PHFinder: Assisted detection of point heteroplasmy in Sanger sequencing chromatograms.

Marcos Suarez-Menendez[1], Vania Rivera-Leon[1], Jooke Robbins[2], Martine Bérubé[1-2] & Per J. Palsbøll[1-2]

1. Marine Evolution and Conservation, Groningen Institute of Evolutionary Life Sciences, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands.

2. Center for Coastal Studies, 5 Holway Avenue, Provincetown, Massachusetts, 02657, United States of America.

Corresponding authors:

Marcos Suárez Menéndez

Marine Evolution and Conservation, Groningen Institute of Evolutionary Life Sciences, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands.

Email: m.suarez.menendez@rug.nl

Abstract

1. Heteroplasmy is the presence of two or more organellar genomes (mitochondrial or plastid DNA) in an organism, tissue, cell or organelle. Heteroplasmy can be detected by visual inspection of Sanger sequencing chromatograms, where it appears as multiple peaks of fluorescence at a single nucleotide position. Visual inspection of chromatograms is both consuming and highly subjective, as heteroplasmy is difficult to differentiate from background noise. Few software solutions are available to automate the detection of point heteroplasmies, and those that are available are typically proprietary, lack customization or are unsuitable for automated heteroplasmy assessment in large datasets.

2. Here, we present PHFinder, a Python-based, open source tool to assist in the detection of point heteroplasmies in large numbers of Sanger chromatograms. PHFinder automatically identifies point heteroplasmies directly from the chromatogram trace data. The program was tested with Sanger sequencing data from 100 humpback whale (*Megaptera novaeangliae*) tissue samples with known heteroplasmies.

3. PHFinder detected most (90%) of the known heteroplasmies thereby greatly reducing the amount of visual inspection required. PHFinder is flexible, enabling explicit specification of key parameters to infer double peaks (i.e., heteroplasmies).

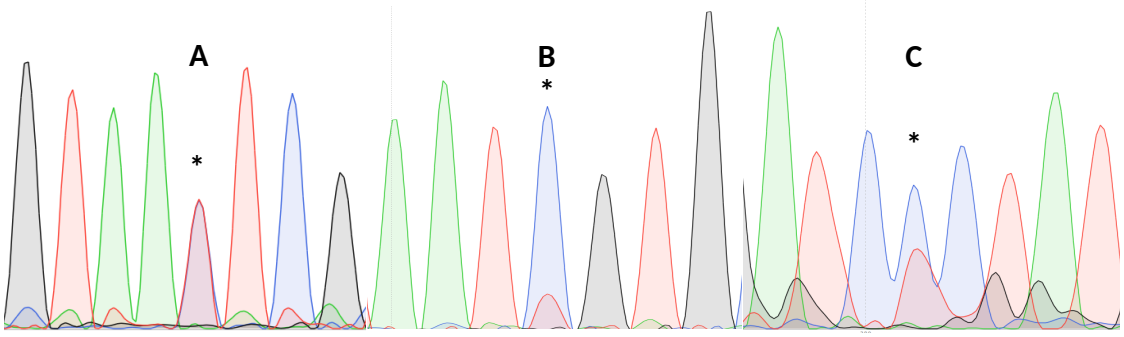
Keywords: AB1, chromatograms, detection, heteroplasmy, Sanger, software,.

Introduction

Heteroplasmy is the presence of multiple organellar (mitochondrial or plastid) genomes in an organism, tissue or cell. Despite advances in so-called next-generation sequencing, Sanger sequencing (Sanger et al., 1977) is still widely employed in studies targeting specific highly-variable organellar DNA regions, such as the mitochondrial control region. The DNA sequence is inferred from the resulting chromatogram, where the base at each nucleotide position is represented by a fluorescent signal of base-specific colour (each representing a different deoxynucleotides). Heteroplasmy, due to point mutations, is apparent as two fluorescent peaks in the same nucleotide position. All other factors being equal, the relative height of each fluorescent peak reflects the relative abundance of each deoxynucleotide, and, by extension, the two mtDNA haplotypes (Irwin et al., 2009). Heteroplasmy due to insertions or deletions results in multiple fluorescent peaks at multiple, consecutive nucleotide positions (length heteroplasmy).

