Evolution of the recombination regulator PRDM9 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, August-Thienemann Str. 2, Plön, D-24306, Germany

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

[#]corresponding author: email: odenthalhesse@evolbio.mpg.de

[§]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*, postzygotic reproductive isolation, meiotic recombination regulation

Abstract

We explored the structure and variability of the *Prdm9* gene, which codes for the PRDM9 protein, in samples of the minke whales from the Atlantic, Pacific and Southern Oceans. The PRDM9 protein controls the reshuffling of parental genomes in most metazoans and we show that minke whale possess all the features characteristic of PRDM9-directed recombination initiation, including complete KRAB, SSXRD and SET domains and a rapidly evolving array of C2H2-type-Zincfingers (ZnF). We uncovered eighteen novel PRDM9 variants and evidence of rapid evolution, particularly at DNA-recognizing positions that evolve under positive selection. At different geographical scales, we observed extensive *Prdm9* allelic 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 [1, 2]; *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 H3K36me3. The fourth, C-terminal, domain comprises an array of type Cystin$_2$Histidin$_2$ zinc-fingers (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 [3, 4], mice [5, 6], non-human primates [7] and other mammals [8, 9]. Even highly domesticated species like equids [8], bovids [10] and ruminants [11, 12] 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 [13].

Although overlapping habitats exist near the Equator, seasonal differences in migration and breeding behavior largely prevent inter-breeding between Common and Antarctic minke whales [14]. Despite this, occasional migration across the Equator has been observed [15] 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 [16, 17]. Thus, hybrids may be both viable and fertile, allowing backcrossing [17]. 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 [17]. More importantly, since both hybrids were female, current data do not exclude postzygotic reproductive isolation through males, as expected under Haldane's rule. Hybrid sterility is a universal phenomenon observed in many eukaryotic inter-species hybrids, including yeast, plants, insects, birds, and mammals [18, 19]. Minke whales offer an unusual opportunity to study *Prdm9*, the only hybrid sterility gene that is known in vertebrates thus far, sometimes called a speciation gene [20].

Material and Methods

PRDM9 occurrence and protein domain prediction in diverse taxa

To infer PRDM9 occurrence and to subsequent investigate protein domain architecture in diverse taxa, the genome sequences were downloaded as given in **Supplementary Table 2**. To infer PRDM9 occurrence in the given species, the annotated protein XP_028019884.1 from *Balaenoptera acutorostrata scammoni* was used as the query protein with exonerate [21] (v2.2.0) and the --protein2genome model

99 to extract the best hit. Subsequently, for each species, the extracted coding sequences
100 (CDS) were translated and investigated with InterProScan [22] (v5.46-81.0) and
101 HMMER3 [23] (v3.3) using KRAB, SSXRD, SET and C_2H_2-ZnF as bait to obtain the
102 protein domain architecture.
103
104 **Phylogenetic Analyses**
105 For the phylogenetic reconstruction across Cetartiodactyla, we used only KRAB,
106 SSXRD and SET domains (Supplementary Figure 1). These protein domains were
107 concatenated and used as input for the software BAli-phy [24] (v3.5.0). BAli-phy was
108 run twice with 10,000 iterations each and the default settings for amino acid input.
109 Subsequently, the majority consensus tree was obtained by skipping the first 10% of
110 trees as burn-in, rooted on the branch outside the Cetacea and visualized in Figtree
111 (http://tree.bio.ed.ac.uk/software/figtree/; v1.4.4).
112
113 The phylogenetic tree from the mitochondrial D-loop region was reconstructed,
114 including several whale species as given in Supplementary Table 3. The
115 corresponding D-loop was aligned with MAFFT version 5 [25](v7.471) with the L-INS-
116 i algorithm and manually curated. The maximum-likelihood tree was calculated under
117 the TN+F+I+G4 model using IQ-TREE [26] (v1.6.12) mid-point rooted and visualised
118 in Figtree (http://tree.bio.ed.ac.uk/software/figtree/; v1.4.4).
119
120 **Long-range sequencing of the Prdm9 gene**
121 We first extracted the full-length DNA sequence of the PRDM9 Protein from the minke
122 whale genome from UCSC genome browser and designed primers to flank the entire
123 protein-coding portion, encompassing the KRAB, SSXDR, PR/SET and C_2H_2-ZnF

124 binding domains using Primer-BLAST [27](Supplementary Table 4). We amplified the
125 *Balaenoptera acutorostrata Prdm9* gene, including all introns and exons across a
126 ~11kb interval by long-range PCR. We then sequenced the entire interval using
127 phased long-range Nanopore Sequencing with MinION (Oxford Nanopore). Whole-
128 length consensus sequences generated from 639 sequencing reads that cover the
129 entire 10582 bp, with an average per base pair coverage of 269x. We used this
130 consensus sequence as reference for *in-silico* predictions of functional domains and
131 mapped human PRDM9 Exons 3 – 11 from ENSEMBL (ENST00000296682.3). By
132 manually splicing all unaligned sequence fragments, we generated an in-silico
133 predicted mRNA of *Balaenoptera acutorostrata* PRDM9. This sequence was then
134 submitted as '.fasta' to the Entrez Conserved Domains Database (CDD) home page,
135 (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) with the following parameters;
136 searching against Database CDD v3.18 – 52910 PSSMs Expect Value threshold: 0,01,
137 without low-complexity filter, composition-based statistics adjustment, rescuing
138 borderline hits (ON) a maximum number of 500 hits and concise result mode).
139

140 **Wild minke whale samples**

141 Performed investigations are not considered to be animal trials in accordance with the
142 German animal welfare act, since samples were obtained from commercial whaling
143 between 1980 and 1984. A total of n=143 DNA samples of minke whale with vague
144 information about the subspecies, from four different defined commercial whaling
145 areas. These include North Atlantic (NA; n=17) and North Pacific (NP; n=4) individuals
146 captured during migration season, as well as Antarctic Ocean Areas IV (AN IV; n=65)
147 and V (AN V; n=57) were obtained during feeding season. A subset of these samples
148 had been used in van Pijlen et al., 1995 [28]. Samples from the Antarctic areas

149 included numerous duplicate samples that were used as internal controls, analyzed for
150 all parameters and excluded from the sample set once identical results had been
151 confirmed (318, 325, 327, 1661, 1663).
152
153 For the detailed DNA-extraction protocol, precise catch data, as well as providers of
154 the samples of most individuals used in this project, see van Pijlen, 1994. Briefly,
155 genomic DNA was extracted from the skin (NA) and muscle biopsies (NP, AN) with
156 Phenol-Chlorophorm and stored in TE at - 80°C in the 1980s. If the DNA was dried out
157 it was first eluted with TE. All DNA concentrations were determined with fluorescent
158 Nanodrop-1000 (Thermo Fisher Scientific). For all performed analysis, 20 ng/µl DNA
159 working stocks were prepared and stored at -20°C.
160
161 **STR genotyping**
162 Nine autosomal, as well as X/Y microsatellite loci with di- or tetramer repeat motifs,
163 were analysed for all samples: EV001, EV037 [29], GATA028, GATA098, GATA417
164 [30], GT023, GT310, GT509, GT575 [31] and sex loci X and Y [32]. Four separate
165 multiplexing reactions were performed for each individual, and each contained 40 ng
166 of DNA, 0.2 µM of each primer, 5 mM Multiplex-Kit (Qiagen) and HPLC water to a total
167 volume of 10 µL per sample. Primers (**Supplementary Table 4**) were purchased from
168 Sigma Aldrich; the reverse Primers were tagged at their 5' end with fluorescent tags
169 (HEX, FAM or JOE). The first multiplexing reaction contained primers GT023 (HEX),
170 EV037 (HEX) (FAM) in forward and reverse direction. The multiplexing conditions were
171 denaturation at 95°C for 15:30 minutes, annealing 1:30 minutes at 59°C, elongation at
172 72°C for 11:30 minutes and held at 12°C. The second multiplexing reaction contained
173 primers GT575 (HEX), GATA028 (FAM) in forward and reverse direction, the

174 multiplexing conditions were denaturation at 95°C for 15:30 minutes, annealing 1:30
175 minutes at 54°C, elongation at 72°C for 11:30 minutes and a final temperature of 12°C.
176 The third multiplexing reaction was set up with primers GATA098 (FAM), GT509 (FAM)
177 and GATA417 (JOE) in forward and reverse direction. The multiplexing conditions
178 were denaturation at 95°C for 15:30 minutes, annealing 1:30 minutes at 54°C,
179 elongation at 72°C for 11:30 minutes and a final temperature of 12°C. The fourth
180 multiplexing reaction contained primers GT310 (HEX) and EV001 (JOE) in forward and
181 reverse direction. The multiplexing conditions were denaturation at 95°C for 15:30
182 minutes, annealing 1:30 minutes at 60°C, elongation at 72°C for 11:30 minutes and a
183 final temperature of 12°C. After amplification, the multiplexing reactions were diluted
184 with 100 µl water (HPLC grade). Genescan ROX^{500} dye size standard (Thermo Fisher
185 Scientific) and HiDi Formamide (Thermo Fisher Scientific) were mixed 1:100 and used
186 for sequencing in a mix of 1 µl of diluted multiplexing product and 10 µl HiDi-ROX^{500}.
187 STR Multiplex fragment analysis was carried out on the 16-capillary electrophoresis
188 system ABI 3300 Genetic Analyser (Applied Biosystems).
189
190 **STR analysis:**
191 We first performed MSA analysis [33] using standard parameters, which calculated
192 Weinberg expectation (Fis), Shannon Index (Hs), allele numbers (A) and allele sizes.
193 To detect both weak and strong population structure, we ran simulations with
194 LOCPRIOR and USEPOPINFO, respectively. For both simulations, the more
195 conservative "correlated allele frequencies" -model was used, which assumes a level
196 of non-independence. To ensure that a sufficient number of steps and runs have been
197 performed we used a burn-in period of 1.000 and runs of 100.000 MCMC repeats for
198 both types of simulations, each for 50 iterations for successive K values from 1 - 10.

