequence as *E. glacialis* in the narrow sense seems justified.

North Pacific right whale (*Eubalaena japonica*)

The identities of the two sequences from this species were confirmed using WFTW.

Humpback whale (*Megaptera novaeangliae*)

The identities of all 433 sequences from this species were confirmed using WFTW. There were 56 sequences for which the evidence was only moderate, all but eight of which were published pre-2000. These sequences were not very long (261bp) which can result in ambiguity in the placement of the sequence in the phylogenetic tree. Seven of these sequences did not fall within the mysticete clade in the first pass using the All Cetaceans v4.3 data set, but they did when they were re-tested using the All Cetaceans v3.1 data set (Table 9). When subsequently tested using the Mysticete data set, their placement was similar to other sequences with similar length from this same era.

Table 9 Sequences having an ambiguous placement in a tree when aligned with the All Cetaceans v4.3 data set.

AF068067
AF068072
AF068077
AF068086
AF068094
AF068105
AF068112

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 10 shows the results of the comparison.

Table 10 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			78	78
Balaenoptera acutorostrata		31	38	69
Balaenoptera acutorostrata scammoni		1		1
Balaenoptera bonaerensis			3	3
Balaenoptera borealis			8	8
Balaenoptera brydei	1	1	55	57
Balaenoptera edeni	7	7	6	20
Balaenoptera musculus	1	1	21	23
Balaenoptera omurai			6	6
Balaenoptera physalus			84	84
Caperea marginata			5	5
Eschrichtius robustus			95	95
Eubalaena australis	7	3	8	18
Eubalaena glacialis			20	20
Eubalaena japonica			2	2
Megaptera novaeangliae	264	55	114	433
Grand Total	**280**	**99**	**543**	**922**

About two-thirds of the sequences had no geographical information recorded in the Genbank record. For the following species, it was possible to compare the recorded location with that of the most similar reference sequence.

Common minke whale (*Balaenoptera acutorostrata*)

Location information was available for about half of the sequences, and was not contradicted by WFTW.

Common minke whale (*Balaenoptera acutorostrata scammoni*)

The single sequence from this species (AY878077) was recorded as being from South Korea, and the closest reference was from the North Pacific Ocean.

Bryde's whale (*Balaenoptera brydei*)

One sequence labelled as *B. brydei*, and identified as a Bryde's (common) whale (AB201259) had its location recorded as 'Japan: Miyagi, Natori', which agreed with the location of the closest reference, from the North Pacific. The other sequence from this species (DQ340979) was reported to be from 'Atlantic Ocean: Canary Islands' but the closest reference was also from the North Pacific.

Bryde's whale (*Balaenoptera edeni*)

Of the species with this label in Genbank, and which subsequently were identified as Bryde's (Kochi), six of the seven had recorded locations which did not match that of the closest reference sequence. Sequences AB201258, AF146378, AF146379, AF146380 and NC_007938 were each recorded as being

from Japan, or 'Japan: Kumamoto, Minamata' but the closest references had ambiguous locations of 'NA' as the ocean basin and 'Japan coast' as the geographical source. This disagreement can simply be explained as typographic errors in the reference data set.

The last case is EF057443 whose location is recorded as India but with a closest reference recorded as from the NA ocean basin and geographical source being the Japan coast (as above). Two lines of evidence suggest that the location is correct. First, both BLAST and a multiple sequence alignment support the species identification. Second, the sequence was submitted by workers at the Central Marine Fisheries Research Institute, Cochin, Kerala, India. Presumably the specimen was sourced in the local region. There are no reference sequences for the Kochi form from the Indian Ocean, by which this can be tested further.

All seven of the sequences which were identified as Bryde's (common form) had recorded locations that agreed with those of the reference sequences.

Blue whale (*Balaenoptera musculus*)

There are two sequences from this species with locations recorded. The first (EF057441) is recorded from India, but the closest reference sequence is from the South Pacific. None of the reference sequences for this species is from the Indian Ocean. As for the preceding species, this sequence was submitted by workers at the Central Marine Fisheries Research Institute, Cochin, Kerala, India from presumably a local source. The identity of the species is not in doubt.

The second sequence (AY235201) was identified as 'blue whale (pygmy)' and its location ('New Caledonia: inshore waters, South West') agrees with that of the closest reference ('SP').