Detecting heteroplasmy is necessary when studying certain mitochondrial diseases (Stewart & Chinnery, 2015), conducting forensic work (Salsa et al., 2001) or estimating mitochondrial mutation rates (Millar et al., 2008). Failing to take heteroplasmy into consideration can also cause difficulty in other studies. For example, in relatedness studies, maternal relatives might appear to have different mitochondrial haplotypes due to a shift in heteroplasmic proportions (Klütsch et al., 2011). Visual inspection of all chromatograms for detection of putative double fluorescent peaks is impractical in large datasets, and is prone to inconsistencies given the subjective nature of the assessment (Fig. 1).

Figure 1. Three examples of point heteroplasmy.



Notes: *Putative heteroplasmies. [A]Completely overlapping fluorescent peaks (likely heteroplasmy). [B]Significantly lower secondary fluorescent peak (likely background noise). [C]Heteroplasmy with background noise, making it more difficult to detect.

Previous studies have applied different criteria to infer heteroplasmies from the ratio of the two fluorescent peaks in a putative heteroplasmic nucleotide position (e.g., > 10%, Brandstätter et al., 2004; Irwin et al., 2009, or > 30%, Baker et al., 2013) or failed to make any objective specification (e.g., Vollmer et al., 2011).

Existing software can facilitate the automatic detection of double fluorescent peaks in chromatograms. These are either proprietary (e.g., SEQUENCHER, GeneCodes Inc., Ann Arbor, MI) or lack customization and tend to disregard some double fluorescent peaks as background noise, or are unable to process large datasets (e.g., SNAPGENE® VIEWER v4.3.7, GSL Biotech LLC). We developed a bioinformatic pipeline (point heteroplasmy finder, PHFinder) as a means to screen DNA chromatograms in the commonly employed AB1 format (Applied Biosystems Inc. 2006); generated by DNA sequencers, such as the Applied BiosystemsTM Genetic Analyzer series (Thermos Scientific Inc.) to detect the double fluorescent peaks in an automated manner. PHFinder facilitates the detection of point heteroplasmies by applying filters using average base call quality scores and the level of background noise in a user-specified target region of the relevant DNA sequence.

Implementation

PHFinder was written in Python v3.6.8 (Van Rossum & Drake Jr, 1995) and BASH (Ramey & Fox, 2016). PHFinder dependencies include; Biopython v1.73 (Cock et al., 2009) and BOWTIE2 (Langmead & Salzberg, 2012).

First, a FASTQ file (sequence of nucleotides and the associated Phred quality scores (Ewing et al., 1998)) is extracted from each AB1 file. FASTQ files are aligned against a reference sequence with BOWTIE2 (Langmead & Salzberg, 2012) and the result is saved in a single Sequence Alignment Map (SAM) format file per alignment. The orientation of the chromatogram (forward or reverse) and starting point of the reference sequence is subsequently extracted from each SAM file in order to position the trace information to the correct region in each chromatogram.

The data elements stored with AB1 files are associated with specific tags. PHFinder uses the information contained in the AB1 tags; DATA9 to DATA12 (trace information for guanine, adenine, thymine and cytosine); PBAS2 (the sequence of base calls); PLOC2 (location of base calls) and PCON2 (per-base call quality score) as specified in the original file fomat (Applied Biosystems Inc. 2006).

