SC/66a/SD/2

Oceanic origins of fin whale products sold in the Japanese market: evidence of an illegal, unreported or undocumented source in the Antarctic

C.S. Baker, D. Steel, N. Funahashi and Frederick Archer

INTERNATIONAL
WHALING COMMISSION

FOR CONSIDERATION BY THE SCIENTIFIC COMMITTEE OF
THE INTERNATIONAL WHALING COMMISSION
SAN DIEGO, USA, 2015

Oceanic origins of fin whale products sold in the Japanese market: evidence of an illegal, unreported or undocumented source in the Antarctic

C.S. BAKER[1], D. STEEL[1], N. FUNAHASHI[2] and FREDERICK ARCHER[3]

1*Marine Mammal Institute, Oregon State University, Newport, Oregon 97365 USA*
2 *International Fund for Animal Welfare, C+ 6F Nishi Shinjuku Well Bldg.,5-24-16 Nishi Shinjuku, Shinjuku-ku, Tokyo, Japan*
3*Southwest Fisheries Science Center, 8901 La Jolla Shores Drive, La Jolla, CA 92037 USA*

ABSTRACT

We report on the inferred oceanic origins of 113 fin whale products purchased in Japanese markets from 1993 to 2009. For this, we used Random Forest, a classification algorithm, based on the largest reference dataset assembled to date for mtDNA sequences of fin whales from the three ocean basins: North Atlantic (n=332, haplotypes=35), North Pacific (n=346, haplotypes=31) and the Southern Hemisphere (n=99, haplotypes=48). The Random Forest model had a high accuracy of classification of reference sequences to known origin (97% correct). We then used this model to classify the mtDNA sequences of the market products to their most likely ocean of origin. We expected the classification of the market products to reflect three oceanic sources, depending on date of purchase: the North Atlantic for special permit hunting by Iceland, which ended in 1989, and commercial whaling by Iceland in 2006 and 2009, with importation after 2008; the Southern Hemisphere for special permit hunting by Japan in the Antarctic (JARPAII), initiated in 2005/06 season; and, the North Pacific for bycatch in Japanese coastal waters.

Of the 44 haplotypes found among the 113 products, 16 were classified as North Atlantic and the remaining 28 as Southern Hemisphere. Most of the products represented by the 16 North Atlantic haplotypes were purchased from 1993 to 1999, roughly consistent with the reported 10-year maximum for storage of products from the Icelandic scientific whaling. Products represented by 19 of the Southern Hemisphere haplotypes were purchased after the hunting of fin whales in the Antarctic was initiated in the austral season of 2005/06. As reported previously (Steel et al. SC/61/BC8), these 19 haplotypes represented at least 19 individuals, exceeding the 15 reported as either JARPAII or as bycatch. Furthermore, products represented by 10 of the Southern Hemisphere haplotypes were purchased before the addition of this species to the JARPAII programme, some dating back to as early as 1993.

Regardless of the inferred oceanic origins, it is difficult to explain the sale of fin whale products from 2000-05, prior to the 2006 JARPAII hunt and after the 10-year storage limit for products from Iceland. Commercial whaling for this species ended in the North Pacific in 1976 and in Antarctic in 1976/77. Taken together with the results of the random forest classification, evidence points to an illegal, unreported or undocumented (IUU) source of fin whales from the Antarctic. Information on the mtDNA haplotypes from the Japanese and Icelandic DNA register would provide greater confidence in excluding bycatch and special permit whaling as sources of these questionable products.

INTRODUCTION

There are a limited number of legitimate sources for fin whale products sold in the Japanese markets. Products from frozen stockpiles of fin whales killed before the 1986 moratorium, or taken in subsequent scientific hunting by Iceland until 1989, were presumably depleted or expired after 10-years, the maximum period of reported storage (CITES, 1997). In July 2001, Japanese regulations were changed to permit the killing and selling of whales that had been accidentally caught in coastal fishing set nets (Anon. 2001). This could be a source of products from North Pacific fin whales, although this species is only infrequently reported as coastal bycatch during the survey period. In the austral season of 2005/06, Japan included fin whales in the species killed under special permit in the Antarctic (JARPA II). The renewal of whale-meat trade between Iceland and Japan, starting in 2008, is a potential source of products from North Atlantic fin whales.

