Journal of Heredity 2003:94(2):111–114
DOI: 10.1093/jhered/esg027

DNA Surveillance: Web-Based Molecular Identification of Whales, Dolphins, and Porpoises

H. A. Ross, G. M. Lento, M. L. Dalebout, M. Goode, G. Ewing, P. McLaren, A. G. Rodrigo, S. Lavery, and C. S. Baker

From the School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand. Development of the site is funded by the University of Auckland Vice Chancellor's Development Fund and by the International Fund for Animal Welfare.

Address correspondence to H. A. Ross at the address above, or e-mail: h.ross@auckland.ac.nz.

Abstract

DNA Surveillance is a Web-based application that assists in the identification of the species and population of unknown specimens by aligning user-submitted DNA sequences with a validated and curated data set of reference sequences. Phylogenetic analyses are performed and results are returned in tree and table format summarizing the evolutionary distances between the query and reference sequences. DNA Surveillance is implemented with mitochondrial DNA (mtDNA) control region sequences representing the majority of recognized cetacean species. Extensions of the system to include other gene loci and taxa are planned. The service, including instructions and sample data, is available at http://www.dna-surveillance.auckland.ac.nz.

Molecular systematic approaches can augment, or even replace, traditional morphology-based techniques for identification of vertebrate species (e.g., Baker and Palumbi 1994; Dalebout et al. 1998; DeSalle and Birstein 1996; Dizon et al. 2000; Henshaw et al. 1997; Roca et al. 2001). These techniques are especially pertinent to biosurveillance: the identification of animals and animal by-products in the context of conservation, wildlife management, and law enforcement.

DNA Surveillance is a service for the application of phylogenetic methods to the identification of species within a particular taxonomic group, such as the currently implemented data set for whales, dolphins, and porpoises (Order: Cetacea). Phylogenetic techniques are implemented in a Web-based program that aligns a user-submitted gene sequence of unknown origin against a set of validated reference sequences, computes the evolutionary distances between the unknown and each of the reference sequences, and then builds a phylogenetic tree to display the affinity of the unknown sequence with the reference sequences. It is important that users of this service ensure that any submitted DNA sequences are derived from a member of the specified taxonomic group. Otherwise, results of the phylogenetic identification could be misleading.

The DNA Surveillance software and the reference data sets have been developed specifically for taxonomic identification. This approach differs from a standard BLAST search of GenBank (Altschul et al. 1990). GenBank entries are not curated and can suffer from species or population misidentification, missing information, and inconsistent terminology. Inconsistent application of keywords also reduces the power of searching GenBank by fields, impeding effective data mining. The reference sequences in DNA Surveillance are prealigned, using a mixture of algorithmic and manual methods, to create an optimized alignment; however, BLAST and related search engines seek locally maximal matches in pairwise comparisons. The taxonomic distribution of sequences in GenBank reflects the sampling protocols of individual research programs rather than phylogenetic diversity. The sequences in DNA Surveillance reflect species and population diversity. The extreme (E) value associated with each sequence hit in a BLAST search is not a rigorous measure of evolutionary distance or genetic similarity, and depends on the size of the database being searched (Karlin and Altschul 1990). The genetic distances and trees in DNA Surveillance are calculated using standard phylogenetic algorithms, as implemented in the Phylogenetic Algorithms Library (Drummond and Strimmer 2001). The model of evolution and model parameters have been chosen for the phylogenetic analyses to provide good discrimination among species; this site does not attempt to provide an

Table 1. Primers used at the University of Auckland in sequencing the mtDNA control region of cetacean species

Primer name	Primer sequence
M13-Dlp1.5-L[a]	5′-TGTAAAACGACGGCCAGTTCACCCAAAGCTGRARTTCTA-3′
Dlp5-H[a]	5′-CCATCGWGATGTCTTATTTAAGRGGAA-3′
Dlp8G-H[b]	5′-GGAGTACTATGTCCTGAACA-3′
Dlp4-H[c]	5′-GCGGGWTRYTGRTTTCACG-3′
Dlp10-L[c]	5′-CCACAGTACTATGTCCGTATT-3′

[a] Dalebout et al. (1998).

[b] Lento et al. (1998) and Pichler et al. (2001).

[c] Dalebout (2002).

accurate estimate of the phylogenetic relationships among the higher-order taxa within the group of interest.

