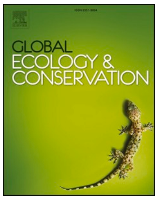
Global Ecology and Conservation 51 (2024) e02947

Contents lists available at ScienceDirect

Global Ecology and Conservation

journal homepage: www.elsevier.com/locate/gecco



Original research article

Baleen whale microsatellite panel for individual identification and parentage assignment in *Mysticeti*

Marcos Suárez-Menéndez[a,*,1], Martine Bérubé[a,b,**,1], Lutz Bachmann[c], Peter Best[d,2], Nick Davison[e], Mads Peter Heide-Jørgensen[f], Véronique Lesage[g], Tom Oosting[a], Rui Prieto[h], Christian Ramp[i,j], Jooke Robbins[b], Richard Sears[i], Mónica A. Silva[h], Mariel T.I. ten Doeschate[e], Marc Tollis[k,l,m], Els Vermeulen[d], Gísli A. Víkingsson[n], Øystein Wiig[c], Per J. Palsbøll[a,b]

[a] *Marine Evolution and Conservation, Groningen Institute of Evolutionary Life Sciences, University of Groningen, Groningen, the Netherlands*
[b] *Center for Coastal Studies, Provincetown, MA, USA*
[c] *Natural History Museum, University of Oslo, Oslo, Norway*
[d] *Mammal Research Institute Whale Unit, University of Pretoria, South Africa*
[e] *Scottish Marine Animal Stranding Scheme (SMASS) Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow, UK*
[f] *Greenland Institute of Natural Resources, Copenhagen, Denmark*
[g] *Maurice Lamontagne Institute, Mont-Joli, Quebec, Canada*
[h] *Institute of Marine Sciences - Okeanos & Institute of Marine Research - IMAR, University of the Azores, Horta, Portugal*
[i] *Mingan Island Cetacean Study, Saint-Lambert, Quebec, Canada*
[j] *Sea Mammal Research Unit, Scottish Ocean Institute, University of St Andrews, United Kingdom*
[k] *Biodesign Institute, Arizona State University, Tempe, AZ, USA*
[l] *School of Life Sciences, Arizona State University, Tempe, AZ, USA*
[m] *School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA*
[n] *Marine and Freshwater Research Institute, Program for Whale Research, Reykjavík, Iceland*

ARTICLE INFO

Keywords:
STR loci
Mysticeti
Multiplex
Paternity
Microsatellite

ABSTRACT

Highly polymorphic single tandem repeat loci (STR, also known as microsatellite loci) remain a familiar, cost efficient class of genetic markers in genetic studies in ecology, behavior and conservation. Here we characterize a new, universal set of ten STR loci in seven species of baleen whales, optimized for PCR amplification in two multiplex reactions along with a Y chromosome marker for sex determination. The optimized, universal set of STR loci provides a convenient starting point for new genetic studies in baleen whales aimed at identifying individuals and populations. Data from the new STR loci were combined with genotypes from previously published STR loci to assess the power to assign parentage using paternity exclusion in four species: fin whale (*Balaenoptera physalus*), humpback whale (*Megaptera novaeangliae*), blue whale (*B. musculus*) and bowhead whale (*Balaena mysticetus*). Our results suggest that parentage studies

* Corresponding author.
** Corresponding author at: Marine Evolution and Conservation, Groningen Institute of Evolutionary Life Sciences, University of Groningen, Groningen, the Netherlands.
E-mail addresses: m.suarez.menendez@rug.nl (M. Suárez-Menéndez), m.berube@rug.nl (M. Bérubé).

[1] Shared first authorships,
[2] deceased

https://doi.org/10.1016/j.gecco.2024.e02947
Received 17 November 2023; Received in revised form 8 April 2024; Accepted 8 April 2024
Available online 9 April 2024

should always be accompanied by a power analysis in order to ascertain that each individual specific study is based upon data with sufficient power to assign parentage with statistical rigor.

1. Introduction

Conservation genetic assessments typically draw inferences from the degree of genetic diversity within and among populations or individuals, e.g., to infer relatedness among individuals, which in turn yield insights into the social organization, mating and population structure of endangered species. During recent years, single nucleotide polymorphisms (SNPs) have gained in popularity relative to short tandem repeat loci (STR), also known as microsatellites (Tautz, 1989), due to their greater abundance in genomes (Morin et al., 2004). However, genotyping STR loci is still common in conservation and ecology, due to higher level of variation per locus than SNPs (Guichoux et al., 2011). STR genotyping is a cost-effective approach and hence a logical first step and a viable alternative when resources are scarce. STR loci can be genotyped using capillary electrophoresis (Butler et al., 2001) or by high throughput sequencing (HTS, Vartia et al., 2016). Although HTS is much less costly on a per-locus basis, HTS requires a substantial initial investment and much higher numbers of loci and individuals to be cost-effective. Accordingly, STR loci genotyping by capillary electrophoresis is suitable for a preliminary assessment, in particular when resources are limited or the research objective only requires genotyping small sample sizes with highly polymorphic loci. The specific set of STR loci to be genotyped is often identified and characterized in the targeted species or from published resources in closely related species. As a result, the specific combination of STR loci typically differs among species and individual studies (Andersen et al., 2003; Rivera-León et al., 2019; Tardy et al., 2023). Although there are obvious downstream advantages in genotyping "universal" sets of STR loci, such optimization is often tedious, due to experimental issues (e.g., cross-species amplification, null alleles and ascertainment bias, Primmer et al., 1996) and varying levels of polymorphism among species. However, the DNA sequences flanking STR loci are often conserved among closely related taxa, which, in principle, should facilitate genotyping of homologous loci in closely related species. We developed a set of ten STR loci and a sexing marker in seven baleen (mysticetes) whale species that can be genotyped in two multiplex polymerase chain reactions (PCR, Saiki et al., 1986). Identifying and optimizing cross-species STR genotyping will facilitate the rapid deployment of new studies in mysticetes. The data generated from the two multiplex PCR reactions enable individual identification (Rew et al., 2011), which is a fundamental first step in population genetics studies (Mesnick et al., 2011; Palsbøll et al., 1997; Rivera-León et al., 2019; Torres-Florez et al., 2014).

STR genotyping is often applied to the identification of related individuals, such as parentage (Blouin, 2003). For example, parentage can be inferred using parentage exclusion (Christie, 2010) or from likelihoods (Gerber et al., 2003; Jones and Wang, 2010; Kalinowski et al., 2007). These two approaches, each have different strengths and weaknesses to be considered in each case, such as the effects of genotyping errors, null alleles, *de novo* mutations and underlying population structure (Jones and Ardren 2003; Pompanon et al., 2005; Pemberton et al., 1995). The number of STR loci and level of polymorphism among loci has a strong effect on the power to infer parentage between individuals, regardless of the specific method employed (Harrison et al., 2013). For example, non-parental individuals closely related to a putative offspring (e.g. siblings or its own offspring) can be incorrectly assigned as a parent if the number of STR loci is insufficient given the sample size and levels of variation among the employed STR loci (Marshall et al., 1998). Most parentage assessments undertaken in wild populations studies have been aimed at terrestrial mammals (e.g., red deer, *Cervus elaphus*, Coulson et al., 1998; or brown bears, *Ursus arctos*, Shimozuru et al., 2022), animals in captivity (e.g., giant groupers, *Epinephelus lanceolatus*, Weng et al., 2021) and marine mammal mass stranding events (Mirimin et al., 2011; Oremus et al., 2013). So far only a limited number of parentage studies have been undertaken in mysticetes, such as, humpback whales (*Megaptera novaeangliae*, Cerchio et al., 2005; Clapham and Palsbøll, 1997; Cypriano-Souza et al., 2010; Nielsen et al., 2001), minke whales (*Balaenoptera acutorostrata*, Skaug et al., 2010), North Atlantic (*Eubalaena glacialis*, Frasier et al., 2007) and southern right whales (*E. australis*, Carroll et al., 2012). Although different methods have been developed to assess the informativeness of STR loci (e.g., Vandeputte, 2012; Wang, 2006) and applied in several studies (Fernandes et al., 2017; Kozfkay et al., 2008), in mysticetes, these assessments are generally not performed. Consequently, there is little or no insight into the error rate (i.e., false positives) and hence the overall rigor of the conclusions based on parentage assignments.

Implementing paternity exclusion methods is straightforward and, with sufficient genotypes, offers a near-conclusive evidence for or against parentage (Chakraborty et al., 1988). Parentage exclusion, at its simplest, is based on Mendelian inheritance, when a putative parent-offspring pair is expected to have one allele that is identical by descent. Due to its simplicity, using parentage exclusion is straightforward to assess the effects of the number of STR loci and their level of variability on the parentage assignment.

Here we present a simple assessment and the associated code to ascertain how many STR loci are necessary to assign parentage rigorously using parentage exclusion. Towards this specific goal, we selected STR loci from previously published sources and characterized an additional 28 new STR loci. These loci were employed to assess the number of STR loci required for rigorous parentage assignment in North Atlantic fin whales (*Balaenoptera physalus*), humpback whales, blue whales (*Balaenoptera musculus*) and bowhead whales (*Balaena mysticetus*).

2. Materials and methods

All tissue samples were collected as skin biopsy from free-ranging whales as described by Palsbøll et al. (1991). Samples were stored at −20 or −80 degrees Celsius (°C) in saturated NaCl with 25% dimethylsulphoxide (Amos and Hoelzel, 1991). Sampling and shipping of whale tissue samples was conducted in accordance with national and international safety, importation guidelines and regulations.