In Baker et al. (2007), we first presented phylogenetic evidence that some fin whale products sold in Japan prior to the scientific hunt of this species in the Antarctic in 2005/06 did not originate from the Icelandic scientific hunt in the North Atlantic, the last legal source of fin whales. By comparing these questionable products with those purchased post-JARPA, we suggested an Illegal, Unreported or Unregulated (IUU) source in the Antarctic or the North Pacific. At that time, however, there were no publicly available reference sequences to characterize the mtDNA diversity of fin whales in the Southern Hemisphere. In Baker et al. (2008) and Steel et al. (2009), we presented further evidence from genotyping of fin whale products purchased from 2006 to 2009 (post-JARPA II). The number of individual fin whales identified by the genotyping was greater than the number reported in the scientific catch or bycatch, suggesting under-reporting of the total numbers of fin whales taken.

Here we report on the inferred oceanic origins of fin whale products purchased in Japan between 1993 and April 2009. For this, we assembled the largest reference dataset assembled to date for mtDNA sequences of fin whales from the three ocean basins (Archer *et al.* 2013; Berubé *et al.* 1998; Sremba *et al.* 2015). We then used the ensemble algorithm, Random Forest (Breiman, 2001) to classify the mtDNA sequences of the market products to their most likely ocean of origin. We also reviewed and corrected the genotype identification of fin whale products and compare mtDNA haplotype identity to that reported from JARPA and JARPA II (Goto *et al.* 2014). The results address previous limitations regarding phylogenetic characterization of oceanic stocks and support our previous inference of an Illegal, Unreported or Unregulated (IUU) source of fin whales from the Antarctic.

MATERIALS AND METHODS

Market surveys and species identification

Fin whale products were identified during surveys of Japanese whale meat markets and other outlets conducted from 1993 to 2009, as described previously (Baker *et al.* 2007; Baker *et al.* 2008; Steel *et al.* 2009). For all surveys DNA extractions from cetacean tissue and subsequent amplifications of the mtDNA control region were conducted on site. All amplified products were then isolated and washed free of the original DNA template before transportation to our home laboratory (Baker and Palumbi 1994). Species identification analysis was based on comparisons of the mtDNA sequences from the market products and reference sequence of known species and geographic origins using the phylogenetic methods described and implemented in *DNA-surveillance* (Ross *et al.* 2003).

Following the inclusion of fin whales in the JARPAII hunt (2005/06), we made a directed effort to sample exhaustively for fin whales. Analysis of these products included microsatellite genotypes and sex, sufficient for identification of individuals sharing the same haplotype (Steel *et al.* 2009).

Individual identification and reported catches of fin whales from 2006-09

We reviewed and corrected the genotype identification and matching of individual fin whale products on the market from 2006 to 2009, as described in (Steel *et al.* 2009). We then compared that to the expected number of individuals from reported catches and bycatch, based on a reviewed annual progress reports to the IWC and the Marine Mammal Stranding Data Base (The National Science Museum, http://svrsh1.kahaku.go.jp/). The JARPA II programme reported killing 10 fin whales (4 males: 6 females) during the 2005/06 season, 3 fin whales during the 2006/07 season (1 males: 2

females) and no fin whales during the 2007/08 season. From 2005 to 2009, there were 7 fin whales reportedly 'stranded' (i.e., found dead as beachcast or on bow of a ship) or taken as bycatch. The five 'stranded' fin whales seem to have been disposed of rather than released to market (22 Jan 2005, Okinawa, 2 Feb 2005, Akita, 19 Mar 2006, Nagasaki, 12 Jan 2009, Nagasaki, 1 Mar 2009, Hokkaido). The two taken as bycatch were reportedly released to market. One was found in a set net alive and then died (or was killed) on 17 Dec 2007, in Iwate. The second also found alive then dead in a set net on 27 Dec 2008, in Wakayama.

