

1 **Title:** Repeated losses of PRDM9-directed recombination despite the conservation of
2 PRDM9 across vertebrates

3
4 Zachary Baker^{1,+,*}, Molly Schumer^{2,3,4+}, Yuki Haba⁵, Lisa Bashkirova⁶, Chris Holland^{4,7},
5 Gil G. Rosenthal^{4,7}, and Molly Przeworski^{1, 2,*}

6
7 ¹ Dept. of Systems Biology, Columbia University

8 ² Dept. of Biological Sciences, Columbia University

9 ³ Harvard Society of Fellows, Harvard University

10 ⁴ Centro de Investigaciones Científicas de las Huastecas “Aguazarca”

11 ⁵ Dept. of Evolution, Ecology and Environmental Biology, Columbia University

12 ⁶ Dept. of Biochemistry and Molecular Biophysics, Columbia University

13 ⁷ Dept. of Biology, Texas A&M University

14 ⁺ Contributed equally

15 ^{*} To whom correspondence should be addressed: ztb2002@columbia.edu and

16 mp3284@columbia.edu

17

18

19 **Abstract**

20

21 Studies of highly diverged species have revealed two mechanisms by which meiotic
22 recombination is directed to the genome—through PRDM9 binding or by targeting
23 promoter-like features—that lead to dramatically different evolutionary dynamics of
24 hotspots. Here, we identify PRDM9 orthologs from genome and transcriptome data in
25 225 species. We find the complete PRDM9 ortholog across distantly related vertebrates
26 but, despite this broad conservation, infer a minimum of six partial and three complete
27 losses. Strikingly, taxa carrying the complete ortholog of PRDM9 are precisely those
28 with rapid evolution of its predicted binding affinity, suggesting that all domains are
29 necessary for directing recombination. Indeed, as we show, swordtail fish carrying only a
30 partial but conserved ortholog share recombination properties with PRDM9 knock-outs.

31

32

33 **Introduction**

34

35 Meiotic recombination is a fundamental genetic process that generates new
36 combinations of alleles on which natural selection can act and, in most sexually-
37 reproducing organisms, plays critical roles in the proper alignment and segregation of
38 homologous chromosomes during meiosis (Coop and Przeworski 2007; de Massy 2013;
39 Lam and Keeney 2014). Meiotic recombination is initiated by a set of double strand
40 breaks (DSBs) deliberately inflicted throughout the genome, whose repair leads to
41 crossover and non-crossover recombination events (Lam and Keeney 2014). Most of the
42 molecular machinery involved in this process in vertebrates has been conserved since the
43 common ancestor of plants, animals and fungi (de Massy 2013). Notably, in all species
44 studied to date, the SPO11 protein generates DSBs, which localize to histone H3 lysine
45 K4 trimethylation marks (H3K4me3) along the genome (Borde et al. 2009; Buard et al.
46 2009; Lam and Keeney 2014). Yet not all features of meiotic recombination are
47 conserved across species. As one example, in many species, including all yeast, plant and
48 vertebrate species studied to date, recombination events are localized to short intervals
49 (of hundreds to thousands of base pairs; Lange et al. 2016) known as recombination
50 hotspots, whereas in others, such as in flies or worms, the recombination landscape seems
51 more uniform, lacking such hotspots (Rockman and Kruglyak 2009; Chan et al. 2012;
52 Heil et al. 2015)

53 Among species with recombination hotspots, there are at least two mechanisms
54 directing their location. In mammalian species, including apes, mice and likely in cattle,
55 the locations of recombination hotspots are specified by PRDM9 binding (Baudat et al.
56 2010; Myers et al. 2010; Parvanov et al. 2010; Sandor et al. 2012; Stevenson et al. 2016).
57 In these species, PRDM9 has four major functional domains: a KRAB, SSXRD and
58 PR/SET domain (referred to as the SET domain in what follows), followed by a C2H2
59 zinc finger (ZF) array (**Figure 1**). During meiosis, PRDM9 binds sites across the genome,
60 as specified by its ZF array (reviewed in Segurel et al. 2011), and the SET domain of
61 PRDM9 makes H3K4me3 and H3K36me3 marks nearby (Eram et al. 2014; Powers et al.
62 2016). These actions ultimately serve to recruit SPO11 to initiate DSBs, by a mechanism
63 that remains unknown but is associated with the presence of both histone marks (Grey et

64 al. 2017; Getun et al. 2017) and may involve KRAB and SSXRD domains (Parvanov et
65 al. 2017).