Southern right whale (*Eubalaena australis*)

Ten sequences had a location recorded as Argentina. Three sequences (AF395050, AF395051, AF395052) had closest matches to reference sequences from SO (Southern Ocean) and these were recorded as agreements. The remaining seven (AF395044, AF395045, AF395046, AF395047, AF395048, AF395049, AF395050, AF395051, AF395052, AF395053) were closest to reference sequences from the South Pacific.

Humpback whale (*Megaptera novaeangliae*)

The majority of sequences with a recorded location were from this species. Fifty-five sequences were recorded as from either 'USA: Prince William Sound' or 'USA: Shumagin Islands' and in each case this agreed with the location of the closest reference sequence (North Pacific).

A set of 11 sequences recorded as from specimens from Eastern Australia were closest to reference sequences from North Pacific (7) or North Atlantic (4). Because no reference sequences were from the Western Pacific Ocean/Tasmen Sea region, then the locations necessarily disagreed.

Locations were also recorded for a larger collection of 253 sequences, from two regions in the Antarctic and one near Brazil (Table 11). Of these, 76 were closest to the reference sequence from the North Atlantic and 177 were closest to the reference from the North Pacific. There are no reference sequences in WFTW for this species from the southern oceans.

Table 11 The number of sequences with recorded locations in the southern oceans which were most similar to reference sequences from the North Atlantic and North Pacific oceans.

	Closest Reference	
Recorded Location	**NA (North Atlantic)**	**NP (North Pacific)**
Antarctica: Area I	17	29
Antarctica: Area II	7	24
Brazil	52	124

Sequence Length and Confidence of Identification

Short sequences may contain less phylogenetic information than longer sequences, resulting in some ambiguity of identification. Table 12 summarizes the lengths of the sequences deposited in Genbank. Sequences with lengths greater than 1000 bp represent whole mitochondrial genomes.

The cases where uncertain or ambiguous identifications occurred are not restricted to very short sequences. Nevertheless, ambiguity did not occur with longer sequences.

Table 12 Frequency distribution of sequence lengths among the years. Size categories containing sequences for which there was uncertainty in making a species identification are outlined with a heavy line. Size categories outlined with a light line contain other cases of uncertainty, such as placement among the mysticetes or subspecies of blue whale. The number of ambiguous cases is given in parentheses.

Sequence Length	pre - 2000	2001	2002	2003	2004	2005	2006	Total
0 - 149	0	0	0	0	0	0	0	0
150 - 199	0	0	0	0	1 (1)	0	0	1
200 - 249	12	0	0	0	0	0	0	12
250 - 299	117 (7)	10	0	0	0	0	51	178
300 - 349	7 (2)	29	0	15	43 (14)	8	0	102
350 - 399	20	0	11	28 (4)	112	30	0	201
400 - 449	2	0	0	1	2	87	1	93
450 - 499	1	70	0	0	2	2	1	76
500 - 549	6	49	0	1	141	0	2	199
550 - 599	11	0	0	0	0	0	0	11
600 - 899	0	0	0	0	0	0	0	0
900 - 949	9	1	0	5	0	0	0	15
950 - 999	0	0	0	1	0	0	0	1
1000+	4	0	0	8	1	16	6	35

Sequence Quality

IUPAC ambiguity codes were found in 12 sequences from Genbank (Table 13). Most of these sequences had at most three sites with ambiguous nucleotides, but one sequence (AB116095) had eight sites where the nucleotide was completely unknown. For each of these sequences, its species identity as recorded in Genbank was confirmed using the WFTW references.

Table 13 Incidence of IUPAC ambiguous nucleotide codes in sequences.

Accession	Species Identity in Genbank Record	Type and Number of Ambiguous Nucleotides
AB116095	*Balaenoptera omurai*	N: 8
AF068071	*Megaptera novaeangliae*	N: 1
AF119961	*Balaenoptera physalus*	D: 1
AF119962	*Balaenoptera physalus*	N: 3
AF119964	*Balaenoptera physalus*	W: 1
AF119966	*Balaenoptera physalus*	Y: 1
AF119967	*Balaenoptera physalus*	H: 1
AF119968	*Balaenoptera physalus*	V: 1
AF119982	*Balaenoptera physalus*	N: 2
AF119983	*Balaenoptera physalus*	R: 1
AF119989	*Balaenoptera physalus*	S: 1
AF119996	*Balaenoptera physalus*	M: 1

Ten sequences had quality scores in the extreme 5% of the distribution for three different measures (Table 14). In one case (AF068105) there had been uncertainty in confirming that the sequence was from a mysticete, although the species identification matched that in Genbank, but in the other cases there was no uncertainty regarding species identity.