It is our understanding that fin whale products imported from Iceland were not released from Japanese customs until October 2008. Only three fin whale products (representing two individuals) in our survey were purchased after this date. One was labeled in store as 'Antarctic fin whale meat' and the first product purchased from the other individual was advertised as from the Wakayama bycatch. Consequently, we expected to find no more than 15 market individuals among products purchased from 2006 to 2009.

Random forest classification and reference datasets

Reference control region sequences were obtained for fin whales from the North Pacific, North Atlantic, and Southern Hemisphere, as described by Berube et al (1998), Archer et al (2013), Sremba et al (2014). Sequences from the market samples were aligned with reference sequences using MAFFT (Katoh et al 2002) and then inspected by eye. Sites with leading or trailing ambiguous bases (Ns) in any sequences were excluded from all sequences. Individual sequences were then collapsed to unique haplotypes.

We used Random Forest (Breiman 2001, Berk 2006, Archer et al. In Review) to create a classification model based on the reference haplotypes using custom code written in R (R Core Team 2014) and the *randomForest* package (Liaw and Wiener 2002). In order to avoid biases in classification due to differences in sample size (Archer et al. In Review), each tree in the forest was grown with a balanced reference set, where the sample size for each stratum was set to the sample size of the smallest stratum, and sampling was done with replacement. A total of 10,000 trees were grown and the remaining parameters in *randomForest* were left to their defaults. Haplotypes from market samples were then classified to the stratum with the greatest proportion of votes in the forest.

RESULTS AND DISCUSSION

Species identification of market products, 1993-2009

From the review of our market surveys, we retrieved mtDNA control region sequences (haplotypes) from a total of 113 fin whale products purchased on Japanese markets from 1993 to 2009 (Table 1). Of these, 33 products were purchased before the release of any products from the JARPA II hunt of fin whales, initiated in the 2005/06 Antarctic season, and 80 product were purchased after this release. We did not include 1 fin whale product purchased from a specialty restaurant in Seoul, South Korea, in 2009, as this was shown by microsatellite genotyping and mtDNA sequencing to be a match to products purchased on the Japanese market (i.e., a replicate sample, (Baker *et al.* 2010). Our summary alignment showed that sequences from the 113 market products represented 44 unique haplotypes based on a 418 bp consensus length.

Individual identification of fin whales, 2006-09

As reported previously (Steel et al. 2009), the 80 fin whale products purchased after the JARPA II hunt were genotyped for individual identification. In our review of these genotypes, we corrected one genotype, reducing the number of individuals in the surveys from 20 to 19. Each of these 19 individuals was represented by a unique haplotype (Table 2), presumably reflecting the high mtDNA diversity in the fin whales of the Southern Hemisphere (Archer *et al.* 2013; Goto *et al.* 2014).

Even with this correction, however, the 19 individual fin whales represented in the market products exceeded the expectation of 15 fin whales from reported sources (Table 2). Of the 12 individual fin whales represented in the 2008-09 survey, 5 were represented in the 2006 survey and these same 5 were also found in the 2007 survey. An additional 2 of the 2008-09 individuals were represented in the 2007 survey, giving a minimum census of 19 market individuals (6 males: 13 females). One of the 12 haplotypes from the 2008-09 survey matched to a haplotype found on the market in 1999 and again in 2000 (see below).

Oceanic origins by Random Forest classification

The reference dataset for the fin whales is the largest yet assembled for this species, with mtDNA control region sequences from 777 individuals. This included the global sample reported in Archer et al. (2013) as well as additional samples from the North Atlantic, as published Berube et al. (1998) and additional samples from bones at the South Georgia whaling station as reported by Sremba et al. (2015). The assembled reference dataset did not include the samples from Antarctic fin whales (catch and biopsy samples) collected by the JARPA and JARPAII program (Goto *et al.* 2014), although samples from the JARPAII catches are assumed to be included in the market products purchased in 2006 and later. Because of differences in the length of sequences from the three studies, the consensus length of the aligned reference dataset was limited to 230 bp. However, even with this short fragment, the variation was sufficient to resolve 113 haplotypes among the 777 individuals. No haplotype was shared between the North Atlantic and other oceans and only one haplotype was shared between the North Pacific and Southern Hemisphere (Table 3).