66 A common feature of the recombination landscape in these species is that
67 recombination tends to be directed away from PRDM9-independent H3K4me3 peaks
68 (Brick et al. 2012) and, in particular, only a small proportion of hotspots occurs at
69 transcription start sites (TSSs; Myers et al. 2005; Coop et al. 2008). In contrast, in yeasts,
70 plants, and vertebrate species (such as birds and canids) that lack functional PRDM9
71 orthologs, recombination events are concentrated at or near promoter-like features,
72 including TSSs and CpG islands (CGIs), perhaps because they are associated with greater
73 chromatin accessibility (Lichten and Goldman 1995; Auton et al. 2013; Choi et al. 2013;
74 Hellsten et al. 2013; Lam and Keeney 2015; Singhal et al. 2015). Similarly, in mouse
75 knockouts for PRDM9, recombination events appear to default to promoter-like features
76 that carry H3K4me3 peaks (Brick et al. 2012; Narasimhan et al. 2016).

77 The mechanisms by which recombination events are targeted to the genome are
78 associated with dramatic differences in the evolution of recombination hotspots. When
79 recombination is directed by PRDM9, hotspot locations are not shared between closely
80 related ape species or between mouse subspecies and differ even among human
81 populations (Ptak et al. 2004; Myers et al. 2005; Ptak et al. 2005; Coop et al. 2008; Hinch
82 et al. 2011; Auton et al. 2012; Steviston et al. 2016). This rapid evolution appears to be
83 driven by two phenomena. First, the binding specificity of the PRDM9 ZF leads to the
84 existence of “hotter” and “colder” alleles, i.e., sequences that are more or less likely to be
85 bound by PRDM9 (Myers et al. 2008). In heterozygotes carrying a colder and a hotter
86 allele, this asymmetry in binding leads to the hotter alleles more often experiencing a
87 DSB (Baker et al. 2015; Davies et al. 2016). Since repair mechanisms use the intact,
88 colder allele as a template, the sequences to which PRDM9 binds are preferentially lost
89 (Boulton et al. 1997; Kauppi et al. 2005). This process of under-transmission of the hotter
90 allele in hot/cold heterozygotes acts analogously to selection for the colder allele
91 (Nagylaki and Petes 1982) and is thus expected to drive the rapid loss of hotspots from
92 the population (leading to the “hotspot paradox”; Pineda-Krch and Redfield 2005; Coop
93 and Myers 2007), consistent with empirical observations in humans and mice (Berg et al.
94 2010; Myers et al. 2010; Baker et al. 2015; Smagulova et al. 2016).

95 In addition to this loss of hotspots in *cis*, changes in the PRDM9 binding domain
96 can also lead to the rapid loss—and gain—of whole sets of hotspots. Interestingly,
97 PRDM9 has the fastest evolving C2H2 ZF array in the mouse and human genomes
98 (Oliver et al. 2009; Myers et al. 2010). More generally, mammalian PRDM9 genes show
99 strong evidence of positive selection at known DNA-binding sites of ZFs (Oliver et al.
100 2009). Thus, in mammals carrying PRDM9, individual hotspots are lost quickly over
101 evolutionary time, but changes in the PRDM9 ZF generate novel sets of hotspots, leading
102 to rapid turnover in the fine-scale recombination landscape between populations and
103 species.