Table 1
Primer sequences and genomic coordinates of the 28 new STR loci used in this study.

Locus	PCR fragment size	Primer designation	Scaffold&	Position&	Oligo-nucleotide sequence (5' - 3')
GATA19406	295		CM020949.2	56,264,761	
		GATA19406F$			AGGTTTAGTGAGGATCCTTCC
		GATA19406R			TGCACATCTTGTAGCTGTGTT
GATA25072	407		CM020942.2	158,399,714	
		GATA25072F			TGGACACATTTAAGGGGATAA
		GATA25072R			AACTTGATTCGCCTTACTTTG
GATA29055	159		CM020942.2	130,013,628	
		GATA29055F			GACTGGTGTTTCTCTGGAGAA
		GATA29055R			GACTTACCAGCCCCTACAAAT
GATA36068	353		CM020947.2	92,217,069	
		GATA36068F			CCAAATTGCTCTCAAGAAAGA
		GATA36068R			CTTTGGAGATCACCGTTTAGA
GATA3635	310		CM020946.2	34,123,516	
		GATA3635F			TCAAATATGGGGAGAAAAACA
		GATA3635R			TATTTATGCTTTTTGCCCATC
GATA38314	331		CM020946.2	42,779,213	
		GATA38314F			AGGAGACAGAAAACACGACTG
		GATA38314R			TACACAGGAACTTGGAGGAAG
GATA43950	368		CM020943.2	2789,293	
		GATA43950F			TGTGGAGAAGATGGGAAATAA
		GATA43950R			CCTAAACATTTCACCCACAAC
GATA52422	321		CM020957.2	12,277,462	
		GATA52422F			TGGGAATCTGCTCTAGAAAAA
		GATA52422R			GTGGACTTGCTGAGGACTTAA
GATA5890064	280		CM020957.2	52,807,970	
		GATA5890064F			ATTACCAGAACTTGGGTCTCC
		GATA5890064R			AGTGGAGTGTCATCTGAAAGC
GATA5890240	273		CM020947.2	14,589,449	
		GATA5890240F			GCACTTTGGACAGAGAACAGT
		GATA5890240R			TAAAAAGGTGACTCGATGAGC
GATA5892687	261		CM020958.2	68,848,906	
		GATA5892687F			ACTTCCTAGCCAAACTGGAAT
		GATA5892687R			ACAGATAATTGGGCCTTAGCT
GATA5943219	252		CM020944.2	122.602,677	
		GATA5943219F			CACCATGAGAGGACTTAAGGA
		GATA5943219R			ATCAAATTAAGTGTGGGCAAA
GATA5946992	283		CM020944.2	110,695,007	
		GATA5946992F			ATCGTATCAGCCACACATTTT
		GATA5946992R			TTTAGAGCACCCTCTTTCAGA
GATA5947654	250		CM020941.2	72,374,148	
		GATA5947654F			CAAAGCATAAAACCAGCAACT
		GATA5947654R			TTATCAGGAATTGGCTTATGC
GATA5984139	280		CM020949.2	40,228,786	
		GATA5984139F			TAGGACACGATGCTTTCACTT
		GATA5984139R			AACAGGGCTGGACTTAGAGAT
GATA6013633	287		CM020944.2	127,372,707	
		GATA6013633F			ACCAGAGATGTGGAACCTGTA
		GATA6013633R			TAAGGTGTTGCCTACAAGAGG
GATA6057581	231		CM020956.2	41,459,017	
		GATA6057581F			CCTAACTATACTGGAGCCCTGA
		GATA6057581R			ATTTCCAGGTCTCTGACACAG
GATA6058119	308		CM020961.2	27,069,836	
		GATA6058119F			GACCAGCTTCTCTTCTCCTCT
		GATA6058119R			TAAGTCAACGATGAGAGGGAG
GATA6058394	335		CM020941.2	15,139,983	
		GATA6058394F			AGCAGTACCCCTCACTAGCTT
		GATA6058394R			AACACTTATCAAGCCCCCTAC
GATA6059012	352		CM020942.2	38,429,421	
		GATA6059012F			CAGGAATCTCAGGGGATTTA
		GATA6059012R			AAAATGAAATGTTGCCTGAAG
GATA6059993	312		CM020950.2	91,833,689	
		GATA6059993F			AATGATCAGCCTCTCATCCTA
		GATA6059993R			GCAAACGAGGACTTTGAAATA
GATA6063318	319		CM020959.2	6562,237	
		GATA6063318F			CCCTAAGTCCTTCTTCAGGAC
		GATA6063318R			GCACTTAGGCATCTGGAAGT
GATA6063862	251		CM020962.2	4337,731	
		GATA6063862F			GGTCAAGCACAGAAAGACTGT

(continued on next page)

Table 1 (*continued*)

Locus	PCR fragment size	Primer designation	Scaffold$^{\&}$	Position$^{\&}$	Oligo-nucleotide sequence (5' - 3')
		GATA6063862R			CTGCTTCATAAGATGGCAGAT
GATA6064765	344		CM020956.2	54,087,188	
		GATA6064765F			CTTTTCTGCTTCTGTAGTGGG
		GATA6064765R			GTTTTGGGGATGAACCTAGAC
GATA6065910	340		CM020943.2	12,718,001	
		GATA6065910F			CAGAACGCTCATCTGAAAAAT
		GATA6065910R			TATGTTAGGCACCCAATAAGC
GATA6237777	116		CM020942.2	169,313,978	
		GATA6327777F			CCCATTCCACTAGATGACAGA
		GATA6237777R			TGTACCCATATCTGCCCATA
GATA91083	183		CM020959.2	54,208,483	
		GATA91083F			CCAAATTGAGACAGCAACTCT
		GATA91083R			ATTGGAAAGGAGAAGGATCAC
GATA97408	179		CM020942.2	166,113,966	
		GATA97408F			GTTGTGTTCCATTGGTTCATT
		GATA97408R			CATGTCGGTCTTTAATCCATC
GT015	179	GT015F	CM020947.2	110,509,213	ACAGAGGCTGTCCTTCCCTCC
		GT015R			TTCCCTATTAGAGGCTCACG

Notes: $^{\$}$F and R denotes forward and reverse orientation, respectively. $^{\&}$Genome coordinates (GenBank assembly accession: GCA_009873245)

2.1. Molecular analyses

Total-cell DNA was extracted using either phenol-chloroform extractions (Sambrook and Russell, 2001) or QIAGEN DNEasy™ extraction columns for animal tissue, following the manufacturer's instructions (QIAGEN Inc.). The quality of the DNA was assessed visually by electrophoresis through 0.7% agarose gel and the amount quantified with a Qubit™ following the manufacturer's instructions (Thermo Fisher Scientific Inc.). DNA extractions were adjusted to a final concentration at 10 ng DNA/μL.

Candidate tetramer STR loci with the repeat motif GATA were identified in the humpback whale genome assembly (Tollis et al., 2019) using the software SCIROKO (ver. 3.4, Kofler et al., 2007). The search for candidate loci was conducted with the following parameter settings: search mode (perfect repeats), minimum number of repeats (5), upper and lower bound of motif length (4), SSR-couple considerations (all).

Oligo-nucleotides were designed for PCR amplification of each STR locus using PRIMER3PLUS (ver. 3.2.6, Untergasser et al., 2012) with default parameter settings, except fixing the annealing temperature at 57 °C and the oligo-nucleotide length at 21 (Table 1). For each pair of oligo-nucleotides, the forward oligo-nucleotide was extended with either a universal T7 or M13 DNA sequence (Schuelke, 2000) to facilitate the labeling of the amplification products with a fluorophore (6-FAM or HEX) of the complementary T7/M13 primer during PCR and detection during capillary electrophoresis on ABI 3730 Genetic Analyzer (Applied Biosystems Inc.).

The genomic coordinates of each STR locus were determined by aligning the oligo-nucleotides against the reference blue whale genome (which is assembled at the chromosome-level, Bukhman et al., 2022) using BOWTIE2 (ver 2.3.5.1, Langmead and Salzberg, 2012) with the parameter settings defined by the preset *–very-sensitive*. PCR amplification of STR loci were conducted in 10 μL volumes, each comprising 10 ng of genomic DNA, 67 mM Tris-HCl, pH 8.8, 2 mM MgCl$_2$, 16.6 mM (NH$_4$)$_2$SO$_4$, 10 mM β-mercaptoethanol, 0.2 mM dNTPs, 1 μM of each oligo-nucleotides as well as 0.4 units of *Taq* DNA polymerase (New England Biolabs Inc.). PCR reactions were subjected to two minutes at 94 degrees Celsius (°C) followed by 35 cycles each comprising 30 seconds at 94 °C, 90 seconds at 57 °C and 30 seconds at 72 °C; followed by 10 minutes at 68 °C. The initial quality of the PCR amplification products was assessed by gel electrophoresis in 2% agarose and 1xTBE at 175 volts for 25 minutes and the products visualized under UV light at 260 nm. All candidate STR loci were amplified as described above except for the three oligo-nucleotides, which were added in the following concentrations: 1 μM of the unlabeled locus-specific oligo-nucleotides, 0.5 μM of the 5'end-labeled (HEX or 6-FAM) oligo-nucleotide and 0.5 μM of the M13/T7-extended, unlabeled locus-specific oligo-nucleotide. PCR reactions were subjected to two minutes at 94 °C followed by 10 cycles each comprising 30 seconds at 94 °C, one minute at 57 °C, and 30 seconds at 72 °C followed by an additional 27 cycles each with 30 seconds at 94 °C, one minute at 55 °C and 30 seconds at 72 °C, followed by 10 minutes at 72 °C. The length of all amplification products were resolved by capillary electrophoresis on an ABI 3730 Genetic Analyzer™ (Applied Biosystems Inc.) using the size standard GeneScan™ 500 ROX™ dye size standard (Applied Biosystems Inc.) and GENEMAPPER™ ver 4.0 (Applied Biosystems Inc.). One of each locus-specific pair of oligo-nucleotide pairs was labeled with either 6-FAM, HEX or NED for the candidate STR loci selected for further analysis and genotyped in 48 individuals in each of seven mysticete species; humpback whale, fin whale, minke whale, sei whale (*B. borealis*), blue whale, southern right whale and bowhead whale.