The Random Forest model constructed with the reference dataset had very low classification error rates (3% overall) for the three ocean basins (Table 4). None of the Southern Hemisphere reference samples were misclassified, while only two of the 332 North Atlantic samples (0.6%) were misclassified. The North Pacific had the highest misclassification error, with 21 of the 346 north Pacific samples (6%) assigned to the Southern Hemisphere. This is presumably due to the close phylogenetic relationship of the North Pacific clade B and C with clades in the Southern Hemisphere (Archer et al 2013). As noted above, however, even with the polyphyletic relationship of the oceanic clades, there was only one haplotype shared between the North Pacific and the Southern Hemisphere.

After trimming to the consensus sequence length, the 44 market haplotypes were reduced to 32 haplotypes. For displaying the temporal distribution of products, however, we retained reference to the 44 market haplotypes in the classification procedure. Of the 44 market haplotypes, 16 were assigned to the north Atlantic and the other 28 were assigned to the Southern Hemisphere (Table 4). Assignments made to the North Atlantic were made with high certainty (as measured by the proportion of trees voting for each ocean basin), with probabilities > 0.97. While nine of the 28 samples assigned to the Southern Hemisphere were assigned with probabilities > 0.99, there were 5 market haplotypes with assignment probabilities as low as 0.75. For 4 of these, the next most likely ocean basin was the North Pacific (20%). For one (JFin22), the next most likely ocean basin was the North Atlantic (20%).

Matching haplotype of market products with JARPA/JARPA II

To validate the Random Forest classification, we made a qualitative comparison of mtDNA hapltoypes of market products with haplotypes reported from JARPA and JARPA II samples using the information shown in Table 2 of (Goto *et al.* 2014). (Goto *et al.* 2014) report 45 haplotypes in the 55 samples collected with a biopsy dart or by lethal sampling. Of these 45 haplotypes, 13 are an apparent match to one of the 44 market haplotypes shown in our Table 5. All 13 of the matching haplotypes were classified by Random Forest as originating from the Southern Hemisphere. In other words, there was 100% agreement with the Random Forest classification of the market haplotypes and the matching of these haplotypes to those reported from JARPA/JARPA II.

We also note that the JARPA25 haplotype is an apparent match to a product purchased before the 2005/06 season. According to (Goto *et al.* 2014), this haplotype was found in only one sample collected in Area IV. The source of this samples (e.g., biopsy dart or lethal take) is not reported but the JARPA/JARPAII program operated in Area IV during 2001-02 and 2005-06, i.e., before and after the addition of fin whales to the catch. The products with matching haplotypes were purchased in June 2002 and February 2003.

CONCLUSIONS

Random Forest shows great promise in assigning fin whales to ocean of origins, despite the polyphyletic relationship of mtDNA lineages among oceans (i.e., an absence of complete lineage sorting, Archer et al. 2013). Although it would be desirable to include information from multiple loci (e.g., microsatellites or SNPs) in a population assignment procedure (Manel *et al.* 2005), this is not always possible given quality of DNA and the lack of standardized loci (and for microsatellites, standardize allele binning) for characterizing population differentiation in wide-spread species. mtDNA, on the other hand, is easy to amplify, even from degraded samples, and the high quality of conventional automated sequencing provides confidence in characterizing haplotype identity, even in different laboratories (Morin *et al.* 2010). The mtDNA control region is also highly variable in most species of whales (e.g., 113 haplotypes in the worldwide dataset used here), and the pattern of substitutions is geographically informative as demonstrated by the Random Forest results.

With the Random Forest classification, there are now three lines of evidence pointing to an Illegal, Unreported or Undocumented (IUU) source of fin whales. First, was the phylogenetic analysis reported by Baker et al. (2007), showing an affinity of post-JARPA II products with pre-JARPA II products. Second was the individual identification of fin whale products, reported by Steel et al., (2009), showing that the number of individuals for sale exceeded the number reported as scientific catch or bycatch. The Random Forest classification, based on the worldwide reference dataset, now provides the most compelling evidence that the source of the IUU products is the Southern Hemisphere. For some products, however, we cannot exclude an alternate source, e.g., unreported coastal bycatch or undocumented import from Iceland (e.g. smuggling).

