Allelic diversity of the PRDM9 coding minisatellite in minke whales

Elena Damm[*§], Kristian K Ullrich[*§], William B Amos[†], Linda Odenthal-Hesse[*#]

Affiliations:

[*]Department Evolutionary Genetics, Research Group Meiotic Recombination and Genome Instability, Max Planck Institute for Evolutionary Biology, Plön D-24306, Germany

[†]Department of Zoology, University of Cambridge, Cambridge, United Kingdom=

[#]corresponding author

[§]These authors contributed equally

ORCID IDs: 0000-0002-6261-6076 (E.D.), 0000-0003-4308-9626 (K.K.U.), 0000-0002-5519-2375 (L.O.-H.)

Running title

PRDM9 diversity in minke whales

Keywords

PRDM9, minke whales, *Balaenoptera acutorostrata*, *Balaenoptera bonarensis,* Microsatellite loci, mtDNA, postzygotic reproductive isolation, meiotic recombination regulation

Corresponding Author:

Dr. Linda Odenthal-Hesse,

Research Group Leader, Group Meiotic Recombination and Genome Instability, Department of Genetics, Max Planck Institute for Evolutionary Biology
August-Thienemann Str. 2,
24306 Plön, Germany

phone number: +49 (0)4522 763 309
email address: odenthalhesse@evolbio.mpg.de (LOH)

Abstract

The PRDM9 protein controls the reshuffling of parental genomes in most metazoans. During meiosis the PRDM9 protein recognizes and binds specific target motifs via its zinc-finger array, which is encoded by a hypervariable minisatellite. To date, PRDM9 diversity has been little studied outside humans, wild mice and some domesticated species where evolutionary constraints may have been relaxed. Here we explore the structure and variability of PRDM9 in samples of the minke whales from the Atlantic, Pacific and Southern Oceans. We show that minke whales possess all the features characteristic of organisms with PRDM9-directed recombination initiation, including complete KRAB, SSXRD and SET domains and a rapidly evolving array of C_2H_2-type-Zincfingers (ZnF). We uncovered eighteen novel PRDM9 variants and evidence of rapid evolution, particularly of DNA-recognizing positions that evolve under positive selection. At different geographical scales, we observed extensive PRDM9 diversity in Antarctic minke whales (*Balaenoptera bonarensis)*, that conversely lack observable population differentiation in mitochondrial DNA and microsatellites. In contrast, a single PRDM9 variant is shared between all Common Minke whales and even across subspecies boundaries of North Atlantic *(B. a. acutorostrata)* and North Pacific *(B. a. scammoni)* minke whale, which do show clear population differentiation. PRDM9 variation of whales predicts distinct recombination initiation landscapes genome wide, which has possible consequences for speciation.

Introduction

The gene *Prdm9* encodes "PR-domain-containing 9" (PRDM9), a meiosis-specific four-domain protein that regulates meiotic recombination in mammalian genomes. The four functional domains of the PRDM9 protein are essential for the placement of double-stranded DNA breaks (DSBs) at sequence-specific target sites. Three of the domains are highly conserved: *i)* the N-terminal Kruppel-associated box- domain (KRAB) that promotes protein-protein binding, for example with EWSR, CXXC1, CDYL and EHMT2 (IMAI *et al.* 2017; PARVANOV *et al.* 2017); *ii)* the SSX-repression-domain (SSXRD) of yet unknown function; *iii)* the central PR/SET domain, a subclass of the SET domain, with methyltransferase activity at H3K4me3 and H3K36e3. The fourth, C-terminal, domain comprises an array of type Cystin$_2$Histidin$_2$ (C_2H_2-ZnF), encoded by a minisatellite-like sequence of 84 bp tandem repeats. This coding minisatellite reveals evidence of positive selection and concerted evolution, with many functional variants having been found in humans (BERG *et al.* 2010; BERG *et al.* 2011), mice (BUARD *et al.* 2014; KONO *et al.* 2014), non-human primates (SCHWARTZ *et al.* 2014) and other mammals (STEINER AND RYDER 2013; AHLAWAT *et al.* 2016b). Even highly domesticated species like equids (STEINER AND RYDER 2013), bovids (AHLAWAT *et al.* 2017) and ruminants (AHLAWAT *et al.* 2016a; AHLAWAT *et al.* 2016c) show high diversity and rapid evolution. In light of this extreme variability between minisatellite-like repeat units in most other mammals, the allele found in the genome of North Pacific minke whale (*Balaenoptera acutorostrata scammoni)* is puzzling, as it comprises of an array of only a single type of C_2H_2-ZnF type repeat unit.

Minke whales are marine mammals of the genus *Balaenoptera*, in the order of baleen whales (*Cetaceans*), that are of particular interest not only because little is known about their population biology, seasonal migration routes and breeding behavior, but also to support future conservation efforts. Minke whales were long considered to be a single species, but have been classified as two distinct species, the Common minke whale (*Balaenoptera acutorostrata)* and the Antarctic minke whale (*Balaenoptera bonaerensis* Burmeister, 1867). The Common minke whale (*B. acutorostrata)* is cosmopolitan in the waters of the Northern Hemisphere. And can be separated into two subspecies, the Atlantic minke whale *(B. acutorostrata acutorostrata)* and the North Pacific minke whale (*B. acutorostrata scammoni*), separated from each other by landmasses and the polar ice cap. Antarctic minke whale*s* inhabit the waters of the Antarctic ocean in the Southern Hemisphere during feeding season but seasonally migrate to the temperate waters near the Equator during the breeding season (BAKKE *et al.* 1996).

Although overlapping habitats exist near the Equator, seasonal differences in migration and breeding behavior largely prevent inter-breeding between Common and Antarctic minke whales (PERRIN AND JR. 2009). Despite this, occasional migration across the Equator has been observed (GLOVER *et al.* 2010) and recent studies have uncovered two instances of hybrid individuals, both females and one with a calf most likely sired by an Antarctic minke whale (GLOVER *et al.* 2013; MALDE *et al.* 2017). Thus, hybrids may be both viable and fertile, allowing backcrossing (MALDE *et al.* 2017). However, it is unclear whether occasional hybridization events have always occurred or whether they are a recent phenomenon driven by anthropogenic changes, including climate change (MALDE *et al.* 2017). More importantly, since both hybrids were female, current

120 data do not exclude postzygotic reproductive isolation through males, as expected
121 under Haldane's rule. Hybrid sterility is a universal phenomenon observed in many
122 eukaryotic inter-species hybrids, including yeast, plants, insects, birds, and mammals
123 (COYNE AND ORR 2004; MAHESHWARI AND BARBASH 2011). Minke whales offer an
124 unusual opportunity to study *Prdm9*, the only hybrid sterility gene that is known in
125 vertebrates thus far, sometimes called a speciation gene (MIHOLA *et al.* 2009).
126

127 **Material and Methods**

128 **PRDM9 occurrence and protein domain prediction in diverse taxa**

129 To infer PRDM9 occurrence and to subsequent investigate protein domain architecture
130 in diverse taxa, the genome sequences were downloaded as given in **Supplementary**
131 **Table 2**. To infer PRDM9 occurrence in the given species, the annotated protein
132 XP_028019884.1 from *Balaenoptera acutorostrata scammoni* was used as the query
133 protein with exonerate (SLATER AND BIRNEY 2005) (v2.2.0) and the --protein2genome
134 model to extract the best hit. Subsequently, for each species, the extracted coding
135 sequences (CDS) were translated and investigated with InterProScan (QUEVILLON *et*
136 *al.* 2005) (v5.46-81.0) and HMMER3 (MISTRY *et al.* 2013) (v3.3) using KRAB, SSXRD,
137 SET and C_2H_2-ZnF as bait to obtain the protein domain architecture.
138

139 **Phylogenetic Analyses**

140 For the phylogenetic reconstruction across *Cetartiodactyla,* we used only KRAB,
141 SSXRD and SET domains (**Supplementary Figure 1**). These protein domains were
142 concatenated and used as input for the software BAli-phy (SUCHARD AND REDELINGS
143 2006) (v3.5.0). BAli-phy was run twice with 10,000 iterations each and the default
144 settings for amino acid input. Subsequently, the majority consensus tree was obtained

145 by skipping the first 10% of trees as burn-in, rooted on the branch outside the *Cetacea*
146 and visualized in Figtree (http://tree.bio.ed.ac.uk/software/figtree/; v1.4.4).
147
148 The phylogenetic tree from the mitochondrial D-loop region was reconstructed,
149 including several whale species as given in **Supplementary Table 3**. The
150 corresponding D-loop was aligned with MAFFT version 5 (KATOH *et al.* 2005)(v7.471)
151 with the L-INS-i algorithm and manually curated. The maximum-likelihood tree was
152 calculated under the TN+F+I+G4 model using IQ-TREE (NGUYEN *et al.* 2015) (v1.6.12)
153 mid-point rooted and visualised in Figtree (http://tree.bio.ed.ac.uk/software/figtree/;
154 v1.4.4).
155
156 **Long-range sequencing of the *Prdm9* gene**
157 We first extracted the full-length DNA sequence of PRDM9 Protein from the minke
158 whale genome from UCSC genome browser and designed primers to flank the entire
159 protein-coding portion, encompassing the KRAB, SSXDR, PR/SET and C_2H_2-ZnF
160 binding domains using Primer-BLAST (YE *et al.* 2012). We amplified the *Balaenoptera*
161 *acutorostrata* Prdm9 gene, including all introns and exons across a ~11kb interval by
162 long-range PCR. We then sequenced the entire interval using phased long-range
163 Nanopore Sequencing with MinION (Oxford Nanopore). Whole-length consensus
164 sequences generated from 639 sequencing reads that cover the entire 10582 bp, with
165 an average per base pair coverage of 269x. We used this consensus sequence as
166 reference for *in-silico* predictions of functional domains and mapped human PRDM9
167 Exons 3 – 11 from ENSEMBL (ENST00000296682.3). By manually splicing all
168 unaligned sequence fragments, we generated an *in-silico* predicted mRNA of
169 *Balaenoptera acutorostrata* PRDM9. This sequence was then submitted as '.fasta' to

170 the Entrez Conserved Domains Database (CDD) home page,
171 (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) with the following parameters;
172 searching against Database CDD v3.18 – 52910 PSSMs Expect Value threshold: 0,01,
173 without low-complexity filter, composition-based statistics adjustment, rescuing
174 borderline hits (ON) a maximum number of 500 hits and concise result mode).
175
176 **Wild minke whale samples**
177 Performed investigations are not considered to be animal trials in accordance with the
178 German animal welfare act, since samples were obtained from commercial whaling
179 between 1980 and 1984. A total of n=143 DNA samples of minke whale with vague
180 information about the subspecies, from four different defined commercial whaling
181 areas. These include North Atlantic (NA; n=17) and North Pacific (NP; n=4) individuals
182 captured during migration season, as well as Antarctic Ocean Areas IV (AN IV; n=65)
183 and V (AN V; n=57) were obtained during feeding season. A subset of these samples
184 had been used in van Pijlen et al., 1995 (VAN PIJLEN *et al.* 1995). Samples from the
185 Antarctic areas included numerous duplicate samples that were used as internal
186 controls, analyzed for all parameters and excluded from the sample set once identical
187 results had been confirmed (318, 325, 327, 1661, 1663).
188
189 For the detailed DNA-extraction protocol, precise catch data, as well as providers of
190 the samples of most individuals used in this project, see van Pijlen, 1994. Briefly,
191 genomic DNA was extracted from the skin (NA) and muscle biopsies (NP, AN) with
192 Phenol-Chlorophorm and stored in TE at - 80°C in the 1980s. If the DNA had dried out
193 due to the long storage time, it was eluted with TE. All DNA concentrations were

194 determined with fluorescent Nanodrop-1000 (Thermo Fisher Scientific). For all
195 performed analysis, 20 ng/µl DNA working stocks were prepared and stored at -20°C.
196

197 **STR genotyping**

198 Nine autosomal, as well as X/Y microsatellite loci with di- or tetramer repeat motifs,
199 were analysed for all samples: EV001, EV037 (VALSECCHI AND AMOS 1996), GATA028,
200 GATA098, GATA417 (PALSBOLL *et al.* 1997), GT023, GT310, GT509, GT575 (BERUBE
201 *et al.* 2000) and sex loci X and Y (BERUBE AND PALSBOLL 1996). Four separate
202 multiplexing reactions were performed for each individual, and each contained 40 ng
203 of DNA, 0.2 µM of each primer, 5 mM Multiplex-Kit (Qiagen) and HPLC water to a total
204 volume of 10 µL per sample. Primers were purchased from Sigma Aldrich; the reverse
205 Primers were tagged at their 5' end with fluorescent tags (HEX, FAM or JOE). The first
206 multiplexing reaction contained primers GT023 (HEX), EV037 (HEX) (FAM) in forward
207 and reverse direction. The multiplexing conditions were denaturation at 95°C for 15:30
208 minutes, annealing 1:30 minutes at 59°C, elongation at 72°C for 11:30 minutes and
209 held at 12°C. The second multiplexing reaction contained primers GT575 (HEX),
210 GATA028 (FAM) in forward and reverse direction, the multiplexing conditions were
211 denaturation at 95°C for 15:30 minutes, annealing 1:30 minutes at 54°C, elongation at
212 72°C for 11:30 minutes and a final temperature of 12°C. The third multiplexing reaction
213 was set up with primers GATA098 (FAM), GT509 (FAM) and GATA417 (JOE) in
214 forward and reverse direction. The multiplexing conditions were denaturation at 95°C
215 for 15:30 minutes, annealing 1:30 minutes at 54°C, elongation at 72°C for 11:30
216 minutes and a final temperature of 12°C. The fourth multiplexing reaction contained
217 primers GT310 (HEX) and EV001 (JOE) in forward and reverse direction. The
218 multiplexing conditions were denaturation at 95°C for 15:30 minutes, annealing 1:30

219 minutes at 60°C, elongation at 72°C for 11:30 minutes and a final temperature of 12°C.
220 After amplification, the multiplexing reactions were diluted with 100 µl water (HPLC
221 grade). Genescan ROX^{500} dye size standard (Thermo Fisher Scientific) and HiDi
222 Formamide (Thermo Fisher Scientific) were mixed 1:100 and used for sequencing in
223 a mix of 1 µl of diluted multiplexing product and 10 µl HiDi-ROX^{500}. STR Multiplex
224 fragment analysis was carried out on the 16-capillary electrophoresis system ABI 3300
225 Genetic Analyser (Applied Biosystems).
226
227 **STR analysis:**
228 We first performed MSA analysis (MUKAJ *et al.* 2020) using standard parameters, which
229 calculated Weinberg expectation (Fis), Shannon Index (Hs), allele numbers (A) and
230 allele sizes. To detect both weak and strong population structure, we ran simulations
231 with LOCPRIOR and USEPOPINFO, respectively. For both simulations, the more
232 conservative "correlated allele frequencies" -model was used, which assumes a level
233 of non-independence. To ensure that a sufficient number of steps and runs have been
234 performed we used a burn-in period of 1.000 and runs of 100.000 MCMC repeats for
235 both types of simulations, each for 50 iterations for successive K values from 1 - 10.
236 (OROZCO-TERWENGEL *et al.* 2011). We used the *a-priori* location to be able to detect
237 even weak population differentiation. In both datasets, we used the web-based
238 STRUCTURE Harvester software (EARL AND VONHOLDT 2011) to determine the rate of
239 change in the log probability between successive K values, via the *ad-hoc* statistic ΔK
240 from (EVANNO *et al.* 2005). Figures were rendered using STRUCTURE PLOT V2.0.
241 (RAMASAMY *et al.* 2014)
242
243 **Sequencing of the Mitochondrial Hypervariable Region on the D-loop**