Multiplex PCRs were performed using the Qiagen™ Multiplex Kit Plus (Qiagen Inc.) in 5 μL reaction volumes following the manufacturer's recommendations. PCR reactions were subjected to two minutes at 94 °C, followed by 35 cycles each comprising 30 seconds at 94 °C, 90 seconds at 57 °C and 30 seconds at 72 °C, followed by 10 minutes at 68 °C. PCR amplification products were separated and detected by capillary gel electrophoresis on an ABI 3730 Genetic Analyzer™ (Applied Biosystems Inc.) including a size standard GeneScan™-500 ROX (Applied Biosystems Inc.). The length of each PCR product was determined with GENEMAPPER™ (ver. 4.1, Applied Biosystems Inc.).

Finally, a random set of 60 DNA extractions collected from different individual fin, humpback, bowhead and blue whales were genotyped at 30, 32, 30 and 30 loci, respectively, and the sex determined by co-amplification of a Y chromosome specific marker by the presence/absence of the Y chromosome specific marker where all STR loci in the multiplex set successfully amplified.

2.2. Data analysis

Allele frequencies, the observed (H_O) and expected (H_E) heterozygosity, as well as the probability of identity for pairs of unrelated individuals (P(I), Paetkau and Strobeck, 1994), corrected for low sample sizes, was estimated as implemented in GIMLET (ver. 1.3.3, Valière, 2002) were estimated in all datasets. Possible deviations from the expected Hardy-Weinberg (Hardy, 1908; Weinberg, 1908) genotype frequencies (HWE) was estimated as implemented in GENALEX (ver. 6.5, Peakall and Smouse, 2012, 2006). The non-exclusion probability of the first parent ($P_{\text{NON-EXCL}}$, Selvin, 1980) was estimated using CERVUS (ver. 3.0.7, Kalinowski et al., 2007) or with a custom script coded in Python (ver. 3, Van Rossum and Drake Jr, 1995). As CERVUS can only analyse one dataset at a time, the same $P_{\text{NON-EXCL}}$ calculation was implemented in the Python script to allow for analysis of large numbers of datasets. $P_{\text{NON-EXCL}}$ denotes the probability that given a set of loci, an unrelated parent will not be excluded as the putative parent.

2.3. Parent-offspring assignment

Putative parent and offspring (PO) pairs were identified using a custom script coded in Python (ver. 3) that identified pairs of multilocus genotypes that segregated according to Mendelian expectations for PO pairs. Sire-dam-offspring (SDO) trios were identified in a similar manner, i.e., by first identifying putative dam and offspring pairs among all possible pairs of multilocus genotypes requiring that the dam is a female. Subsequently, putative sires were identified among the male genotypes where the multi-locus STR genotype was consistent with the putative dam's and offspring's genotypes and Mendelian segregation. The assessment did not allow for mismatches, i.e., putative PO pairs with loci that did not segregate according to Mendelian expectations (e.g., due to genotyping errors, null-alleles or missing data) were not considered.

KININFOR (ver. 2, Wang, 2006) was employed to assess the informativeness of the STR loci. Since most relatedness estimators are highly correlated, the "informativeness of relationship" (IR) criterion was used to rank STR loci in terms of statistical power to discern among different degrees of relatedness. Two assessments were conducted to assess the power of a given number of STR loci to infer PO pairs. The first assessment was based upon either the most or least informative STR loci ranked by their IR value (see above); the second assessment was based upon random samples of loci chosen among the STR loci genotyped without replacement. In both assessments, data sets from eight to the maximum number of loci genotyped for each data set were generated and PO pairs and SDO trios were identified as described above.

2.4. Assessments of paternity assignments

Close male relatives to the individuals in putative PO pairs are among the individuals most likely to be incorrectly assigned as the sire. Although a slight deviation from random mating has been detected (Cerchio et al., 2005), behavioral observations (e.g., Frasier et al., 2007) and genetic studies (e.g., Clapham and Palsbøll, 1997) suggest that baleen whales are promiscuous and that mating can be considered random. In a finite population there is always a possibility that a dam mate with the same sire multiple times thus producing offspring related as full siblings. However, such instances are likely rare unless the overall breeding population size is very small. This implies that first (the offspring's own offspring) and second-order relatives (the offspring's grandparents, grandchildren and half-siblings) are the individuals that are most likely to be incorrectly assigned as the sire. Since each offspring has more second-order than first-order relatives, we focused our assessment of the probability of incorrectly assigning paternity to second-order relatives. Using half-siblings as representing second-order relatives, we estimated the probability of incorrectly assigning a half-sibling as the sire by generating simulated data sets. We generated *in silico* pairs of half-siblings by randomly sampling alleles at each STR and sex locus from *in silico* pairs of sires and dams. The probability of an incorrect paternity assignment in each assessment was estimated as the proportion of half-sibling pairs in which one allele was shared with the sire at all STR loci. In this assessment up to two mismatches (i.e., loci at which the half-sibling and sire did not share an allele) were allowed. The assessment was conducted with a maximum of 50 STR loci sampled at random, with replacement, from the observed data. For each set of STR loci, 5000 calves and half-siblings were simulated, and the assessment was repeated 100 times. The median probability of an incorrect paternity assignment for each set of STR loci was smoothed using a Savitzky-Golay filter (Savitzky and Golay, 1964). The "knee of the curve", or critical number of loci (i.e. the point where adding additional STR loci had a diminished effect) was estimated using the Python package *kneed* (ver. 0.8.2, Satopaa et al., 2011).

3. Results

3.1. Newly characterized str loci

A total of 5047 STR loci with the tetramer repeat motif GATA and a minimum of five repeats were detected in the humpback genome assembly. We chose and tested 44 STR loci at random among which 28 STR loci (Table 1) were deemed of sufficient quality in terms of the (a) "cleanness" of the PCR amplification product, and (b) indications of multiple alleles in some mysticete species inferred from the agarose gel electrophoresis (above). The 28 selected STR loci were genotyped in eight DNA extractions in each mysticete species. The majority of the STR loci amplified in the rorquals species, and fewer in the right and bowhead whale (Table S1). Among the 28 candidate loci, ten STR loci that were polymorphic in most baleen whales were selected for further characterization. The ten selected candidate STR loci and a sex specific Y chromosome-specific locus were optimized for multiplexed PCR amplification. PCR amplification was conducted in two reactions, each with five STR loci. The Y chromosome-specific locus was co-amplified in one of the two multiplex PCR panels (Table 2). In total, 48 DNA extractions from each mysticete species were genotyped with these two multiplex PCR panels. Among the ten newly selected loci, three loci were monomorphic in some species: GATA25072 (bowhead whale), GATA5947654 (southern right whale) and GATA6237777 (blue whale). H_E was similar among species (Table 3). The range of allele lengths at locus GATA5947654 differed in the blue whale, and locus GATA5984139 in the bowhead whale and the southern right whale relative to the remaining species (Table S1). P(I) ranged from 5.1×10^{-12} (*B. physalus*) to 2.5×10^{-8} (*B. acutorostrata*) among the seven sets of 48 samples (Table 3), yielding an expectation of minimum 5.7×10^{-9} (*B. physalus*) and 2.8×10^{-5} (*B. acutorostrata*) sample pairs of unrelated individuals having identical genotypes due by chance.

3.2. Parentage assignments

In order to evaluate the impact of both the quantity and informativeness of the genetic markers on the parentage inference, a total of 1300 datasets per species were generated by selecting different combinations of loci from that observed datasets of 60 whale individuals (30 females and 30 males), all genotyped at either 30 (blue, bowhead and fin whale) or 32 loci (humpback whale). Only the blue and bowhead whale datasets contained missing data, missing a total of 14 and five genotypes, respectively. No sample was missing data at more than two genotypes, nor were there any samples with identical multi-locus genotypes.

$P_{\text{NON-EXCL}}$ was estimated with CERVUS from the above sets of 60 individuals in each of the four above-mentioned species (Table S2). The highest mean diversity and power to exclude an unrelated individual as a parent was estimated in fin whales (30 loci, mean $H_E = 0.77$, $P_{\text{NON-EXCL}} = 7\times10^{-8}$). The lowest power was observed in the blue whale sample (30 loci, mean $H_E = 0.70$, $P_{\text{NON-EXCL}} = 3.6\times10^{-6}$). The variation in the humpback (32 loci, mean $H_E = 0.75$; and $P_{\text{NON-EXCL}} = 2.9\times10^{-7}$) and bowhead whale (30 loci, mean $H_E = 0.73$, $P_{\text{NON-EXCL}} = 5.2\times10^{-7}$, respectively) samples were in between the two extremes (Table S2).

PO pairs were identified in each species as pairs of samples with the expected Mendelian segregation of alleles. Requiring no missing data or mismatches; 20, 12, six and five PO pairs in total were detected in humpback, fin, bowhead and blue whales, respectively. No additional (incorrect) PO pair assignments were detected if the assessment was based on as few as 17, 11, 27 and 17 of the STR loci with the highest IR values. In contrast, 25, 23, 28 and 30 of the least informative loci were required to achieve the same result (Fig. 1, Table S2).