The problems associated with using GenBank for species identification are particularly pertinent to cetaceans, as the diagnostic morphological features used to distinguish species can be subtle, and the deposited sequences are usually not associated with identifiable reference material (e.g., as with skin biopsy samples obtained from free-swimming animals). The reference data sets comprise sequences from the highly variable mitochondrial DNA (mtDNA) control region, which has proven to be an effective tool for the species identification of test specimens and for differentiating intra- and interspecific relationships among closely related cetacean species (Baker et al. 1996; Dalebout 2002; Dalebout et al. 1998, 2002). Reference sequences have been selected to reflect the generic, specific, or geographic diversity observed at a taxonomic level and to maximize the discriminatory power of the analysis. Sequences were included only if the specimen had been expertly identified and diagnostic skeletal material or photographic records were collected (Dizon et al. 2000). Sequences were either retrieved from GenBank or were analyzed at the University of Auckland using the primers indicated in Table 1 and Figure 1.

The reference data sets consist of a total of 121 sequences and provide coverage of the taxonomic and geographic diversity of cetaceans (67 of the 81 recognized species, in 11 of 14 families; Table 2). A list of the species represented in each data set is available on the website. Data sets are arranged in hierarchical order, allowing initial family-level identification of cetaceans, and subsequently more detailed analysis within the suborders Mysticeti (baleen whales) and Odontoceti (toothed whales). The diverse (20+ species) odontocete family Ziphiidae (beaked whales) is represented by a comprehensive validated data set. The phylogenetic trees created are rooted using an appropriate outgroup: the sperm whale (*Physeter macrocephalus*) for the mysticete, odontocete, and general cetacean reference data sets, and the pygmy sperm whale (*Kogia breviceps*) for the ziphiid reference data set. The latter outgroup was chosen to reduce outgroup branch length in the resultant trees. DNA Surveillance also supports phylogeographic searches: a reference set of sequences from humpback whale (*Megaptera novaeangliae*) populations allows the identification of the geographic origin of samples from this species. Fin whale (*Balaenoptera physalus*) and blue whale (*Balaenoptera musculus*) are used as outgroups for this reference data set.

In a typical analysis, the user pastes a DNA sequence (in FASTA or text format) into a data input window and chooses the appropriate reference data set. A standardized phylogenetic analysis is then performed, using parameter values and a model of evolution which have proven effective in species identification. The model and parameters used are the simplest required to provide the level of discrimination required to differentiate species identity. The query sequence

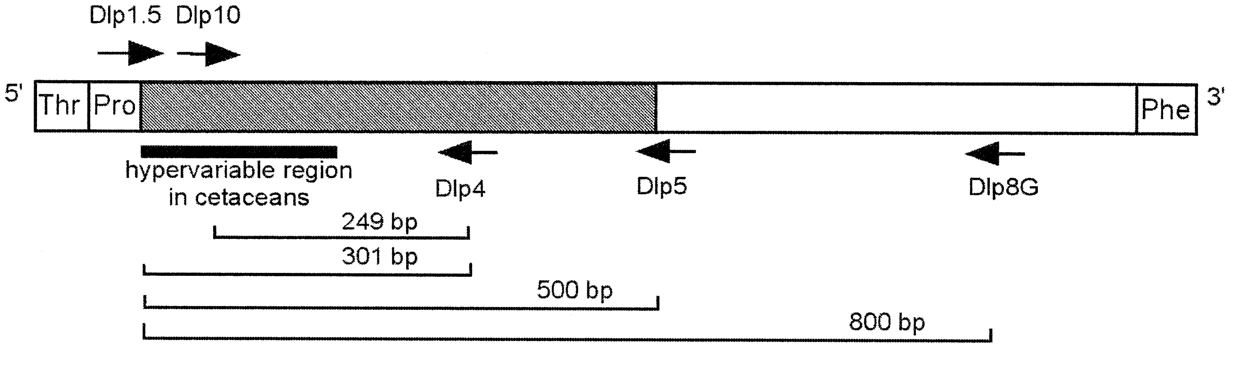

Figure 1. A schematic map of the mtDNA control region and the binding sites and orientation of the primers used in sequencing cetacean DNA. The shaded region represents the portion of the control region covered by most sequences in the reference data sets. Position 1 of the control region alignment corresponds to position 15891 of the fin whale (*Balaenoptera physalus*) mtDNA genome (Arnason et al. 1991).

Table 2. Representation of cetacean families in the reference data sets based on mtDNA control region sequences

Suborder	Family	No. of species represented/in family	No. of sequences
Mysticeti	Balaenidae (right whales)[a]	1/3	2
	Neobalaenidae (pygmy right whale)	1/1	2
	Eschrichtiidae (gray whales)	1/1	2
	Balaenopteridae (rorquals)	7/8	17
Odontoceti[b]	Physeteridae (sperm whale)	1/1	2
	Kogiidae (pygmy sperm whales)	2/2	3
	Ziphiidae (beaked whales)[c]	20/20	39
	Pontoporiidae (La Plata dolphin)	1/1	2
	Monodontidae (beluga and narwhal)	2/2	4
	Delphinidae (oceanic dolphins)	25/36	39
	Phocoenidae (porpoises)	6/6	11
Total		67/81	121

Geographically diverse species are represented by multiple sequences. The classification of species follows Rice (1998).