Finally, we note the parallels between the evidence for an IUU source of fin whales products from the Southern Hemisphere, with evidence for an IUU source of sei whales products from the Southern Hemisphere (Baker *et al.* 2014; Baker *et al.* 2015; Yoshida *et al.* 2015). To improve understanding of oceanic population structure and to help identify sources of IUU takes of both species, we **recommend** that information from the Japanese and Icelandic DNA registers be made available through the data availability procedure of the IWC Scientific Committee (IWC 2004).

ACKNOWLEDGMENTS

Funding for market surveys in Japan was provided by the International Fund for Animal Welfare. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author(s) and do not necessarily reflect the views of NOAA or the Department of Commerce.

REFERENCES

Anonymous 2001. Japanese Ministry of Agriculture, Forestry and Fisheries 20 April 2001 revisions to its Ministerial Ordinance No. 92, to take effect 1 July 2001, http://www.maff.go.jp/mud/410.html .

Archer, F.I., P.A. Morin, B.L. Hancock-Hanser, K.M. Robertson, M.S. Leslie, M. Bérubé, S. Panigada and B.L. Taylor. 2013. Mitogenomic Phylogenetics of Fin Whales (*Balaenoptera physalus* spp.): Genetic Evidence for Revision of Subspecies. PLoS ONE 8:e63396.

Archer, F.I., K.K. Martien and B.L. Taylor. (In Review). Diagnosability of mtDNA with Random Forests: Using sequence data to delimit subspecies. Marine Mammal Science

Baker, C.S., N. Funahashi and D. Steel. 2007. Market surveys of whales 2006 via internet purchases, with reference to oceanic origins of fin whale products. Report (SC/59/BC9) to the Scientific Committee of the International Whaling Commission Report (SC/59/BC9).

Baker, C.S. and S.R. Palumbi. 1994. Which whales are hunted? A molecular genetic approach to monitoring whaling. Science 265:1538-1539.

Baker, C.S., D. Steel, Y. Choi, H. Lee, K.S. Kim, S.-K. Choi, Y.-U. Ma, C. Hambleton, L. Psihoyos, R.L. Brownell and N. Funahashi. 2010. Genetic evidence of illegal trade in protected whales links Japan with the U.S. and South Korea. Biology Letters 6:647-650.

Baker, C.S., D. Steel and N. Funahashi. 2008. Market surveys of whale meat in Japan, 2007 – 2008, with reference to the number of fin whales for sale. Report (SC60/BC2) to the Scientific Committee of the International Whaling Commission.

Baker, C.S., D. Steel and N. Funahashi. 2014. Unknown stock origins of sei whales sold in Japanese markets, 1997 to 2009. Report (SC/65b/IA08) to the Scientific Committee of the International Whaling Commission.

Baker, C.S., D. Steel, P. Wade and N. Funahashi. 2015. Sei whales sold in Japanese markets do not match DNA register of JARPNII catches. Report (SC/66a/IAxx) to the Scientific Committee of the International Whaling Commission.

Berk, R. 2006. An introduction to ensemble methods for data analysis. Sociological Methods and Research 34:263-295.

Berubé, M., A. Aguilar, D. Dendanto, F. Larsen and G. Notarbartolo-Di-Sciara. 1998. Population genetic structure of North Atlantic, Mediterranean Sea and Sea of Cortez fin whales, Balaenoptera physalus (Linnaeus 1758): Analysis of mitochondrial and nuclear loci. Molecular Ecology 7:585-599.

Bérubé, M., J. Urbán, A.E. Dizon, R.L. Brownell and P.J. Palsbøll. 2002. Genetic identification of a small and highly isolated population of fin whales (*Balaenoptera physalus*) in the Sea of Cortez, México. Conservation Genetics 3:183-190.

Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

CITES. 1997. Doc 10.40 Interpretation and Implementation of the Convention – ILLEGAL TRADE IN WHALE MEAT. http://www.cites.org/sites/default/files/eng/cop/10/doc/E10-40.pdf.

CITES. 1973. Convention on International Trade in Endangered Species of Wild Flora and Fauna, part of the U.S. Endangered Species Act. Public Law 93-205, Title 50.