104 The mechanism driving the rapid evolution of the PRDM9 ZF is unclear. One
105 hypothesis is that the under-transmission of hotter alleles eventually leads to the erosion
106 of a sufficient number of hotspots that the proper alignment or segregation of homologs
107 during meiosis is jeopardized, strongly favoring new ZF alleles (Coop and Myers 2007;
108 Myers et al. 2010; Ubeda and Wilkins 2011). Whether hotspot loss would exert a
109 sufficiently strong and immediate selection pressure to explain the very rapid evolution of
110 the PRDM9 ZF remains unclear. An alternative explanation has emerged recently from
111 the finding that in mice, widespread asymmetric binding by PRDM9 on the two
112 homologs is associated with hybrid sterility (Davies et al. 2016; Smagulova et al. 2016).
113 Since older PRDM9 motifs are more likely to have experienced erosion and hence to be
114 found in heterozygotes for hotter and colder alleles, there may be an immediate
115 advantage to new alleles that lead to greater symmetry in PRDM9 binding (Davies et al.
116 2016). Regardless of the explanation, the rapid evolution of the PRDM9 ZF is likely tied
117 to its role in recombination.

118 Conversely, in species that do not use PRDM9 to direct meiotic recombination
119 events, the rapid evolution of recombination hotspots is not seen. In birds that lack an
120 ortholog of PRDM9, the locations of recombination hotspots are conserved over long
121 evolutionary time scales. Similarly, both the location and heats of recombination hotspots
122 are conserved across highly diverged yeast species, in which H3K4me3 marks are made
123 by a single gene without a DNA binding domain (Lam and Keeney 2015). In these taxa,
124 it remains unknown whether the coincidence of recombination with functional genomic
125 elements, such as TSSs and CGIs, is facilitated by specific binding motifs or simply by

126 greater accessibility of the recombination machinery to these genomic regions (Brick et
127 al. 2012; Auton et al. 2013; Choi et al. 2013; Lam and Keeney 2015; Singhal et al.
128 2015b). Even if there are specific motifs that increase rates of recombination near
129 functional genomic elements, they are likely to have important, pleiotropic consequences
130 on gene regulation (Nicolas et al. 1989). Thus, there may be a strong countervailing force
131 to the loss of hotspots by under-transmission of hotter alleles, leading to the evolutionary
132 stability of hotspots.

133 These observations sketch the outline of a general pattern, whereby species that
134 do not use PRDM9 to direct recombination target promoter-like features and have stable
135 fine-scale recombination landscapes, whereas those that employ PRDM9 tend to
136 recombine away from promoters and experience rapid turnover of hotspot locations. This
137 dramatic difference in the localization of hotspots and their evolutionary dynamics has
138 important evolutionary consequences for genome structure and base composition, for
139 linkage disequilibrium (LD) levels along the genome, as well as for introgression patterns
140 in naturally occurring hybrids (Fullerton et al. 2001; McVean et al. 2004; Duret and
141 Galtier 2009; Janousek et al. 2015). It is therefore important to establish the generality of
142 these two mechanisms and characterize their distribution across species.

143 To date, studies of fine-scale recombination are limited to a handful of organisms.
144 In particular, although it has been previously reported that the PRDM9 gene arose early
145 in metazoan evolution (Oliver et al. 2009), direct evidence of its role in recombination is
146 limited to placental mammals (mice, primates and more circumstantially cattle). It
147 remains unknown which species carry an intact ortholog and, more broadly, when
148 PRDM9-directed recombination is likely to have arisen. To address these questions, we
149 investigated the PRDM9 status of 225 species of vertebrates, using a combination of
150 genome sequences and RNAseq data.