Table 14 Sequences with three extreme measures of sequence quality.

Accession	Species Identity in Genbank Record
AF068105	*Megaptera novaeangliae*
AF119962	*Balaenoptera physalus*
AF119994	*Balaenoptera physalus*
AF120004	*Balaenoptera physalus*
AF120005	*Balaenoptera physalus*
AF120006	*Balaenoptera physalus*
AY329960	*Megaptera novaeangliae*
AY329963	*Megaptera novaeangliae*
AY330094	*Megaptera novaeangliae*
AY390274	*Balaenoptera musculus*

There were seven sequences with an ambiguous species identity, excluding the blue whales (Tables 3). None of these sequences had any measure of sequence quality in the extreme 5% of the distribution.

Of the 42 sequences which were identified as belonging to a different species, only two (AB116099 and AB201258) had any measures of species quality in the extreme 5% of the distribution. In both cases they had three such extreme measures (Table 14).

DISCUSSION

Accurate species identification using genetic information depends upon many things. The procedure is based on the premise of the accumulation of genetic differences in parallel with speciation and other manifestations of evolutionary differentiation. The technique depends upon reference data sets comprising sequences derived from authoritatively identified specimens representing the naturally occurring genetic variation Uncertain or mistaken identification will arise when the unknown sequence is technically unreliable, when the taxonomy is imperfect or when demographic factors and evolutionary history have left species incompletely differentiated (Ross et al., 2008).

Of the 922 sequences, the species identity of only seven could not be determined unequivocally. The subspecies of an additional 14, all blue whales, could not be determined. The blue whale sequences involved were not unusually short. On the estimated phylogenetic trees including the query sequences, the blue whale subspecies were not reciprocally monophyletic. This suggests that more work is required to determine if there is a mtDNA region which reliably distinguishes these two forms of blue whale.

The significant proportion of the sequences labelled as *Balaenoptera acutorostrata* were identified as belonging to a different subspecies or species. This is because WFTW recognises the subspecies *B. a. acutorostrata*, when Genbank does not, and because there appears to have been a taxonomic revision, creating *B. a. scammoni* and *B. bonaerensis* which has not been reflected retrospectively in the sequence labels in Genbank. The sequences labelled as *Balaenoptera edeni* contained representatives of three species recognised by WFTW (the common form, Kochi form and omurai form) indicating that retrospective relabelling may be required. WFTW indicated that the sequences labelled as *Eubalaena glacialis* contained a mixture of species. This discrepancy may represent a combination of misidentification (*E. australis* X72199) and taxonomic revision (*E. japonica*). Overall there is little evidence for misidentification *per se*. Of greater significance appears to be the need for retrospective relabelling of sequences in Genbank following a taxonomic revision.

The phylogenetic methodology can be used to validate the origin of a specimen if there is sufficiently marked genetic differentiation among geographic populations and if that genetic variation is sampled. Most of the disagreements reported here regarding the sampling location arose because of limited geographic representation in the reference datasets. Haplotypes having a wide distribution are represented from a single location. When a disagreement occurred, it was difficult to determine whether a sampling site had been incorrectly reported or if the data set lacked sufficient geographic resolution. If sample location is to be validated using the WFTW references, then species-specific alignments containing representative haplotypes from all of the major populations will be needed.

The sample locations are recorded in Genbank records in the '/country' and '/note' sequence features. The 'country' feature is ambiguous in meaning, cannot record samples taken at sea and is meaningless in the case of countries with coasts on multiple oceans. The 'notes' feature is used for many other purposes. Perhaps a biogeographic database, such as OBIS (www.iobis.org/) is a more appropriate place to record sequence collection details.

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.

ACKNOWLEDGMENTS

Matthew Goode provided assistance with the computing.

REFERENCES

Baker, C. S., Dalebout, M. L., Lavery, S., and Ross, H. A. 2003. www.DNA-surveillance: applied molecular taxonomy for species conservation and discovery. *Trends Ecol. Evol.* 18:271-272.

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

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

Ross, H. A., and Murugan, S. 2006. Using phylogenetic analyses and reference datasets to validate the species identities of cetacean sequences in GenBank. *Mol. Phylogenet. Evol.* 40:866-871.

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

Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.