199 [34]. We used the *a-priori* location to be able to detect even weak population
200 differentiation. In both datasets, we used the web-based STRUCTURE Harvester
201 software [35] to determine the rate of change in the log probability between successive
202 K values, via the *ad-hoc* statistic ΔK from [36]. Figures were rendered using
203 STRUCTURE PLOT V2.0. [37]
204
205 **Sequencing of the Mitochondrial Hypervariable Region on the D-loop**
206 The noncoding mtDNA-D-Loop region of 143 individuals was amplified in two
207 overlapping PCR-reactions. We PCR amplified two different lengths fragments for
208 each individual: 1066 bp and 331 bp and sequenced the longer product in forward and
209 the shorter in the reverse direction. The used primers (found in **Supplementary Table**
210 **4**) were MT4 (M13F) and MT3 (M13R) for the longer product, and BP15851 (M13F)
211 and MN312 (M13R) for the shorter PCR product from [38]. The 10 µL reaction was
212 carried out with 40 ng genomic DNA and 0.2 µM of each primer and 5 mM Multiplex
213 PCR Kit (Qiagen). Cycling conditions were identical for both directions with 95°C 15:30
214 min, 53°C 1:30 min, 72°C 13:30 min and hold at 4°C in a Veriti Thermal Cycler (Applied
215 Biosystems). The PCR products were purified with 3 µl Exo/SAP and then cycle-
216 sequenced with the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher
217 Scientific) according to manufacturer instructions with BP15851 (M13F) for the forward
218 PCR product and MN312 (M13R) for the reverse PCR product, respectively. The mixes
219 were then purified with BigDye X-Terminator™ Purification Kit (Thermo Fisher
220 Scientific) and sequenced by capillary electrophoresis on an ABI 16-capillary 3130xl
221 Analyzer (Applied Biosystems). The sequences were *de-novo* assembled, and
222 consensus sequences were generated with Geneious Software 10.2.3 [39].
223

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

225 To characterise the minisatellite-coding for the C_2H_2-ZnF array in more detail, we
226 designed primers nested between the first two conserved ZnFs, and the reverse primer
227 identical to the long-range amplification primer distal to the coding sequence for the
228 C_2H_2-ZnF array. The minisatellite coding for the C_2H_2-ZnF array of PRDM9 of 143
229 *Balaenoptera* individuals was amplified from 20 ng genomic DNA, in a 20 µl PCR
230 reaction, optimised for the amplification of minisatellites. With 0.5 mM primers that
231 were designed for this study using the *Balaenoptera. acutorostrata scammoni*
232 *(*XM_007172595) reference and included the PRDM9 minisatellite-like C_2H_2-ZnF array
233 and additional 100bp flanking regions at 5' and 3'. Primers Prdm9ZnFA_Bal_R: and
234 Prdm9ZnFA_Bal_F (**Supplementary Table 4**) and 1x AJ-PCR Buffer described in [40]
235 and 0.025 U/µl Taq-Polymerase and 0.0033 U/µl Pfu-Polymerase. Cycling conditions
236 were: initial denaturation at 95°C for 1:30 min followed by 33 cycles including 96°C 15
237 s, 61°C 20 s and 70 °C 2:00 min and finally 70°C 5 min, and hold at 4°C in a Veriti
238 Thermal Cycler (Applied Biosystems). Agarose gel electrophoresis (1.5% Top Vision
239 Low Melting Point Agarose gel (Thermo Fisher Scientific) with SYBR Safe (Thermo
240 Fisher Scientific) was used to visualise allele sizes as well as zygosity. All bands were
241 excised from the gel (Molecular Imager® Gel Doc™ XR System with Xcita-Blue™
242 Conversion Screen (Biorad), and PCR products were recovered from the gel fragments
243 with 2 U/100 mg Agarase. If the individuals were homozygous, the extracted DNA was
244 directly Sanger sequenced from in 5' and 3' directions for each sample with the
245 BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific) according
246 to the manufacturer's protocol with the same primers as for the amplification
247 (Prdm9ZnFA_Bal_R/Prdm9ZnFA_Bal_F). The sequencing-reaction was carried out in

248 the ABI 16-capillary 3130xl Analyzer (Applied Biosystems). Heterozygous samples
249 were subcloned before sequencing.
250
251 **Subcloning of heterozygous alleles of identical lengths**
252 Subcloning was performed for a subset of samples when Sanger sequencing revealed
253 heterozygous alleles of the same length, that could not be distinguished by
254 electrophoresis, but revealed heterozygous nucleotides in the chromatogram. Instead
255 of inferring haplotypes, we wanted to experimentally ascertain both alleles. Thus, we
256 cloned the remaining PCR product into the TOPO TA (Invitrogen) vector and
257 transferred the vectors into OneShotTop10 chemically competent cells (Invitrogen). All
258 steps were carried out according to the manufacturer's manuals. Eight positive clones
259 were picked for each sample, and the DNA was extracted by boiling in HPLC-grade
260 water at 96°C for 10 Minutes. A centrifugation step then removed the cell debris, and
261 the supernatant was directly used for PCR, gel-purification and Sanger-sequencing as
262 described in the section above.
263
264 **Sequence Analysis of the PRDM9 C_2H_2-ZnF array**
265 Sequences determined different alleles and numbers of repeat units per array, which
266 were *de novo* annotated and the resulting consensus sequences were translated into
267 the corresponding protein variants of each C_2H_2-ZnF by Geneious Software 10.2.3
268 [39]. All PCR products contained the complete minisatellite-like coding sequence for
269 the C_2H_2 domain and additional 100 base pairs flanking regions at the 5' and 3' ends.
270
271 **PRDM9 C_2H_2-ZnF array coding sequence dN/dS analysis**

272 The internal C_2H_2-ZnFs were obtained as described for the phylogenetic
273 reconstruction. Subsequently, all repeat units were stacked using only non-unique
274 alleles. The resulting codon alignment was used to determine basic population
275 parameters for each population (population theta using segregating sites and
276 nucleotide diversity) by either including or excluding hypervariable amino acid positions
277 (-1, +3, +6) with the R package "pegas" [41]. Episodic diversifying selection among
278 Zinc-finger sites using was determined by a mixed-effects model of evolution (MEME)
279 at https://www.datamonkey.org as described earlier [42].
280

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

282 The gene tree of PRDM9 was inferred either using only the repeats coding for the
283 internal C_2H_2-ZnFs or the concatenated KRAB, SSXRD and SET domain. For the
284 internal C_2H_2-ZnFs based phylogenetic reconstruction (**Figure 1**), internal C_2H_2-ZnFs
285 were extracted, and pairwise distances were calculated as in Vara et al. 2019 [43] with
286 the developmental R package repeatR (https://gitlab.gwdg.de/mpievolbio-it/repeatr). In
287 brief, only species were used with one leading C_2H_2-ZnF and at least eight internal
288 C_2H_2-ZnFs (see supplementary Figure 1). For each possible repeat combination (r, r')
289 the hamming distances of the corresponding repeat units $r = (r_1; r_2; r_3; \ldots)$ and $r' = (r'_1;$
290 $r'_2; r'_3; \ldots)$ were used to derive the edit distance between r and r'. Before calculating
291 the edit distance, the codons coding for the hypervariable amino acid positions (-1, +3,
292 +6) were removed for each repeat unit and weighting cost of $w_{mut} = 1$, $w_{indel} = 3.5$ and
293 $w_{slippage} = 1.75$ as given in [43]. The resulting pairwise edit distance matrix was used to
294 calculate a neighbour-joining tree with the *bionj* function of the R package ape [44],
295 rooted on the branch leading to the bottlenose dolphin (*Tursiops truncatus)* and
296 visualized in Figtree (http://tree.bio.ed.ac.uk/software/figtree/ v1.4.4).

297

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

299 DNA-binding Specificities of the different Cys_2His_2 Zinc Finger Protein variants were
300 predict *in-silico* using the SVM polynomial kernel method within "Princeton ZnF"
301 (http://zf.princeton.edu/) [45].

302

303 **Results**

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

305 For a broad view on the evolution of PRDM9 in *Cetartiodactyla*, we identified PRDM9
306 orthologs from all publicly available genomic resources (
307 **Supplementary Table 2**). To first establish an evolutionary context based on the
308 protein-coding regions for the proximal conserved domains, we used the *Balaenoptera*
309 *acutorostrata scammoni* protein as a query and extracted coding sequences of the N-
310 terminal (conserved) domains. Previous studies of PRDM9 in vertebrates found rapid
311 evolution of DNA binding sites only when the KRAB and SSXRD domains were present
312 [46]. Consequently, we used InterProScan to create a curated dataset of
313 *Cetartiodactyla* PRDM9 orthologues which contained the KRAB, SSXRD, and SET
314 domains, including a PRDM9 orthologue in Antarctic minke whale. Our initial analyses
315 revealed that minke whales, and *Cetartiodactyla* in general, possessed all the features
316 characteristic of organisms with PRDM9-directed recombination initiation, including
317 complete KRAB, SSXRD and SET domains. However, organisms with PRDM9
318 directed recombination, also display a rapidly-evolving PRDM9 C_2H_2-ZnF array [46]
319 With the sole exception of Hippopotamus, all *Cetartiodactyla* also contained the C_2H_2-
320 ZnF array, with large variation in the number of C_2H_2-ZnF between species.

321

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

323 To characterize the sequence and structure of the *Prdm9* gene, we began by
324 amplifying and sequencing it from a pooled sample, containing DNA from six
325 individuals, five Antarctic minke whales and one common minke whale. The consensus
326 sequence was subjected to *in-silico* prediction that successfully recovered all relevant
327 PRDM9 protein domains with high-confidence: the KRAB domain (E-value: $1.84e^{-11}$);
328 the SSXRD motif (E-value: $4.46e^{-10}$); the PR/SET domain (E-value: $9.69e^{-05}$); and
329 several Zinc-Fingers (**Figure 1**). The complete C_2H_2-domain comprises an array of
330 C_2H_2-ZnFs (the C_2H_2-ZnF-array), as well as a single C_2H_2-ZnF (ZnF#1) that is located
331 proximally (E-value: $6.52e^{-04}$). This PRDM9-ZnF#1 is identical to that of mice [47] as
332 well as rats, elephants, human, chimpanzee, macaque and orangutan [48], and thus
333 appears to be conserved across large evolutionary timescales. The second C_2H_2-ZnF
334 (E-value: $1.65e^{-11}$) appears conserved across *Balaenoptera*. The same second C_2H_2-
335 ZnF is seen in our pooled sample, as well as the genome reference of the North Pacific
336 minke whale (*Balaenoptera acutorostrata scammoni)* and Antarctic minke whale
337 (*Balaenoptera bonarensis)* as well as the blue whale (*Balaenoptera musculus).*
338 However, from the third ZnF onwards, sequencing across pooled samples exhibited
339 high variability, which we could not resolve.