Goto, M., N. Kanda and L.A. Pastene. 2014. Genetic analysis on stock structure whales of fin whales in the Antarctic based on mitochondrial and microsatellite DNA. Report (SC/F14/J32) presented at the JARPAII review by the Scientific Committee of the International Whaling Commission, February 2014.

IWC. 2004. Annex T: Report of the data availability working group. Journal of Cetacean Research and Management (Suppl.):406-408.

Katoh, Misawa, Kuma, Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059-3066

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R News 2/3:18-22.

Manel, S., O.E. Gaggiotti and R.S. Waples. 2005. Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology and Evolution 20:136-142.

Morin, P.A., K.K. Martien, F.I. Archer, F. Cipriano, D. Steel, J. Jackson and B.L. Taylor. 2010. Applied conservation genetics and the need for quality control and reporting of genetic data used in fisheries and wildlife management. Journal of Heredity 101:1-10.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ross, H.A., G.M. Lento, M.L. Dalebout, M. Goode, 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.

Sremba, A.L., A.R. Martin and C. Scott Baker. 2015. Species identification and likely catch time period of whale bones from South Georgia. Marine Mammal Science 31:122-132.

Steel, D., N. Funahashi, R.M. Hamner and C.S. Baker. 2009. Market surveys of whale meat in Japan 2008/2009: How many fin whales are for sale? Report (SC61/BC8) to the Scientific Committee of the International Whaling Commission.

Yoshida, H., M. Goto and L.A. Pastene. 2015. Genetic analyses of market samples provide no evidence for additional stock structure of sei whales in the North Pacific. Report (SC/66a/IAxx) to the Scientific Committee for the International Whaling Commission.

Table 1: Number of fin whale products purchased in commercial markets of Japan between 1993 and April 2009, as determined by phylogenetic reconstruction of sequences from the mtDNA control region. Sample sizes indicate number of products identified to species, not necessarily the number of unique individuals represented by the products.

Survey periods	Japan 1993/2004	Japan 2006	Japan 2007	Japan 2008/09
Fin whale, *B. physalus*	33	15	38	27

Table 2: Corrected individual identification of fin whales in the three market surveys (2006, 2007 and 2008/09) and the number of products of each individual that were sampled in each survey, as first reported by (Steel *et al.* 2009). The individual formerly referred to as Jmarket16 was collapsed into Jmarket4 (now JFin4) following review of DNA profiles.

Individual information		Market survey		
Haplotype code	Sex	2006	2007	2008-09
JFin1	F	3	1	4
JFin2	F	2		
JFin3	M	3	7	1
JFin4	F	2	3	2
JFin5	M	2	2	2
JFin6	F	1		
JFin7	F	1		
JFin8	M	1	1	1
JFin9	F		14	11
JFin10	F		1	
JFin11	F		1	
JFin12	M		5	1
JFin13	F		1	
JFin14	M		1	
JFin15	F		1	
JFin17	F			1
JFin18*	M			1
JFin19	F			2
JFin20	F			1
Cumulative number of products		15	53	80
Cumulative observed individuals		8	15	19
Cumulative expected individuals		10	13^	15#

* Haplotype JMarket18 was also found in samples collected in 1999 and 2000

^ Excludes the bycaught animal reported 17 Dec 2007 based on mtDNA identity and sampling dates (see Baker et al 2008)

# Includes 2 bycaught animals, excludes JARPA II 2008/09 catch of 1 fin whale based on sampling dates. Also excludes import of product from Iceland (see text for details).

Table 3: Summary of mtDNA datasets used in the Random Forest classification for oceanic origins of fin whales. The NA, NP and SH represent 'reference' datasets of known origins (Archer *et al.* 2013; Berubé *et al.* 1998; Sremba *et al.* 2015). "Market" represents the 'unknowns' to be classified by Random Forest.