244 The noncoding mtDNA-D-Loop region of 143 individuals was amplified in two
245 overlapping PCR-reactions. We PCR amplified two different lengths fragments for
246 each individual: 1066 bp and 331 bp and sequenced the longer product in forward and
247 the shorter in the reverse direction. The used primers were MT4 (M13F:
248 GTAAAACGACGGCCAGTCCTCCCTAAGACTCAAGGAAG) and MT3 (M13R:
249 CAGGAAACAGCTATGACCCATCTAGACATTTTCAGTG for the longer product, and
250 BP15851 (M13F: GTAAAACGACGGCCAGTGAAGAAGTATTACACTCCACCAT) and
251 MN312 (M13R: CAGGAAACAGCTATGACCCGTGATCTAATGGAGCGGCCA) for the
252 shorter PCR product from (GLOVER *et al.* 2012). The 10 µL reaction was carried out
253 with 40 ng genomic DNA and 0.2 µM of each primer and 5 mM Multiplex PCR Kit
254 (Qiagen). Cycling conditions were identical for both directions with 95°C 15:30 min,
255 53°C 1:30 min, 72°C 13:30 min and hold at 4°C in a Veriti Thermal Cycler (Applied
256 Biosystems). The PCR products were purified with 3 µl Exo/SAP and then cycle-
257 sequenced with the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher
258 Scientific) according to manufacturer instructions with BP15851 (M13F) for the forward
259 PCR product and MN312 (M13R) for the reverse PCR product, respectively. The mixes
260 were then purified with BigDye X-Terminator™ Purification Kit (Thermo Fisher
261 Scientific) and sequenced by capillary electrophoresis on an ABI 16-capillary 3130xl
262 Analyzer (Applied Biosystems). The sequences were *de-novo* assembled, and
263 consensus sequences were generated with Geneious Software 10.2.3 (KEARSE *et al.*
264 2012).
265

266 **Prdm9 coding minisatellite array PCR and Sequencing**

267 To characterise the minisatellite-coding for the C_2H_2-ZnF array in more detail, we
268 designed primers nested between the first two conserved ZnFs, and the reverse primer

269 identical to the long-range amplification primer distal to the coding sequence for the
270 C_2H_2-ZnF array. The minisatellite coding for the C_2H_2-ZnF array of PRDM9 of 143
271 *Balaenoptera* individuals was amplified from 20 ng genomic DNA, in a 20 µl PCR
272 reaction, optimised for the amplification of minisatellites. With 0.5 mM primers that
273 were designed for this study using the *Balaenoptera. acutorostrata scammoni*
274 *(*XM_007172595) reference and included the PRDM9 minisatellite-like C_2H_2-ZnF array
275 and additional 100bp flanking regions at 5' and 3'. Prdm9ZnFA_Bal_R:
276 GAGCTGAGGTATGTGGTCGC and Prdm9ZnFA_Bal_F:
277 ACTCATGGGGGCAGGAGTAT and 1x AJ-PCR Buffer described in (JEFFREYS *et al.*
278 1990) and 0.025 U/µl Taq-Polymerase and 0.0033 U/µl Pfu-Polymerase. Cycling
279 conditions were: initial denaturation at 95°C for 1:30 min followed by 33 cycles
280 including 96°C 15 s, 61°C 20 s and 70 °C 2:00 min and finally 70°C 5 min, and hold at
281 4°C in a Veriti Thermal Cycler (Applied Biosystems). Agarose gel electrophoresis
282 (1.5% Top Vision Low Melting Point Agarose gel (Thermo Fisher Scientific) with SYBR
283 Safe (Thermo Fisher Scientific) was used to visualise allele sizes as well as zygosity.
284 All bands were excised from the gel (Molecular Imager® Gel Doc™ XR System with
285 Xcita-Blue™ Conversion Screen (Biorad), and PCR products were recovered from the
286 gel fragments with 2 U/100 mg Agarase. If the individuals were homozygous, the
287 extracted DNA was directly Sanger sequenced from in 5' and 3' directions for each
288 sample with the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher
289 Scientific) according to the manufacturer's protocol with the same primers as for the
290 amplification (Prdm9ZnFA_Bal_R/Prdm9ZnFA_Bal_F). The sequencing-reaction was
291 carried out in the ABI 16-capillary 3130xl Analyzer (Applied Biosystems).
292 Heterozygous samples were subcloned before sequencing.
293

294 **Subcloning of heterozygous alleles of identical lengths**

295 Subcloning was performed for a subset of samples when Sanger sequencing revealed
296 heterozygous alleles of the same length, that could not be distinguished by
297 electrophoresis, but revealed heterozygous nucleotides in the chromatogram. Instead
298 of inferring haplotypes, we wanted to experimentally ascertain both alleles. Thus, we
299 cloned the remaining PCR product into the TOPO TA (Invitrogen) vector and
300 transferred the vectors into OneShotTop10 chemically competent cells (Invitrogen). All
301 steps were carried out according to the manufacturer's manuals. Eight positive clones
302 were picked for each sample, and the DNA was extracted by boiling in HPLC-grade
303 water at 96°C for 10 Minutes. A centrifugation step then removed the cell debris, and
304 the supernatant was directly used for PCR, gel-purification and Sanger-sequencing as
305 described in the section above.

306

307 **Sequence Analysis of the PRDM9 C_2H_2-ZnF array**

308 Sequences determined different alleles and numbers of repeat units per array, which
309 were *de novo* annotated and the resulting consensus sequences were translated into
310 the corresponding protein variants of each C_2H_2-ZnF by Geneious Software 10.2.3
311 (KEARSE *et al.* 2012). All PCR products contained the complete minisatellite-like coding
312 sequence for the C_2H_2 domain and additional 100 base pairs flanking regions at the 5'
313 and 3' ends.

314

315 **PRDM9 C_2H_2-ZnF array coding sequence dN/dS analysis**

316 The internal C_2H_2-ZnFs were obtained as described for the phylogenetic
317 reconstruction. Subsequently, all repeat units were stacked using only non-unique
318 alleles. The resulting codon alignment was used to determine basic population

319 parameters for each population (population theta using segregating sites and
320 nucleotide diversity) by either including or excluding hypervariable amino acid positions
321 (-1, +3, +6) with the R package "pegas" (PARADIS 2010). Episodic diversifying selection
322 among Zinc-finger sites using was determined by a mixed-effects model of evolution
323 (MEME) at https://www.datamonkey.org as described earlier (MURRELL *et al.* 2012).
324

325 **Phylogenetic analyses of the PRDM9 hypervariable domain**

326 The gene tree of PRDM9 was inferred either using only the repeats coding for the
327 internal C_2H_2-ZnFs or the concatenated KRAB, SSXRD and SET domain. For the
328 internal C_2H_2-ZnFs based phylogenetic reconstruction (**Figure 1**), internal C_2H_2-ZnFs
329 were extracted, and pairwise distances were calculated as in Vara et al. 2019 (VARA
330 *et al.* 2019) with the developmental R package repeatR
331 (https://gitlab.gwdg.de/mpievolbio-it/repeatr). In brief, only species were used with one
332 leading C_2H_2-ZnF and at least eight internal C_2H_2-ZnFs (see supplementary Figure 1).
333 For each possible repeat combination (r, r') the hamming distances of the
334 corresponding repeat units $r = (r_1; r_2; r_3; \ldots)$ and $r' = (r'_1; r'_2; r'_3; \ldots)$ were used to derive
335 the edit distance between r and r'. Before calculating the edit distance, the codons
336 coding for the hypervariable amino acid positions (-1, +3, +6) were removed for each
337 repeat unit and weighting cost of $w_{mut} = 1$, $w_{indel} = 3.5$ and $w_{slippage} = 1.75$ as given in
338 (VARA *et al.* 2019). The resulting pairwise edit distance matrix was used to calculate a
339 neighbour-joining tree with the *bionj* function of the R package ape (PARADIS *et al.*
340 2004), rooted on the branch leading to the bottlenose dolphin (*Tursiops truncatus)* and
341 visualized in Figtree (http://tree.bio.ed.ac.uk/software/figtree/ v1.4.4).
342

343 **Prediction of DNA-binding motifs of different PRDM9 variants**

344 DNA-binding Specificities of the different Cys_2His_2 Zinc Finger Protein variants were
345 predict *in-silico* using the SVM polynomial kernel method within "Princeton ZnF"
346 (http://zf.princeton.edu/) (PERSIKOV AND SINGH 2014).
347

348 **Results**

349 **The evolutionary context of PRDM9 in *Cetartiodactyla***

350 For a broad view on the evolution of PRDM9 in *Cetartiodactyla*, we identified PRDM9
351 orthologs from all publicly available genomic resources (**Supplementary Table 2**). To
352 first establish an evolutionary context based on the protein-coding regions for the
353 proximal conserved domains, we used the *Balaenoptera acutorostrata scammoni*
354 protein as a query and extracted coding sequences of the N-terminal (conserved)
355 domains. Previous studies of PRDM9 in vertebrates found rapid evolution of DNA
356 binding sites only when the KRAB and SSXRD domains were present (BAKER *et al.*
357 2017). Consequently, we used InterProScan to create a curated dataset of
358 *Cetartiodactyla* PRDM9 orthologues which contained the KRAB, SSXRD, and SET
359 domains, including a PRDM9 orthologue in Antarctic minke whale. Our initial analyses
360 revealed that minke whales, and *Cetartiodactyla* in general, possessed all the features
361 characteristic of organisms with PRDM9-directed recombination initiation, including
362 complete KRAB, SSXRD and SET domains. However, organisms with PRDM9
363 directed recombination, also display a rapidly-evolving PRDM9 C_2H_2-ZnF array (BAKER
364 *et al.* 2017) With the sole exception of Hippopotamus, all *Cetartiodactyla* also
365 contained the C_2H_2-ZnF array, with large variation in the number of C_2H_2-ZnF between
366 species.
367

368 **Characterizing the *Prdm9* gene in minke whales**

369 To characterize the sequence and structure of the *Prdm9* gene, we began by
370 amplifying and sequencing it from a pooled sample, containing DNA from six
371 individuals, five Antarctic minke whales and one common minke whale. The consensus
372 sequence was subjected to *in-silico* prediction that successfully recovered all relevant
373 PRDM9 protein domains with high-confidence: the KRAB domain (E-value: $1.84e^{-11}$);
374 the SSXRD motif (E-value: $4.46e^{-10}$); the PR/SET domain (E-value: $9.69e^{-05}$); and
375 several Zinc-Fingers (**Figure 1**). The complete C_2H_2-domain comprises an array of
376 C_2H_2-ZnFs (the C_2H_2-ZnF-array), as well as a single C_2H_2-ZnF (ZnF#1) that is located
377 proximally (E-value: $6.52e^{-04}$). This PRDM9-ZnF#1 is identical to that of mice (PAIGEN
378 AND PETKOV 2018) as well as rats, elephants, human, chimpanzee, macaque and
379 orangutan (MYERS *et al.* 2010), and thus appears to be conserved across large
380 evolutionary timescales. The second C_2H_2-ZnF (E-value: $1.65e^{-11}$) appears conserved
381 across *Balaenoptera*. The same second C_2H_2-ZnF is seen in our pooled sample, as
382 well as the genome reference of the North Pacific minke whale (*Balaenoptera*
383 *acutorostrata scammoni)* and Antarctic minke whale (*Balaenoptera bonarensis)* as well
384 as the blue whale (*Balaenoptera musculus).* However, from the third ZnF onwards,
385 sequencing across pooled samples exhibited high variability, which we could not
386 resolve.
387

388 **Variation of the PRDM9 coding minisatellite in minke whales**

389 We assessed the variability of the C_2H_2-ZnF-array by analyzing the coding minisatellite
390 in a large number of minke whales including Antarctic minke whale (*Balaenoptera*
391 *bonaerensis* Burmeister, 1867) and two Common minke whale subspecies, Atlantic
392 minke whale and North Pacific minke whale. Amplification and subsequent
393 electrophoresis on agarose gels resolved six different allele sizes of the last exon of

394 the *Prdm9* gene. A single size was observed in all Common minke whales from
395 sampled populations in the North Atlantic as well as North Pacific. In contrast, five
396 different sizes were identified across Antarctic minke whale populations sampled in the
397 Southern Hemisphere. All common minke whales have a size consistent with eleven
398 84 bp repeats while the most prevalent genotype in the Antarctic minke whales
399 corresponds to nine repeats, with variation between six and ten repeats. Therefore,
400 *Prdm9* shows size variation length of the coding minisatellite, that result from variation
401 in the number of repeat units.
402

403 Sequencing of the DNA purified from gel fragments, as well as cloning of apparently
404 homozygous bands, revealed a remarkable diversity of minisatellite alleles. Over 80%
405 of animals possessed repeat-arrays of nine repeats; however, these comprised six
406 different alleles. Because in addition to variation in repeat-number, there is nucleotide
407 variation between 84bp-minisatellite repeats. And finally, not only which repeat-types
408 are present, but also their order varies between individuals.
409

410 **Diversity and diversifying selection on C_2H_2-zinc fingers of PRDM9**

411 Each 84bp repeat unit translates into 28 amino-acids coding for a single C_2H_2-ZnF. To
412 explore diversity between the ZnF, we extracted all 84bp repeat units, which we used
413 to predict C_2H_2-ZnFs through an HMMER algorithm (PERSIKOV AND SINGH 2014). This
414 yielded 26 different versions with HMMER bit scores >17 (**Figure 1** and
415 **Supplementary Figure 4**). Based on amino acid variation within each predicted C_2H_2-
416 ZnFs, we identified 14 "common" variants found in multiple individuals as well as twelve
417 "rare" variants found in only a single individual (**Supplementary Figure 4**). The most
418 variable amino acids are 13, 16 and 19, which correspond to positions -1, +3 and +6

419 of the alpha-helices that are responsible for DNA binding specificity. However, five of
420 the fourteen "common" variants have the same amino sequence LNG at these
421 positions, differing instead at three positions in beta-helices flanking the cysteine
422 residues that bind the zinc-ion (positions 3, 10 and 12). Three of the five LNG variants
423 are also found in the reference sequences of both the Antarctic minke whale and the
424 North Atlantic minke whale. We next used the fourteen common C_2H_2-ZnFs to test for
425 signals of selection, finding Evidence that amino acid 16, is under episodic diversifying
426 selection (**Supplementary Figure 5**). This amino acid is located at position +3 of the
427 alpha helix, one of the three positions responsible for DNA-binding specificity.
428

429 **Evolutionary turnover of PRDM9 in minke whales**

430 Using all minisatellite-coding sequences for the array, we determined C_2H_2-ZNF
431 domain variants that each individual possessed. Minisatellite size homoplasy equates
432 to an identical PRDM9 C_2H_2-ZnF array in Common minke whales, but not Antarctic
433 minke whales. The only C_2H_2 domain variant found in all samples from the Northern
434 Hemisphere NH12_A consists of eleven C_2H_2-ZnF in the array and the C_2H_2-ZnF, that
435 is located proximally and appears conserved across *Balaenoptera*. In contrast,
436 samples from the Southern Hemisphere displayed seventeen different C_2H_2-Domains
437 of PRDM9 (**Figure 1**).
438

439 We want to understand the evolutionary relationships of the different alleles that we
440 observed in minke whale species. However, phylogenetic analysis of minisatellite
441 repeat-structures is challenging due to their rapid rate of evolution. Minisatellites,
442 including that of *Prdm9*, evolve mainly by unequal crossing-over and gene conversion
443 in meiosis (JEFFREYS *et al.* 2013). Stepwise mutation models that are commonly

444 assumed are based on microsatellites, and thus give only a poor fit to minisatellite
445 evolution (JIN AND CHAKRABORTY 1995). So, instead of comparing full-length-*Prdm9*
446 allelic variants, we separated each array into 84bp repeat types. We then calculated
447 pairwise Hamming (i.e. minimum edit) distances between all repeats, after removing
448 the nucleotides coding for the hypervariable amino acid positions (-1, +3, +6), and
449 applying specific weighting costs as given in (VARA *et al.* 2019). In addition to the
450 repeats present in our minke whale samples, we also included repeats coding for the
451 internal C_2H_2-ZnFs of all *Cetartiodactyla* species into our phylogenetic analyses (see
452 **Supplementary Figure 1**) Here, we included only those species that possessed the
453 complete N-terminal domain architecture, the proximal C_2H_2-ZnF and at least eight
454 successive C_2H_2-ZnF in the C-terminal DNA-binding domain. We restricted our
455 phylogenetic analyses on the internal minisatellite repeats since our analyses revealed
456 that the DNA-binding positions of the most proximal ZnF in the array (ZnF#2) are not
457 only conserved in minke whales but also generally within *Cetacea* (**Supplementary**
458 **Figure 2**). In fact, we observed only a single DNA-binding amino-acid change between
459 C_2H_2-ZnF #2 of all *Cetacea* compared to all *Ruminantia* (**Supplementary Figure 3**).
460
461 The PRDM9 phylogenetic analyses in **Figure 2** show that all species cluster into
462 distinct phylogenetic groups. Reference alleles cluster with their own subspecies, even
463 including the unusual reference allele of *Balaenoptera acutorostrata scammoni*, which
464 nevertheless clusters with our common minke whale samples. Similarly, the genome
465 reference allele of *Balaenoptera bonarensis* (Antarctic minke whale) clusters within the
466 group of SH_8A and SH-8B alleles found in our Antarctic minke whale samples, which
467 appear to be more ancestral than other minke whale alleles. The phylogenetic
468 reconstruction also suggests that common minke whale alleles have appeared more