340

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

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

347 the *Prdm9* gene. A single size was observed in all Common minke whales from
348 sampled populations in the North Atlantic as well as North Pacific. In contrast, five
349 different sizes were identified across Antarctic minke whale populations sampled in the
350 Southern Hemisphere. All common minke whales have a size consistent with eleven
351 84 bp repeats while the most prevalent genotype in the Antarctic minke whales
352 corresponds to nine repeats, with variation between six and ten repeats. Therefore,
353 *Prdm9* shows size variation length of the coding minisatellite, that result from variation
354 in the number of repeat units.
355

356 Sequencing of the DNA purified from gel fragments, as well as cloning of apparently
357 homozygous bands, revealed a remarkable diversity of minisatellite alleles. Over 80%
358 of animals possessed repeat-arrays of nine repeats; however, these comprised six
359 different alleles. Because in addition to variation in repeat-number, there is nucleotide
360 variation between 84bp-minisatellite repeats. And finally, not only which repeat-types
361 are present, but also their order varies between individuals.
362

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

364 Each 84bp repeat unit translates into 28 amino-acids coding for a single C_2H_2-ZnF. To
365 explore diversity between the ZnF, we extracted all 84bp repeat units, which we used
366 to predict C_2H_2-ZnFs through an HMMER algorithm [45]. This yielded 26 different
367 versions with HMMER bit scores >17 (**Figure 1** and **Supplementary Figure 4**). Based
368 on amino acid variation within each predicted C_2H_2-ZnFs, we identified 14 "common"
369 variants found in multiple individuals as well as twelve "rare" variants found in only a
370 single individual (**Supplementary Figure 4**). The most variable amino acids are 13,
371 16 and 19, which correspond to positions -1, +3 and +6 of the alpha-helices that are

372 responsible for DNA binding specificity. However, five of the fourteen "common"
373 variants have the same amino sequence LNG at these positions, differing instead at
374 three positions in beta-helices flanking the cysteine residues that bind the zinc-ion
375 (positions 3, 10 and 12). Three of the five LNG variants are also found in the reference
376 sequences of both the Antarctic minke whale and the North Atlantic minke whale. We
377 next used the fourteen common C_2H_2-ZnFs to test for signals of selection, finding
378 Evidence that amino acid 16, is under episodic diversifying selection (**Supplementary**
379 **Figure 5**). This amino acid is located at position +3 of the alpha helix, one of the three
380 positions responsible for DNA-binding specificity.
381

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

383 Using all minisatellite-coding sequences for the array, we determined C_2H_2-ZNF
384 domain variants that each individual possessed. Minisatellite size homoplasy equates
385 to an identical PRDM9 C_2H_2-ZnF array in Common minke whales, but not Antarctic
386 minke whales. The only C_2H_2 domain variant found in all samples from the Northern
387 Hemisphere NH12_A consists of eleven C_2H_2-ZnF in the array and the C_2H_2-ZnF, that
388 is located proximally and appears conserved across *Balaenoptera*. In contrast,
389 samples from the Southern Hemisphere displayed seventeen different C_2H_2-Domains
390 of PRDM9 (**Figure 1**).
391

392 We want to understand the evolutionary relationships of the different alleles that we
393 observed in minke whale species. However, phylogenetic analysis of minisatellite
394 repeat-structures is challenging due to their rapid rate of evolution. Minisatellites,
395 including that of *Prdm9*, evolve mainly by unequal crossing-over and gene conversion
396 in meiosis [49]. Stepwise mutation models that are commonly assumed are based on

397 microsatellites, and thus give only a poor fit to minisatellite evolution [50]. So, instead
398 of comparing full-length-*Prdm9* allelic variants, we separated each array into 84bp
399 repeat types. We then calculated pairwise Hamming (i.e. minimum edit) distances
400 between all repeats, after removing the nucleotides coding for the hypervariable amino
401 acid positions (-1, +3, +6), and applying specific weighting costs as given in [43]. In
402 addition to the repeats present in our minke whale samples, we also included repeats
403 coding for the internal C_2H_2-ZnFs of all *Cetartiodactyla* species into our phylogenetic
404 analyses (see **Supplementary Figure 1**) Here, we included only those species that
405 possessed the complete N-terminal domain architecture, the proximal C_2H_2-ZnF and
406 at least eight successive C_2H_2-ZnF in the C-terminal DNA-binding domain. We
407 restricted our phylogenetic analyses on the internal minisatellite repeats since our
408 analyses revealed that the DNA-binding positions of the most proximal ZnF in the array
409 (ZnF#2) are not only conserved in minke whales but also generally within *Cetacea*
410 (**Supplementary Figure 2**). In fact, we observed only a single DNA-binding amino-
411 acid change between C_2H_2-ZnF #2 of all *Cetacea* compared to all *Ruminantia*
412 (**Supplementary Figure 3**).
413
414 The PRDM9 phylogenetic analyses in **Figure 2** show that all species cluster into
415 distinct phylogenetic groups. Reference alleles cluster with their own subspecies, even
416 including the unusual reference allele of *Balaenoptera acutorostrata scammoni*, which
417 nevertheless clusters with our common minke whale samples. Similarly, the genome
418 reference allele of *Balaenoptera bonarensis* (Antarctic minke whale) clusters within the
419 group of SH_8A and SH-8B alleles found in our Antarctic minke whale samples, which
420 appear to be more ancestral than other minke whale alleles. The phylogenetic
421 reconstruction also suggests that common minke whale alleles have appeared more

422 recently and evolved mainly by an increase in repeat-copy after splitting from the
423 Antarctic minke whale. Our phylogenetic analysis thus fits well with an evolutionary
424 history where common minke whales split from Antarctic minke whales approximately
425 5 million years ago.
426
427 **Diversity of PRDM9 at different geographical scales**
428 To explore the allelic diversity of *Prdm9* at different geographical scales, we partitioned
429 the data by sampling location wherever accurate catch-locations were available
430 (**Figure 2**). Allelic diversity varied between sampling locations, even within the
431 Antarctic sampling sites. The most common alleles are SH_10A, with SH_10B, and
432 these occur at frequencies of 10-20% in all sampling locations. However, all other
433 alleles differ between different sampling populations, each of which carries at least one
434 allele found exclusively in a single population. Sample size does not predict allele
435 number, with the highest allele number (ten) being associated with an average-sized
436 sample (n=34).
437
438 **Population structure of minke whales**
439 We are interested in the possible evolutionary consequences of PRDM9 for minke
440 whale speciation in the face of what may be recent secondary mixing, and this requires
441 a context of current levels of population isolation. We thus explored variability at the
442 84bp *Prdm9* minisatellite repeat across the four sample regions for population genetic
443 analyses (**Table 1**). Taking each repeat type as an individual allele, we computed the
444 mutation rate per base pair per generation (population θ), as well as the average
445 pairwise nucleotide diversity. The latter was calculated in two ways: (i) using the entire
446 minisatellite-like repeat sequence; (ii) after removing nucleotides coding for the

447 hypervariable sites that translate into the DNA binding positions of individual C_2H_2-
448 ZnF, which are known to be under positive selection. We observed the highest
449 population θ in Antarctic Area IV, compared to all other sampling locations, as seen in
450 **Table 1**. Across the microsatellite-repeat types, the Average Pairwise Nucleotide
451 Diversity (APND, including nucleotides coding for the hypervariable sites) is similar
452 across all sampled populations and between common minke whale and Antarctic
453 minke whales. After excluding the nucleotides coding for hypervariable sites (-1, +3,
454 +6), the APND is decreased roughly 2.5-fold, across minke whales.
455
456 We used two approaches to quantify genetic differentiation of PRDM9 between the
457 different minke whale populations: the per-site distance for multiple alleles [51] and
458 Jost's D (D_J), the fraction of allelic variation among populations [52]. Both measures
459 indicate weak population differentiation, the highest differentiation being between the
460 North Atlantic and North Pacific (G_{ST} = –0.0039, D_J = 0.0077). The second highest
461 population differentiation was seen between the North Pacific and both Antarctic
462 regions, G_{ST} = –0.0035, D_J = 0.0070, with slightly lower differentiation between Atlantic
463 and Antarctic, G_{ST} = –0.0011, D_J = 0.0022. The lowest population differentiation was
464 between Antarctic Areas IV and V (ANV vs ANIV, G_{ST} = –0.0004, D_J = 0.0009).
465
466 To explore population structure differentiation further, we also analysed polymorphic
467 microsatellites and the hypervariable region of the mitochondrial D-Loop (mtDNA HVR)
468 (**Figure 3**). Despite the low *Prdm9* diversity in the Northern Hemisphere, genetic
469 structuring was revealed by the mtDNA HVR, with a clear distinction between the North
470 Atlantic and North Pacific whales. We also examined nine unlinked autosomal
471 microsatellite loci, chosen for their high information content in minke whales [17, 29]

472 and sex-chromosome markers ZFX/ZFY, from [38]. Microsatellite summary statistics
473 reveal high heterozygosity of effectively zero F_{IS} (Supplementary Table 1), so we
474 continued with a Bayesian analysis of population structure. We used *STRUCTURE*
475 with a recessive allele model because any null alleles were due to polymorphisms,
476 rather than missing data [53] and the "admixture" model. However, given that minke
477 whales occur in geographically isolated groups, are separated by hemispheres, and
478 have asynchronous breeding seasons, we nevertheless assumed a low probability for
479 incomplete lineage sorting, which could otherwise confound structure inference,
480 particularly for weak population differentiation. The most likely number of clusters
481 returned by ΔK is $k = 2$, which distinguishes the two hemispheres. Increasing to $k = 3$
482 and $k = 4$, further separates the common minke whale subspecies (as shown in **Figure**
483 **3**), a result that holds even without *a priori* location data (NOLOCS) where only strong
484 population structure should be detected (**Supplementary Figure 6**).
485

486 **Putative minke whale Recombination initiation motifs**

487 Different C_2H_2-ZnF domains target different recombination hotspots [3, 4, 54, 55].
488 Target hotspots in humans [4, 48] and mice [55] can be predicted from C_2H_2 ZnF
489 sequences using the SVM polynomial kernel model for *de-novo* binding prediction [45].
490 To understand the situation in minke whales, we identified a number of "core motifs",
491 based on ZnFs #3-#6, that appear particularly important for DNA binding [56, 57]
492 (**Figure 4**). As must be true, the single PRDM9 variant found in common minke whales
493 results in a single "core motif". However, a large majority of Antarctic minke whales
494 also share a single "core motif", despite a much larger diversity of PRDM9 (**Figure 4**).
495 Core motifs show no overlap between common minke whales and Antarctic minke
496 whales, a trend that is particularly evident at C_2H_2-ZnF#3 where it creates distinctive

497 DNA binding motifs. Thus, it appears as if core motifs are shared between the majority
498 of minke whale PRDM9 variants within a given species, but not between species
499 inhabiting different Hemispheres.
500
501 **Discussion:**
502 **PRDM9 shows high diversity, especially in Antarctic minke whales**
503 We characterized the PRDM9 C_2H_2-ZnF domain diversity of 134 individuals from four
504 natural populations of minke whales and discovered a total of eighteen variants. The
505 minisatellite encoding the DNA-binding C_2H_2-ZnF array apparently evolves rapidly in
506 minke whales, with at least one of the DNA-recognizing position likely under positive
507 selection. This high diversity is similar to that observed at the *Prdm9* gene in other
508 vertebrate species, such as humans, mice, bovines and primates [4-6, 12, 58].