The presence of a double fluorescent peak (i.e., a potential heteroplasmy) was inferred from the values of three *ad hoc* indexes calculated from the above data:

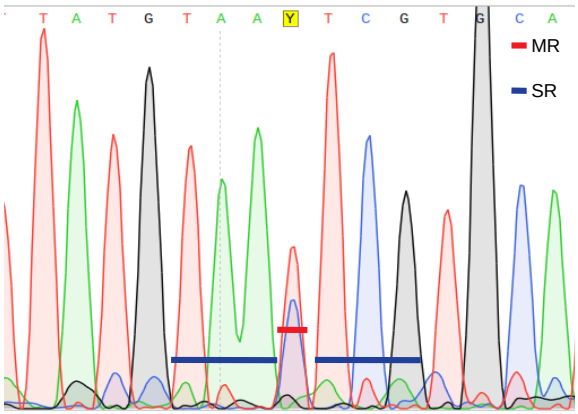
I. Average base call quality (AQ) of the bases in the targeted DNA sequence region (measured as Phred quality scores, ranging from 0 to 93).

II. Main ratio (MR) of a double fluorescent peak, i.e., the height of the second highest peak as a fraction of the highest peak (Fig.2).

III. Secondary ratios (SR) of the three down- and upstream nucleotide positions flanking the putative heteroplasmic position; estimated as the height of the second highest peak as a fraction of the highest peak (Fig. 2).

A position is inferred to be heteroplasmic by PHFinder if the three indexes described above exceeds user determined threshold values (Fig. 2).

Figure 2. Main ratio (MR) and secondary ratio (SR) indexes.



Material and methods

A test set of mitochondrial control region DNA sequences were extracted from 100 skin samples collected from a long-term study of individual humpback whales (*Megaptera novaeangliae)* in the Gulf of Maine (North Atlantic). DNA sequence data were randomly selected from 30 samples with known heteroplasmies as well as from 70 samples that appeared homoplasmic. Heteroplasmies were identified based on comparison to samples from close maternal relatives (known through longitudinal studies of individuals or microsatellite markers) or experimental confirmation using the dCAPS technique (Neff et al., 1998) (data not shown).

The presence of heteroplasmy in a sample was assumed to be true if either one of two conditions was met; if another directly related whale, i.e., mother or sibling, (confirmed by long-term population studies and microsatellite markers) was heteroplasmic in that nucleotide position, or if the heteroplasmy was subsequently experimentally confirmed using the dCAPS technique (Neff et al., 1998) (data not shown).

Skin samples were collected by biopsy techniques (Palsbøll et al., 1991) and stored in 5 M NaCl with 25% DMSO (dimethyl sulfoxide, Amos & Hoelzel, 1991) at -20/-80 degrees Celsius (ºC) prior

to DNA extraction. Total-cell DNA was extracted by standard phenol/chloroform extractions as described by Russel and Sambrook (2001) or using QIAGEN DNEasyTM Blood and Tissue Kit (QIAGEN Inc.) following the manufacturer's instructions. Extracted DNA was stored in 1xTE buffer (10 mM Tris–HCl, 1mM EDTA, pH 8.0) at -20°C.

The sequence of the first 500 base pairs (bps) of the 5' end of the mitochondrial control region was determined as described previously by Palsbøll et al. (1995) using the oligo-nucleotide primers BP16071R (Drouot et al., 2004) and MT4F (Arnason et al., 1993). Unincorporated nucleotides and primers were removed from the Polymerase Chain Reactions (PCR, Mullis & Faloona, 1987) with Shrimp Alkaline Phosphatase and Exonuclease I, as described by Werle et al. (1994). Subsequent cycle sequencing conducted with the above-mentioned nucleotide primers and the BigDye® Terminator v3.1 Cycle Sequencing kit (Applied BiosystemsTM Inc.) following the manufacturer's protocol. The cycle sequencing products were precipitated with by ethanol and sodium (Russel and Sambrook, 2001). The order of sequencing fragments was resolved by electrophoresis on an ABI 3730 DNA AnalyzerTM or and ABI PRISM® 377 DNA Sequencer (Applied Biosystems Inc.).