[a] Additional sequences from other species will be added shortly.

[b] Missing are river dolphin families Platanistidae, Iniidae, and Lipotidae.

[c] New species (Dalebout et al. 2002) to be added shortly.

is aligned by a simple profile alignment (Gribskov and Veretnik 1996; Gribskov et al. 1987, 1990) against the prealigned data set of reference sequences using the penalty values: transitions = 1, transversions = 2, gap creation = 3, and gap extension = 1. The evolutionary distances among all of the aligned sequences are calculated using the F84 model of evolution (Felsenstein 1984; Kishino and Hasegawa 1989), with a transition/transversion ratio of 2 and equilibrium nucleotide frequencies as calculated empirically across all reference sequences. A neighbor-joining (NJ) tree is built from the table of evolutionary distances (Saitou and Nei 1987) and rooted using an outgroup appropriate for each data set. Negative branch lengths are allowed in the estimation of the tree, and then they are set to zero for purposes of presentation. The phylogenetic tree, in both graphic and Newick text format, and a table of distances are displayed and can be downloaded to disk. Taxa are color-coded at an appropriate taxonomic level in the graphic format of the tree to assist the user in judging taxonomic affinities. An optional bootstrap analysis using 100, 500, or 1,000 pseudoreplicates (Felsenstein 1986) can be performed to assess the robustness of the resulting phylogenetic tree. Resulting bootstrap scores $\geq$50% are displayed on the relevant nodes of the NJ tree. Also, the user can choose that a computationally more intensive full alignment of the query and reference sequences be performed as part of the analyses.

The reliability of DNA Surveillance was tested by submitting an unaligned copy of each reference sequence as a query sequence to each data set in which it occurs. The test was judged a success (1) if the query sequence was the shortest evolutionary distance to a member of the same taxon and (2) if it was monophyletic with respect to the other sequences of the same taxon. The relevant taxon for comparison was the family for the cetacean database, the species for the mysticete, odontocete, and ziphiid databases, and the population for the humpback database. DNA Surveillance correctly identified the taxon for 100% of the sequences in the cetacean, mysticete, ziphiid, and humpback databases, and for 90% of the sequences in the odontocete database. In six cases, mostly members of the porpoises (Phocoenidae), one or both of the criteria for successful identification were not met. Deviations in the placement of a sequence could occur as a result of the profile alignment differing from the original, manually adjusted alignment. The homology of many individual nucleotide sites in porpoise DNA sequences is problematic and multiple alignments are plausible. The ability to perform a full alignment of the query and reference sequences or a bootstrap analysis are provided as advanced search options that can overcome such uncertainty. Nevertheless, some taxa are naturally weakly differentiated and, as indicated in the online documentation, the user must employ other evidence in determining the species identity. Discriminating power should increase with the use of reference databases of a narrower taxonomic scope, such as that for the ziphiids or greater geographic representation, such as for the humpback.

Because DNA Surveillance has been implemented with mtDNA control region sequences from cetaceans, unreliable or misleading species identification could arise if the query sequence is from a noncetacean, or from a different gene locus. DNA Surveillance will necessarily attempt to align the submitted sequence with the chosen reference database and then to perform a phylogenetic analysis on that alignment. To reduce the chance of misidentification, a warning message is displayed if the divergence of the query sequence is judged to be outside the range found among cetacean species for this locus or if the length of the sequence is insufficient for a confident match. The current criteria for this warning are (1) if the query sequence is less than 60% of the length of the shortest sequence in the reference database, or (2) if the average cost, using the penalty values given above, of aligning the query sequence with each of the reference sequences is more than 25% greater than the highest average cost of aligning each reference sequence with each of the other reference sequences. In the second case, only that section of the query sequence which overlaps the reference sequences is considered. If there are multiple query sequences, then a warning is issued on the basis of the poorest-matching sequence. Users of DNA Surveillance are strongly encouraged to perform a BLAST search against GenBank to exclude the possibility that the origin of the sequence is noncetacean.

Reliance on the topology of the phylogenetic tree in identification of the sperm whale (*P. macrocephalus*) is problematic at present. Given its distinctiveness, this species is used as the outgroup for trees constructed using the cetacean, mysticete, and odontocete databases. Our tree-building algorithm forces all sequences except the outgroup into a monophyletic clade. This has the consequence of separating a submitted sperm whale sequence from the reference sperm whale sequences. A family-level data set for the Physeteridae, which is under development, should solve this problem. At present, however, identity of sperm whale sequences can be evaluated using the evolutionary distances.