509 However, PRDM9 diversity is not equally abundant in both hemispheres. In contrast to
510 the extensive PRDM9 diversity we observe in Antarctic minke whales (*Balaenoptera*
511 *bonarensis)*, only a single PRDM9 variant is found in Common minke whales. This
512 variant is shared between two subspecies, the North Atlantic minke whale *(B. a.*
513 *acutorostrata)* and North Pacific minke whales *(B. a. scammoni).* In most mammals,
514 extensive PRDM9 variation typically is seen between, and within, a given subspecies.
515 This holds true also for Antarctic minke whales, thus the observation of an identical
516 PRDM9 variant shared between two subspecies of common minke whales is
517 unexpected. The PRDM9 protein level diversity observed within each Hemisphere
518 contrasted with the levels of population differentiation seen in microsatellite as well as
519 mitochondrial haplotypes. However, diversity between repeat types of the *Prdm9*
520 minisatellite can reveal even weak population structure and is consistent with data on

521 mitochondrial haplotypes and microsatellites. This contrasting pattern between
522 protein-level conservation and nucleotide differentiation is fascinating and points to
523 functional constraints operating on different levels of PRDM9 evolution.

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

525 Taxonomists have separated the common minke whale into two subspecies, the North
526 Atlantic minke whale *(B. a. acutorostrata)* and the North Pacific minke whale (*B. a.*
527 *scammoni*), which diverged approximately 1.5 million years ago. variability between
528 84bp *Prdm9* minisatellite repeat units, as well as microsatellite markers and mtDNA all
529 reveal weak population structure and differentiation between minke whales in North
530 Atlantic and North Pacific. Our population genetic data support this classification and
531 suggest that these populations may be in the process of speciating. Yet fascinatingly,
532 all individuals of both Northern Hemisphere populations had identical C_2H_2 domains,
533 which suggests that these subspecies should still be able to interbreed if given a
534 chance. As geographical isolation of the two subspecies is currently still upheld by the
535 permanent polar ice, allopatric speciation may be promoted. However, due to global
536 warming, the two subspecies might come into secondary contact again in the future.
537 The removal of the current geographical barrier could therefore disrupt speciation,
538 especially in the light of identical C_2H_2 domains encoded by the mammalian hybrid
539 sterility gene *Prdm9.*

540

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

542 Analysis of the hypervariable region (HVR) on the mitochondrial D-Loop suggests that
543 Antarctic minke whale and Common minke whale evolved from a common ancestor in
544 the Southern Hemisphere during a period of global warming approximately 5 million

545 years ago [59]. Common minke whales and Antarctic minke whales are now separated
546 by both geography and seasonality. However, while their winter habitats and breeding
547 grounds remain unknown, it is assumed that *B. a. acutorostrata* migrates south
548 between November and March to give birth in warmer waters and occasional
549 individuals are seen as far south as the Gulf of Mexico [60]. Interbreeding between the
550 populations of the Northern and Southern Hemisphere appears unlikely at the time our
551 samples were collected (the 1980s) because our study confirms previous findings [28]
552 that common minke whales and Antarctic minke whales exhibit extensive genetic
553 differences in their mtDNA and microsatellites. However, recent reports have
554 uncovered rare hybridization events between Common minke whale and Antarctic
555 minke whales [16], which shows that interbreeding is unlikely but possible. The
556 question remains whether known reproductive isolation mechanism exists in hybrids
557 between the two minke whale species.
558
559 The *Prdm9* gene remains the only hybrid sterility gene known in vertebrates. Large
560 evolutionary turnover of the coding minisatellite is seen in all mammals characterised
561 to date- including minke whales (*this study*). One consequence is genetic variation in
562 the DNA-binding C_2H_2-ZnF-domain that in turn introduces variation in species'
563 recombination landscapes, that differ even within populations of the same species [4,
564 61]. Based on computed PRMD9 C_2H_2-ZnF binding motifs found in Common minke
565 whale and Antarctic minke whale, we would speculate that the recombination initiation
566 landscapes of the two variants would not overlap. The associated variation in
567 recombination landscapes is implicated in generating reproductive isolation in
568 mammals, which has been well characterized in mice. Here, historical erosion of
569 genomic binding sites of PRDM9 ZnF domains has been shown to lead to the

570 introduction of asymmetric sets of DSBs in evolutionary divergent homologous genomic
571 sequences [61, 62]. This asymmetry is likely caused because PRDM9 initiation can
572 erode its own binding via biased gene conversion over long evolutionary timescales.
573
574 As a result, in hybrid genomes, the variant of one species preferentially binds the
575 ancestral, binding sites on the homologue of the other species, that have not been
576 eroded, and vice versa. The resulting asymmetry of recombination initiation sites is
577 believed to be responsible for the inefficient DSB repair, defective pairing, asynapsis
578 and segregation of the chromosomes that are observed in intersubspecific mouse
579 hybrids [63-65]. The low PRDM9 diversity in the Northern Hemisphere would predict
580 extensive historical erosion of PRDM9 binding sites in the genomes of Common Minke
581 Whales. A degree of erosion may also be true across the Southern Hemisphere, where
582 most animals share identical core motifs. Thus, the predicted high level of PRDM9
583 binding site erosion, coupled with non-overlapping minke whale PRDM9 binding motifs
584 would predict asymmetric PRDM9 recombination imitation in minke whale hybrids
585 (**Supplementary Figure 7**), which is implicated in hybrid male sterility.
586
587 Despite this potential incompatibility, two viable female minke whale hybrids have been
588 observed, one of which showed signs of successful backcrossing [15, 16]. This
589 observation alone is not sufficient to clarify whether postzygotic isolation mechanisms
590 do or do not generally operate. After all, Haldane's rule predicts that the sterile sex will
591 be the heterogametic males, rather than the females on which observations were
592 made, an expectation that holds in PRDM9-mediated hybrid sterility in mice [20, 66,
593 67]. Therefore, it still remains to be understood whether sporadically occurring
594 hybridization events will generate infertile males or, instead, signify that start of a

595 breakdown of genetic isolation. Further studies on minke whale hybrids are necessary
596 to elucidate this matter.
597

598 **Conclusion**

599 We uncovered that Minke whales, and related species, possess coding sequences for
600 a complete set of KRAB, SSXRD and SET domains, a necessary feature of organisms
601 with PRDM9-regulated recombination [46]. Sequencing revealed evolutionary turnover
602 of the PRDM9 coding minisatellite. Our phylogenetic analysis revealed a good
603 association of groups with their geographical origin, a high level of size homoplasy in
604 common minke whales and a lower level of size homoplasy in Antarctic minke whales.
605 In Antarctic minke whales (*Balaenoptera bonarensis*), we observe extensive diversity
606 within the C_2H_2-ZNF domain of PRDM9, even at fine geographical scale. In direct
607 opposition, we observe no genetic diversity within the C_2H_2-ZNF in common minke
608 whales. In each Hemisphere, the PRDM9 protein diversity is contrasting strongly with
609 the underlying population structure we observe between and within subspecies.
610 However, on the nucleotide level, variation between *Prdm9* minisatellite repeat types
611 mirrored the variation in mitochondrial DNA and microsatellites in respective animals.
612 This apparent contradiction points to maintenance of conserved protein sequence,
613 even across minke whale subspecies boundaries.
614

615 **Acknowledgements**

616 We thank the Max Planck Society for funding this study.
617 We thank Nicole Thomsen and Olga Eitel for technical support. We would also like to
618 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:

1. Imai Y, Baudat F, Taillepierre M, Stanzione M, Toth A, de Massy B: **The PRDM9 KRAB domain is required for meiosis and involved in protein interactions**. *Chromosoma* 2017, **126**(6):681-695.
2. Parvanov ED, Tian H, Billings T, Saxl RL, Spruce C, Aithal R, Krejci L, Paigen K, Petkov PM: **PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis**. *Mol Biol Cell* 2017, **28**(3):488-499.
3. Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ: **PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans**. *Nature genetics* 2010, **42**(10):859-863.
4. Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ: **Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(30):12378-12383.
5. Kono H, Tamura M, Osada N, Suzuki H, Abe K, Moriwaki K, Ohta K, Shiroishi T: **Prdm9 polymorphism unveils mouse evolutionary tracks**. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2014, **21**(3):315-326.
6. Buard J, Rivals E, Dunoyer de Segonzac D, Garres C, Caminade P, de Massy B, Boursot P: **Diversity of Prdm9 zinc finger array in wild mice unravels new facets of the evolutionary turnover of this coding minisatellite**. *PloS one* 2014, **9**(1):e85021.
7. Schwartz JJ, Roach DJ, Thomas JH, Shendure J: **Primate evolution of the recombination regulator PRDM9**. *Nat Commun* 2014, **5**:4370.
8. Steiner CC, Ryder OA: **Characterization of Prdm9 in equids and sterility in mules**. *PloS one* 2013, **8**(4):e61746.
9. Ahlawat S, Sharma P, Sharma R, Arora R, Verma NK: **Evidence of positive selection and concerted evolution in the rapidly evolving PRDM9 zinc finger domain in goats and sheep**. 2016, **47**(6):740-751.
10. Ahlawat S, De S, Sharma P, Sharma R, Arora R, Kataria RS, Datta TK, Singh RK: **Evolutionary dynamics of meiotic recombination hotspots regulator PRDM9 in bovids**. *Mol Genet Genomics* 2017, **292**(1):117-131.
11. Ahlawat S, Sharma P, Sharma R, Arora R, De S: **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* 2016, **11**(5):e0156159.
12. Ahlawat S, Sharma P, Sharma R, Arora R, Verma NK, Brahma B, Mishra P, De S: **Evidence of positive selection and concerted evolution in the rapidly evolving PRDM9 zinc finger domain in goats and sheep**. *Anim Genet* 2016, **47**(6):740-751.

13. Bakke I, Johansen S, Bakke Ø, El-Gewely MR: **Lack of population subdivision among the minke whales (Balaenoptera acutorostrata) from Icelandic and Norwegian waters based on mitochondrial DNA sequences**. *Marine Biology* 1996, **125**(1):1-9.
14. Perrin WF, Jr. RLB: **Minke Whales: Balaenoptera acutorostrata and B. bonaerensis**. In: *Encyclopedia of Marine Mammals (Second Edition).* Edited by Perrin WF, Würsig B, Thewissen JGM, Second edn. Academic Press; 2009: 733-735.
15. Glover KA, Kanda N, Haug T, Pastene LA, Oien N, Goto M, Seliussen BB, Skaug HJ: **Migration of Antarctic minke whales to the Arctic**. *PloS one* 2010, **5**(12):e15197.
16. Glover KA, Kanda N, Haug T, Pastene LA, Oien N, Seliussen BB, Sorvik AG, Skaug HJ: **Hybrids between common and Antarctic minke whales are fertile and can back-cross**. *BMC Genet* 2013, **14**:25.
17. Malde K, Seliussen BB, Quintela M, Dahle G, Besnier F, Skaug HJ, Oien N, Solvang HK, Haug T, Skern-Mauritzen R *et al*: **Whole genome resequencing reveals diagnostic markers for investigating global migration and hybridization between minke whale species**. *BMC genomics* 2017, **18**(1):76.
18. Maheshwari S, Barbash DA: **The genetics of hybrid incompatibilities**. *Annual review of genetics* 2011, **45**:331-355.
19. Coyne JA, Orr HA: **Speciation**: Sinauer Associates, Inc.; 2004.
20. Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J: **A mouse speciation gene encodes a meiotic histone H3 methyltransferase**. *Science* 2009, **323**(5912):373-375.
21. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC Bioinformatics* 2005, **6**:31.
22. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W116-120.
23. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions**. *Nucleic Acids Res* 2013, **41**(12):e121.
24. Suchard MA, Redelings BD: **BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny**. *Bioinformatics* 2006, **22**(16):2047-2048.
25. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic Acids Res* 2005, **33**(2):511-518.
26. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Molecular biology and evolution* 2015, **32**(1):268-274.
27. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL: **Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction**. *BMC Bioinformatics* 2012, **13:134**.
28. van Pijlen IA, Amos B, Burke T: **Patterns of genetic variability at individual minisatellite loci in minke whale Balaenoptera acutorostrata populations from three different oceans**. *Molecular biology and evolution* 1995, **12**(3):459-472.
29. Valsecchi E, Amos W: **Microsatellite markers for the study of cetacean populations**. *Mol Ecol* 1996, **5**(1):151-156.