469 recently and evolved mainly by an increase in repeat-copy after splitting from the
470 Antarctic minke whale. Our phylogenetic analysis thus fits well with an evolutionary
471 history where common minke whales split from Antarctic minke whales approximately
472 5 million years ago.
473

474 **Diversity of PRDM9 at different geographical scales**

475 To explore the allelic diversity of *Prdm9* at different geographical scales, we partitioned
476 the data by sampling location wherever accurate catch-locations were available
477 (**Figure 2**). Allelic diversity varied between sampling locations, even within the
478 Antarctic sampling sites. The most common alleles are SH_10A, with SH_10B, and
479 these occur at frequencies of 10-20% in all sampling locations. However, all other
480 alleles differ between different sampling populations, each of which carries at least one
481 allele found exclusively in a single population. Sample size does not predict allele
482 number, with the highest allele number (ten) being associated with an average-sized
483 sample (n=34).
484

485 **Population structure of minke whales**

486 We are interested in the possible evolutionary consequences of PRDM9 for minke
487 whale speciation in the face of what may be recent secondary mixing, and this requires
488 a context of current levels of population isolation. We thus explored variability at the
489 84bp *Prdm9* minisatellite repeat across the four sample regions for population genetic
490 analyses (**Table 1**). Taking each repeat type as an individual allele, we computed the
491 mutation rate per base pair per generation (population θ), as well as the average
492 pairwise nucleotide diversity. The latter was calculated in two ways: (i) using the entire
493 minisatellite-like repeat sequence; (ii) after removing nucleotides coding for the

494 hypervariable sites that translate into the DNA binding positions of individual C_2H_2-
495 ZnF, which are known to be under positive selection. We observed the highest
496 population θ in Antarctic Area IV, compared to all other sampling locations, as seen in
497 ***Table 1***. Across the microsatellite-repeat types, the Average Pairwise Nucleotide
498 Diversity (APND, including nucleotides coding for the hypervariable sites) is similar
499 across all sampled populations and between common minke whale and Antarctic
500 minke whales. After excluding the nucleotides coding for hypervariable sites (-1, +3,
501 +6), the APND is decreased roughly 2.5-fold, across minke whales.
502
503 We used two approaches to quantify genetic differentiation of PRDM9 between the
504 different minke whale populations: the per-site distance for multiple alleles (NEI 1973)
505 and Jost's D (D_J), the fraction of allelic variation among populations (JOST 2008). Both
506 measures indicate weak population differentiation, the highest differentiation being
507 between the North Atlantic and North Pacific ($G_{ST} = -0.0039$, $D_J = 0.0077$). The
508 second highest population differentiation was seen between the North Pacific and both
509 Antarctic regions, $G_{ST} = -0.0035$, $D_J = 0.0070$, with slightly lower differentiation
510 between Atlantic and Antarctic, $G_{ST} = -0.0011$, $D_J = 0.0022$. The lowest population
511 differentiation was between Antarctic Areas IV and V (ANV vs ANIV, $G_{ST} = -0.0004$,
512 $D_J = 0.0009$).
513
514 To explore population structure differentiation further, we also analysed polymorphic
515 microsatellites and the hypervariable region of the mitochondrial D-Loop (mtDNA
516 HVR). Despite the low *Prdm9* diversity in the Northern Hemisphere, genetic structuring
517 was revealed by the mtDNA HVR, with a clear distinction between the North Atlantic
518 and North Pacific whales. We also examined nine unlinked autosomal microsatellite

519 loci, chosen for their high information content in minke whales (VALSECCHI AND AMOS
520 1996; MALDE *et al.* 2017) and sex-chromosome markers ZFX/ZFY, from (GLOVER *et al.*
521 2012). Microsatellite summary statistics reveal high heterozygosity of effectively zero
522 F_{IS} (Supplementary Table 1), so we continued with a Bayesian analysis of population
523 structure. We used *STRUCTURE* with a recessive allele model because any null
524 alleles were due to polymorphisms, rather than missing data (PORRAS-HURTADO *et al.*
525 2013) and the "admixture" model. However, given that minke whales occur in
526 geographically isolated groups, are separated by hemispheres, and have
527 asynchronous breeding seasons, we nevertheless assumed a low probability for
528 incomplete lineage sorting, which could otherwise confound structure inference,
529 particularly for weak population differentiation. The most likely number of clusters
530 returned by ΔK is $k = 2$, which distinguishes the two hemispheres. Increasing to $k = 3$
531 and $k = 4$, further separates the common minke whale subspecies, a result that holds
532 even without *a priori* location data (NOLOCS) where only strong population structure
533 should be detected (**Supplementary Figure 6**).
534

535 **Putative minke whale Recombination initiation motifs**

536 Different C_2H_2-ZnF domains target different recombination hotspots (BERG *et al.* 2010;
537 PARVANOV *et al.* 2010; BERG *et al.* 2011; COLE *et al.* 2014). Target hotspots in humans
538 (MYERS *et al.* 2010; BERG *et al.* 2011) and mice (COLE *et al.* 2014) can be predicted
539 from C_2H_2 ZnF sequences using the SVM polynomial kernel model for *de-novo* binding
540 prediction (PERSIKOV AND SINGH 2014). To understand the situation in minke whales,
541 we identified a number of "core motifs", based on ZnFs #3-#6, that appear particularly
542 important for DNA binding (BILLINGS *et al.* 2013; STRIEDNER *et al.* 2017) (**Figure 4**). As
543 must be true, the single PRDM9 variant found in common minke whales results in a

544 single "core motif". However, a large majority of Antarctic minke whales also share a
545 single "core motif", despite a much larger diversity of PRDM9 (**Figure 4**). Core motifs
546 show no overlap between common minke whales and Antarctic minke whales, a trend
547 that is particularly evident at C_2H_2-ZnF#3 where it creates distinctive DNA binding
548 motifs. Thus, it appears as if core motifs are shared between the majority of minke
549 whale PRDM9 variants within a given species, but not between species inhabiting
550 different Hemispheres.
551

552 **Discussion:**

553 **PRDM9 shows high diversity, especially in Antarctic minke whales**

554 We characterized the PRDM9 C_2H_2-ZnF domain diversity of 134 individuals from four
555 natural populations of minke whales and discovered a total of eighteen variants. The
556 minisatellite encoding the DNA-binding C_2H_2-ZnF array apparently evolves rapidly in
557 minke whales, with at least one of the DNA-recognizing position likely under positive
558 selection. This high diversity is similar to that observed at the *Prdm9* gene in other
559 vertebrate species, such as humans, mice, bovines and primates (BERG *et al.* 2011;
560 GROENEVELD *et al.* 2012; BUARD *et al.* 2014; KONO *et al.* 2014; AHLAWAT *et al.* 2016c).

561 However, PRDM9 diversity is not equally abundant in both hemispheres. In contrast to
562 the extensive PRDM9 diversity we observe in Antarctic minke whales (*Balaenoptera*
563 *bonarensis)*, only a single PRDM9 variant is found in Common minke whales. This
564 variant is shared between two subspecies, the North Atlantic minke whale *(B. a.*
565 *acutorostrata)* and North Pacific minke whales *(B. a. scammoni).* In most mammals,
566 extensive PRDM9 variation typically is seen between, and within, a given subspecies.
567 This holds true also for Antarctic minke whales, thus the observation of an identical

568 PRDM9 variant shared between two subspecies of common minke whales is
569 unexpected. The PRDM9 protein level diversity observed within each Hemisphere
570 contrasted with the levels of population differentiation seen in microsatellite as well as
571 mitochondrial haplotypes. However, diversity between repeat types of the *Prdm9*
572 minisatellite can reveal even weak population structure and is consistent with data on
573 mitochondrial haplotypes and microsatellites. This contrasting pattern between
574 protein-level conservation and nucleotide differentiation is fascinating and points to
575 functional constraints operating on different levels of PRDM9 evolution.

576 **Population genetic analyses of minke whales and potential speciation**

577 Taxonomists have separated the common minke whale into two subspecies, the North
578 Atlantic minke whale *(B. a. acutorostrata)* and the North Pacific minke whale (*B. a.*
579 *scammoni*), which diverged approximately 1.5 million years ago. variability between
580 84bp *Prdm9* minisatellite repeat units, as well as microsatellite markers and mtDNA all
581 reveal weak population structure and differentiation between minke whales in North
582 Atlantic and North Pacific. Our population genetic data support this classification and
583 suggest that these populations may be in the process of speciating. Yet fascinatingly,
584 all individuals of both Northern Hemisphere populations had identical C_2H_2 domains,
585 which suggests that these subspecies should still be able to interbreed if given a
586 chance. As geographical isolation of the two subspecies is currently still upheld by the
587 permanent polar ice, allopatric speciation may be promoted. However, due to global
588 warming, the two subspecies might come into secondary contact again in the future.
589 The removal of the current geographical barrier could therefore disrupt speciation,
590 especially in the light of identical C_2H_2 domains encoded by the mammalian hybrid
591 sterility gene *Prdm9.*

592

593 **Hybrids between Antarctic minke whale and Common minke whale**

594 Analysis of the hypervariable region (HVR) on the mitochondrial D-Loop suggests that
595 Antarctic minke whale and Common minke whale evolved from a common ancestor in
596 the Southern Hemisphere during a period of global warming approximately 5 million
597 years ago (PASTENE *et al.* 2007). Common minke whales and Antarctic minke whales
598 are now separated by both geography and seasonality. However, while their winter
599 habitats and breeding grounds remain unknown, it is assumed that *B. a. acutorostrata*
600 migrates south between November and March to give birth in warmer waters and
601 occasional individuals are seen as far south as the Gulf of Mexico (ROSEL *et al.* 2016).
602 Interbreeding between the populations of the Northern and Southern Hemisphere
603 appears unlikely at the time our samples were collected (the 1980s) because our study
604 confirms previous findings (VAN PIJLEN *et al.* 1995) that common minke whales and
605 Antarctic minke whales exhibit extensive genetic differences in their mtDNA and
606 microsatellites. However, recent reports have uncovered rare hybridization events
607 between Common minke whale and Antarctic minke whales (GLOVER *et al.* 2013),
608 which shows that interbreeding is unlikely but possible. The question remains whether
609 known reproductive isolation mechanism exists in hybrids between the two minke
610 whale species.

611

612 The *Prdm9* gene remains the only hybrid sterility gene known in vertebrates. Large
613 evolutionary turnover of the coding minisatellite is seen in all mammals characterised
614 to date- including minke whales (*this study*). One consequence is genetic variation in
615 the DNA-binding C_2H_2-ZnF-domain that in turn introduces variation in species'
616 recombination landscapes, that differ even within populations of the same species

617 (BERG *et al.* 2011; PRATTO *et al.* 2014). Based on computed PRMD9 C_2H_2-ZnF binding
618 motifs found in Common minke whale and Antarctic minke whale, we would speculate
619 that the recombination initiation landscapes of the two variants would not overlap. The
620 associated variation in recombination landscapes is implicated in generating
621 reproductive isolation in mammals, which has been well characterized in mice. Here,
622 historical erosion of genomic binding sites of PRDM9 ZnF domains has been shown to
623 lead to the introduction of asymmetric sets of DSBs in evolutionary divergent
624 homologous genomic sequences (PRATTO *et al.* 2014; SMAGULOVA *et al.* 2016). This
625 asymmetry is likely caused because PRDM9 initiation can erode its own binding via
626 biased gene conversion over long evolutionary timescales.
627
628 As a result, in hybrid genomes, the variant of one species preferentially binds the
629 ancestral, binding sites on the homologue of the other species, that have not been
630 eroded, and vice versa. The resulting asymmetry of recombination initiation sites is
631 believed to be responsible for the inefficient DSB repair, defective pairing, asynapsis
632 and segregation of the chromosomes that are observed in intersubspecific mouse
633 hybrids (DAVIES *et al.* 2016; FOREJT 2016; ZELAZOWSKI AND COLE 2016). The low
634 PRDM9 diversity in the Northern Hemisphere would predict extensive historical erosion
635 of PRDM9 binding sites in the genomes of Common Minke Whales. A degree of
636 erosion may also be true across the Southern Hemisphere, where most animals share
637 identical core motifs. Thus, the predicted high level of PRDM9 binding site erosion,
638 coupled with non-overlapping minke whale PRDM9 binding motifs would predict
639 asymmetric PRDM9 recombination imitation in minke whale hybrids, which is
640 implicated in hybrid male sterility.
641

642 Despite this potential incompatibility, two viable female minke whale hybrids have been
643 observed, one of which showed signs of successful backcrossing (GLOVER *et al.* 2010;
644 GLOVER *et al.* 2013). This observation alone is not sufficient to clarify whether
645 postzygotic isolation mechanisms do or do not generally operate. After all, Haldane's
646 rule predicts that the sterile sex will be the heterogametic males, rather than the
647 females on which observations were made, an expectation that holds in PRDM9-
648 mediated hybrid sterility in mice (MIHOLA *et al.* 2009; DZUR-GEJDOSOVA *et al.* 2012;
649 BHATTACHARYYA *et al.* 2014). Therefore, it still remains to be understood whether
650 sporadically occurring hybridization events will generate infertile males or, instead,
651 signify that start of a breakdown of genetic isolation. Further studies on minke whale
652 hybrids are necessary to elucidate this matter.
653

654 **Conclusion**

655 We uncovered that Minke whales, and related species, possess coding sequences for
656 a complete set of KRAB, SSXRD and SET domains, a necessary feature of organisms
657 with PRDM9-regulated recombination (BAKER *et al.* 2017). Sequencing revealed
658 evolutionary turnover of the PRDM9 coding minisatellite. Our phylogenetic analysis
659 revealed a good association of groups with their geographical origin, a high level of
660 size homoplasy in common minke whales and a lower level of size homoplasy in
661 Antarctic minke whales. In Antarctic minke whales (*Balaenoptera bonarensis*), we
662 observe extensive diversity within the C_2H_2-ZNF domain of PRDM9, even at fine
663 geographical scale. In direct opposition, we observe no genetic diversity within the
664 C_2H_2-ZNF in common minke whales. In each Hemisphere, the PRDM9 protein diversity
665 is contrasting strongly with the underlying population structure we observe between
666 and within subspecies. However, on the nucleotide level, variation between *Prdm9*

667 minisatellite repeat types mirrored the variation in mitochondrial DNA and
668 microsatellites in respective animals. This apparent contradiction points to
669 maintenance of conserved protein sequence, even across minke whale subspecies
670 boundaries.
671
672 **Declarations**
673
674 **Consent for publication**
675 Not applicable
676
677 **Availability of data and materials**
678 Most data generated or analyzed during this study are included in this published article
679 and its supplementary information files, Genetic data generated and analyzed during
680 the current study are available in the Zenodo repository,
681 https://doi.org/10.5281/zenodo.4309437, and R Scripts are available from
682 https://gitlab.gwdg.de/mpievolbio-it/repeatr.
683
684 **Competing Interest**
685 The authors declare no competing interests
686
687 **Funding**
688 We thank the Max Planck Society for funding this study.
689
690 **Author Contributions**

ED designed the experiments, analyzed data and wrote parts of the manuscript. KKU designed experiments and analyzed the data, WA provided the samples and wrote parts of the manuscript and L.O.-H. conceptualized the study, designed the experiments, analyzed data and wrote the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

none

Acknowledgements

We thank Nicole Thomsen and Olga Eitel for technical support. We would also like to thank Kevin Glover and Bjørghild B. Seliussen for sharing primer sequences and genotyping conditions for microsatellites markers and the hypervariable segments of the mtDNA D-loop.