Reference dataset	**# samples**	**# haplotypes**	**haplotype diversity**	**frequency singletons**
North Atlantic (NA)	332	35	0.8336	0.0421
North Pacific (NP)	346	31	0.6382	0.0231
S. Hemisphere (SH)	99	48	0.9715	0.3030
Total	777	113	--	--
Market	44	32	0.9757	0.5777

Table 4: Random Forest classification matrix for reference samples from North Atlantic, North Pacific and Southern Hemisphere fin whales. Columns are ocean basins of origin, rows are classification ocean basin. Overall error rate = (2 + 21) / 777 = 0.0296.

Ocean	**NA**	**NP**	**SH**	**error**
N. Atlantic (NA)	330	0	2	0.0060
N. Pacific (NP)	0	325	21	0.0607
S. Hemisphere (SH)	0	0	99	0.0000

Table 5: The classification probability of fin whale products originating from the North Atlantic (NA), North Pacific (NP) or Southern Hemisphere (SH) based on Random Forest analysis of 44 mtDNA haplotypes. The frequency of products is shown in three sampling periods: Icelandic, 1993-99, assuming a 10-year persistence of products from this scientific hunting; Pre-JARPAII, 2000-04, when there was no know source of fin whale products; and, Post-JARPAII, 2006-09, when products from 13 southern and 2 North Pacific fin whales (bycatch) were available on the market.

Market haplotype code	NA	NP	SH	Icelandic, 1993-99	Pre-JARPAII, 2000-04	Post-JARPAII, 2006-09	reference match	JARPA match
JFin26	1.000	0.000	0.000	1				
JFin27	1.000	0.000	0.000	1				
JFin28	1.000	0.000	0.000	1				
JFin35	1.000	0.000	0.000	1				
JFin36	1.000	0.000	0.000	1			Mitogenome	
JFin37	1.000	0.000	0.000	3			AY822094	
JFin24	0.991	0.001	0.007	1				
JFin25	0.991	0.001	0.007	1				
JFin30	0.991	0.001	0.007	1				
JFin40	0.991	0.001	0.007		1		AY582748	
JFin29	0.999	0.000	0.001	1				
JFin31	0.979	0.017	0.004	1				
JFin32	0.985	0.007	0.008	2				
JFin33	0.995	0.001	0.004	1				
JFin38	0.992	0.004	0.003		1			
JFin39	0.998	0.000	0.002	1				
JFin21	0.340	0.059	0.601	1				
JFin23	0.187	0.065	0.748	1				
JFin1	0.004	0.001	0.995			8		JARPA19
JFin42	0.004	0.001	0.995		2			
JFin43	0.004	0.001	0.995		1			
JFin2	0.001	0.060	0.938			2	ANT91319	JARPA26
JFin3	0.000	0.001	0.999			11		
JFin4	0.000	0.001	0.999			7		JARPA18
JFin12	0.000	0.001	0.999			6		JARPA38
JFin15	0.000	0.001	0.999			1		
JFin10	0.000	0.001	0.999			1	ANT91313	JARPA22
JFin44	0.000	0.001	0.999		2		ANT91310	JARPA25
JFin5	0.047	0.203	0.750			6		JARPA8
JFin19	0.047	0.203	0.750			2		
JFin34	0.047	0.203	0.750	1				
JFin6	0.002	0.092	0.907			1		
JFin11	0.002	0.092	0.907			1		JARPA23
JFin7	0.029	0.064	0.907			1		JARPA27
JFin8	0.009	0.202	0.789			3		JARPA20
JFin9	0.047	0.203	0.750			25		JARPA41
JFin13	0.002	0.102	0.896			1		
JFin14	0.025	0.084	0.892			1	SGeorgia	JARPA21
JFin17	0.045	0.006	0.949			1		
JFin18	0.082	0.151	0.767	1	1	1		
JFin20	0.000	0.028	0.972			1	ANT91312	JARPA24
JFin22	0.198	0.028	0.774	1				
JFin41	0.006	0.074	0.920	1				
JFin45	0.114	0.086	0.800		2			
Totals				**23**	**10**	**80**		