714 30. Palsboll PJ, Berube M, Larsen AH, Jorgensen H: **Primers for the amplification**
715 **of tri- and tetramer microsatellite loci in baleen whales**. *Mol Ecol* 1997,
716 **6**(9):893-895.
717 31. Berube M, Jorgensen H, McEwing R, Palsboll PJ: **Polymorphic di-nucleotide**
718 **microsatellite loci isolated from the humpback whale, Megaptera**
719 **novaeangliae**. *Mol Ecol* 2000, **9**(12):2181-2183.
720 32. Berube M, Palsboll P: **Identification of sex in cetaceans by multiplexing with**
721 **three ZFX and ZFY specific primers**. *Mol Ecol* 1996, **5**(2):283-287.
722 33. Mukaj A, Pialek J, Fotopulosova V, Morgan AP, Odenthal-Hesse L, Parvanov
723 ED, Forejt J: **Prdm9 inter-subspecific interactions in hybrid male sterility**
724 **of house mouse**. *Molecular biology and evolution* 2020.
725 34. Orozco-terWengel P, Corander J, Schlotterer C: **Genealogical lineage sorting**
726 **leads to significant, but incorrect Bayesian multilocus inference of**
727 **population structure**. *Mol Ecol* 2011, **20**(6):1108-1121.
728 35. Earl DA, vonHoldt BM: **STRUCTURE HARVESTER: a website and program**
729 **for visualizing STRUCTURE output and implementing the Evanno method**.
730 *Conservation Genetics Resources* 2011, **4**(2):359-361.
731 36. Evanno G, Regnaut S, Goudet J: **Detecting the number of clusters of**
732 **individuals using the software STRUCTURE: a simulation study**. *Mol Ecol*
733 2005, **14**(8):2611-2620.
734 37. Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG: **STRUCTURE PLOT: a**
735 **program for drawing elegant STRUCTURE bar plots in user friendly**
736 **interface**. *Springerplus* 2014, **3**:431.
737 38. Glover KA, Haug T, Øien N, Walløe L, Lindblom L, Seliussen BB, Skaug HJ:
738 **The Norwegian minke whale DNA register: a data base monitoring**
739 **commercial harvest and trade of whale products**. *Fish and Fisheries* 2012,
740 **13**(3):313-332.
741 39. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton
742 S, Cooper A, Markowitz S, Duran C *et al*: **Geneious Basic: an integrated and**
743 **extendable desktop software platform for the organization and analysis of**
744 **sequence data**. *Bioinformatics* 2012, **28**(12):1647-1649.
745 40. Jeffreys AJ, Neumann R, Wilson V: **Repeat unit sequence variation in**
746 **minisatellites: a novel source of DNA polymorphism for studying variation**
747 **and mutation by single molecule analysis**. *Cell* 1990, **60**(3):473-485.
748 41. Paradis E: **pegas: an R package for population genetics with an integrated-**
749 **modular approach**. *Bioinformatics* 2010, **26**(3):419-420.
750 42. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL:
751 **Detecting individual sites subject to episodic diversifying selection**. *PLoS*
752 *genetics* 2012, **8**(7):e1002764.
753 43. Vara C, Capilla L, Ferretti L, Ledda A, Sanchez-Guillen RA, Gabriel SI, Albert-
754 Lizandra G, Florit-Sabater B, Bello-Rodriguez J, Ventura J *et al*: **PRDM9**
755 **diversity at fine geographical scale reveals contrasting evolutionary**
756 **patterns and functional constraints in natural populations of house mice**.
757 *Molecular biology and evolution* 2019.
758 44. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and**
759 **Evolution in R language**. *Bioinformatics* 2004, **20**(2):289-290.
760 45. Persikov AV, Singh M: **De novo prediction of DNA-binding specificities for**
761 **Cys2His2 zinc finger proteins**. *Nucleic Acids Res* 2014, **42**(1):97-108.

762 46. Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG,
763 Przeworski M: **Repeated losses of PRDM9-directed recombination despite**
764 **the conservation of PRDM9 across vertebrates**. *Elife* 2017, **6**.
765 47. Paigen K, Petkov PM: **PRDM9 and Its Role in Genetic Recombination**.
766 *Trends Genet* 2018, **34**(4):291-300.
767 48. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean
768 G, Donnelly P: **Drive against hotspot motifs in primates implicates the**
769 **PRDM9 gene in meiotic recombination**. *Science* 2010, **327**(5967):876-879.
770 49. Jeffreys AJ, Cotton VE, Neumann R, Lam KW: **Recombination regulator**
771 **PRDM9 influences the instability of its own coding sequence in humans**.
772 *Proceedings of the National Academy of Sciences of the United States of*
773 *America* 2013, **110**(2):600-605.
774 50. Jin L, Chakraborty R: **Population structure, stepwise mutations,**
775 **heterozygote deficiency and thei implications in DNA forensics**. *Heredity*
776 1995, **74**:274-285.
777 51. Nei M: **Analysis of Gene Diversity in Subdivided Populations**. *Proceedings*
778 *of the National Academy of Sciences of the United States of America* 1973,
779 **70**(12):pp. 3321-3323.
780 52. Jost L: **G(ST) and its relatives do not measure differentiation**. *Mol Ecol*
781 2008, **17**(18):4015-4026.
782 53. Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, Lareu MV: **An**
783 **overview of STRUCTURE: applications, parameter settings, and**
784 **supporting software**. *Front Genet* 2013, **4**:98.
785 54. Parvanov ED, Petkov PM, Paigen K: **Prdm9 controls activation of**
786 **mammalian recombination hotspots**. *Science* 2010, **327**(5967):835.
787 55. Cole F, Baudat F, Grey C, Keeney S, de Massy B, Jasin M: **Mouse tetrad**
788 **analysis provides insights into recombination mechanisms and hotspot**
789 **evolutionary dynamics**. *Nature genetics* 2014, **46**(10):1072-1080.
790 56. Billings T, Parvanov ED, Baker CL, Walker M, Paigen K, Petkov PM: **DNA**
791 **binding specificities of the long zinc-finger recombination protein PRDM9**.
792 *Genome biology* 2013, **14**(4):R35.
793 57. Striedner Y, Schwarz T, Welte T, Futschik A, Rant U, Tiemann-Boege I: **The**
794 **long zinc finger domain of PRDM9 forms a highly stable and long-lived**
795 **complex with its DNA recognition sequence**. *Chromosome Res* 2017,
796 **25**(2):155-172.
797 58. Groeneveld LF, Atencia R, Garriga RM, Vigilant L: **High diversity at PRDM9 in**
798 **chimpanzees and bonobos**. *PloS one* 2012, **7**(7):e39064.
799 59. Pastene LA, Goto M, Kanda N, Zerbini AN, Kerem D, Watanabe K, Bessho Y,
800 Hasegawa M, Nielsen R, Larsen F *et al*: **Radiation and speciation of pelagic**
801 **organisms during periods of global warming: the case of the common**
802 **minke whale, Balaenoptera acutorostrata**. *Mol Ecol* 2007, **16**(7):1481-1495.
803 60. Rosel PE, Wilcox LA, Monteiro C, Tumlin MC: **First record of Antarctic minke**
804 **whale, Balaenoptera bonaerensis, in the northern Gulf of Mexico**. *Marine*
805 *Biodiversity Records* 2016, **9**(1).
806 61. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD:
807 **DNA recombination. Recombination initiation maps of individual human**
808 **genomes**. *Science* 2014, **346**(6211):1256442.

62. Smagulova F, Brick K, Pu Y, Camerini-Otero RD, Petukhova GV: **The evolutionary turnover of recombination hot spots contributes to speciation in mice**. *Genes Dev* 2016, **30**(3):266-280.
63. Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R *et al*: **Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice**. *Nature* 2016, **530**(7589):171-176.
64. Forejt J: **Genetics: Asymmetric breaks in DNA cause sterility**. *Nature* 2016, **530**(7589):167-168.
65. Zelazowski MJ, Cole F: **X marks the spot: PRDM9 rescues hybrid sterility by finding hidden treasure in the genome**. *Nat Struct Mol Biol* 2016, **23**(4):267-269.
66. Dzur-Gejdosova M, Simecek P, Gregorova S, Bhattacharyya T, Forejt J: **Dissecting the genetic architecture of F1 hybrid sterility in house mice**. *Evolution* 2012, **66**(11):3321-3335.
67. Bhattacharyya T, Reifova R, Gregorova S, Simecek P, Gergelits V, Mistrik M, Martincova I, Pialek J, Forejt J: **X chromosome control of meiotic chromosome synapsis in mouse inter-subspecific hybrids**. *PLoS genetics* 2014, **10**(2):e1004088.
68. Tamura K, Nei M: **Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees**. *Molecular biology and evolution* 1993, **10**(3):512-526.