References:

Ahlawat, S., S. De, P. Sharma, R. Sharma, R. Arora *et al.*, 2017 Evolutionary dynamics of meiotic recombination hotspots regulator PRDM9 in bovids. Mol Genet Genomics 292**:** 117-131.

Ahlawat, S., P. Sharma, R. Sharma, R. Arora and S. De, 2016a Zinc Finger Domain of the PRDM9 Gene on Chromosome 1 Exhibits High Diversity in Ruminants but Its Paralog PRDM7 Contains Multiple Disruptive Mutations. PLoS One 11**:** e0156159.

Ahlawat, S., P. Sharma, R. Sharma, R. Arora and N. K. Verma, 2016b Evidence of positive selection and concerted evolution in the rapidly evolving PRDM9 zinc finger domain in goats and sheep. 47**:** 740-751.

Ahlawat, S., P. Sharma, R. Sharma, R. Arora, N. K. Verma *et al.*, 2016c Evidence of positive selection and concerted evolution in the rapidly evolving PRDM9 zinc finger domain in goats and sheep. Anim Genet 47**:** 740-751.

Baker, Z., M. Schumer, Y. Haba, L. Bashkirova, C. Holland *et al.*, 2017 Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. Elife 6.

Bakke, I., S. Johansen, Ø. Bakke and M. R. El-Gewely, 1996 Lack of population subdivision among the minke whales (Balaenoptera acutorostrata) from Icelandic and Norwegian waters based on mitochondrial DNA sequences. Marine Biology 125: 1-9.

Berg, I. L., R. Neumann, K. W. Lam, S. Sarbajna, L. Odenthal-Hesse *et al.*, 2010 PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42: 859-863.

Berg, I. L., R. Neumann, S. Sarbajna, L. Odenthal-Hesse, N. J. Butler *et al.*, 2011 Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc Natl Acad Sci U S A 108: 12378-12383.

Berube, M., H. Jorgensen, R. McEwing and P. J. Palsboll, 2000 Polymorphic di-nucleotide microsatellite loci isolated from the humpback whale, Megaptera novaeangliae. Mol Ecol 9: 2181-2183.

Berube, M., and P. Palsboll, 1996 Identification of sex in cetaceans by multiplexing with three ZFX and ZFY specific primers. Mol Ecol 5: 283-287.

Bhattacharyya, T., R. Reifova, S. Gregorova, P. Simecek, V. Gergelits *et al.*, 2014 X chromosome control of meiotic chromosome synapsis in mouse inter-subspecific hybrids. PLoS Genet 10: e1004088.

Billings, T., E. D. Parvanov, C. L. Baker, M. Walker, K. Paigen *et al.*, 2013 DNA binding specificities of the long zinc-finger recombination protein PRDM9. Genome Biol 14: R35.

Buard, J., E. Rivals, D. Dunoyer de Segonzac, C. Garres, P. Caminade *et al.*, 2014 Diversity of Prdm9 zinc finger array in wild mice unravels new facets of the evolutionary turnover of this coding minisatellite. PLoS One 9: e85021.

Cole, F., F. Baudat, C. Grey, S. Keeney, B. de Massy *et al.*, 2014 Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. Nat Genet 46: 1072-1080.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Inc.

Davies, B., E. Hatton, N. Altemose, J. G. Hussin, F. Pratto *et al.*, 2016 Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature 530: 171-176.

Dieringer, D., and C. Schlötterer, 2003 microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Molecular Ecology Notes.

Dzur-Gejdosova, M., P. Simecek, S. Gregorova, T. Bhattacharyya and J. Forejt, 2012 Dissecting the genetic architecture of F1 hybrid sterility in house mice. Evolution 66: 3321-3335.

Earl, D. A., and B. M. vonHoldt, 2011 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4: 359-361.

Evanno, G., S. Regnaut and J. Goudet, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14: 2611-2620.

Forejt, J., 2016 Genetics: Asymmetric breaks in DNA cause sterility. Nature 530: 167-168.

Glover, K. A., T. Haug, N. Øien, L. Walløe, L. Lindblom *et al.*, 2012 The Norwegian minke whale DNA register: a data base monitoring commercial harvest and trade of whale products. Fish and Fisheries 13: 313-332.

Glover, K. A., N. Kanda, T. Haug, L. A. Pastene, N. Oien *et al.*, 2010 Migration of Antarctic minke whales to the Arctic. PLoS One 5: e15197.

Glover, K. A., N. Kanda, T. Haug, L. A. Pastene, N. Oien *et al.*, 2013 Hybrids between common and Antarctic minke whales are fertile and can back-cross. BMC Genet 14: 25.

Groeneveld, L. F., R. Atencia, R. M. Garriga and L. Vigilant, 2012 High diversity at PRDM9 in chimpanzees and bonobos. PLoS One 7: e39064.

Imai, Y., F. Baudat, M. Taillepierre, M. Stanzione, A. Toth *et al.*, 2017 The PRDM9 KRAB domain is required for meiosis and involved in protein interactions. Chromosoma 126: 681-695.

Jeffreys, A. J., V. E. Cotton, R. Neumann and K. W. Lam, 2013 Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. Proc Natl Acad Sci U S A 110: 600-605.

Jeffreys, A. J., R. Neumann and V. Wilson, 1990 Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60: 473-485.

Jin, L., and R. Chakraborty, 1995 Population structure, stepwise mutations, heterozygote deficiency and thei implications in DNA forensics. Heredity (Edinb) 74: 274-285.

Jost, L., 2008 G(ST) and its relatives do not measure differentiation. Mol Ecol 17: 4015-4026.

Katoh, K., K. Kuma, H. Toh and T. Miyata, 2005 MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung *et al.*, 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649.

Kono, H., M. Tamura, N. Osada, H. Suzuki, K. Abe *et al.*, 2014 Prdm9 polymorphism unveils mouse evolutionary tracks. DNA Res 21: 315-326.

Maheshwari, S., and D. A. Barbash, 2011 The genetics of hybrid incompatibilities. Annu Rev Genet 45: 331-355.

Malde, K., B. B. Seliussen, M. Quintela, G. Dahle, F. Besnier *et al.*, 2017 Whole genome resequencing reveals diagnostic markers for investigating global migration and hybridization between minke whale species. BMC Genomics 18: 76.

Mihola, O., Z. Trachtulec, C. Vlcek, J. C. Schimenti and J. Forejt, 2009 A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323: 373-375.

Mistry, J., R. D. Finn, S. R. Eddy, A. Bateman and M. Punta, 2013 Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41: e121.

Mukaj, A., J. Pialek, V. Fotopulosova, A. P. Morgan, L. Odenthal-Hesse *et al.*, 2020 Prdm9 inter-subspecific interactions in hybrid male sterility of house mouse. Mol Biol Evol.

Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler *et al.*, 2012 Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8: e1002764.

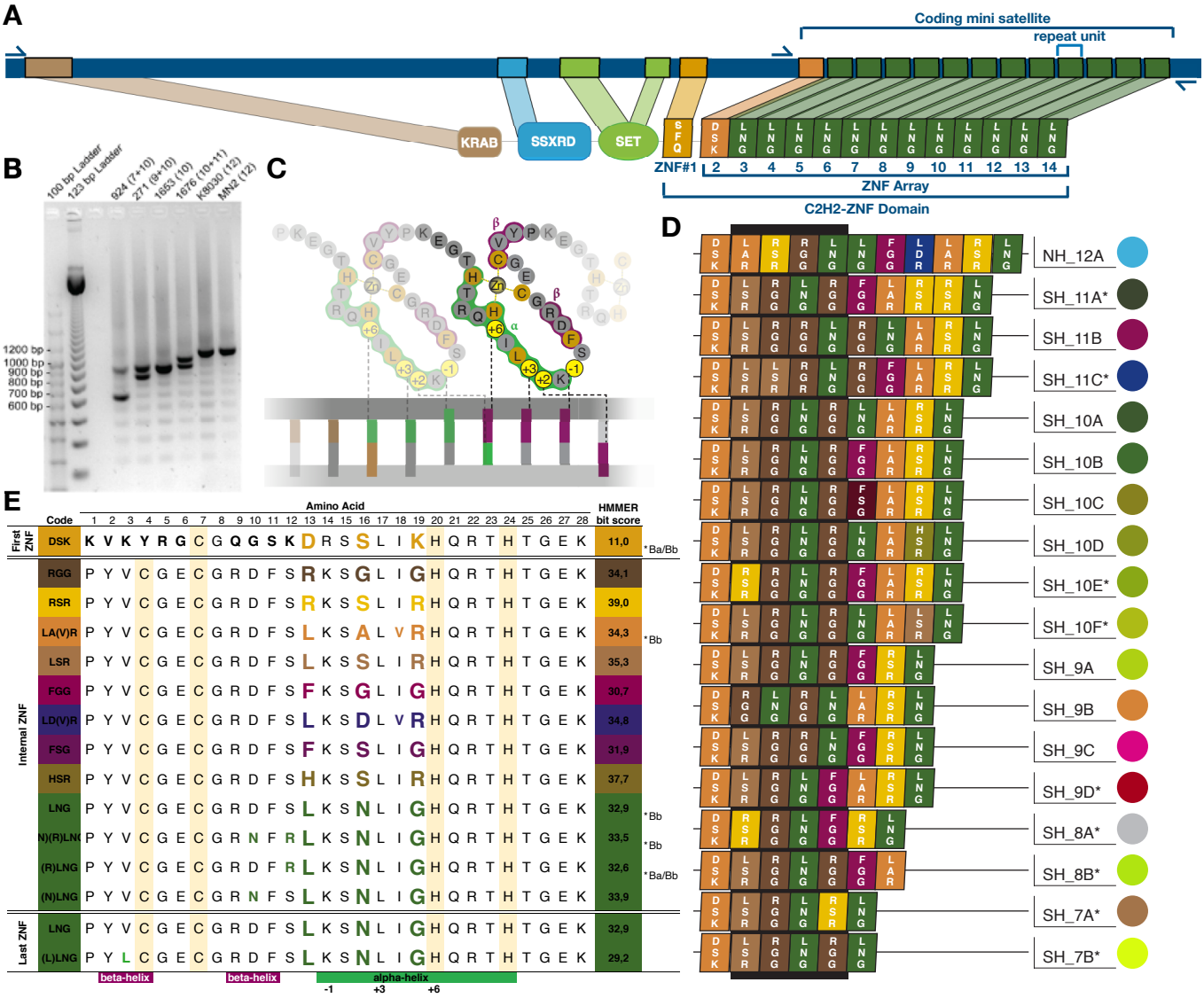
Myers, S., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman *et al.*, 2010 Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876-879.

821 Nei, M., 1973 Analysis of Gene Diversity in Subdivided Populations. Proc Natl Acad
822 Sci U S A 70: pp. 3321-3323.
823 Nguyen, L. T., H. A. Schmidt, A. von Haeseler and B. Q. Minh, 2015 IQ-TREE: a fast
824 and effective stochastic algorithm for estimating maximum-likelihood
825 phylogenies. Mol Biol Evol 32: 268-274.
826 Odenthal-Hesse, L., I. L. Berg, A. Veselis, A. J. Jeffreys and C. A. May, 2014
827 Transmission distortion affecting human noncrossover but not crossover
828 recombination: a hidden source of meiotic drive. PLoS Genet 10: e1004106.
829 Orozco-terWengel, P., J. Corander and C. Schlotterer, 2011 Genealogical lineage
830 sorting leads to significant, but incorrect Bayesian multilocus inference of
831 population structure. Mol Ecol 20: 1108-1121.
832 Paigen, K., and P. M. Petkov, 2018 PRDM9 and Its Role in Genetic Recombination.
833 Trends Genet 34: 291-300.
834 Palsboll, P. J., M. Berube, A. H. Larsen and H. Jorgensen, 1997 Primers for the
835 amplification of tri- and tetramer microsatellite loci in baleen whales. Mol Ecol
836 6: 893-895.
837 Paradis, E., 2010 pegas: an R package for population genetics with an integrated-
838 modular approach. Bioinformatics 26: 419-420.
839 Paradis, E., J. Claude and K. Strimmer, 2004 APE: Analyses of Phylogenetics and
840 Evolution in R language. Bioinformatics 20: 289-290.
841 Parvanov, E. D., P. M. Petkov and K. Paigen, 2010 Prdm9 controls activation of
842 mammalian recombination hotspots. Science 327: 835.
843 Parvanov, E. D., H. Tian, T. Billings, R. L. Saxl, C. Spruce *et al.*, 2017 PRDM9
844 interactions with other proteins provide a link between recombination hotspots
845 and the chromosomal axis in meiosis. Mol Biol Cell 28: 488-499.
846 Pastene, L. A., M. Goto, N. Kanda, A. N. Zerbini, D. Kerem *et al.*, 2007 Radiation and
847 speciation of pelagic organisms during periods of global warming: the case of
848 the common minke whale, Balaenoptera acutorostrata. Mol Ecol 16: 1481-
849 1495.
850 Perrin, W. F., and R. L. B. Jr., 2009 Minke Whales: Balaenoptera acutorostrata and B.
851 bonaerensis, pp. 733-735 in *Encyclopedia of Marine Mammals (Second*
852 *Edition)*, edited by W. F. Perrin, B. Würsig and J. G. M. Thewissen, Academic
853 Press.
854 Persikov, A. V., and M. Singh, 2014 De novo prediction of DNA-binding specificities
855 for Cys2His2 zinc finger proteins. Nucleic Acids Res 42: 97-108.
856 Porras-Hurtado, L., Y. Ruiz, C. Santos, C. Phillips, A. Carracedo *et al.*, 2013 An
857 overview of STRUCTURE: applications, parameter settings, and supporting
858 software. Front Genet 4: 98.
859 Pratto, F., K. Brick, P. Khil, F. Smagulova, G. V. Petukhova *et al.*, 2014 DNA
860 recombination. Recombination initiation maps of individual human genomes.
861 Science 346: 1256442.
862 Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder *et al.*, 2005 InterProScan:
863 protein domains identifier. Nucleic Acids Res 33: W116-120.
864 Ramasamy, R. K., S. Ramasamy, B. B. Bindroo and V. G. Naik, 2014 STRUCTURE
865 PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly
866 interface. Springerplus 3: 431.
867 Rosel, P. E., L. A. Wilcox, C. Monteiro and M. C. Tumlin, 2016 First record of Antarctic
868 minke whale, Balaenoptera bonaerensis, in the northern Gulf of Mexico. Marine
869 Biodiversity Records 9.