The number of putative PO pairs detected varied greatly among species, number of loci and the specific combination of STR loci. The largest variation in the number of PO pairs detected was observed in data sets comprising only eight STR loci (selected at random among all genotyped loci). Among the 100 eight-loci datasets, the highest degree of variation in the number of observed putative PO pairs was found in the blue whale for which the dataset that yielded the lowest number of PO pairs only accounted for 2.7% of the pairs identified in the dataset that yielded the highest number (seven and 261 PO pairs respectively). Humpback whales presented the lowest variation from 27 to 137 PO pairs observed (19.7% variation) among the 100 eight-loci datasets (Fig. 2, top graph, range). Thus, a large variability was observed in the number of PO pairs obtained from the same number of STR loci depending on the species and the informativeness of STR loci in each dataset.

Similar variability was observed in the detection of SDO trios. Among the 100 eight STR loci datasets, the dataset with the fewest observed SDO trios accounted for only 1.2% and 17.9% of the highest number of SDO trios detected for blue whales and humpback whales, respectively (Fig. 2, middle graph, range). These variability patterns, described above, were also reflected in the ranges of the $P_{\text{NON-EXCL}}$ estimated with the custom Python script for the same datasets, (Fig. 2, bottom graph).

Table 2
Multiplex oligo-nucleotide panels for selected mysticete species.

Panel A					
Locus	Primer Reverse	Primer Forward	Dye	Range	
GATA6237777	R	F	HEX	108	152
GATA6064765*	R2	F2	HEX	182	230
GATA6063318	R	F	HEX	300	350
GATA91083	R	F	FAM	149	220
GATA38314	R	F	FAM	303	367
Panel B					
Locus	Primer Reverse	Primer Forward	Dye	Range	
GATA5947654*	R3	F3	HEX	130	160
GATA5984139	R	F	HEX	204	280
GATA52422RF	R	F	HEX	320	479
GATA25072*	R2	F2	FAM	76	148
GATA97408	R2	F2	FAM	164	215
SRY	R	F	FAM	332	332

* New primers were re-designed in order to enable multiplexing: GATA25072F2 (5'-CACCTGCTTTAAACTGTGTATAGT-3'), GATA25072R2 (5'-GATCTAGCAACTCTTTCTAGGC-3'), GATA6064765F2 (5'-GTACAAATGCACTTTCTCCCG-3') and GATA6064765R2 (5'-AGGCACTTATCAGTTCCAAGT-3'), SRYF, (5'-TGTGAACGGTGAGGATTA-3') and SRYR, (5'-GTGCATGGCTCGTAGTCT-3') GATA5947654R3 (5'-TCAGCCTCCATAATTGCATAAG-3') and GATA5947654F3 (5'-GTTATTAGATAGGGTTCTCTGCAG-3'), GATA97408F2 (5'-CTCCCACCACTGATTTGTAATA-3') and GATA97408R2 (5'-AGCCTAGTTTTATGGTACCTCT-3').

Table 3
Summary statistics for ten microsatellite loci in 48 individuals in a selection of mysticete species.

Species	Locus	N_A	H_O	H_E	Pr(*I*)	HWE Signif
B. borealis	GATA25072	6	0.83	0.75	9.55E-02	ns
	GATA38314	7	0.63	0.73	1.04E-01	***
	GATA52422	4	0.65	0.64	1.57E-01	ns
	GATA5947654	7	0.83	0.77	7.75E-02	ns
	GATA5984139	8	0.79	0.72	8.91E-02	ns
	GATA6063318	9	0.88	0.83	4.04E-02	ns
	GATA6064765	10	0.94	0.86	2.71E-02	ns
	GATA6237777	2	0.13	0.12	7.76E-01	ns
	GATA91083	4	0.81	0.67	1.64E-01	*
	GATA97408	7	0.69	0.76	7.81E-02	ns
	All loci				1.17E-10	
E. australis	GATA25072	3	0.42	0.41	4.06E-01	***
	GATA38314	20	0.63	0.83	3.74E-02	**
	GATA52422	11	0.25	0.76	6.81E-02	***
	GATA5947654	1	0	0	1.00E+00	-
	GATA5984139	5	0.21	0.35	4.31E-01	***
	GATA6063318	8	0.5	0.79	6.82E-02	***
	GATA6064765	11	0.54	0.88	1.99E-02	ns
	GATA6237777	7	0.81	0.8	5.57E-02	ns
	GATA91083	3	0.46	0.43	3.68E-01	ns
	GATA97408	8	0.79	0.79	6.56E-02	ns
	All loci				8.13E-10	
B. acutorostrata	GATA25072	4	0.58	0.59	2.04E-01	ns
	GATA38314	7	0.77	0.79	6.80E-02	ns
	GATA52422	6	0.65	0.66	1.57E-01	ns
	GATA5947654	4	0.23	0.28	5.18E-01	ns
	GATA5984139	8	0.48	0.53	2.43E-01	***
	GATA6063318	6	0.75	0.7	1.36E-01	ns
	GATA6064765	7	0.77	0.76	8.68E-02	ns
	GATA6237777	5	0.63	0.61	2.03E-01	ns
	GATA91083	6	0.63	0.61	2.18E-01	ns
	GATA97408	4	0.6	0.67	1.74E-01	ns
	All loci				2.49E-08	
M. novaeangliae	GATA25072	9	0.85	0.84	3.65E-02	ns
	GATA38314	6	0.77	0.67	1.43E-01	ns
	GATA52422	6	0.75	0.72	1.19E-01	ns
	GATA5947654	5	0.73	0.67	1.38E-01	ns
	GATA5984139	11	0.81	0.82	4.27E-02	***
	GATA6063318	9	0.9	0.81	5.17E-02	ns
	GATA6064765	6	1	0.77	8.21E-02	**
	GATA6237777	8	0.9	0.85	3.29E-02	ns
	GATA91083	7	0.77	0.77	7.30E-02	ns
	GATA97408	7	0.73	0.75	7.96E-02	ns
	All loci				2.97E-12	
B. physalus	GATA25072	12	0.88	0.85	2.94E-02	***
	GATA38314	10	0.75	0.75	8.93E-02	***
	GATA52422	15	0.81	0.76	6.12E-02	**
	GATA5947654	8	0.67	0.69	1.36E-01	***
	GATA5984139	13	0.85	0.86	2.76E-02	***
	GATA6063318	12	0.73	0.73	9.80E-02	***
	GATA6064765	10	0.88	0.85	3.20E-02	ns
	GATA6237777	4	0.31	0.47	3.62E-01	***
	GATA91083	5	0.69	0.69	1.30E-01	ns
	GATA97408	8	0.85	0.8	5.72E-02	ns
	All loci				5.10E-12	
B. mysticetus	GATA25072	1	0	0	1.00E+00	-
	GATA38314	12	0.79	0.81	4.18E-02	ns
	GATA52422	6	0.75	0.73	1.01E-01	ns
	GATA5947654	7	0.4	0.71	1.21E-01	***
	GATA5984139	3	0.44	0.5	3.49E-01	ns
	GATA6063318	7	0.67	0.61	1.90E-01	ns
	GATA6064765	12	0.79	0.78	5.51E-02	***
	GATA6237777	6	0.81	0.74	9.40E-02	ns
	GATA91083	5	0.46	0.75	9.70E-02	***
	GATA97408	4	0.08	0.12	7.66E-01	***
	All loci				1.30E-08	
B. musculus	GATA25072	8	0.83	0.78	6.49E-02	ns
	GATA38314	15	0.77	0.88	1.55E-02	***

(*continued on next page*)

Table 3 (*continued*)

Species	Locus	N_A	H_O	H_E	Pr(*I*)	HWE Signif
	GATA52422	3	0.29	0.5	2.82E-01	ns
	GATA5947654	3	0	0.1	8.10E-01	***
	GATA5984139	13	0.56	0.88	1.91E-02	***
	GATA6063318	14	0.65	0.86	2.45E-02	*
	GATA6064765	28	0.73	0.93	4.17E-03	***
	GATA6237777	1	0	0	1.00E+00	-
	GATA91083	11	0.81	0.78	6.84E-02	ns
	GATA97408	9	0.6	0.75	8.55E-02	ns
	All loci				2.62E-12	

Notes. N_A denotes the number of alleles at each locus and H_O and H_E, observed and expected heterozygosity, respectively. HWE signif, denotes the level of significance for Hardy-Weinberg equilibrium. The probability of identity, P(*I*) (for a population where individuals randomly mate with correction for small samples of individuals), are estimated as described in Paetkau et Strobeck (1994) and Kendall, Stuart, (1977). Finally, ns=not significant, * P<0.05, ** P<0.01, *** P<0.001.