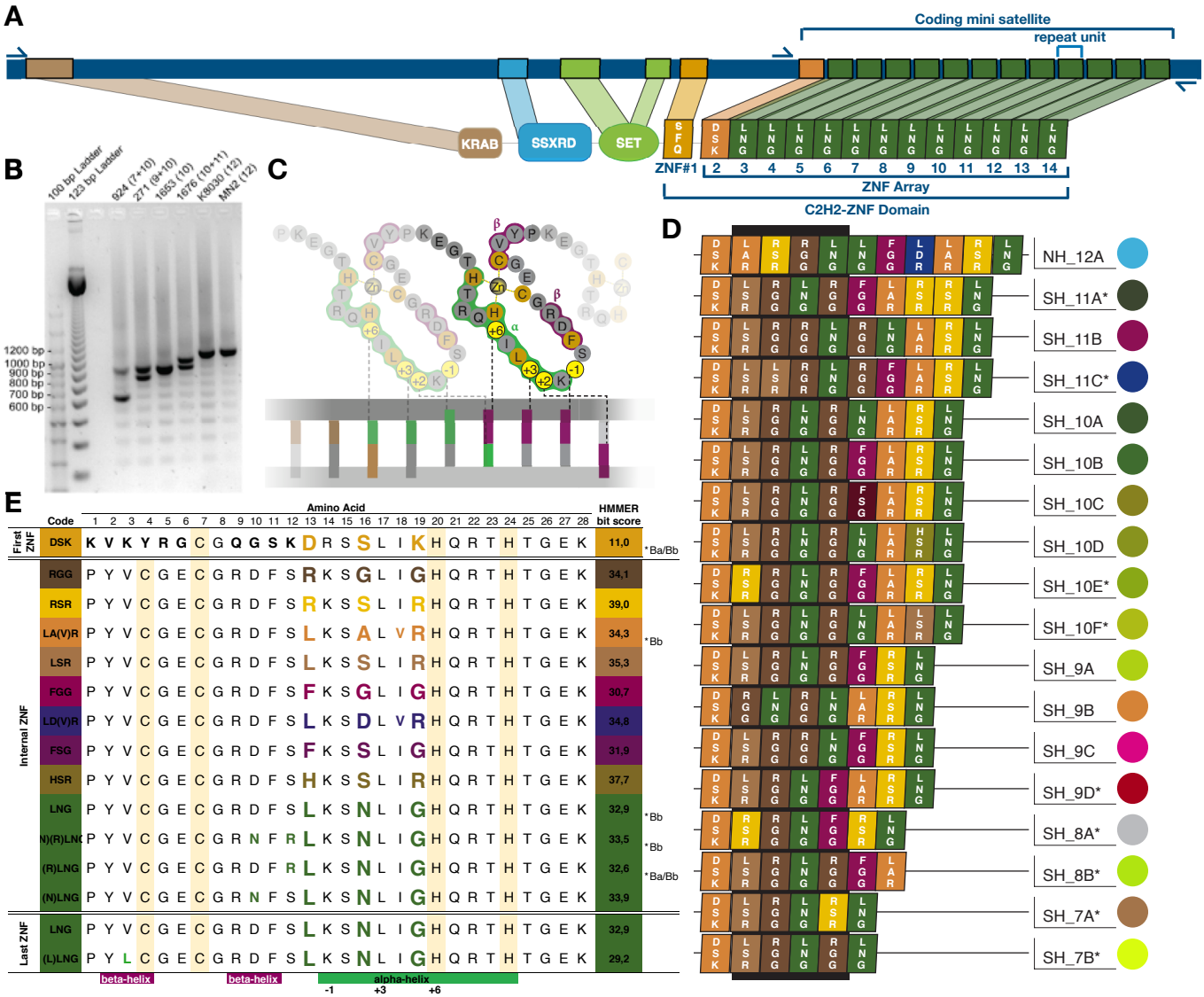
Availability of data and materials

Most data generated or analyzed during this study are included in this published article and its supplementary information files, Genetic data generated and analyzed during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.4309437, and R Scripts are available from https://gitlab.gwdg.de/mpievolbio-it/repeatr.

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.

845 **Figures and Tables**



846

847 **Figure 1 Diversity in the Cys2His2-ZnF domain in minke whales.** (A) *Prdm9* gene
848 annotation in *Balaenoptera acutorostrata scammoni* genome reference**,** with primer
849 site annotations. We predicted the PRDM9 protein (B) Representative PCR products
850 of DNAs from individuals from AN IV (294,271), AN V (1653), NP (K8030) and NA
851 (MN2) showing variation in the number of minisatellite repeats units (number of ZnFs
852 in brackets) (C) Stylized structure of C_2H_2-ZnF binding to DNA, with nucleotide-
853 specificity conferred by amino-acids in positions -1, +2, +3, and +6 of alpha-helices.
854 (D) PRDM9 C_2H_2-ZnF arrays identified in minke whales, named with broad-scale
855 sampling location, Northern Hemisphere (NH) and Southern Hemisphere (SH) and
856 total number of ZnFs in the C_2H_2-ZnFs domain. (E) Types of PRDM9 C_2H_2-ZnFs in
857 minke whales. We generate three-letter-codes using the IUPAC nomenclature of
858 amino-acids involved in DNA-binding. All variable amino acids are colored, and
859 asterisks label C_2H_2-ZnFs also present in genome references of *Balaenoptera*
860 *acutorostrata scammoni* (*Ba) or *Balaenoptera bonarensis* (*Bb), see also
861 Supplementary Figure 4.

862

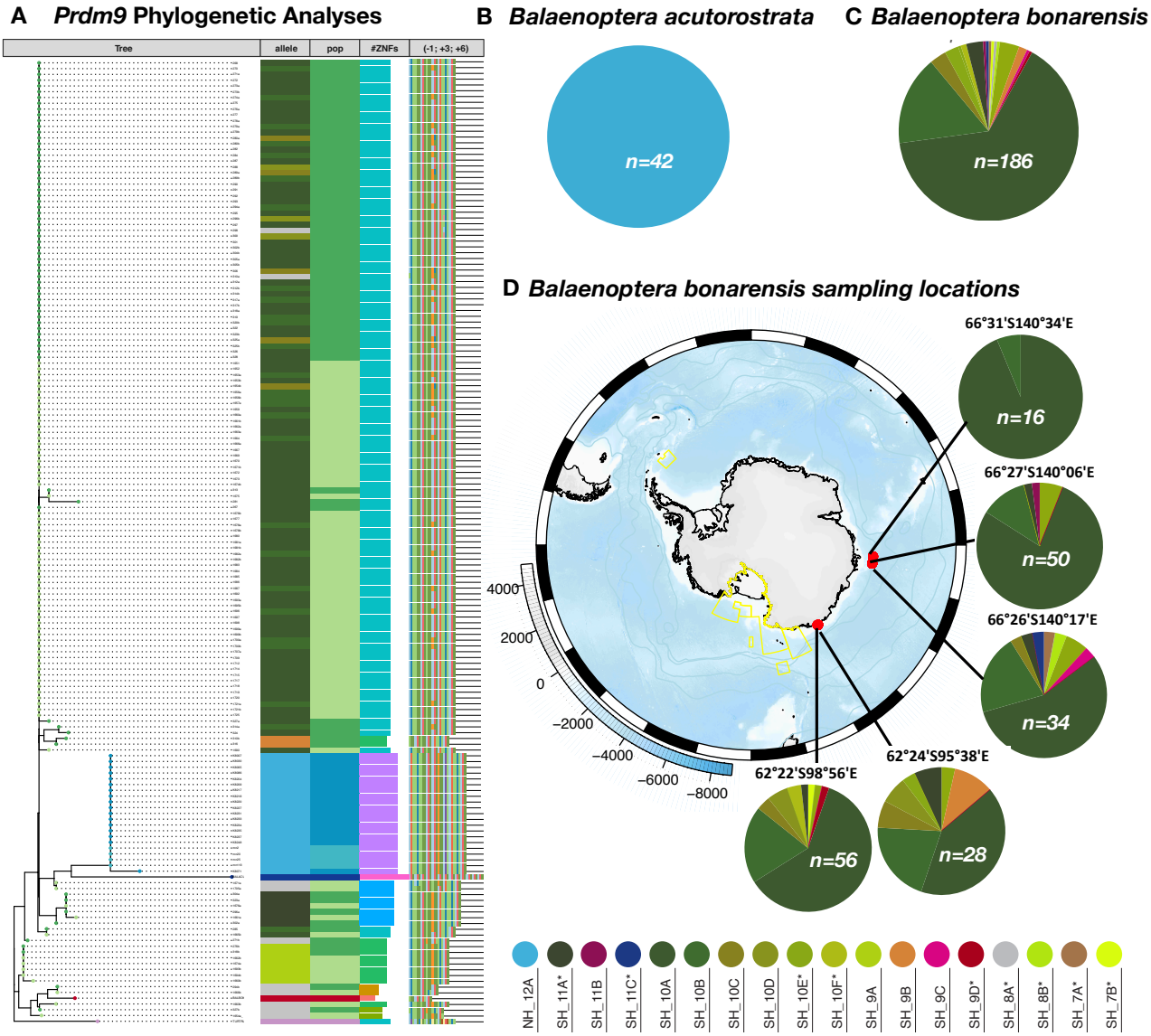


Figure 2 *Prdm9* Phylogenetic analyses and allele frequencies at different geographical scales. (A) PRDM9 phylogenetic analyses across all geographical regions, including several outgroups, from top to bottom: **-1, +3, +6**: PRDM9 variants as in Figure 1. **# ZnFs:** number of ZnFs **pop:** assigned 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 Common minke whales (C) allelic diversity in Antarctic minke whales (D) fine-scale diversity by Sampling locations of Antarctic minke whales on a map of the Southern Ocean (yellow delineations show protected areas).

Table 1 Nucleotide diversity [68] of the minisatellite coding for the PRDM9-C_2H_2-ZnF array, analyzed per sample region. Average pairwise nucleotide diversity was analyzed for the all sites, and also excluding the nucleotides coding for amino-acids at hypervariable sites (-1, +3, +6) in the alpha-helix of C_2H_2-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

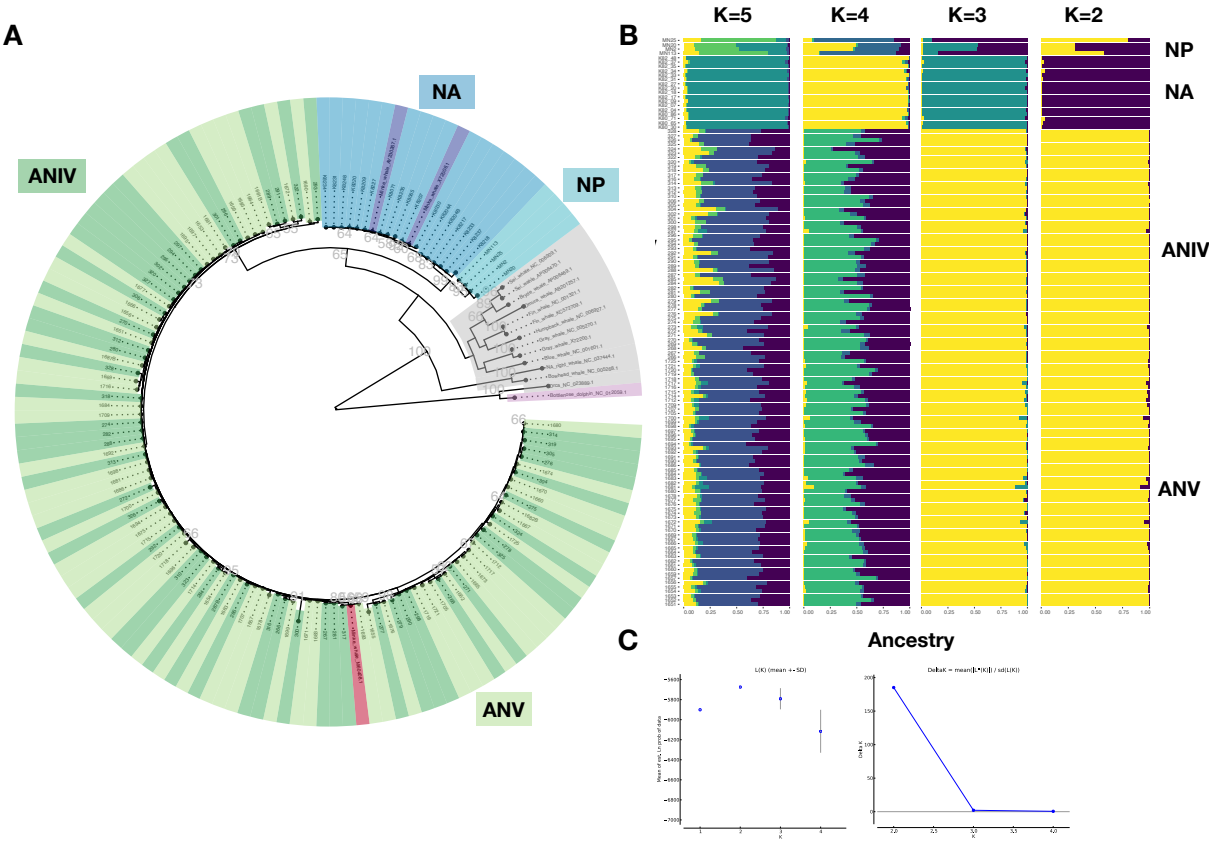


Figure 3 Population structure of minke whales (NP) North Pacific, (NA) North Atlantic (ANIV) Antarctic Area IV (ANV) Antarctic Area V (A) Phylogeny of minke whale mitochondrial D-loop region, containing the hypervariable segments (HVS). (B) Population STRUCTURE analyses of minisatellites with *a priori*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