870 Schwartz, J. J., D. J. Roach, J. H. Thomas and J. Shendure, 2014 Primate evolution
871 of the recombination regulator PRDM9. Nat Commun 5**:** 4370.
872 Slater, G. S., and E. Birney, 2005 Automated generation of heuristics for biological
873 sequence comparison. BMC Bioinformatics 6**:** 31.
874 Smagulova, F., K. Brick, Y. Pu, R. D. Camerini-Otero and G. V. Petukhova, 2016 The
875 evolutionary turnover of recombination hot spots contributes to speciation in
876 mice. Genes Dev 30**:** 266-280.
877 Steiner, C. C., and O. A. Ryder, 2013 Characterization of Prdm9 in equids and sterility
878 in mules. PLoS One 8**:** e61746.
879 Striedner, Y., T. Schwarz, T. Welte, A. Futschik, U. Rant *et al.*, 2017 The long zinc
880 finger domain of PRDM9 forms a highly stable and long-lived complex with its
881 DNA recognition sequence. Chromosome Res 25**:** 155-172.
882 Suchard, M. A., and B. D. Redelings, 2006 BAli-Phy: simultaneous Bayesian inference
883 of alignment and phylogeny. Bioinformatics 22**:** 2047-2048.
884 Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in
885 the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol
886 Evol 10**:** 512-526.
887 Valsecchi, E., and W. Amos, 1996 Microsatellite markers for the study of cetacean
888 populations. Mol Ecol 5**:** 151-156.
889 van Pijlen, I. A., B. Amos and T. Burke, 1995 Patterns of genetic variability at individual
890 minisatellite loci in minke whale Balaenoptera acutorostrata populations from
891 three different oceans. Mol Biol Evol 12**:** 459-472.
892 Vara, C., L. Capilla, L. Ferretti, A. Ledda, R. A. Sanchez-Guillen *et al.*, 2019 PRDM9
893 diversity at fine geographical scale reveals contrasting evolutionary patterns
894 and functional constraints in natural populations of house mice. Mol Biol Evol.
895 Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen *et al.*, 2012 Primer-BLAST:
896 A tool to design target-specific primers for polymerase chain reaction. BMC
897 Bioinformatics 13:134.
898 Zelazowski, M. J., and F. Cole, 2016 X marks the spot: PRDM9 rescues hybrid sterility
899 by finding hidden treasure in the genome. Nat Struct Mol Biol 23**:** 267-269.
900

901

902 **Figures and Tables**



903

904 **Figure 1 Diversity in the Cys2His2-ZnF domain in minke whales.** (A) *Prdm9* gene
905 annotation in *Balaenoptera acutorostrata scammoni* genome reference, with primer
906 site annotations. We predicted the PRDM9 protein (B) Representative PCR products
907 of DNAs from individuals from AN IV (294,271), AN V (1653), NP (K8030) and NA
908 (MN2) showing variation in the number of minisatellite repeats units (number of ZnFs
909 in brackets) (C) Stylized structure of C_2H_2-ZnF binding to DNA, with nucleotide-
910 specificity conferred by amino-acids in positions -1, +2, +3, and +6 of alpha-helices.
911 (D) PRDM9 C_2H_2-ZnF arrays identified in minke whales, named with broad-scale
912 sampling location, Northern Hemisphere (NH) and Southern Hemisphere (SH) and

913 total number of ZnFs in the C_2H_2-ZnFs domain. (E) Types of PRDM9 C_2H_2-ZnFs in
914 minke whales. We generate three-letter-codes using the IUPAC nomenclature of
915 amino-acids involved in DNA-binding. All variable amino acids are colored, and
916 asterisks label C_2H_2-ZnFs also present in genome references of *Balaenoptera*
917 *acutorostrata scammoni* (*Ba) or *Balaenoptera bonarensis* (*Bb), see also
918 Supplementary Figure 4.
919

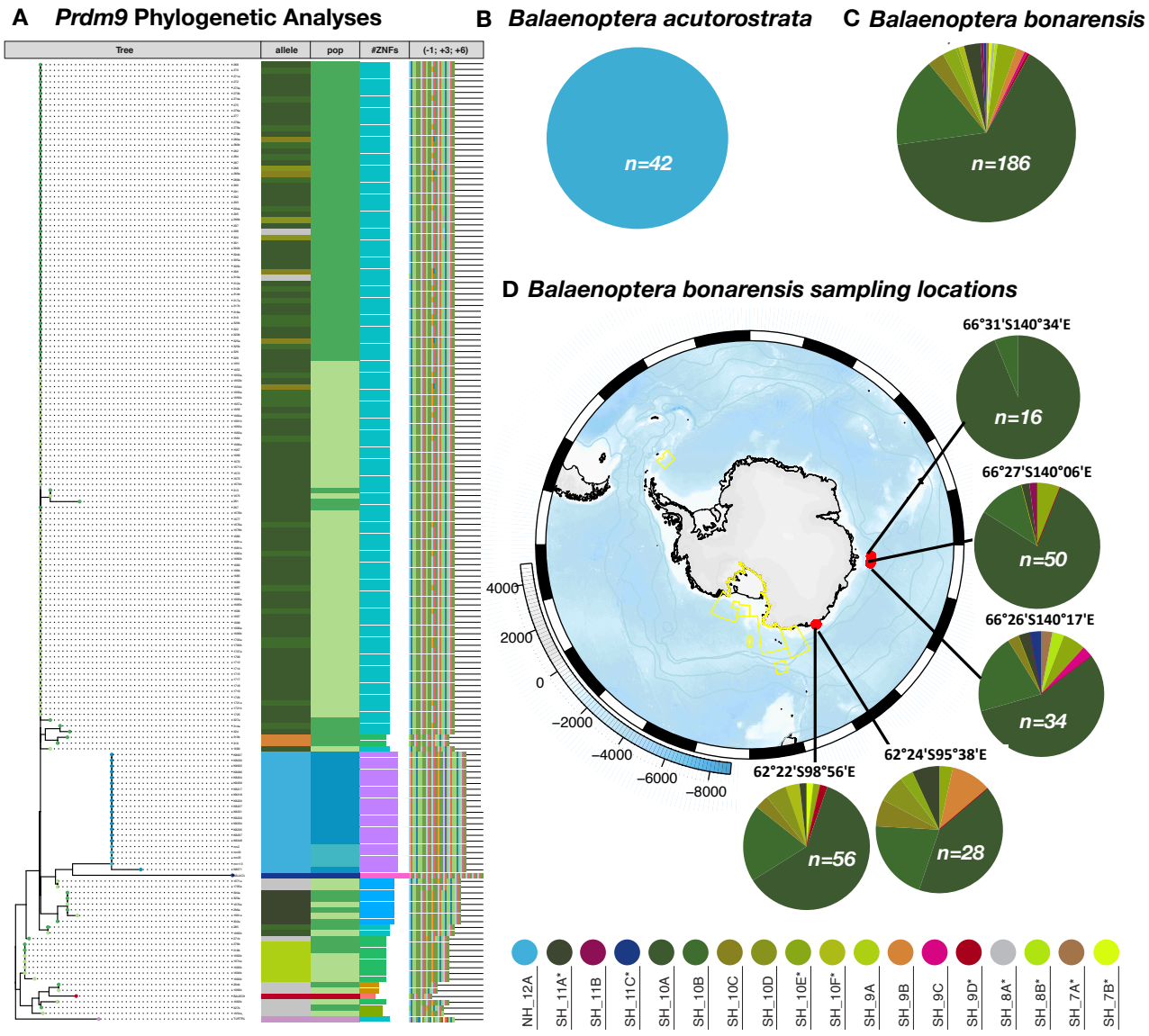
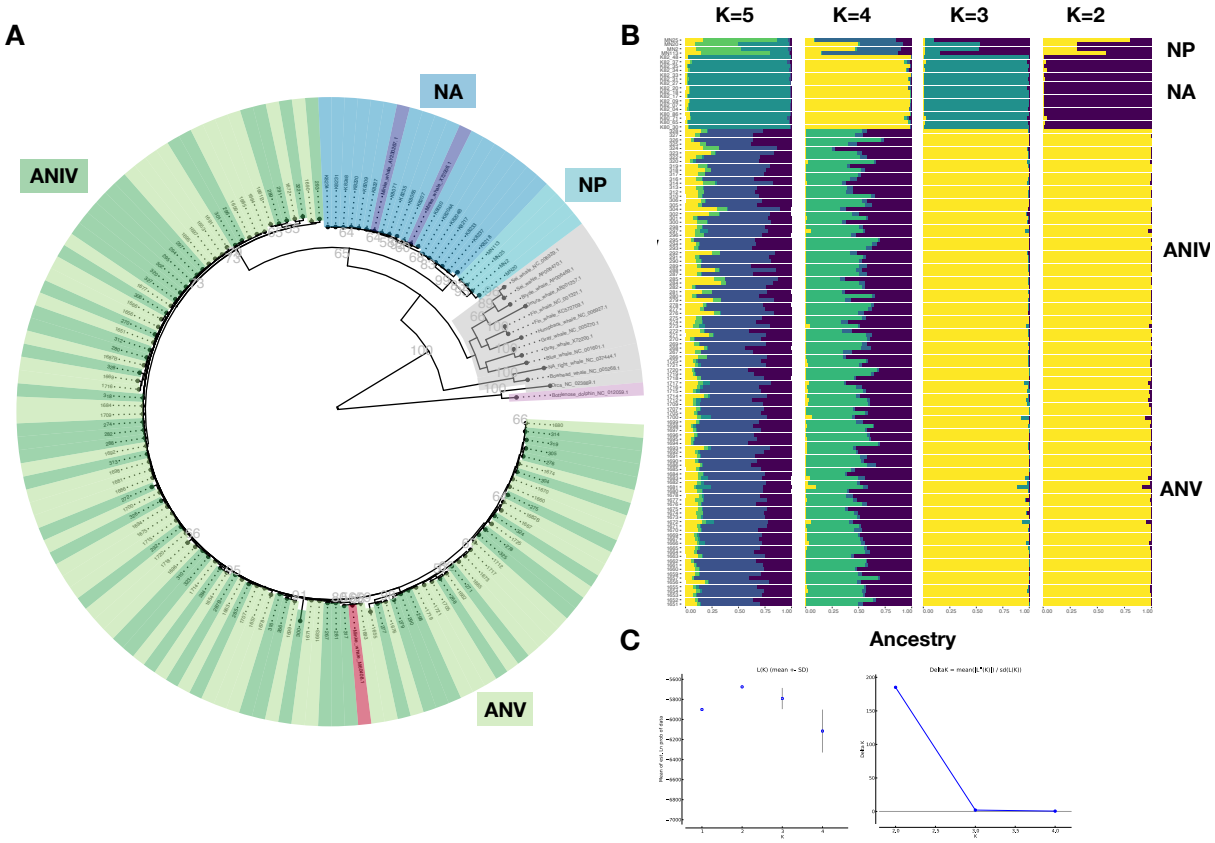


Figure 2 *Prdm9* Phylogenetic analyses and allele frequencies at different geographical scales. (A) PRDM9 phylogenetic analyses across all geographical regions, including several outgroups, from top to bottom: **-1, +3, +6**: PRDM9 variants as in Figure 1. **# ZnFs:** number of ZnFs **pop:** assigned population of individuals (dark blue): North Pacific, (light blue) North Atlantic, (light green) Antarctic Area IV, (dark green) Antarctic Area V, as well as Genome Reference alleles: (blue*) Balaenoptera acutorostrata scammoni*, (red) *Balaenoptera bonarensis* and outgroups (pink) *Tursiops Truncatus* **allele:** PRDM9 coding minisatellite allele colored as the translated variant from Figure 1. **Tree:** Nucleotide Phylogeny of PRDM9 alleles. (B) allelic diversity in

930 Common minke whales (C) allelic diversity in Antarctic minke whales (D) fine-scale
931 diversity by Sampling locations of Antarctic minke whales on a map of the Southern
932 Ocean (yellow delineations show protected areas).
933
934 **Table 1 Nucleotide diversity (TAMURA AND NEI 1993) of the minisatellite coding for**
935 **the PRDM9-C_2H_2-ZnF array**, analyzed per sample region. Average pairwise
936 nucleotide diversity was analyzed for the all sites, and also excluding the nucleotides
937 coding for amino-acids at hypervariable sites (-1, +3, +6) in the alpha-helix of C_2H_2-
938 ZnFs.

Sample region	Area V	Area IV	North Atlantic	North Pacific
Number of analyzed 84bp repeat units per population	554	577	170	40
Population theta (segregating sites)	2,031	1,731	2,102	2,116
Average pairwise nucleotide diversity (excluding hypervariable sites)	0.0148±0.0001	0.0143±0.0001	0.0184±0.0002	0.0183± 0.0002
Average pairwise nucleotide diversity (including hypervariable sites)	0.0382±0.0005	0.0374±0.0005	0.0434±0.0006	0.0437±0.0006

939



940

941 **Figure 3 Population structure of minke whales** (NP) North Pacific, (NA) North
942 Atlantic (ANIV) Antarctic Area IV (ANV) Antarctic Area V (A) Phylogeny of minke whale
943 mitochondrial D-loop region, containing the hypervariable segments (HVS). (B)
944 Population STRUCTURE analyses of minisatellites with *a priori* location information.
945 (C) Log probability of data L(K) as a function of K (mean ± SD) (D) Magnitude of ΔK as
946 a function of K (mean ± SD over 50 replicates).

947

948

949

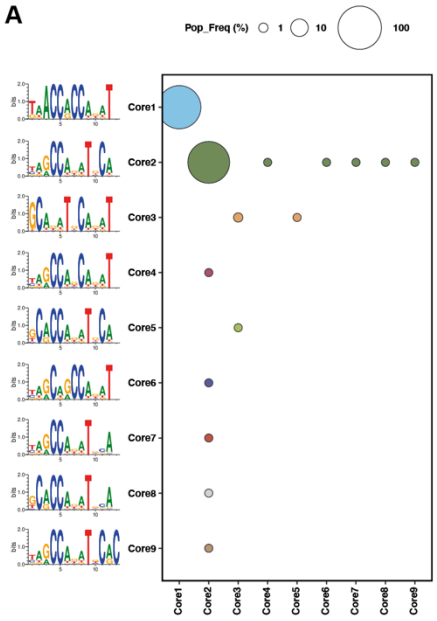
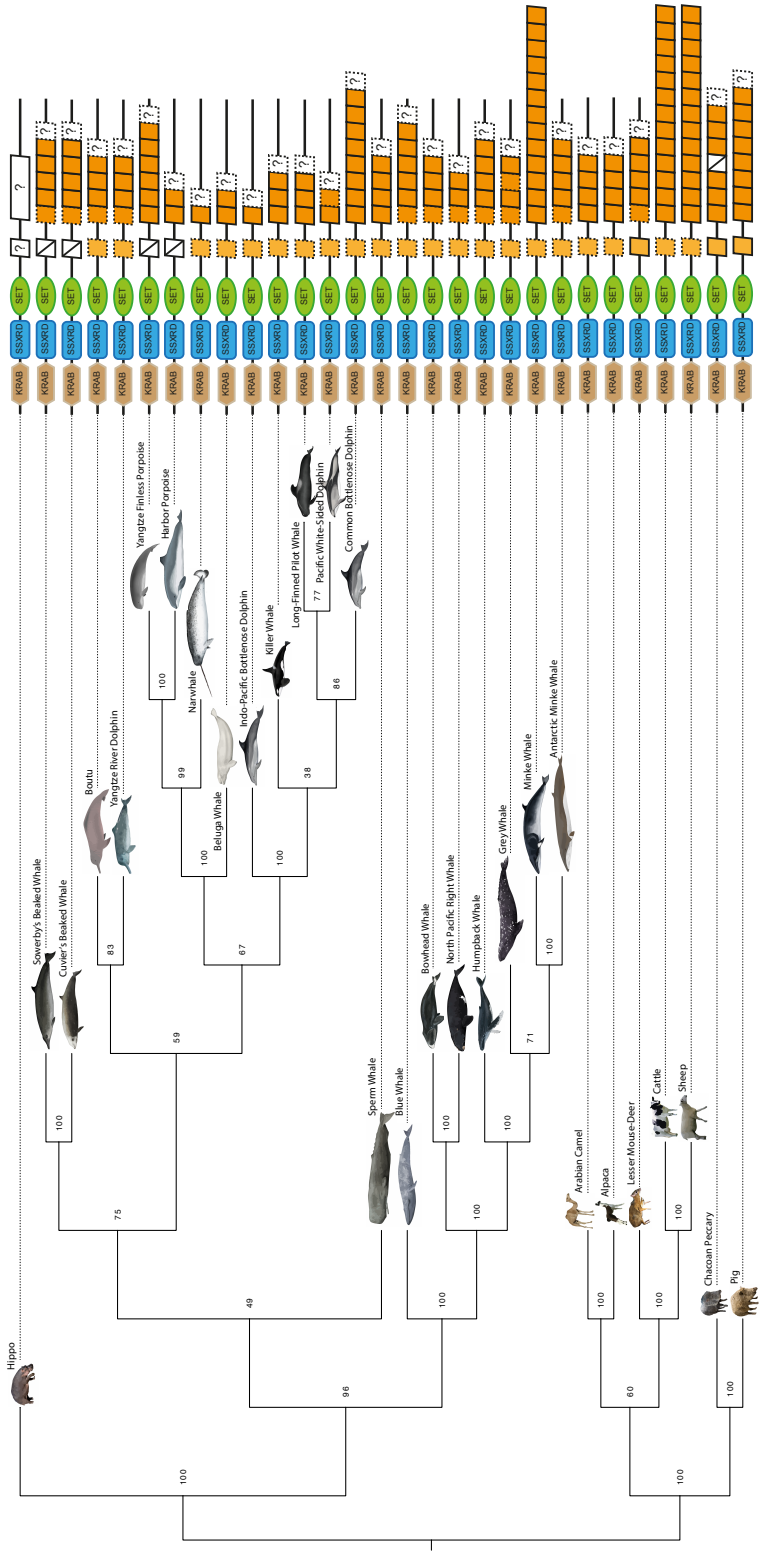
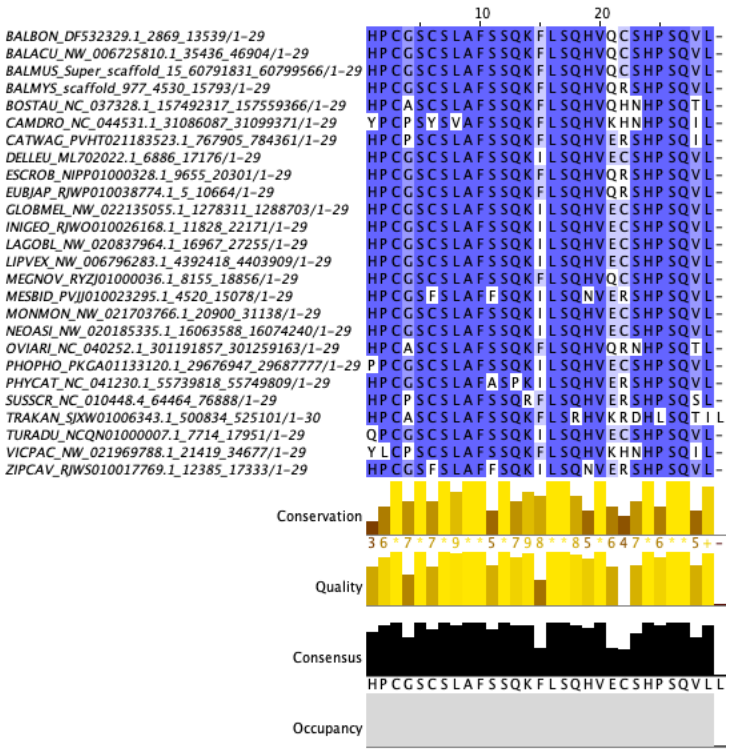