In recognition of the proprietary nature of some gene sequences, user-submitted query sequences are neither captured nor stored, except in temporary caching. DNA Surveillance can also protect the privacy of reference data sets while allowing their use for identification. Details of reference sequences are revealed to users at the discretion of the data administrator or owner.

Note that while the locus (mtDNA control region) and method of analysis (NJ tree) used were chosen specifically to address questions of species or population identity, they may not be as well suited to the robust reconstruction of higher-level relationships among more distantly related cetacean species. Some of the higher-level relationships suggested by DNA Surveillance are not well supported by bootstrap simulation and should not be considered an accurate reflection of the evolutionary relationships among these taxa (e.g., for the family Ziphiidae, in which reconstructions suggest that the genus *Mesoplodon* is not monophyletic; see Dalebout [2002] for further discussion regarding higher-level relationships in this family).

Anticipated developments in DNA Surveillance include (1) data sets of other taxonomically informative cetacean gene sequences (i.e., mtDNA cytochrome *b*), (2) additional family- and species-level reference data sets, (3) tools for better delegated administration of reference data sets, and (4) improved statistical confirmation of species identification through the use of maximum likelihood. We invite experts on such taxonomic groups as carnivores, marine and freshwater turtles, commercially valuable fish, and sharks to contact us to explore implementation of DNA Surveillance for these taxa.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Arnason U, Gullbedr A, and Widegren B, 1991. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. J Mol Evol 33:556–568.

Baker CS, Cipriano F, and Palumbi SR, 1996. Molecular genetic identification of whale and dolphin products from commercial markets in Korea and Japan. Mol Ecol 5:671–685.

Baker CS and Palumbi SR, 1994. Which whales are hunted? A molecular genetic approach to monitoring whaling. Science 265:1538–1539.

Dalebout ML, 2002. Species identity, genetic diversity and molecular systematic relationships among the Ziphiidae (beaked whales) (PhD dissertation). Auckland, New Zealand: University of Auckland.

Dalebout ML, Mead JG, Baker CS, Baker AN, and Van Helden AL, 2002. A new species of beaked whale *Mesoplodon perrini* sp. n. (Cetacea: Ziphiidae) discovered through phylogenetic analyses of mitochondrial DNA sequences. Mar Mamm Sci 18:577–608.

Dalebout ML, Van Helden A, Van Waerebeek K, and Baker CS, 1998. Molecular genetic identification of southern hemisphere beaked whales (Cetacea: Ziphiidae). Mol Ecol 7:687–694.

DeSalle R and Birstein VJ, 1996. PRC identification of black caviar. Nature 381:197–198.

Dizon A, Baker CS, Cipriano F, Lento G, Palsboll P, and Reeves R, 2000. Molecular genetic identification of whales, dolphins and porpoises: proceedings of a workshop on the forensic use of molecular techniques to identify wildlife products in the market place. NOAA Technical Memorandum NMFS NOAA-TM-NMFS-SWFSC-286. La Jolla, CA: National Oceanic and Atmospheric Administration.

Drummond A and Strimmer K, 2001. PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics 17:662–663.

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

Felsenstein J, 1986. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Gribskov M, Lüthy R, and Eisenberg D, 1990. Profile analysis. Meth Enzymol 183:146–159.

Gribskov M, McLachlan AD, and Eisenberg D, 1987. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84:4355–4358.

Gribskov M and Veretnik S, 1996. Identification of sequence patterns with profile analysis. Meth Enzymol 266:198–212.

Henshaw MD, LeDuc RG, Chivers SJ, and Dizon AE, 1997. Identification of beaked whales (family Ziphiidae) using mtDNA sequences. Mar Mamm Sci 13:487–495.

Karlin S and Altschul SF, 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268.

Kishino H and Hasegawa M, 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179.

Lento GM, Dalebout ML, and Baker CS, 1998. Species and individual identification of whale and dolphin products for sale in Japan by mtDNA sequences and nuclear microsatellite profiles. SC/50/08. Cambridge: Scientific Committee of the International Whaling Commission.

Pichler FB, Dalebout ML and Baker CS, 2001. Non-destructive DNA extraction from sperm whale teeth and scrimshaw. Mol Ecol Notes 1: 106–109.

Rice DW, 1998. Marine mammals of the world: systematics and distribution. Special Publication no. 4. Lawrence, KS: Society for Marine Mammalogy.

Roca AL, Georgiadis N, Pecon-Slattery J, and O'Brien SJ, 2001. Genetic evidence for two species of elephant in Africa. Science 293:1473–1475.

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

Received June 27, 2002
Accepted January 7, 2003

Corresponding Editor: Stephen J. O'Brien