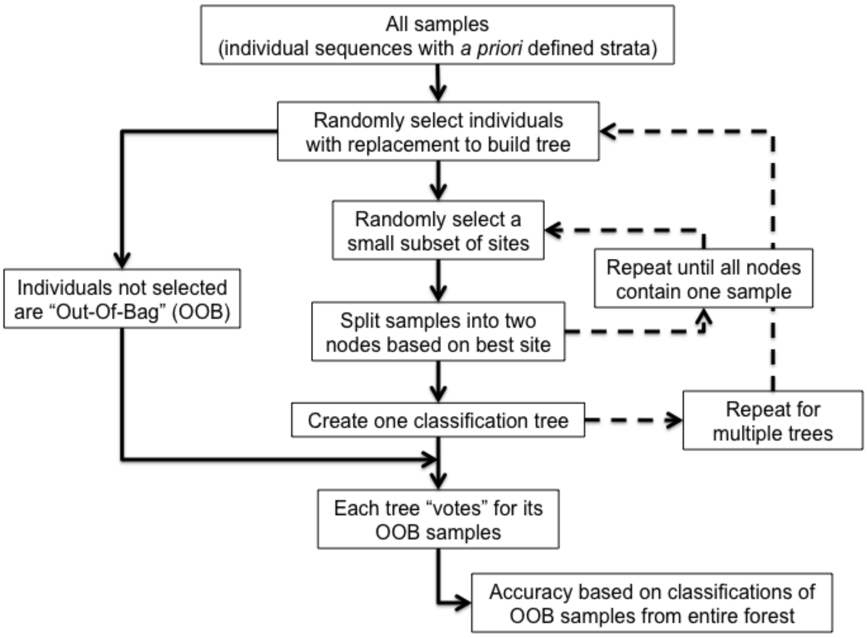
Appendix. Description of Random Forest algorithm as applied to DNA sequence data. Excerpted from Archer, F.I., K.K. Martien and B.L. Taylor. (In Review). Diagnosability of mtDNA with Random Forests: Using sequence data to delimit subspecies. Marine Mammal Science

Random Forests (Breiman, 2001) is an ensemble-based classification algorithm that extends the more familiar method of CART by adding several layers of stochasticity to the tree growing process. This permits the algorithm to fully explore the predictive capability of all variables, as well as producing an internally validated classifier. A Random Forest is a collection of bifurcating CART-like decision trees. In our implementation, the initial data comprise a set of samples represented by aligned mtDNA sequences, each grouped into their *a priori* defined taxa (e.g., putative subspecies). All samples in the dataset are used, rather than reducing the sequences to unique haplotypes, so that the frequencies of haplotypes, and hence that of their constituent nucleotide substitutions are properly represented. Additionally, sequences can be reduced to just variable sites so the analysis does not waste time evaluating conserved sites that have no classification information.

The process of building a Random Forests model is illustrated in the figure below For each tree in the forest, the first step is to select a random set of samples that are used as the training set for the tree. Those samples not selected (the out-of-bag or OOB samples) are set aside for cross-validation of the tree's prediction accuracy.   The tree is then grown in the following iterative manner:

1) Choose a random subset of nucleotide sites from all available sites.
2) For each site chosen, create a rule that best splits the sequences into two groups.
3) Choose the site that produces the best split and create two daughter nodes of sequences based on that split.
4) For each of these daughter nodes, return to step 1 and repeat until a predefined stopping point is reached, such as all nodes containing a single sample.

The sequences for the OOB samples are then sent through the decision tree based on its splitting rules and classified according to the stratum of the sample in the final node they end up in. In this manner, a tree produces a single "vote" to a given stratum for each OOB sample. Steps 1 through 4 are repeated multiple times to produce many trees (the "forest"), each of which votes for the strata of their own respective OOB samples. The probability (p) that a sample is classified to a given stratum is the fraction of trees voting for that stratum in the subset of trees in the forest where the sample was OOB. Thus, a sample is predicted to belong to the stratum with the largest p. In the simple case of two strata, this would be the stratum for which $p > 0.5$.


All samples
(individual sequences with *a priori* defined strata)
Randomly select individuals with replacement to build tree
Randomly select a small subset of sites
Repeat until all nodes contain one sample
Individuals not selected are "Out-Of-Bag" (OOB)
Split samples into two nodes based on best site
Create one classification tree
Repeat for multiple trees
Each tree "votes" for its OOB samples
Accuracy based on classifications of OOB samples from entire forest