Figure 4 PRDM9 Array binding predictions and population frequency based on ZnFs#3-#6 "core motif" combinations across all individuals

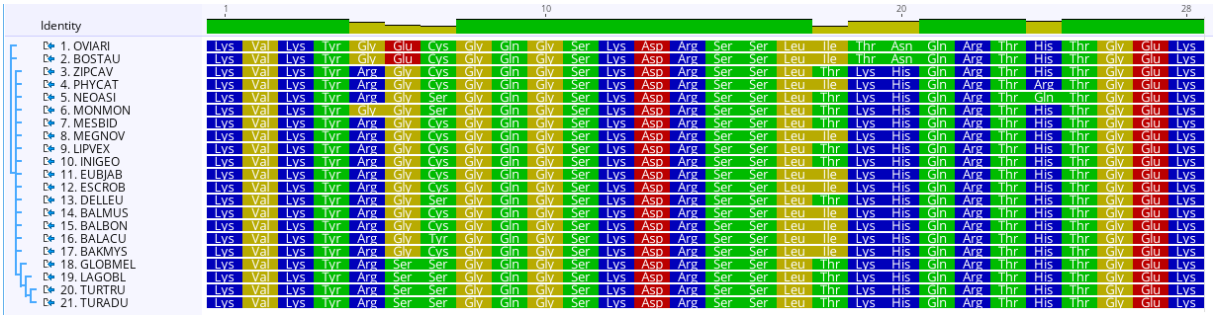
Supplementary Information



Supplementary Figure 1: PRDM9 Phylogenetic analysis across *Cetartiodactyla*, using N-terminal KRAB, SSXRD and SET domains, with data from public genomic databases collected in (**Supplementary Table 2**)



Supplementary Figure 2: **Alignment of proximal ZnFs of *Cetartiodactyla***, in which this ZnF (outside the ZnF array) was present in the public genomic databases (Supplementary Table 2). The proximal ZnF is fully conserved across *Balaenoptera.* A single amino-acid change at position 22 in bowhead whale (*Balaena mysticetus)* is observed, compared to *Balaenoptera*, with the DNA contacting residues remaining conserved.



Supplementary Figure 3: Alignment of second ZnFs of all *Cetartiodactyla*, where this ZnF was present in public genomic databases (Supplementary Table 2). Across *Cetacea*, ZnF#2 are conserved at the amino-acids 13, 16 and 19, which code for the DNA-binding positions of the alpha-helix of each ZnF. Amino-acids 19 is changed between *Cetacea* and *Ruminantia*, and additional changes are observed at amino-acids 5, 6, 18, and 24.

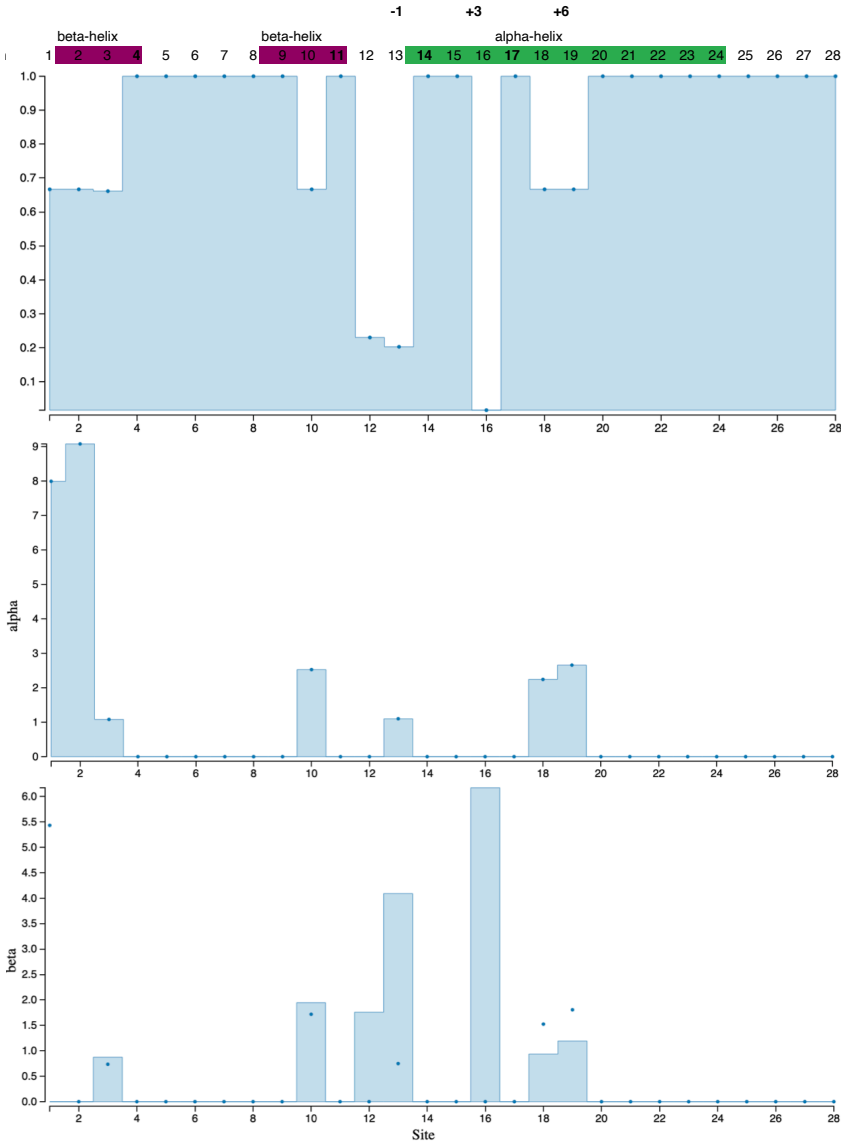
beta-helix | beta-helix | alpha-helix (-1, +3, +6)

	Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	HMMER bit score	
First ZNF	DSK	K	V	K	Y	R	G	C	G	Q	G	S	K	D	R	S	S	L	I	K	H	Q	R	T	H	T	G	E	K	11,0	*Ba/Bb
Internal ZNF	RGG	P	Y	V	C	G	E	C	G	R	D	F	S	R	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	K	34,1	
	RSR	P	Y	V	C	G	E	C	G	R	D	F	S	R	K	S	S	L	I	R	H	Q	R	T	H	T	G	E	K	39,0	
	LA(V)R	P	Y	V	C	G	E	C	G	R	D	F	S	L	K	S	A	L	V	R	H	Q	R	T	H	T	G	E	K	34,3	*Bb
	LSR	P	Y	V	C	G	E	C	G	R	D	F	S	L	K	S	S	L	I	R	H	Q	R	T	H	T	G	E	K	35,3	
	FGG	P	Y	V	C	G	E	C	G	R	D	F	S	F	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	K	30,7	
	LD(V)R	P	Y	V	C	G	E	C	G	R	D	F	S	L	K	S	D	L	V	R	H	Q	R	T	H	T	G	E	K	34,8	
	FSG	P	Y	V	C	G	E	C	G	R	D	F	S	F	K	S	S	L	I	G	H	Q	R	T	H	T	G	E	K	31,9	
	HSR	P	Y	V	C	G	E	C	G	R	D	F	S	H	K	S	S	L	I	R	H	Q	R	T	H	T	G	E	K	37,7	
	LNG	P	Y	V	C	G	E	C	G	R	D	F	S	L	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	32,9	*Bb
	N)(R)LNG	P	Y	V	C	G	E	C	G	R	N	F	R	L	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	33,5	*Bb
	(R)LNG	P	Y	V	C	G	E	C	G	R	D	F	R	L	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	32,6	*Ba/Bb
	(N)LNG	P	Y	V	C	G	E	C	G	R	N	F	S	L	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	33,9	
Last ZNF	LNG	P		V	C	G	E	C	G	R	D	F	S	L	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	32,9	
	(L)LNG	P		L	C	G	E	C	G	R	D	F	S	L	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	29,2	
	RGG	P		V	C	A	E	C	G	R	D	F	S	R	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	K	33,5	
	RGG	P		V	C	G	Q	C	G	R	D	F	S	R	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	K	33,5	
	RGG	P		V	C	G	E	C	G	G	D	F	S	R	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	K	29,2	
	RSR	P		V	C	R	E	C	G	R	D	F	S	R	K	S	S	L	I	R	H	Q	R	T	H	T	G	E	K	37,8	
	LS(V)R	P		V	C	G	E	C	G	R	D	F	S	L	K	S	S	L	V	R	H	Q	R	T	R	T	G	E	K	19,3	
	FGG(E)	P		V	C	G	E	C	G	R	D	F	S	F	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	E	30,7	
	FNG	P		V	C	G	E	C	G	R	D	F	S	F	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	33,4	
	HGG	P		V	C	G	E	C	G	R	D	F	S	H	K	S	G	L	I	G	H	Q	R	T	H	T	G	E	K	33,4	
	LSG	P		V	C	G	E	C	G	R	D	F	S	L	K	S	S	L	I	G	H	Q	R	T	H	T	G	E	K	31,5	
	PNG	P		V	C	G	E	C	G	R	D	F	S	P	K	S	N	L	I	G	H	Q	R	T	H	T	G	E	K	32,1	
	LHG	P		V	C	G	E	C	E	R	D	F	S	L	K	S	H	L	I	G	H	Q	R	T	H	T	G	E	K	29,5	
	RA(V)R	P		V	C	G	E	C	G	R	D	F	S	R	K	S	A	L	V	R	H	Q	R	T	H	T	G	E	K	38,2	

Supplementary Figure 4: All PRDM9 C_2H_2-ZnFs types in minke whales. We generate three-letter-codes for each C_2H_2-ZnF using the IUPAC nomenclature of amino-acids involved in DNA-binding. All variable amino acids are colored, and

982 asterisks label C_2H_2-ZnFs also present in genome references of *Balaenoptera*

983 *acutorostrata scammoni* (*Ba) or *Balaenoptera bonarensis* (*Bb).



984

985 **Supplementary Figure 5: Signals of selection on 84bp nucleotide repeats coding**

986 **for each ZnF identified in this study. "**Site" shows zinc-fingers amino-acid positions,

987 from top to bottom: a site is classified based on whether α>β (negative selection) or

988 α<β (positive selection). (A) p-value is significant when <0,05 (B) fel.α instantaneous

989 synonymous site rate (C) fel.β instantaneous non-synonymous site rate.

990

991 Supplementary Table 1: Descriptive statistics of microsatellites used in this study

992 analyzed using (DIERINGER AND SCHLÖTTERER 2003). The richness of allelic variants

993 (number of observed alleles) (**A**) Expected Heterozygosity (**H_E**) observed

994 Heterozygosity (**H_O**), excess or deficiency from the average number of heterozygotes

995 (**F_{IS}**) and (**H_S**), the probability that two random alleles from the same, randomly

996 chosen subpopulation are different by state.

Locus	Origin	A	Allele sizes	H_E	H_o	FIS	H_s
GATA028	AN IV	20	148-272	0,93	0,74	0,21	2,73
	AN V	20	148-274	0,92	0,88	0,05	2,67
	NA	-	-	-	-	-	-
	NP	-	-	-	-	-	-
GT575	AN IV	26	190-262	0,95	0,95	0	3,00
	AN V	29	185-262	0,95	0,97	-0,02	3,06
	NA	8	212-240	0,85	0,89	-0,06	1,89
	NP	5	206-216	0,89	1,00	-0,22	1,56
GT310	AN IV	18	81-135	0,88	0,78	0,11	2,41
	AN V	18	81-137	0,90	0,82	0,09	2,46
	NA	5	113-127	0,59	0,33	0,43	1,01
	NP	6	107-121	0,93	0,75	0,14	1,73
EV001	AN IV	13	103-147	0,83	0,95	-0,14	1,97
	AN V	14	105-141	0,82	0,78	0,04	1,97
	NA	5	139-153	0,72	0,56	0,22	1,35
	NP	4	131-151	0,86	0	1,00	1,39
GATA098	AN IV	9	76-108	0,83	0,89	-0,08	1,85
	AN V	10	70-108	0,84	0,97	-0,16	1,94
	NA	5	80-96	0,73	0,72	-0,01	1,35
	NP	3	84-92	0,68	1,00	-0,65	0,97
GT509	AN IV	19	179-219	0,92	0,89	0,03	2,66
	AN V	20	179-217	0,93	0,92	0,00	2,70
	NA	10	191-213	0,85	1,00	-0,19	1,97
	NP	5	195-211	0,89	1,00	-0,22	1,56
GATA417	AN IV	17	188-264	0,92	0,93	-0,02	2,58
	AN V	19	184-260	0,92	0,92	-0,01	2,60
	NA	8	212-240	0,86	0,94	-0,11	1,91
	NP	5	204-242	0,89	0,75	0,11	1,56
GT023	AN IV	14	90-118	0,87	0,81	0,07	2,24
	AN V	15	86-118	0,88	0,95	-0,08	2,36
	NA	8	92-108	0,77	0,78	-0,02	1,64
	NP	5	94-110	0,86	1,00	-0,27	1,49

EV037	AN IV	18	184-222	0,91	0,49	0,46	2,53
	AN V	16	184-222	0,93	0,52	0,44	2,59
	NA	7	192-208	0,72	0,72	-0,02	1,52
	NP	5	176-196	0,86	1,00	-0,27	1,49
	Origin	**A**		**H_e**	**H_o**	**FIS**	**H_s**
Average	AN IV	17,1		0,89	0,82	0,07	2,44
	AN V	17,9		0,90	0,86	0,04	2,48
	NA	7		0,76	0,74	0,03	1,58
	NP	4,75		0,86	0,81	-0,05	1,47

Supplementary Table 2: Public genome resources for PRDM9 occurrence and subsequent protein domain prediction with InterProScan. Note, that a 'normal' PRDM9 domain structure consists of a single C_2H_2 ZnF domain followed by an C_2H_2-ZnF-array. C_2H_2-ZnF prediction is based on http://zf.princeton.edu/fingerSelect.php. According to HMMER, most confident ZnF domains should have scores > 17.7, C_2H_2 ZnF domains with lower scores are indicated in brackets and were excluded from all downstream analyses.