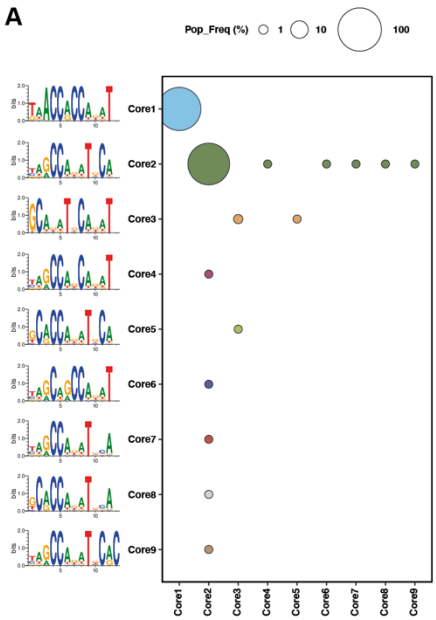
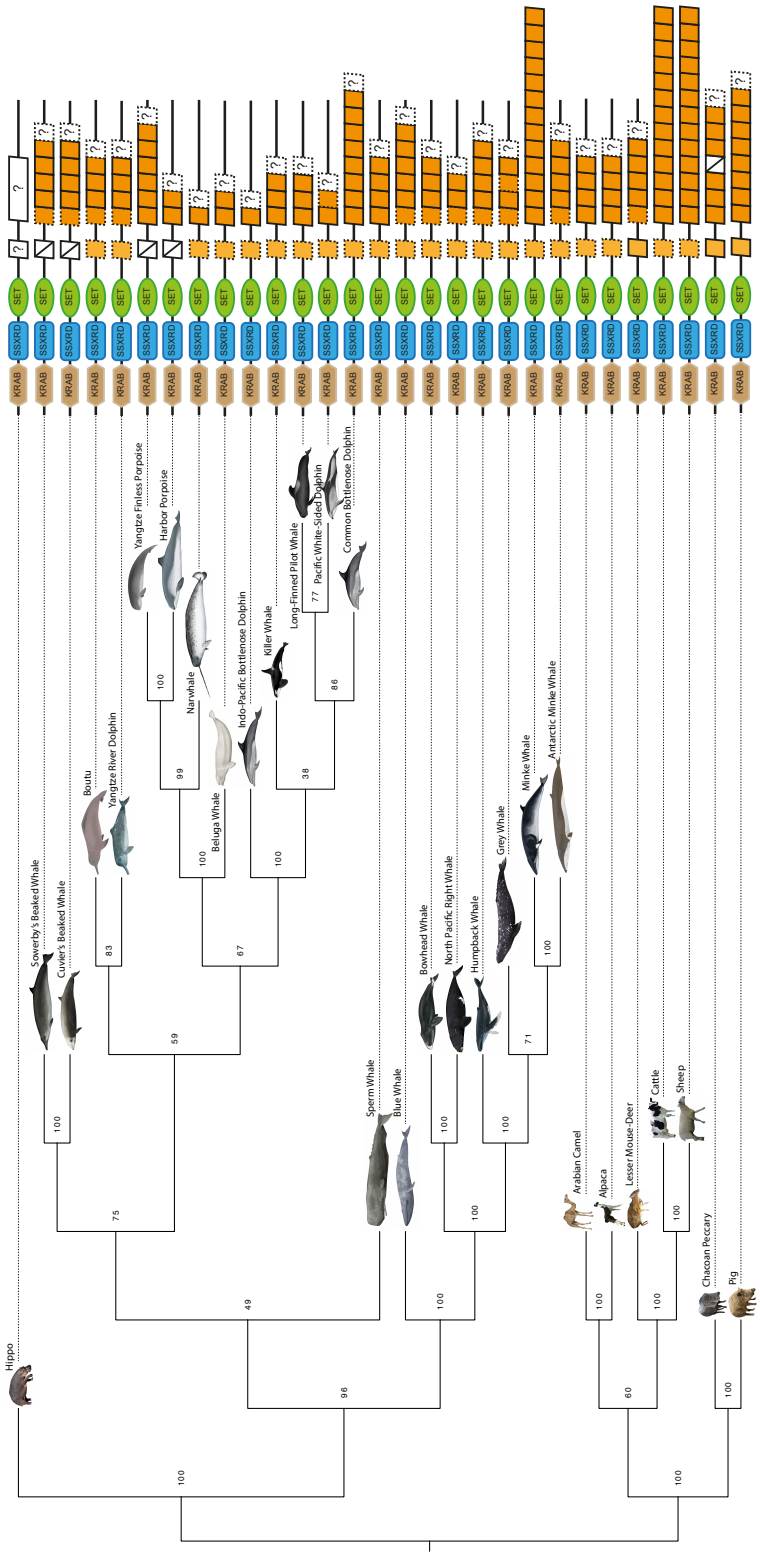
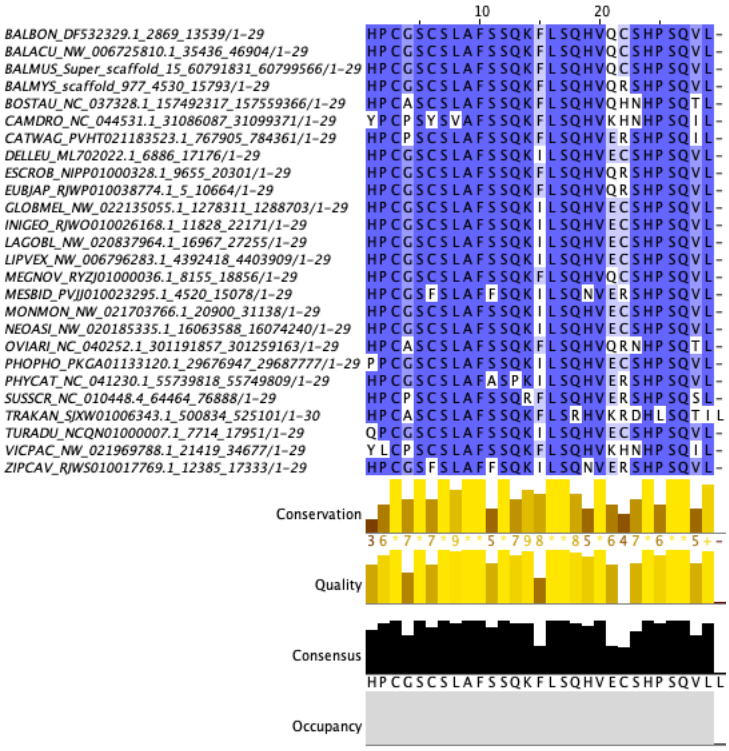


Figure 4 PRDM9 Array binding predictions and population frequency based on ZnFs#3-#6 "core motif" combinations across all individuals

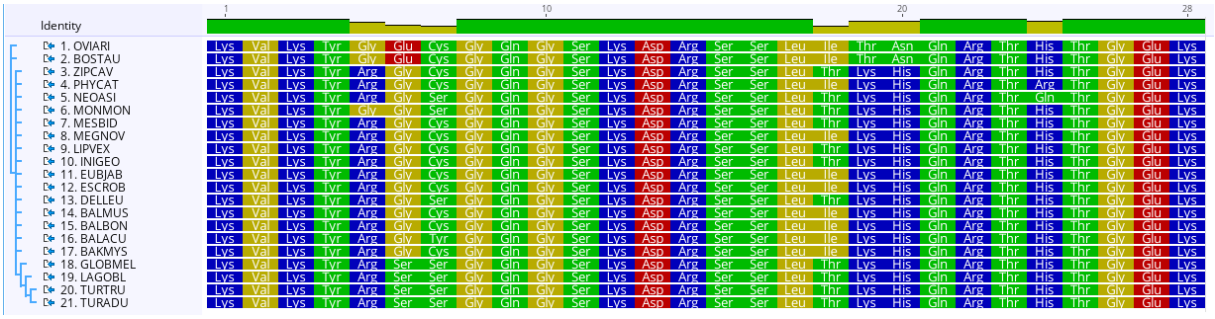
Supplementary Information



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.