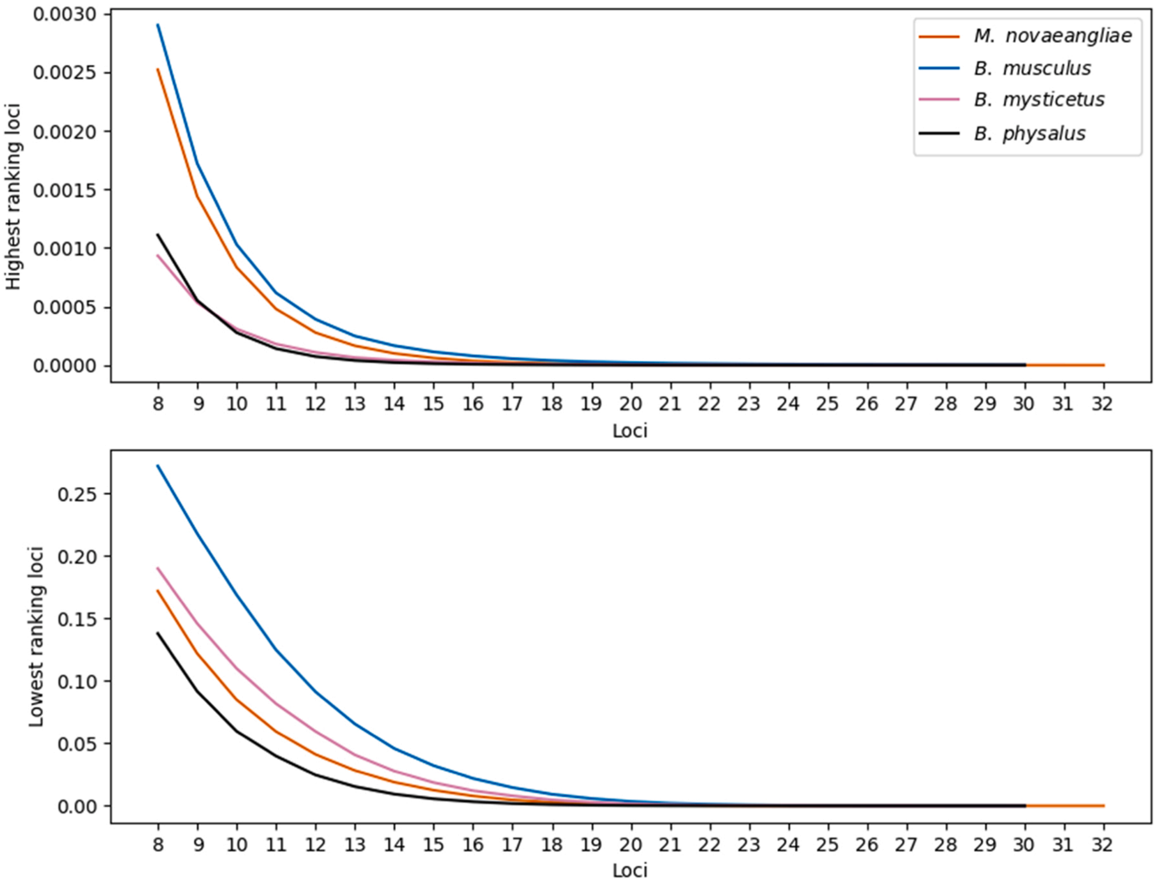


Fig. 1. Combined $P_{\text{NON-EXCL}}$ estimated with the custom Python script for STR loci ranked by their IR value, starting with the first eight ranked loci (by increasing and decreasing IR values respectively) adding an additional locus consecutively.

As expected, the median (and range) in the number of putative PO pairs and SDO trios declined with an increasing number of STR loci (i.e., when the power to exclude untrue parents increased). The minimum number of loci to detect the same number of PO pairs and SDO trios observed in the full data set differed among species (i.e., when the power to exclude untrue parents was "sufficient", Fig. 2, vertical lines).

The relative percentage of half-siblings assigned as a sire was consistent with the combined non-exclusion probabilities for the different datasets in each species. The probabilities of incorrect assignments from highest to lowest were detected in blue, humpback, fin and bowhead whales (Fig. 3). The critical number of loci was 15 and 17 loci for zero or one mismatch, respectively in the bowhead whale; 16 and 17 in the fin whale; 16 and 18 in the humpback whale; and 17 and 19 in the blue whale. Allowing for two mismatches (or missing data) the critical number of loci increased to 21 for the humpback and bowhead whales, 22 for blue whales and 19 for the fin

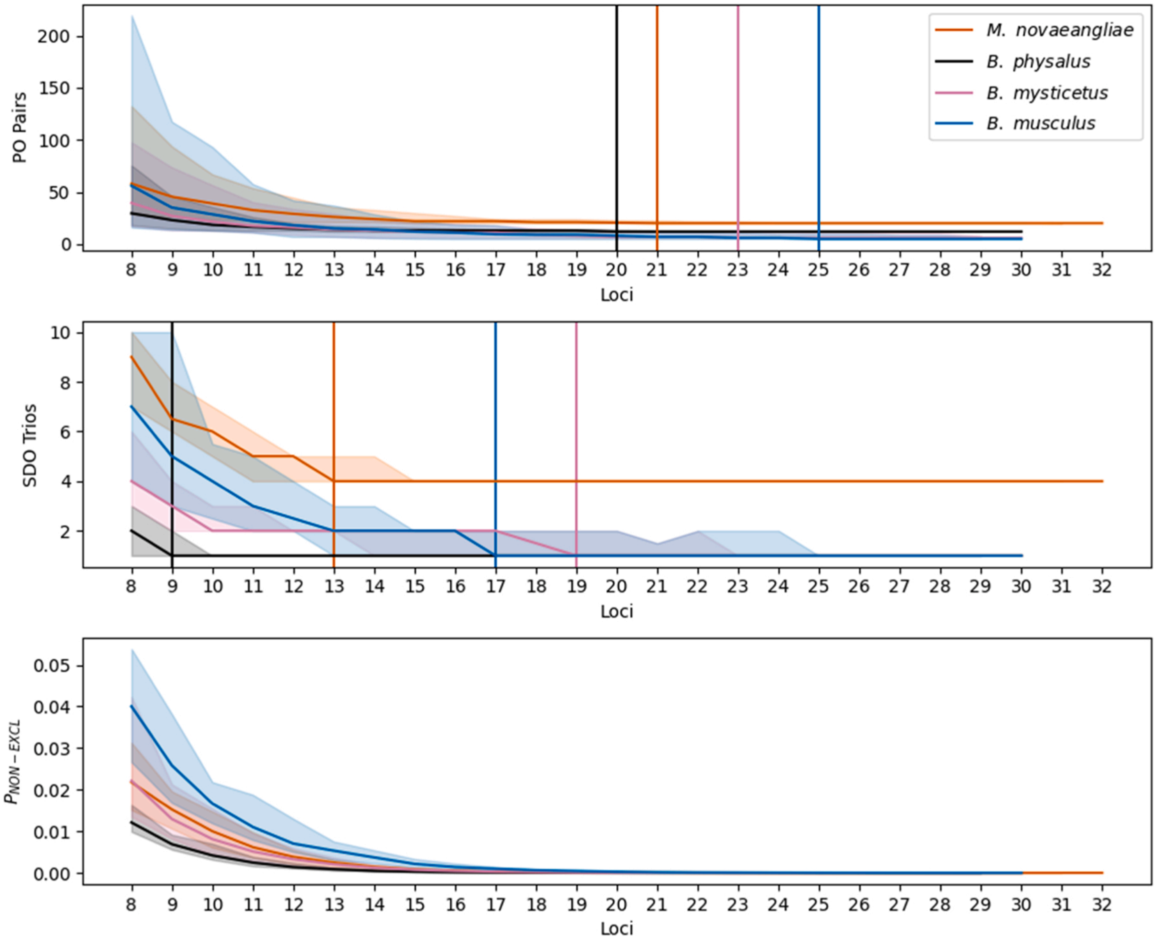


Fig. 2. Effects of the number and informativeness (IR) of STR loci on PO pairs (top), SDO trio assignment (middle) and $P_{\text{NON-EXCL}}$ estimated with the custom Python script (bottom). Vertical lines indicate at which number of STR loci the median begins to equal number of PO/SDOs observed in the full data set. Shaded areas indicate the full range of values obtained per number of loci.

whales. More generally, allowing mismatches greatly impacted the probability of incorrect assignments, e.g., between a 28 and 139-fold difference between zero and two mismatches at 18 loci.

4. Discussion and conclusions

Multi-allelic STR loci remain a cost-efficient alternative to genomic approaches compared to whole genome sequencing and reduced representation methods, such as genotype-by-sequencing (e.g., Peterson et al., 2012). The low cost per sample of STR genotyping enables efficient screening of samples using either HTS or capillary electrophoresis, depending on available resources and sample size, e.g., for duplicate samples and pairs of first order relatives. PO pairs, and especially full pedigrees (i.e., SDO trios, Suárez-Menéndez et al., 2023), are interesting targets for many conservation and ecological aspects, such as kinship mark-recapture (Palsbøll, 1999; Skaug, 2001).

The increasing access to reference genomes makes identification of suitable STR loci simple and straightforward. In this study, 28 “novel” STR loci were identified among ~5000 candidate STR loci in the humpback reference genome from which a STR genotyping assay comprising ten STR loci and a Y chromosome-specific locus was developed that amplified in two multiplex PCR amplifications in seven baleen whale species. Standardization of a basic set of STR loci applicable to a group of closely related species, simplifies the deployment of new studies (incl. species identification as shown here).

Individual identification is an indispensable step in many studies, from mark-recapture (Frasier et al., 2020; Smith et al., 1999) to population genetics (Mesnick et al., 2011; Rivera-León et al., 2019; Torres-Florez et al., 2014). The combined P(*I*) obtained for the seven species in the 10 STR loci assay was sufficiently low to carry out individual identification with a high level of confidence (Rew et al., 2011). To address genotype errors and missing data, a minimum number of matching STR loci is typically established for individual identification (Rew et al., 2011). Alternatively, a combined P(*I*) threshold can be implemented where samples are deemed duplicates if the combined P(*I*) of the matching STR loci meets the threshold, irrespective of the number of STR loci.