SPECIES	COMMON	ALIAS	TAXID	KRAB	SSXRD	SET	PROXIMAL C_2H_2 ZNF	C_2H_2-ZNF ARRAY	URL-GENOME
BALAENOPTERA ACUTOROSTRATA	minke whale	BALACU	9767	x	x	x	x	13(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/0 /GCF_000493695.1_BalAcu1.0/GCF_0004936 cu1.0_genomic.fna.gz
BALAENOPTERA BONARENSIS	Antarctic minke whale	BALBON	33556	x	x	x	x	4(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/97 A_000978805.1_ASM97880v1/GCA_000978805.1_ 0v1_genomic.fna.gz
BALAENOPTERA MUSCULUS	blue whale	BALMUS	9771	-	x	x	x	5(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/009/87 A_009873245.2_mBalMus1.v2/GCA_009873245.2_ .v2_genomic.fna.gz
BALAENA MYSTICETUS	bowhead whale	BALMYS	27602	x	x	x	x	3(2)	http://alfred.liv.ac.uk/downloads/bowhead_whale/bo ale_scaffolds.zip
BOS TAURUS	cattle	BOS	9913	x	x	x	x	13(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/26 _002263795.1_ARS-UCD1.2/GCF_002263795.1_A UCD1.2_genomic.fna.gz

		TAU							
CAMELUS DROMEDARIUS	arabian camel	CAMDRO	9838	x	x	x	x	4(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/80 _000803125.2_CamDro3/GCF_000803125.2_CamD mic.fna.gz
CATAGONUS WAGNERI	chacoan peccary	CATWAG	51154	x	x	x	x	7	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/02 A_004024745.2_CatWag_v2_BIUU_UCD/GCA_004 CatWag_v2_BIUU_UCD_genomic.fna.gz
DELPHINAPTERUS LEUCAS	beluga whale	DELLEU	9749	x	x	x	x	2(1)	ftp://ftp.ensembl.org/pub/release-100/fasta/delphinapterus_leucas/dna/Delphinapterus SM228892v3.dna.toplevel.fa.gz
ESCHRICHTIUS ROBUSTUS	grey wahle	ESCROB	9764	x	x	x	x	2(3)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/002/18 A_002189225.1_ASM218922v1/GCA_002189225.1 922v1_genomic.fna.gz
EUBALAENA JAPONICA	north pacific right whale	EUBJAP	302098	x	x	x	x	2(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/36 A_004363455.1_EubJap_v1_BIUU/GCA_00436345 p_v1_BIUU_genomic.fna.gz
GLOBICEPHALA MELAS	long-finned pilot whale	GLOMEL	9731	x	x	x	x	3(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/006/54 _006547405.1_ASM654740v1/GCF_006547405.1_ 0v1_genomic.fna.gz
HIPPOPOTAMUS AMPHIBIUS	hippopotamus	HIPAMP	9833	x	x	x	-	-	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/02 A_004027065.2_HipAmp_v2_BIUU_UCD/GCA_004 HipAmp_v2_BIUU_UCD_genomic.fna.gz
INIA GEOFFRENSIS	boutu	INIGEO	9725	x	x	x	x	3(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/36 A_004363515.1_IniGeo_v1_BIUU/GCA_004363515 _v1_BIUU_genomic.fna.gz
LAGENORHYNCHUS OBLIQUIDENS	pacific white-sided dolphin	LAGOBL	90247	x	x	x	x	1(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/67 _003676395.1_ASM367639v1/GCF_003676395.1_ 9v1_genomic.fna.gz
LIPOTES VEXILLIFER	yangtze river dolphin	LIPVEX	118797	x	x	x	x	3(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/44 _000442215.1_Lipotes_vexillifer_v1/GCF_0004422 es_vexillifer_v1_genomic.fna.gz
MEGAPTERA NOVAEANGLIAE	humpback whale	MEG	9773	x	x	x	x	4(2)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/32 A_004329385.1_megNov1/GCA_004329385.1_meg omic.fna.gz

		NOVV							
MESOPLODON BIDENS	sowerby's beaked whale	MESBID	48745	x	x	x	-	4(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/02 A_004027085.1_MesBid_v1_BIUU/GCA_00402708 d_v1_BIUU_genomic.fna.gz
MONODON MONOCEROS	narwhal	MONMON	40151	x	x	x	x	1(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/005/19 _005190385.1_NGI_Narwhal_1/GCF_005190385.1_ whal_1_genomic.fna.gz
NEOPHOCAENA ASIAEORIENTALIS ASIAEORIENTALIS	yangtze finless porpoise	NEOASI	1706337	x	x	x	x	6	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/003/03 _003031525.1_Neophocaena_asiaeorientalis_V1/G 1525.1_Neophocaena_asiaeorientalis_V1_genomic
ORCINUS ORCA	killer whale	ORCORC	9733	x	x	x	x	3(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/33 _000331955.2_Oorc_1.1/GCF_000331955.2_Oorc_ mic.fna.gz
OVIS ARIES	sheep	OVIARI	9940	x	x	x	x	13(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/74 _002742125.1_Oar_rambouillet_v1.0/GCF_002742 _rambouillet_v1.0_genomic.fna.gz
PHOCOENA PHOCOENA	harbor porpoise	PHOPHO	9742	x	x	x	x	2	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/07 A_003071005.1_ASM307100v1/GCA_003071005.1 100v1_genomic.fna.gz
PHYSETER CATODON	sperm whale	PHYCAT	9755	x	x	x	x	4(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/83 _002837175.2_ASM283717v2/GCF_002837175.2_ 7v2_genomic.fna.gz
SUS SCROFA	pig	SUSSCR	9823	x	x	x	x	9	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/00 _000003025.6_Sscrofa11.1/GCF_000003025.6_Ss genomic.fna.gz
TRAGULUS KANCHIL	lesser mouse-deer	TRAKAN	1088131	x	x	x	x	5(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/006/40 A_006408655.1_LMD/GCA_006408655.1_LMD_ge gz
TURSIOPS ADUNCUS	indopacific bottlenose dolphin	TURADU	79784	x	x	x	x	1(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/22 A_003227395.1_ASM322739v1/GCA_003227395.1 739v1_genomic.fna.gz
TURSIOPS TRUNCATUS	common bottlenose dolphin	TUR	9739	x	x	x	x	8(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/011/76 _011762595.1_mTurTru1.mat.Y/GCF_011762595.1 1.mat.Y_genomic.fna.gz

		TRU							
VICUGNA PACOS	alpaca	VICPAC	30538	x	x	x	x	4(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/16 _000164845.3_VicPac3.1/GCF_000164845.3_VicPa omic.fna.gz
ZIPHIUS CAVIROSTRIS	cuvier's beaked whale	ZIPCAV	9760	x	x	x	-	4(1)	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/36 A_004364475.1_ZipCav_v1_BIUU/GCA_00436447 _v1_BIUU_genomic.fna.gz

1005

1006 **Supplementary Table 3: Public mitochondrial resources.**

SPECIES	COMMON	ALIAS	TAXID	URL-SEQUENCE
BALAENOPTERA ACUTOROSTRATA	minke whale	BALACU	9767	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA ACUTOROSTRATA	minke whale	BALACU	9767	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA BONAERENSIS	Antarctic minke whale	BALBON	33556	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA BOREALIS	sei whale	BALBOR	9768	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA BOREALIS	sei whale	BALBOR	9768	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA BRYDEI	bryde whale	BALBRY	255365	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA MUSCULUS	blue whale	BALMUS	9771	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA OMURAI	omura whale	BALOMU	255217	https://www.ncbi.nlm.nih.gov/
BALAENOPTERA PHYSALUS	fin whale	BALPHY	9770	https://www.ncbi.nlm.nih.gov/

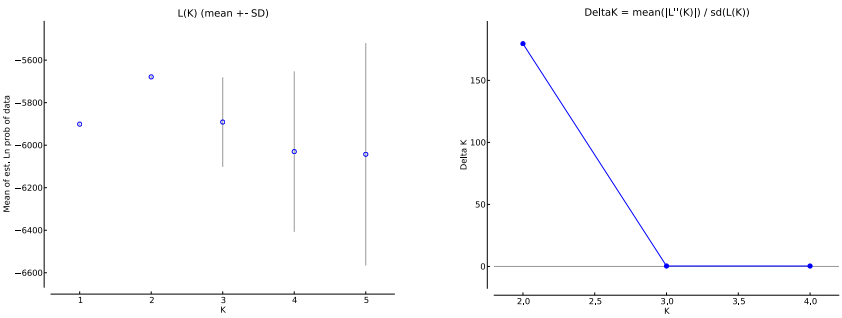
BALAENOPTERA PHYSALUS	fin whale	BALPHY	9770	https://www.ncbi.nlm.nih.gov/
BALAENA MYSTICETUS	bowhead whale	BALMYS	27602	https://www.ncbi.nlm.nih.gov/
ESCHRICHTIUS ROBUSTUS	grey whale	ESCROB	9764	https://www.ncbi.nlm.nih.gov/
ESCHRICHTIUS ROBUSTUS	grey whale	ESCROB	9764	https://www.ncbi.nlm.nih.gov/
EUBALAENA GLACIALIS	north atlantic right whale	EUBGLA	27606	https://www.ncbi.nlm.nih.gov/
LIPOTES VEXILLIFER	yangtze river dolphin	LIPVEX	118797	https://www.ncbi.nlm.nih.gov/
MEGAPTERA NOVAEANGLIAE	humpback whale	MEGNOV	9773	https://www.ncbi.nlm.nih.gov/
ORCINUS ORCA	killer whale	ORCORC	9733	https://www.ncbi.nlm.nih.gov/
PHYSETER CATODON	sperm whale	PHYCAT	9755	https://www.ncbi.nlm.nih.gov/
TURSIOPS TRUNCATUS	common bottlenose dolphin	TURTRU	9739	https://www.ncbi.nlm.nih.gov/

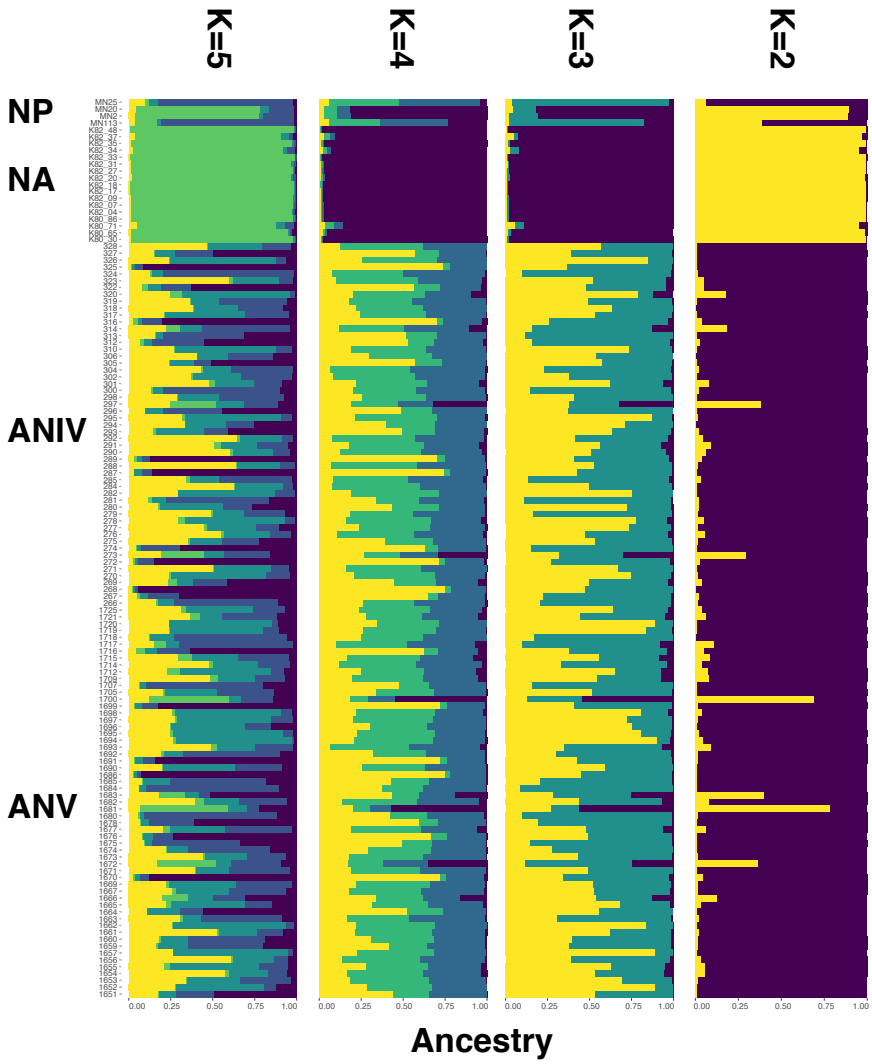
1007

1008 **Supplementary Table 4: Primers used in this study**

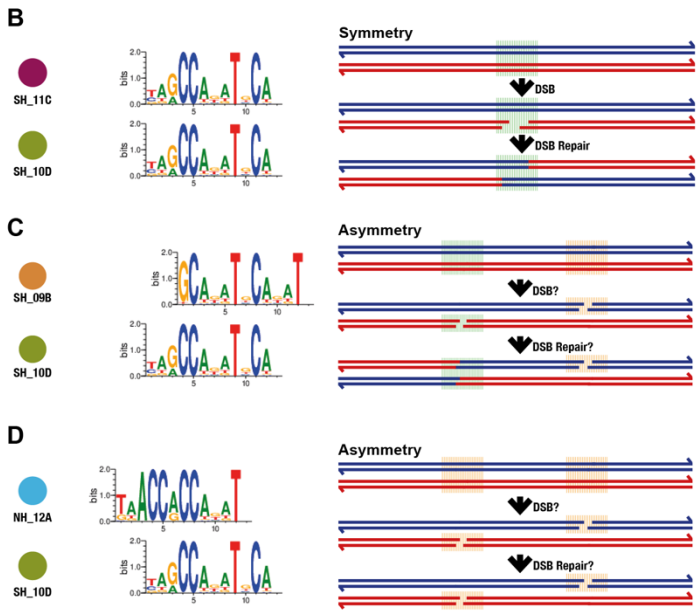
Method	Name	Sequence
Whole length Prdm9 amplification	Prdm9Bal_full_F	CCA CGG CCA AAG ATG CTT TC
PRDM9-Zinc-Finger Typing	Prdm9Bal_ZnFA_F	GAG CTG AGG TAT GTG GTC GC
	Prdm9Bal_full+ZnFA_R	ACT CAT GGG GGC AGG AGT AT
STR genotyping	EV001_F	[JOE] CCC TGC TCC CCA TTC TC
	EV001_R	ATA AAC TCT AAT ACA CTT CCT CCA AC
	EV037_F	[HEX] AGC TTG ATT TGG AAG TCA TGA
	EV037_R	TAG TAG AGC CGT GAT AAA GTG C
	GATA028_F	[6FAM] AAA GAC TGA GAT CTA TAG TTA
	GATA028_R	CGC TGA TAG ATT AGT CTA GG
	GATA098_F	[6FAM] TGT ACC CTG GAT GGA TAG ATT
	GATA098_R	TCA CCT TAT TTT GTC TGT CTG
	GATA417_F	[JOE] CTG AGA TAG CAG TTA CAT GGG
	GATA417_R	TCTGCTCAGGAAATTTTCAAG
	GT023_F	[HEX] GTT CCC AGG CTC TGCACT CTG
	GT023_R	CATTTCCTACCCACCTGTCAT
	GT211_F	[HEX] CTG CTC TAT TCT ATG AAA GCA

	GT211_R	CTCCAGTATACCTATCTTGTC
	GT310_F	[6FAM] GAA TAC TCC CAG TAG TTT CTC
	GT310_R	TAA CTT GTG GAA GAT GCC AAC
	GT509_F	[6FAM] CAG CTG CAA AAC CTT GAC ATT
	GT509_R	GTA AAA TGT TTC CAG TGC ATC
	GT575_F	[HEX] TAT AAG TGA ATA CAA AGA CCC
	GT575_R	ACC ATC AAC TGG AAG TCT TTC
Sex-Determination	ZFX_R	CAC TTA TGG GGG TAG TCC TTT
	ZFY_R	ATT ACA TGT CGT TTC AAA TCA
	ZFYX_F	[6FAM] ATA GGT CTG CAG ACT CTT CTA
Mitochondrial D-loop Sequencing	BP15851(M13F)	GTAAAACGACGGCCAGTGAAGAAGTATTACACTCCACCAT
	MN312(M13R)	CAGGAAACAGCTATGACCCGTGATCTAATGGAGCGGCCA
	MT3(M13Rev)	CAGGAAACAGCTATGACCCATCTAGACATTTTCAGTG
	MT4(M13F)	GTAAAACGACGGCCAGTCCTCCCTAAGACTCAAGGAAG