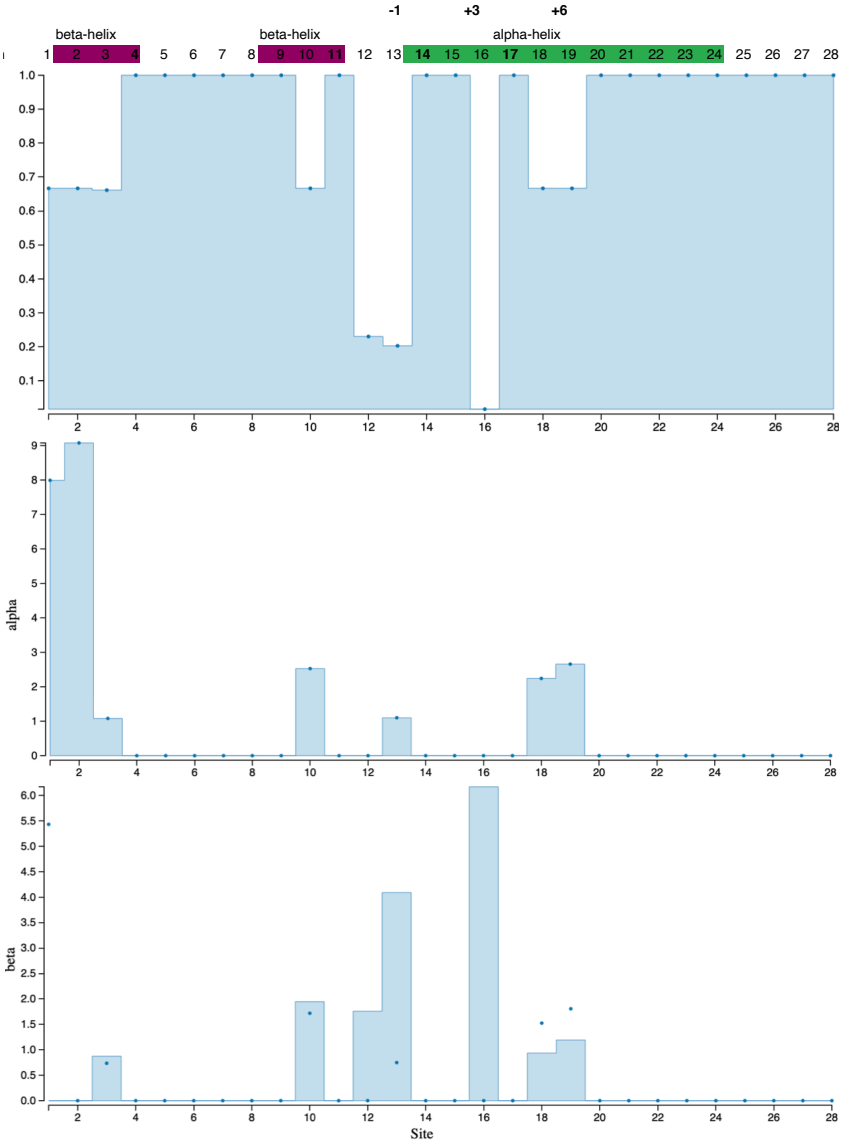


Supplementary Figure 3: Alignment of second ZnFs of all *Cetartiodactyla*, where this ZnF was present in public genomic databases (Supplementary Table **2**). Across *Cetacea*, ZnF#2 are conserved at the amino-acids 13, 16 and 19, which code for the DNA-binding positions of the alpha-helix of each ZnF. 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 asterisks label C_2H_2-ZnFs also present in genome references of *Balaenoptera acutorostrata scammoni* (*Ba) or *Balaenoptera bonarensis* (*Bb).



Supplementary Figure 5: Signals of selection on 84bp nucleotide repeats coding for each ZnF identified in this study. "Site" shows zinc-fingers amino-acid positions, from top to bottom: a site is classified based on whether α>β (negative selection) or α<β (positive selection). (A) p-value is significant when <0,05 (B) fel.α instantaneous synonymous site rate (C) fel.β instantaneous non-synonymous site rate.

933 Supplementary Table 1: Descriptive statistics of microsatellites used in this study
934 analyzed using (Dieringer and Schlötterer 2003). The richness of allelic variants
935 (number of observed alleles) (**A**) Expected Heterozygosity (**H_E**) observed
936 Heterozygosity (**H_O**), excess or deficiency from the average number of heterozygotes
937 (**F_{IS}**) and (**H_S**), the probability that two random alleles from the same, randomly
938 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	BOSTAU	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
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	MEGNOV	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
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	TURTRU	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
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_VicP 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

948

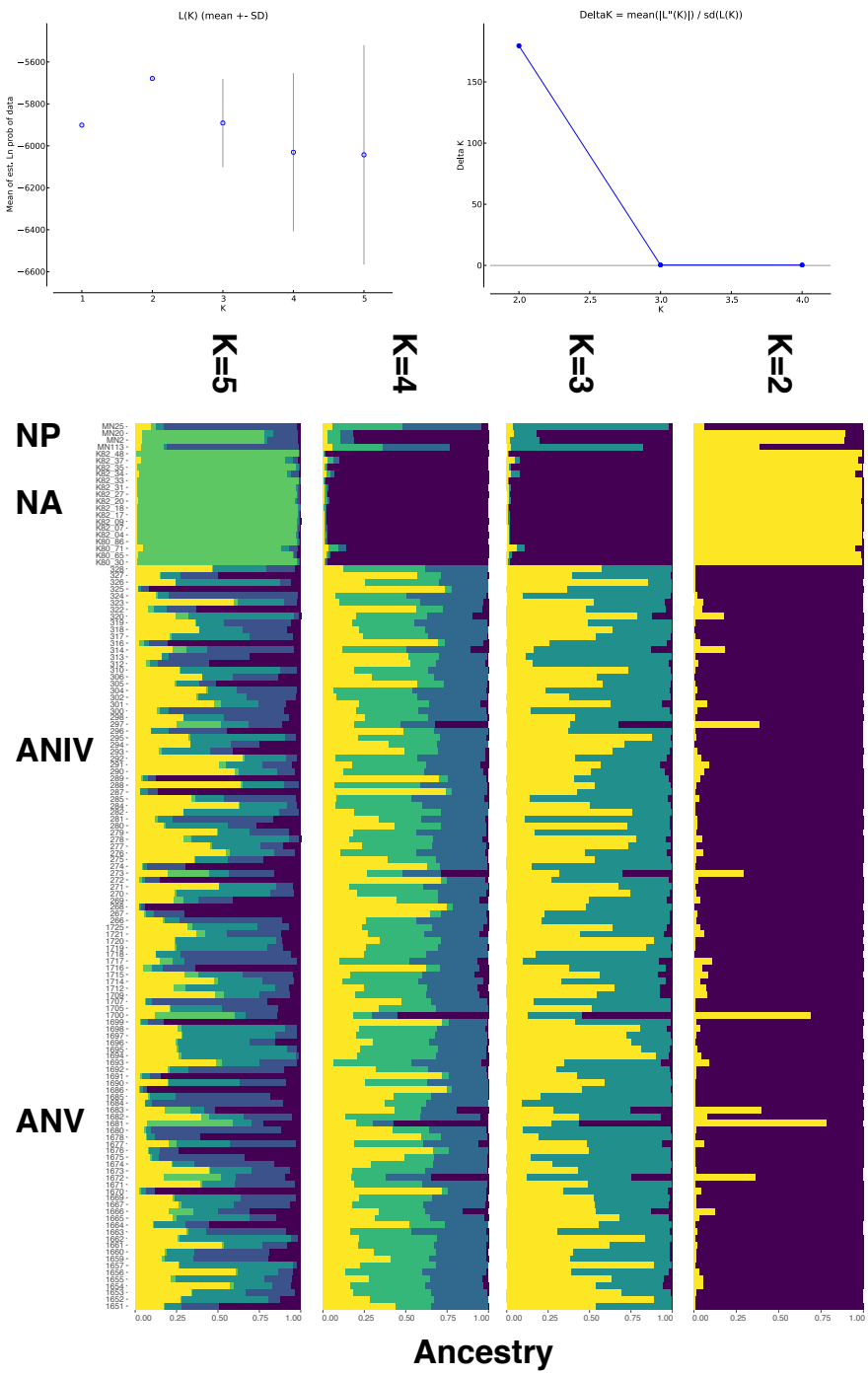
949 **Supplementary Table 3: Public mitochondrial resources.**

SPECIES	COMMON	ALIAS	TAXID	URL-SEQUENCE
BALAENOPTERA ACUTOROSTRATA	minke whale	BALACU	9767	https://www.ncbi.nlm.nih.gov/nuccore/X7200
BALAENOPTERA ACUTOROSTRATA	minke whale	BALACU	9767	https://www.ncbi.nlm.nih.gov/nuccore/AY230
BALAENOPTERA BONAERENSIS	Antarctic minke whale	BALBON	33556	https://www.ncbi.nlm.nih.gov/nuccore/M6040
BALAENOPTERA BOREALIS	sei whale	BALBOR	9768	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
BALAENOPTERA BOREALIS	sei whale	BALBOR	9768	https://www.ncbi.nlm.nih.gov/nuccore/AP006
BALAENOPTERA BRYDEI	bryde whale	BALBRY	255365	https://www.ncbi.nlm.nih.gov/nuccore/AP006
BALAENOPTERA MUSCULUS	blue whale	BALMUS	9771	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
BALAENOPTERA OMURAI	omura whale	BALOMU	255217	https://www.ncbi.nlm.nih.gov/nuccore/AB201
BALAENOPTERA PHYSALUS	fin whale	BALPHY	9770	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
BALAENOPTERA PHYSALUS	fin whale	BALPHY	9770	https://www.ncbi.nlm.nih.gov/nuccore/KC572
BALAENA MYSTICETUS	bowhead whale	BALMYS	27602	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
ESCHRICHTIUS ROBUSTUS	grey whale	ESCROB	9764	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
ESCHRICHTIUS ROBUSTUS	grey whale	ESCROB	9764	https://www.ncbi.nlm.nih.gov/nuccore/X7220
EUBALAENA GLACIALIS	north atlantic right whale	EUBGLA	27606	https://www.ncbi.nlm.nih.gov/nuccore/NC_03
LIPOTES VEXILLIFER	yangtze river dolphin	LIPVEX	118797	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
MEGAPTERA NOVAEANGLIAE	humpback whale	MEGNOV	9773	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
ORCINUS ORCA	killer whale	ORCORC	9733	https://www.ncbi.nlm.nih.gov/nuccore/NC_02
PHYSETER CATODON	sperm whale	PHYCAT	9755	https://www.ncbi.nlm.nih.gov/nuccore/NC_00
TURSIOPS TRUNCATUS	common bottlenose dolphin	TURTRU	9739	https://www.ncbi.nlm.nih.gov/nuccore/NC_01

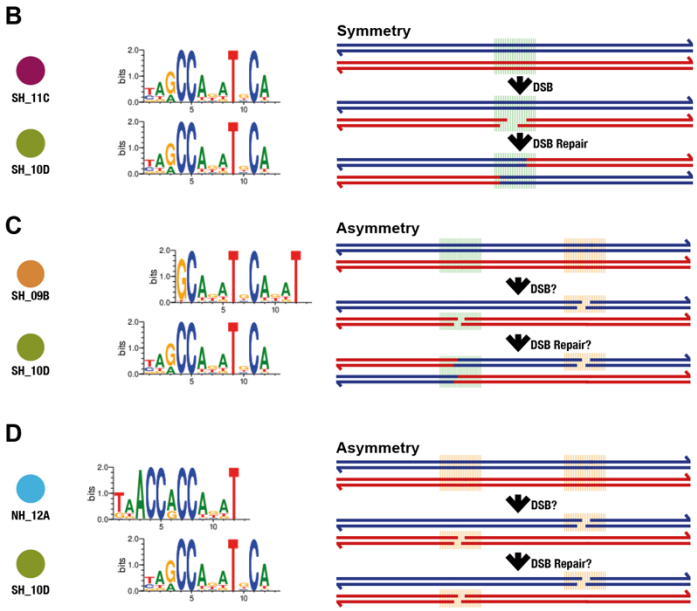
950

951 **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).