All chromatograms were visually inspected for point heteroplasmies using SNAPGENE® VIEWER (v4.3.7, GSL Biotech LLC). PHFinder was tested and validated (on GNU/Linux systems) by analysing the dataset with different threshold values for each index (MR: 15, 25, 35, 45; SR: 0.2, 0.3, 0.4, 0.5; and AQ: 30, 40, 50, 60) resulting in 64 different combinations of index threshold values. The results of each index threshold combination were visualized to assess which combination was most efficient in detecting the known heteroplasmies; specifically, the number of detected heteroplasmic AB1 files versus the number of false positives (nucleotide positions that were incorrectly deemed as heteroplasmic).

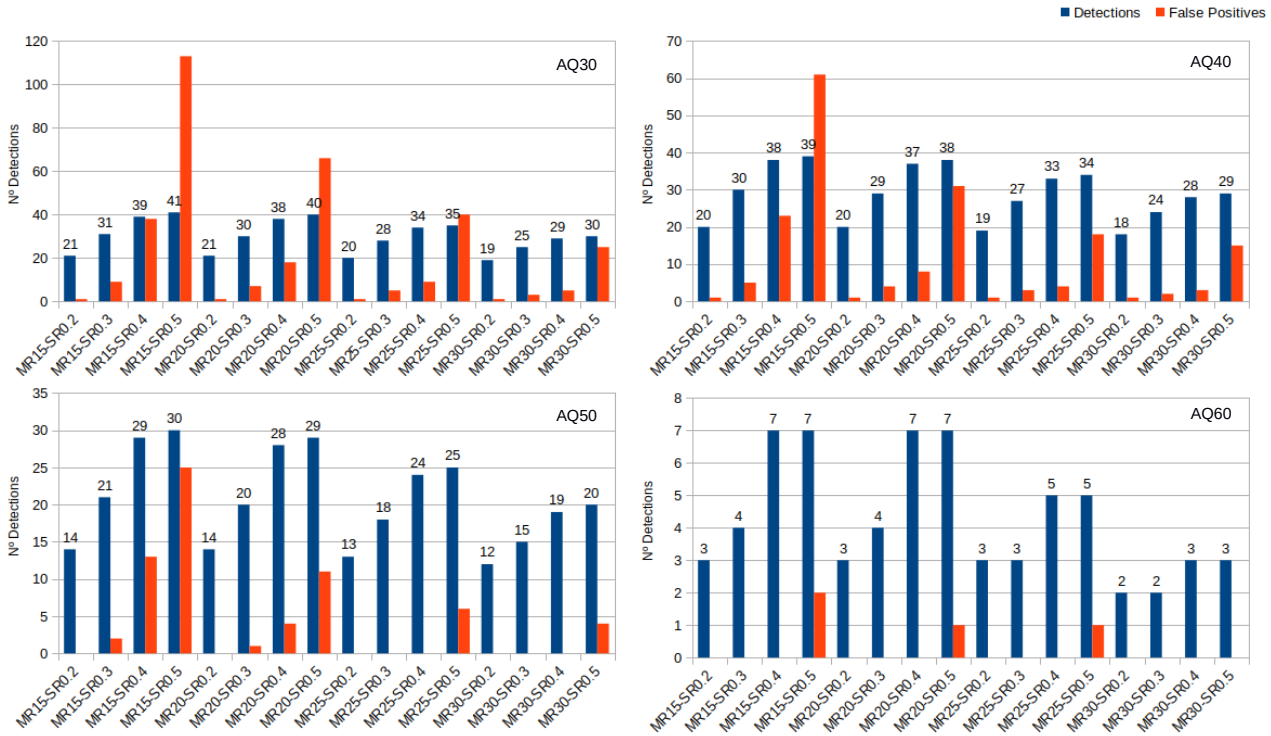
Results

The data from the 100 samples comprised 189 AB1 files (tissue samples were sequenced between 1 -6 times). The first 500 bps of the mitochondrial control region of *M. novaeangliae*, started at

position 15,490 and end at position 15,970 according to the reference mitochondrial genome sequence NC_006927.1 (NCBI Reference Sequence) published by Sasaki et al. (2005), which was also used as reference for the alignment. The analyses conducted here targeted the region from position 15,540 to position 15,815 (275 bps).