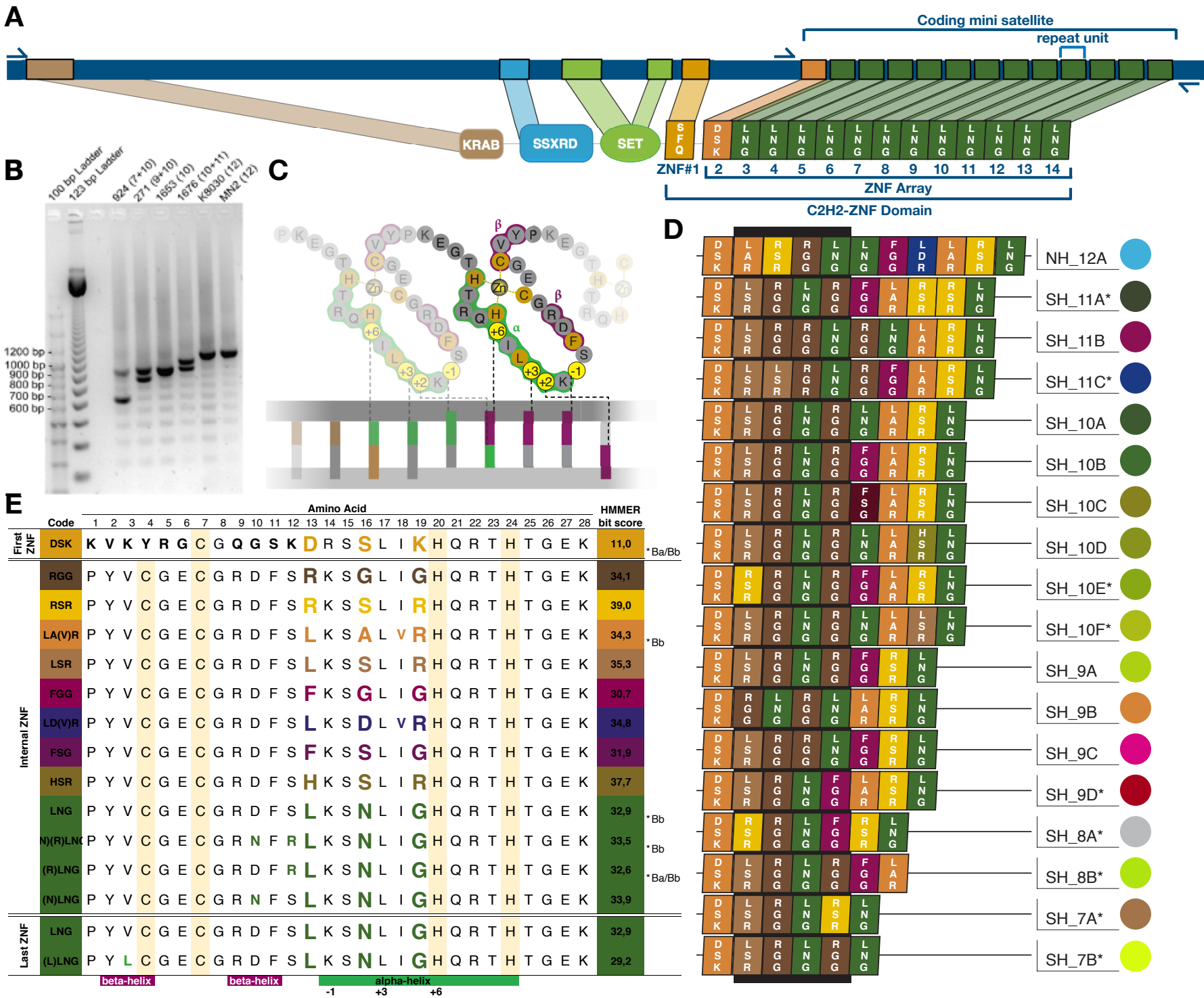


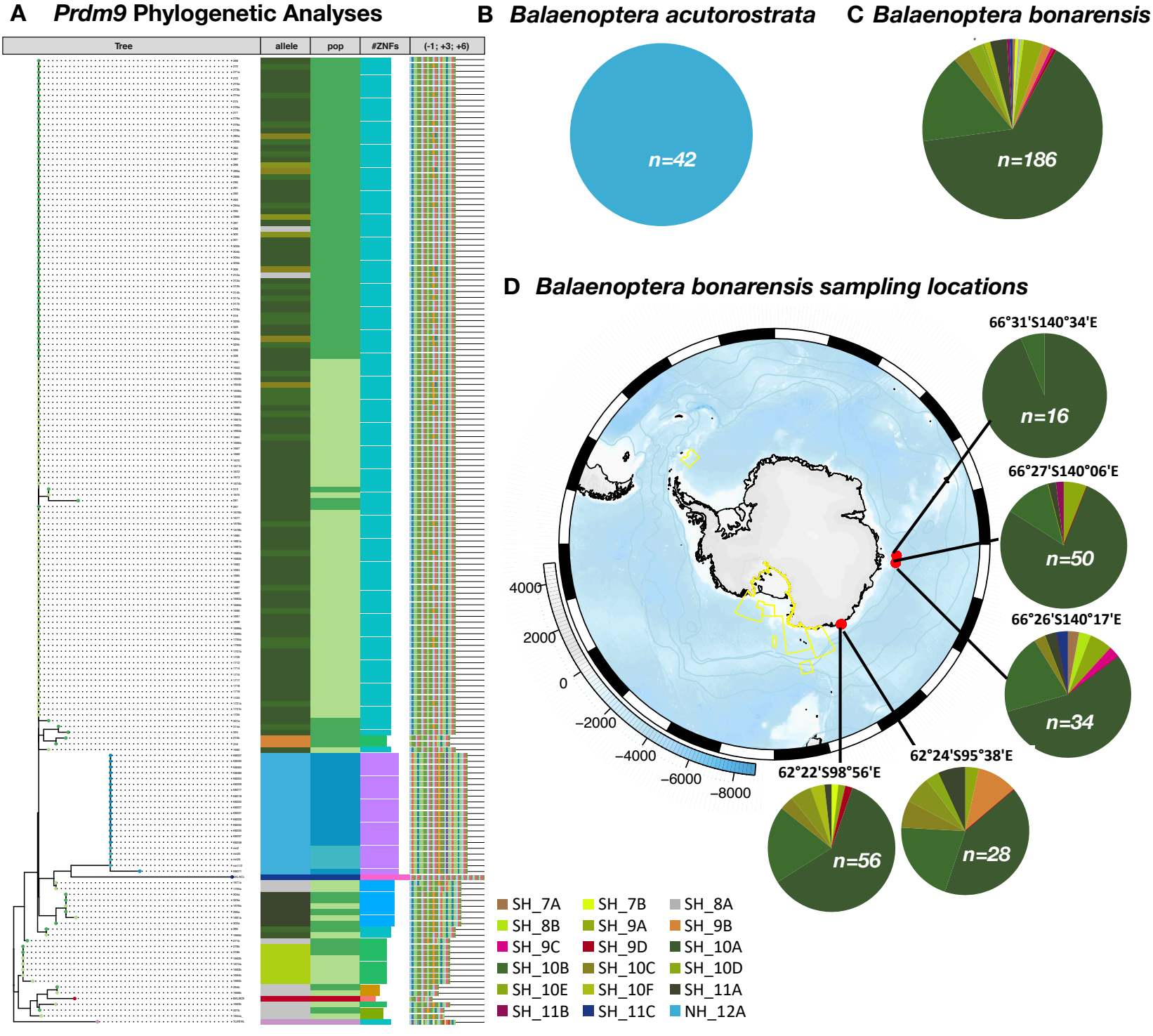


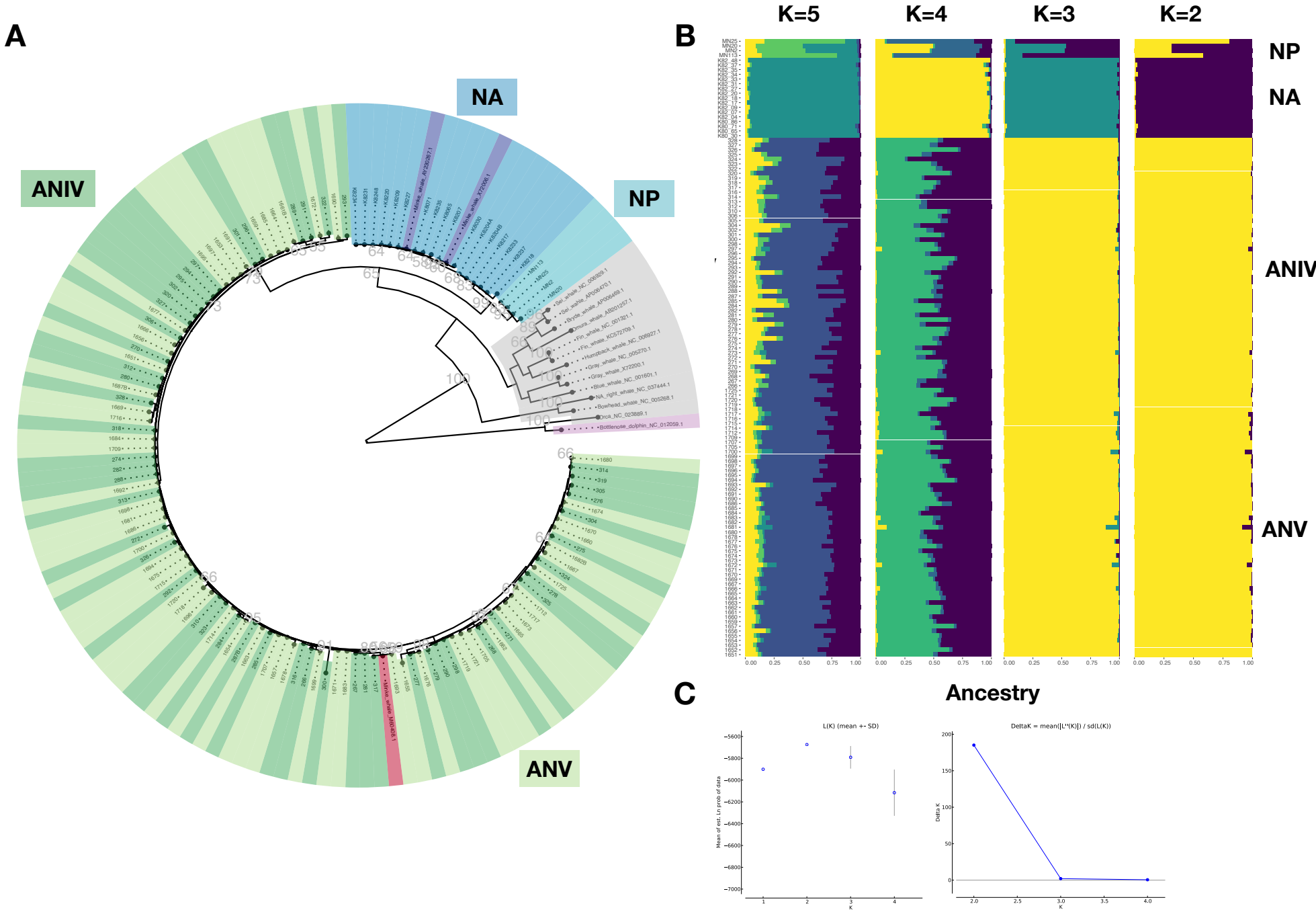
Supplementary Figure 6 Population structure analyses on a set of ten hypervariable microsatellite loci using STRUCTURE without *a priori* location information (NOLOCS)



Supplementary Figure 7 Principle of PRDM9 binding symmetry and asymmetry (A) Identical "core motif" combinations are predicted to result in fully symmetric binding and efficient DSB formation (BERG *et al.* 2010; BERG *et al.* 2011) (B) variants with somewhat dissimilar "core motifs", are predicted to result in some degree of asymmetry. In humans a similar degree of motif-match still allowed DSBs necessary for successful recombination (BERG *et al.* 2010; BERG *et al.* 2011; ODENTHAL-HESSE *et al.* 2014) (C) hypothetical combination of the most common variant of minke whale species, of two hemispheres, generating a putative interspecies hybrid combination. Here we would predict asymmetric positioning of recombination initiation sites, which is implicated in F_1-hybrid male sterility in mammals (FOREJT 2016).


A
Coding mini satellite
repeat unit
KRAB
SSXRD
SET
ZNF#1
ZNF Array
C2H2-ZNF Domain
B
100 bp Ladder
123 bp Ladder
624 (7+10)
271 (9+10)
1653 (10)
1676 (10+11)
K9030 (12)
MN2 (12)
1200 bp
1000 bp
900 bp
800 bp
700 bp
600 bp
C
D
NH_12A
SH_11A*
SH_11B
SH_11C*
SH_10A
SH_10B
SH_10C
SH_10D
SH_10E*
SH_10F*
SH_9A
SH_9B
SH_9C
SH_9D*
SH_8A*
SH_8B*
SH_7A*
SH_7B*
E
Amino Acid
Code
HMMER bit score
First ZNF
Internal ZNF
Last ZNF
DSK K V K Y R G C G Q G S K D R S S L I K H Q R T H T G E K 11,0 *Ba/Bb
RGG P Y V C G E C G R D F S R K S G L I G H Q R T H T G E K 34,1
RSR P Y V C G E C G R D F S R K S S L I R H Q R T H T G E K 39,0
LA(V)R P Y V C G E C G R D F S L K S A L v R H Q R T H T G E K 34,3 *Bb
LSR P Y V C G E C G R D F S L K S S L I R H Q R T H T G E K 35,3
FGG P Y V C G E C G R D F S F K S G L I G H Q R T H T G E K 30,7
LD(V)R P Y V C G E C G R D F S L K S D L v R H Q R T H T G E K 34,8
FSG P Y V C G E C G R D F S F K S S L I G H Q R T H T G E K 31,9
HSR P Y V C G E C G R D F S H K S S L I R H Q R T H T G E K 37,7
LNG P Y V C G E C G R D F S L K S N L I G H Q R T H T G E K 32,9 *Bb
N)(R)LNG P Y V C G E C G R N F R L K S N L I G H Q R T H T G E K 33,5 *Bb
(R)LNG P Y V C G E C G R D F R L K S N L I G H Q R T H T G E K 32,6 *Ba/Bb
(N)LNG P Y V C G E C G R N F S L K S N L I G H Q R T H T G E K 33,9
LNG P Y V C G E C G R D F S L K S N L I G H Q R T H T G E K 32,9
(L)LNG P Y L C G E C G R D F S L K S N L I G H Q R T H T G E K 29,2
beta-helix
beta-helix
alpha-helix
-1
+3
+6


A Prdm9 Phylogenetic Analyses
Tree
allele
pop
#ZNFs
(-1; +3; +6)
B Balaenoptera acutorostrata
n=42
C Balaenoptera bonarensis
n=186
D Balaenoptera bonarensis sampling locations
66°31'S140°34'E
n=16
66°27'S140°06'E
n=50
66°26'S140°17'E
n=34
62°24'S95°38'E
n=28
62°22'S98°56'E
n=56
4000
2000
0
−2000
−4000
−6000
−8000
SH_7A
SH_7B
SH_8A
SH_8B
SH_9A
SH_9B
SH_9C
SH_9D
SH_10A
SH_10B
SH_10C
SH_10D
SH_10E
SH_10F
SH_11A
SH_11B
SH_11C
NH_12A


A
NA
NP
ANIV
ANV
B
K=5
K=4
K=3
K=2
NP
NA
ANIV
ANV
Ancestry
C
L(K) (mean +- SD)
Mean of est. Ln prob of data
DeltaK = mean(|L''(K)|) / sd(L(K))
Delta K
K