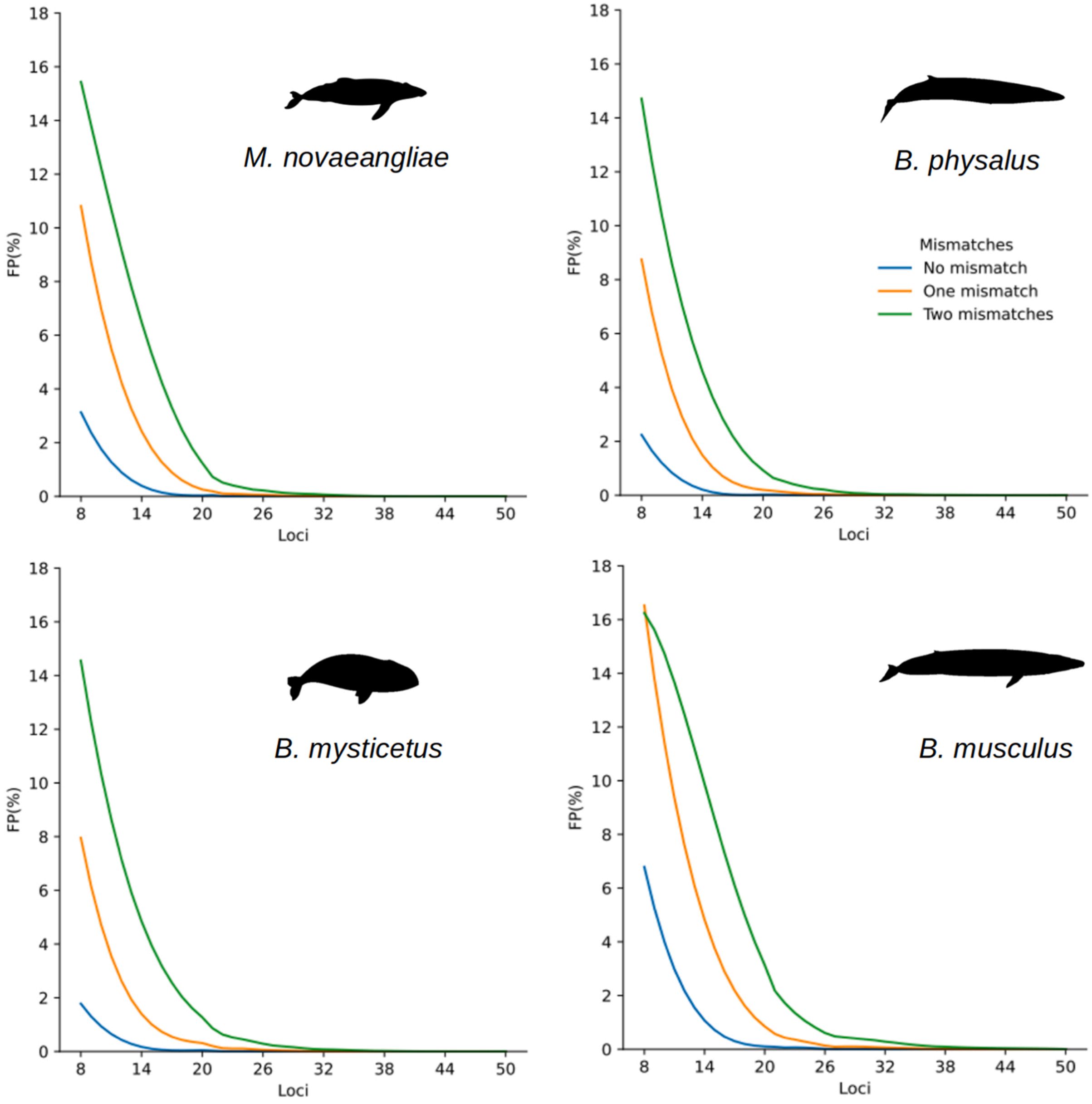


Fig. 3. Percentage of simulated half-siblings assigned as a sire (false positives, FP) per number of loci within each species.

The second goal of this study was to determine how many STR loci are necessary to identify PO pairs and SDO trios with a high degree of confidence. The power of STR loci to discern between PO pairs and other close relationships is mostly a function of the level of polymorphism at each included locus, the mating system of the targeted species, and reproductive rates. Some studies have inferred parentage employing as few as three STR loci (Zane et al., 1999) and among baleen whales, parentage assignments were based on as few as nine to 13 STR loci (Carroll et al., 2012; Cypriano-Souza et al., 2010). In this study, we demonstrate that the number of STR loci required to identify PO pairs with statistical rigor (using parentage exclusion probabilities) can be assessed in a relatively straightforward manner, hence maximizing the power and reliability of downstream parentage analysis. Unsurprisingly, we found that it is necessary to consider both the number of STR loci and the degree of variation at each locus. The difference, in terms of how many loci that must be genotyped for a rigorous paternity assignment, was substantial between the loci with the lowest and highest IR values. Homologous STR loci differ in the degree of informativeness among species and populations, e.g., populations with lower genetic diversity, which, in turn, lowers the power to identify PO pairs and SDO trios. Thus, parentage-based studies should be accompanied by a power analysis of the kind conducted here, applied to the specific dataset and the results reported. Among mysticetes, only a few parentage assessments have undertaken such a power analysis. Some studies accepted high rates of false positives (e.g., the 80% or 95% "confidence" options available in CERVUS). Lastly, some studies applied *ad hoc* criteria to reject PO pairs and SDO trios, e.g., discarding PO pairs if two or more putative males were assigned as the putative sire to an individual, but accepting PO pairs and SDO trios (in the same data set) when only a single male was assigned as the putative sire (e.g., Carroll et al., 2012; Frasier et al., 2007), thus, ignoring the unsampled part of the population as well as incorrectly assigning a close relative as the putative sire.

Genotyping errors might result in incorrect parentage/paternity exclusions and thus it is tempting to allow a few mismatches when assigning parentage/paternity. However, allowing even a few mismatches may lead to a significant increase in incorrect parentage/ paternity assignments (Wang, 2010), especially when the number of loci or degree of polymorphism is low, as observed in this study. Therefore, the consequences of permitting an arbitrary number of mismatches on the rate of incorrect parentage/paternity assignment should be assessed. Ideally, after identifying putative PO pairs with some Mendelian violations (as many as possible), those putative genotyping errors (which would result in missing some parentage assignments) should be re-genotyped, either to correct a possible genotype error or confirm the Mendelian violation (Hoffman and Amos, 2005). In addition, null alleles (Paetkau and Strobeck, 1995) could also cause loci violating the expected Mendelian segregation in a PO pair. Null alleles at a STR locus are alleles that fail to amplify, likely due to a sequence polymorphism in the priming site. This results in both individuals of a putative PO pair being homozygous but for different alleles (i.e., they share the non-amplifying allele). In order to minimize the number of possible null alleles, new oligo-nucleotides for the locus in question must be re-designed and homozygous samples genotyped at the locus in question. Initially, this extra effort may seem costly in terms of additional experimental effort but should be viewed against the gain of having a more reliable dataset.

In the same vein, missing genotypes are common in most data sets and often not reported or under-reported. For example, missing genotypes are usually reported as the overall percentage of missing genotypes in the dataset, or as the minimum number of loci used in assignments (e.g., Gerber et al., 2022). Since the effect of employing a reduced number of loci is likely to increase the fraction of incorrect parentage/paternity assignments, the effect of the inclusion of samples with the minimum allowed number of genotypes (and their level of polymorphism) should be assessed as well to ascertain that the pre-set minimum number of genotypes actually is sufficient to identify PO pairs and SDO trios with a reasonable power.

Even though other paternity assessments methods, such those based on likelihoods (Gerber et al., 2003; Jones and Wang, 2010; Kalinowski et al., 2007) sometimes factor in the presence of genotype errors, null alleles or missing data, these factors can still affect the results and should not be ignored (Harrison et al., 2013). Thus, an evaluation of the STR loci and dataset employed is necessary, regardless of the paternity assessment method.

In conclusion, we presented a new set of 10 STR loci and one Y chromosome marker that are amplified in two multiplexes in baleen whales. This set of STR loci (including a sex marker) serves as an easy, optimized starting point to conduct individual-based studies in mysticetes. We also provide extended sets of STR loci (from previously published sources) with which to conduct rigorous parentage/ paternity assignments along with the necessary Python scripts to assess the statistical power of specific sets of STR loci in order to ascertain the power to exclude non-parental genotypes.

CRediT authorship contribution statement

Christian Ramp: Writing – review & editing, Resources. **Jooke Robbins:** Writing – review & editing, Resources. **Richard Sears:** Writing – review & editing, Resources. **Marcos Suárez-Menéndez:** Writing – review & editing, Writing – original draft, Software, Formal analysis, Conceptualization. **Mónica A. Silva:** Writing – review & editing, Resources. **Martine Bérubé:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Mariel T.I ten Doeschate:** Writing – review & editing, Resources. **Lutz Bachmann:** Writing – review & editing, Resources. **Marc Tollis:** Writing – review & editing, Resources. **Peter Best:** Resources. **Els Vermeulen:** Writing – review & editing, Resources. **Nick Davison:** Writing – review & editing, Resources. **Gísli A. Víkingsson:** Resources. **Mads Peter Heide-Jørgensen:** Writing – review & editing, Resources. **Øystein Wiig:** Writing – review & editing, Resources. **Véronique Lesage:** Writing – review & editing, Resources. **Per J. Palsbøll:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Conceptualization. **Tom Oosting:** Writing – review & editing, Data curation. **Rui Prieto:** Writing – review & editing, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and scripts are available in Zenodo https://zenodo.org/doi/10.5281/zenodo.10948827

Acknowledgements

MSM was supported by a doctorate fellowship from the University of Groningen. PJP and MB were supported by the University of Groningen, University of Copenhagen, University of California Berkeley and Stockholm University. Additional funding was provided by the Aage V. Jensen Foundation (PJP), US National Marine Fisheries Service (PJP and MB), the International Whaling Commission Scientific Committee (MB and PJP), WWF-DK (PJP), the Commission for Scientific Research in Greenland (PJP) and the Greenland Home Rule (PJP). MAS and RP were supported by AZORES 2020, through the EU Fund 01–0145-FEDER-000140. The authors would like to thank Andrew Brownlow for access to samples from stranded whales as well as Wensi Hao for her technical assistance.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.gecco.2024.e02947.