Among of the 188 AB1 files, PHFinder was unable to process five and another 19 files contained no data. Among the 30 samples (42 AB1 files) with known heteroplasmies included a point heteroplasmy at seven different nucleotide positions with MRs ranging from 18 to 88 (Table S1). Fig. 3 shows the fraction of these known heteroplasmies detected by PHFinder for each combination of index threshold values as well as the number of false positives.

Figure 3. Number of detected heteroplasmic AB1 files vs. false positives for each combination of index threshold values divided according to AQ indexes.



The number of samples and AB1 files included in each analysis varied with the AQ index: AQ 30, 100 samples and 143 AB1 files. AQ 40, 88 samples and 115 AB1 files. AQ50, 68 samples and 79 AB1 files. AQ60, 37 samples and 15 AB1 files. Detailed results for each combination of index threshold values are shown in Table S2.

Discussion

PHFinder was developed to assist the detection of point heteroplasmies in large data sets. The program automates a first pass of the data, reducing the number of AB1 files that need to be visually inspected. Although PHFinder detected most of the point heteroplasmies present in the dataset (95.3%), the present analysis revealed some limitations.

The samples were randomly extracted from a large dataset in order to include a wide range of DNA and DNA sequences in terms of overall DNA sequence quality, length, DNA strand sequenced, as well as corrupted or older AB1 files and different MR values in order to test PHFinder under realistic conditions. The samples were represented by different numbers of AB1 files as some samples were re-sequenced several times (i.e. because of bad quality or to sequence both DNA strands). Among the 100 samples, 30 samples contained a known heteroplasmic nucleotide position to ensure sufficient data for testing PHFinder.

PHFinder was unable to process five AB1 files due to Biopython's v1.73 (Cock et al., 2009) inability to access AB1 files generated by older DNA sequencers (ABI PRISM® 377 DNA Sequencer in this instance). Potential compatibility issues could be due to differences in the tags in AB1 files, and hence resolved by modifying the tag names in the PHFinder main script.

The PHFinder assessment targeted the region from position 15,540 to position 15,815 (275 bps) in order to avoid regions of chromatograms with elevated background noise (i.e., the 5' and 3' end). This strategy reduced the proportion of false positives of heteroplasmies and increased the average quality of chromatograms.

Unsurprisingly, the main limiting factor was the average base call quality of the AB1 files. Low average base call quality is mainly due to elevated background noise which, in turn, may be erroneously inferred as putative heteroplasmies. These kinds of false positives were easily recognised as artefacts during subsequent visual inspection of the chromatograms. Since all putative heteroplasmies highlighted by PHFinder should be visually confirmed, the most efficient approach is to employ a combination of index threshold values that yields the highest number of heteroplasmies and lowest number of false positives (in this case; MR: 20, SR: 0.4 and AQ: 40, Fig. 3). In this study the aforementioned combination of index threshold values identified 27 (out of 30 known heteroplasmies in 37 AB1 files) and only eight false positives.

Employing low AQ index threshold values increased the number of AB1 files (and the corresponding samples) in an assessment, without a similar increase in heteroplasmy detection, i.e., the overall frequency of heteroplasmy detections was reduced (e.g., Fig. 3, AQ30 vs AQ40). We observed a clear trade-off between the detected heteroplasmies and the number of false positives. Low PHFinder index threshold values (i.e., low MR, SR, and AQ) improved the identification of heteroplasmies but also the number of false positives. Since all samples with putative heteroplasmies require visual inspection, lowering the threshold index values led to increasing amounts of visual inspection. Applying higher index threshold values (i.e., a high MR, SR and AQ) had the opposite effect, i.e., fewer false positives and les visual inspection accompanied with a lower detection rate of heteroplasmies.