References

Amos, W., Hoelzel, A.R., 1991. Long-term preservation of whale skin for DNA analysis. Rep. Int. Whal. Comm. Spec. Issue 13, 99–104.

Andersen, L.W., Born, E.W., Dietz, R., Haug, T., Øien, N., Bendixen, C., 2003. Genetic population structure of minke whales *Balaenoptera acutorostrata* from Greenland, the North East Atlantic and the North Sea probably reflects different ecological regions. Mar. Ecol. Prog. Ser. 247, 263–280. https://doi.org/10.3354/meps247263.

Blouin, M.S., 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. 18, 503–511. https://doi.org/10.1016/S0169-5347(03)00225-8.

Bukhman, Y.V., Morin, P.A., Meyer, S., Chu, L.-F. (Jack), Jacobsen, J.K., Antosiewicz-Bourget, J., Mamott, D., Gonzales, M., Argus, C., Bolin, J., Berres, M.E., Fedrigo, O., Steill, J., Swanson, S.A., Jiang, P., Rhie, A., Formenti, G., Phillippy, A.M., Harris, R.S., Wood, J.M.D., Howe, K., Kirilenko, B., Munegowda, C., Hiller, M., Jain, A., Kihara, D., Johnston, J.S., Ionkov, A., Raja, K., Toh, H., Lang, A., Wolf, M., Jarvis, E., Thomson, J.A., Chaisson, M.J.P., Stewart, R., 2022. A high-quality blue whale genome. Segm. Duplic., Hist. Demogr. https://doi.org/10.21203/rs.3.rs-1910240/v1.

Butler, J.M., Ruitberg, C.M., Vallone, P.M., 2001. Capillary electrophoresis as a tool for optimization of multiplex PCR reactions. Fresenius J. Anal. Chem. 369, 200–205. https://doi.org/10.1007/s002160000641.

Carroll, E.L., Childerhouse, S.J., Christie, M., Lavery, S., Patenaude, N., Alexander, A., Constantine, R., Steel, D., Boren, L., Scott Baker, C., 2012. Paternity assignment and demographic closure in the New Zealand southern right whale. Mol. Ecol. 21, 3960–3973. https://doi.org/10.1111/j.1365-294X.2012.05676.x.

Cerchio, S., Jacobsen, J.K., Cholewiak, D.M., Falcone, E.A., Merriwether, D.A., 2005. Paternity in humpback whales, *Megaptera novaeangliae*: assessing polygyny and skew in male reproductive success. Anim. Behav. 70, 267–277. https://doi.org/10.1016/j.anbehav.2004.10.028.

Chakraborty, R., Meagher, T.R., Smouse, P.E., 1988. Parentage Analysis with Genetic Markers in Natural Populations. I. the Expected Proportion of Offspring with Unambiguous Paternity. Genetics 118, 527–536.

Christie, M.R., 2010. Parentage in natural populations: novel methods to detect parent-offspring pairs in large data sets. Mol. Ecol. Resour. 10, 115–128. https://doi.org/10.1111/j.1755-0998.2009.02687.x.

Clapham, P.J., Palsbøll, P.J., 1997. Molecular analysis of paternity shows promiscuous mating in female humpback whales (*Megaptera novaeangliae*, Borowski). Proc. R. Soc. Lond. Ser. B 264, 95–98.

Coulson, T.N., Pemberton, J.M., Albon, S.D., Beaumont, M., Marshall, T.C., Slate, J., Guinness, F.E., Clutton-Brock, T.H., 1998. Microsatellites reveal heterosis in red deer. Proc. R. Soc. B Biol. Sci. 265, 489–495.

Cypriano-Souza, A.L., Fernández, G.P., Lima-Rosa, C.A.V., Engel, M.H., Bonatto, S.L., 2010. Microsatellite Genetic Characterization of the Humpback Whale (*Megaptera novaeangliae*) Breeding Ground off Brazil (Breeding Stock A). J. Hered. 101, 189–200. https://doi.org/10.1093/jhered/esp097.

Fernandes, T., Herlin, M., Belluga, M.D.L., Ballón, G., Martinez, P., Toro, M.A., Fernández, J., 2017. Estimation of genetic parameters for growth traits in a hatchery population of gilthead sea bream (*Sparus aurata* L.). Aquac. Int. 25, 499–514. https://doi.org/10.1007/s10499-016-0046-5.

Frasier, T.R., Hamilton, P.K., Brown, M.W., Conger, L.A., Knowlton, A.R., Marx, M., Slay, C.K., Kraus, S.D., White, B.N., 2007. Patterns of male reproductive success in a highly promiscuous whale species: the endangered North Atlantic right whale. Mol. Ecol. 16, 5277–5293. https://doi.org/10.1111/j.1365-294X.2007.03570.x.

Frasier, T.R., Petersen, S.D., Postma, L., Johnson, L., Heide-Jørgensen, M.P., Ferguson, S.H., 2020. Abundance estimation from genetic mark-recapture data when not all sites are sampled: An example with the bowhead whale. Glob. Ecol. Conserv. 22, e00903 https://doi.org/10.1016/j.gecco.2020.e00903.

Gerber, L., Connor, R.C., Allen, S.J., Horlacher, K., King, S.L., Sherwin, W.B., Willems, E.P., Wittwer, S., Krützen, M., 2022. Social integration influences fitness in allied male dolphins. Curr. Biol. 32, 1664–1669.e3. https://doi.org/10.1016/j.cub.2022.03.027.

Gerber, S., Chabrier, P., Kremer, A., 2003. famoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. Mol. Ecol. Notes 3, 479–481. https://doi.org/10.1046/j.1471-8286.2003.00439.x.

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., Lepoittevin, C., Malausa, T., Revardel, E., Salin, F., Petit, R. j, 2011. Current trends in microsatellite genotyping. Mol. Ecol. Resour. 11, 591–611. https://doi.org/10.1111/j.1755-0998.2011.03014.x.

Hardy, G.H., 1908. Mendelian Proportions in a Mixed Population. Science 28, 49–50. https://doi.org/10.1126/science.28.706.49.

Harrison, H.B., Saenz-Agudelo, P., Planes, S., Jones, G.P., Berumen, M.L., 2013. Relative accuracy of three common methods of parentage analysis in natural populations. Mol. Ecol. 22, 1158–1170. https://doi.org/10.1111/mec.12138.

Hoffman, J.I., Amos, W., 2005. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Mol. Ecol. 14, 599–612. https://doi.org/10.1111/j.1365-294X.2004.02419.x.

Jones, O.R., Wang, J., 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol. Ecol. Resour. 10, 551–555. https://doi.org/10.1111/j.1755-0998.2009.02787.x.

Kalinowski, S.T., Taper, M.L., Marshall, T.C., 2007. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16, 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x.

Kendall, M.G., Stuart, A., 1977. The Advanced Theory of Statistics. Macmillan, New York.

Kofler, R., Schlötterer, C., Lelley, T., Affiliations, 2007. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23, 1683–1685 https://doi.org/doi: 10.1093/bioinformatics/btm157.

Kozfkay, C.C., Campbell, M.R., Heindel, J.A., Baker, D.J., Kline, P., Powell, M.S., Flagg, T., 2008. A genetic evaluation of relatedness for broodstock management of captive, endangered Snake River sockeye salmon, *Oncorhynchus nerka*. Conserv. Genet. 9, 1421–1430. https://doi.org/10.1007/s10592-007-9466-0.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.

Marshall, T.C., Slate, J., Kruuk, L.E.B., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639–655. https://doi.org/10.1046/j.1365-294x.1998.00374.x.

Mesnick, S.L., Taylor, B.L., Archer, F.I., Martien, K.K., Treviño, S.E., Hancock-Hanser, B.L., Moreno Medina, S.C., Pease, V.L., Robertson, K.M., Straley, J.M., Baird, R. W., Calambokidis, J., Schorr, G.S., Wade, P., Burkanov, V., Lunsford, C.R., Rendell, L., Morin, P.A., 2011. Sperm whale population structure in the eastern and central North Pacific inferred by the use of single-nucleotide polymorphisms, microsatellites and mitochondrial DNA. Mol. Ecol. Resour. 11, 278–298. https://doi.org/10.1111/j.1755-0998.2010.02973.x.

Mirimin, L., Banguera-Hinestroza, E., Dillane, E., Hoelzel, A.R., Cross, T.F., Rogan, E., 2011. Insights into Genetic Diversity, Parentage, and Group Composition of Atlantic White-Sided Dolphins (*Lagenorhynchus acutus*) off the West of Ireland Based on Nuclear and Mitochondrial Genetic Markers. J. Hered. 102, 79–87. https://doi.org/10.1093/jhered/esq106.