Heteroplasmies with a MR as low as 15% could be potentially detected, based on the detection limit of Sanger sequencing (Tsiatis et al., 2010). Although the lowest MR in this study was inferred at 18%, (Table S1), the lowest MR index threshold applied was 15% in order to assess the effect on the rate of false positives.

The optimal index threshold value combination will likely depend upon the aim of a study and the overall quality of the DNA sequence chromatograms. If the goal is to compare heteroplasmy

frequencies, then a fairly strict index threshold value combination can be employed to reduce the number of false positives. In more detailed assessments that are aimed at identifying as many heteroplasmies as possible (e.g., to detect, novel deleterious mutations), a lower index threshold value combination will facilitate a higher heteroplasmy detection rate, but require elevated levels of post-analysis visual inspection. If the targets are specific known heteroplasmies (i.e., in a known nucleotide position), a lowered index threshold value combination can be employed as only putative heteroplasmies in the targeted positions would require post-analysis visual inspection.

The main advantage of PHFinder is the ability to customize assessments on a case-by-case basis. Written in Python, the source code can easily be changed and improved to fit specific research needs. PHFinder provides the ability to implement and apply explicit assessment criteria (in terms of base quality and fluorescent peak ratios), thereby facilitating direct objective comparisons among different data sets. Although a final visual inspection of chromatograms with putative heteroplasmic sites will always be required, PHFinder greatly reduces the number of chromatograms requiring visual inspection which is especially valuable in large datasets.

Author contributions

The research project was conceived and designed by M.S., V.R., M.B. and P.J.P. M.S. performed the data analyses, developed the software and wrote the manuscript with V. R., M.B. and P.J.P. Tissue samples and pedigree data were provided by J.R. The laboratory work was performed by M.B. All authors reviewed and partook in manuscript revisions. The authors have no competing interests.

Data accessibility

PHFinder scripts and all data used in this paper are available online from https://github.com/MSuarezMenendez/PHFinder, as well as instructions of how to use PHFinder on the command line.

References

Amos, B., & Hoelzel, A. . (1991). Long-term skin preservation of whale skin for DNA analysis. *Rep Int Whaling Comm Spec. Issue*, *13*, 99–103.

Applied Biosystems *Genetic Analysis Data File Format SUBJECT: ABIF File Format Specification and Sample File Schema*. (2006). Retrieved from https://projects.nfstc.org/workshops/resources/articles/ABIF_File_Format.pdf

Arnason, U., Gullberg, A., & Widegren, B. (1993). Cetacean mitochondrial DNA control region: sequences of all extant baleen whales and two sperm whale species. *Molecular Biology and Evolution*, *10*(5), 960–970. https://doi.org/10.1093/oxfordjournals.molbev.a040061

Baker, S. ., Steel, D., Calambokidis, J., Falcone, E., González-Peral, U., Barlow, J., … Yamaguchi, M. (2013). Strong maternal fidelity and natal philopatry shape genetic structure in North Pacific humpback whales. *Marine Ecology Progress Series*, *494*, 291–306. https://doi.org/10.3354/meps10508

Brandstätter, A., Niederstätter, H., & Parson, W. (2004). Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. *International Journal of Legal Medicine*, *118*(1), 47–54. https://doi.org/10.1007/s00414-003-0418-z

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Drouot, V., Bérubé, M., Gannier, A., Goold, J. C., Reid, R. J., & Palsbøll, P. J. (2004). A note on genetic isolation of Mediterranean sperm whales (Physeter macrocephalus) suggested by mitochondrial DNA. *Journal of Cetacean Research and Management*, *6*(1), 29–32.

248 Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer
249 traces usingPhred. I. Accuracy assessment. *Genome research*, 8(3), 175-185.
250 https://doi.org/10.1101%2Fgr.8.3.175

251 Irwin, J. A., Saunier, J. L., Niederstätter, H., Strouss, K. M., Sturk, K. A., Diegoli, T. M., …
252 Parsons, T. J. (2009). Investigation of Heteroplasmy in the Human Mitochondrial DNA
253 Control Region: A Synthesis of Observations from More Than 5000 Global Population
254 Samples. *Journal of Molecular Evolution*, *68*(5), 516–527. https://doi.org/10.1007/s00239-
255 009-9227-4

256 Klütsch, C. F. C., Seppälä, E. H., Uhlén, M., Lohi, H., & Savolainen, P. (2011). Segregation of point
257 mutation heteroplasmy in the control region of dog mtDNA studied systematically in deep
258 generation pedigrees. International Journal of Legal Medicine, 125(4), 527–535.
259 https://doi.org/10.1007/s00414-010-0524-7

260 Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature*
261 *Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

262 Millar, C. D., Dodd, A., Anderson, J., Gibb, G. C., Ritchie, P. A., Baroni, C., … Lambert, D. M.
263 (2008). Mutation and evolutionary rates in adélie penguins from the antarctic. *PLoS Genetics*,
264 *4*(10). https://doi.org/10.1371/journal.pgen.1000209

265 Mullis, K. B., & Faloona, F. A. (1987). [21] Specific synthesis of DNA in vitro via a polymerase-
266 catalyzed chain reaction. *Methods in Enzymology*, *155*, 335–350. https://doi.org/10.1016/0076-
267 6879(87)55023-6

268 Neff, M., Neff, J., Chory, J., & Pepper, A. (1998). dCAPS, a simple technique for the genetic
269 analysis of single nucleotide polymorphisms: experimental applications in Arabidopsis thaliana
270 genetics. In *The Plant Journal* (Vol. 14). https://doi.org/https://doi.org/10.1046/j.1365-
271 313X.1998.00124.x

Palsbøll, P. J., Clapham, P. J., Mattila, D. K., Larsen, F., Sears, R., Siegismund, H. R., … Arctander, P. (1995). Distribution of mtDNA haplotypes in North Atlantic humpback whales: the influence of behaviour on population structure. *Marine Ecology Progress Series*, *116*, 1–10. https://doi.org/10.2307/44634989

Palsbøll, P. J., Larsen, F., & Hansen, E. S. (1991). Sampling of skin biopsies from free-ranging large cetaceans at West Greenland: development of new designs. *IWC Workshop Paper*, (13), 1–8.

Ramey, C., & Fox, B. (2016). Reference Manual: Reference Documentation for Bash Edition 4.4. *Free Software Foundation*.

Russel, D. W., & Sambrook, J. (2001). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Salas, A., Lareu, M. V., & Carracedo, A. (2001). Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: A case report. International Journal of Legal Medicine, 114(3), 186–190. https://doi.org/10.1007/s004140000164

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors. 74*(12), 5463–5467. https://doi.org/https://doi.org/10.1073/pnas.74.12.5463

Sasaki, T., Nikaido, M., Hamilton, H., Goto, M., Kato, H., Kanda, N., … Okada, N. (2005). Mitochondrial Phylogenetics and Evolution of Mysticete Whales. *Systematic Biology*, *54*(1), 77–90. https://doi.org/10.1080/10635150590905939

Stewart, J. B., & Chinnery, P. F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature Reviews Genetics*, *16*(9), 530–542. https://doi.org/10.1038/nrg3966

Tsiatis, A. C., Norris-Kirby, A., Rich, R. G., Hafez, M. J., Gocke, C. D., Eshleman, J. R., & Murphy, K. M. (2010). Comparison of Sanger Sequencing, Pyrosequencing, and Melting Curve Analysis for the Detection of KRAS Mutations: Diagnostic and Clinical Implications. *The*

Journal of Molecular Diagnostics, *12*(4), 425–432. https://doi.org/10.2353/JMOLDX.2010.090188

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Amsterdam: Centrum voor Wiskunde en Informatica.

Werle, E., Schneider, C., Renner, M., Völker, M., & Fiehn, W. (1994). Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Research*, *22*(20), 4354–4355. https://doi.org/https://doi.org/10.1093/nar/22.20.4354