Morin, P.A., Luikart, G., Wayne, R.K., 2004. SNPs in ecology, evolution and conservation. Trends Ecol. Evol. 19, 208–216. https://doi.org/10.1016/j.tree.2004.01.009.

Nielsen, R., Mattila, D.K., Clapham, P.J., Palsbøll, P.J., 2001. Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. Genetics 157, 1673–1682.

Oremus, M., Gales, R., Kettles, H., Baker, C.S., 2013. Genetic Evidence of Multiple Matrilines and Spatial Disruption of Kinship Bonds in Mass Strandings of Long-finned Pilot Whales, *Globicephala melas*. J. Hered. 104, 301–311. https://doi.org/10.1093/jhered/est007.

Paetkau, D., Strobeck, C., 1994. Microsatellite analysis of genetic variation in black bear populations. Mol. Ecol. 3, 489–495. https://doi.org/10.1111/j.1365-294X.1994.tb00127.x.

Paetkau, D., Strobeck, C., 1995. The molecular basis and evolutionary history of a microsatellite null allele in bears. Mol. Ecol. 4, 519–520. https://doi.org/10.1111/j.1365-294X.1995.tb00248.x.

Palsbøll, P., Larsen, F., Sigurd-Hansen, E., 1991. Sampling of skin biopsies from free-ranging large cetaceans at West Greenland: Development of biopsy tips and new designs of bolts. Rep. Int. Whal. Comm. Spec. Issue 13, 71–79.

Palsbøll, P., Allen, J., Berube, M., Clapham, P., Feddersen, T., Hammond, P., Hudson, R., Jørgensen, H., Katona, S., Larsen, A., Larsen, F., Lien, J., Mattila, D., Sigurjónsson, J., Sears, R., Smith, T., Sponer, R., Stevick, P., Øien, N., 1997. Genetic tagging of humpback whales. Nature 388, 767–769. https://doi.org/10.1038/42005.

Palsbøll, P.J., 1999. Genetic tagging: contemporary molecular ecology. Biol. J. Linn. Soc. 68, 3–22. https://doi.org/10.1111/j.1095-8312.1999.tb01155.x.

Peakall, R., Smouse, P.E., 2006. genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes 6, 288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x.

Peakall, R., Smouse, P.E., 2012. GenAlEx 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research - an update. Bioinforma. Adv. Access.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. Plos One 7, e37135. https://doi.org/10.1371/journal.pone.0037135.

Primmer, C.R., Ellegren, H., Saino, N., Møller, A.P., 1996. Directional evolution in germline microsatellite mutations. Nat. Genet. 13, 391–393. https://doi.org/10.1038/ng0896-391.

Rew, M.B., Robbins, J., Mattila, D., Palsbøll, P.J., Bérubé, M., 2011. How many genetic markers to tag an individual? An empirical assessment of false matching rates among close relatives. Ecol. Appl. 21, 877–887. https://doi.org/10.1890/10-0348.1.

Rivera-León, V.E., Urbán, J., Mizroch, S., Brownell, R.L., Oosting, T., Hao, W., Palsbøll, P.J., Bérubé, M., 2019. Long-term isolation at a low effective population size greatly reduced genetic diversity in Gulf of California fin whales. Sci. Rep. 9, 12391 https://doi.org/10.1038/s41598-019-48700-5.

Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B., Erlich, H.A., 1986. Analysis of enzymatically amplified ß-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. Nature 324, 163–166.

Sambrook, J., Russell, D.W., 2001. Molecular Cloning. A laboratory manual, Third. ed. Cold Spring Harbor Laboratory Press, New York.

Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B., 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. 2011 31st International Conference on Distributed Computing Systems Workshops. Presented at the 2011 31st International Conference on Distributed Computing Systems Workshops (ICDCS Workshops). IEEE, Minneapolis, MN, USA, pp. 166–171. https://doi.org/10.1109/ICDCSW.2011.20.

Savitzky, Abraham, Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.

Schuelke, M., 2000. An economic method for the fluorescent labeling of PCR fragments. Nat. Biotechnol. 18, 1–2 https://doi.org/DOI:10.1038/71855.

Selvin, S., 1980. Probability of nonpaternity determined by multiple allele codominant systems. Am. J. Hum. Genet. 32, 276–278.

Shimozuru, M., Jimbo, M., Adachi, K., Kawamura, K., Shirane, Y., Umemura, Y., Ishinazaka, T., Nakanishi, M., Kiyonari, M., Yamanaka, M., Amagai, Y., Ijuin, A., Sakiyama, T., Kasai, S., Nose, T., Shirayanagi, M., Tsuruga, H., Mano, T., Tsubota, T., Fukasawa, K., Uno, H., 2022. Estimation of breeding population size using DNA-based pedigree reconstruction in brown bears. Ecol. Evol. 12, e9246 https://doi.org/10.1002/ece3.9246.

Skaug, H.J., 2001. Allele-sharing methods for estimation of population size. Biometrics 57, 750–756.

Skaug, H.J., Bérubé, M., Palsbøll, P.J., 2010. Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate. Mol. Ecol. Resour. 10, 693–700. https://doi.org/10.1111/j.1755-0998.2010.02833.x.

Smith, T.D., Allen, J., Clapham, P.J., Hammond, P.S., Katona, S., Larsen, F., Lien, J., Mattila, D., Palsbøll, P.J., Sigurjónsson, J., Stevick, P.T., ØIen, N., 1999. An Ocean-Basin-Wide Mark-Recapture Study of the North Atlantic Humpback Whale (*Megaptera novaeangliae*). Mar. Mammal. Sci. 15, 1–32. https://doi.org/10.1111/j.1748-7692.1999.tb00779.x.

Suárez-Menéndez, M., Bérubé, M., Furni, F., Rivera-León, V.E., Heide-Jørgensen, M.-P., Larsen, F., Sears, R., Ramp, C., Eriksson, B.K., Etienne, R.S., Robbins, J., Palsbøll, P.J., 2023. Wild pedigrees inform mutation rates and historic abundance in baleen whales. Science 381, 990–995. https://doi.org/10.1126/science.adf2160.

Tardy, C., Ody, D., Gimenez, O., Planes, S., 2023. Abundance of fin whales (*Balaenoptera physalus*) in the north-western Mediterranean Sea, using photo-identification and microsatellite genotyping. Mar. Ecol. 44, e12737 https://doi.org/10.1111/maec.12737.

Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17, 6463–6471. https://doi.org/10.1093/nar/17.16.6463.

Tollis, M., Robbins, J., Webb, A.E., Kuderna, L.F.K., Caulin, A.F., Garcia, J.D., Berube, M., Pourmand, N., Marques-Bonet, T., O'Connell, M.J., Palsboll, P.J., Maley, C. C., 2019. Return to the sea, get huge, beat cancer: an analysis of cetacean genomes including an assembly for the humpback whale (*Megaptera novaeangliae*). Mol. Biol. Evol. 36, 1746–1763. https://doi.org/10.1093/molbev/msz099.

Torres-Florez, J.P., Hucke-Gaete, R., LeDuc, R., Lang, A., Taylor, B., Pimper, L.E., Bedriñana-Romano, L., Rosenbaum, H.C., Figueroa, C.C., 2014. Blue whale population structure along the eastern South Pacific Ocean: evidence of more than one population. Mol. Ecol. 23, 5998–6010. https://doi.org/10.1111/mec.12990.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3–new capabilities and interfaces. Nucleic Acids Res 40.

Valière, N., 2002. gimlet: a computer program for analysing genetic individual identification data. Mol. Ecol. Notes 2, 377–379. https://doi.org/10.1046/j.1471-8286.2002.00228.x-i2.

Van Rossum, G., Drake Jr, F.L., 1995. Python reference manual. Centrum voor Wiskunde en Informatica, Amsterdam.

Vandeputte, M., 2012. An accurate formula to calculate exclusion power of marker sets in parentage assignment. Genet. Sel. Evol. 44, 1–4. https://doi.org/10.1186/1297-9686-44-36.

Vartia, S., Villanueva-Cañas, J.L., Finarelli, J., Farrell, E.D., Collins, P.C., Hughes, G.M., Carlsson, J.E.L., Gauthier, D.T., McGinnity, P., Cross, T.F., FitzGerald, R.D., Mirimin, L., Crispie, F., Cotter, P.D., Carlsson, J., 2016. A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. R. Soc. Open Sci. 3, 150565 https://doi.org/10.1098/rsos.150565.

Wang, J., 2006. Informativeness of genetic markers for pairwise relationship and relatedness inference. Theor. Popul. Biol. 70, 300–321. https://doi.org/10.1016/j.tpb.2005.11.003.

Wang, J., 2010. Effects of genotyping errors on parentage exclusion analysis. Mol. Ecol. 19, 5061–5078. https://doi.org/10.1111/j.1365-294X.2010.04865.x.

Weinberg, W., 1908. Uber den Nachweis der Vererbung beim Menschen. Jh Ver. Vater Nat. Wurttemb 64, 369–382.

Weng, Z., Yang, Y., Wang, X., Wu, L., Hua, S., Zhang, H., Meng, Z., 2021. Parentage Analysis in Giant Grouper (*Epinephelus lanceolatus*) Using Microsatellite and SNP Markers from Genotyping-by-Sequencing Data. Genes 12, 1042. https://doi.org/10.3390/genes12071042.

Zane, Nelson, Jones, Avise, 1999. Microsatellite assessment of multiple paternity in natural populations of a live-bearing fish, *Gambusia holbrooki*. J. Evol. Biol. 12, 61–69. https://doi.org/10.1046/j.1420-9101.1999.00006.x.