151
152

153 **Results**

154

155 ***Initial identification of PRDM9 orthologs in vertebrates***

156 In order to identify which species have PRDM9 orthologs, we searched publically
157 available nucleotide and whole genome sequences to create a curated dataset of
158 vertebrate PRDM9 sequences. To this end, we implemented a *blastp*-based approach
159 against the *RefSeq* database, using human PRDM9 as a query sequence (see **Methods** for
160 details). We supplemented this dataset with 44 genes strategically identified from 30
161 whole genome assemblies and seven genes identified from de novo assembled
162 transcriptomes from testis of five species lacking genome assemblies (see **Methods** for
163 details). Neighbor joining (NJ) and maximum likelihood trees were built using identified
164 SET domains to distinguish *bona fide* PRDM9 orthologs from members of paralogous
165 gene families and to characterize the distribution of PRDM9 duplication events (**Figure**
166 **1- Figure Supplement 1; Figure 1- Figure Supplement 2**). Since the placement of the
167 major taxa used in our analysis is not controversial, in tracing the evolution of PRDM9
168 orthologs, we assumed that the true phylogenetic relationships between taxa are those
169 reported by several recent papers (synthesized by the TimeTree project; Hedges et al.
170 2015).

171 This approach identified 227 PRDM9 orthologs (**Supplementary File 1A;**
172 **Supplementary File 1B**), found in jawless fish, cartilaginous fish, bony fish,
173 coelacanths, turtles, snakes, lizards, and mammals. We confirmed the absence of PRDM9
174 in all sampled birds and crocodiles (Oliver et al. 2009; Singhal et al. 2015), the absence
175 of non-pseudogene copies in canids (Oliver et al. 2009; Munoz-Fuentes et al. 2011), and
176 additionally were unable to identify PRDM9 genes in amphibians (**Figure 1**), despite
177 targeted searches of whole genome sequences (**Supplementary File 1B**).

178 We further inferred an ancient duplication of PRDM9 in the common ancestor of
179 teleost fish, apparently coincident with the whole genome duplication that occurred in
180 this group (**Figure 1, Figure 2**). We used both phylogenetic methods and analysis of the
181 ZF structure to distinguish these copies (see **Figure 2- Figure Supplement 1, Methods**)
182 and refer to them as PRDM9 α and PRDM9 β in what follows. While PRDM9 β orthologs
183 were identified in each species of teleost fish examined, we were unable to identify

184 PRDM9 α type orthologs within three major teleost taxa, suggesting at minimum three
185 losses of PRDM9 α type orthologs within teleost fish (**Figure 2, Supplementary File**
186 **1A**). Several additional duplication events appear to have occurred more recently in other
187 vertebrate groups, including in jawless fish, cartilaginous fish, bony fish, and mammals
188 (**Supplementary File 1A**).

189

190 *Expression of PRDM9 in the germline of major vertebrate groups*

191 Since a necessary condition for PRDM9 to play a role in meiotic recombination is
192 for it to be expressed in the germline, we looked for PRDM9 in expression data from
193 testis tissues in order to confirm its presence. We focused on testis expression rather than
194 ovaries because although both obviously contain germline cells, preliminary analyses
195 suggested that meiotic gene expression is more reliably detected in testes (see **Methods**).
196 We selected 23 representative species, spanning each major vertebrate group, with
197 publically available testis expression or testis RNA-seq (**Supplementary File 2A**); we
198 also generated testis RNA-seq data for two species of bony fish (see **Methods**). In teleost
199 fish with both PRDM9 α and PRDM9 β genes, we were able to detect either the expression
200 of both orthologs or only expression of PRDM9 α orthologs. In species of teleost fish with
201 only PRDM9 β genes, we consistently identified expression of PRDM9 β genes. More
202 generally, we were able to identify PRDM9 expression in nearly all RNA-seq datasets
203 from species in which the genome carried a putative ortholog, the elephant shark
204 (*Callorhinus milii*) being the sole exception (**Supplementary File 2B; Supplementary**
205 **File 2C**).

206

207 *Confirmation of PRDM9 loss events*

208 Concerned that absences of PRDM9 observed in some species could reflect lower
209 quality genome assemblies rather than true loss events, we also used testis RNAseq data
210 to investigate putative losses of PRDM9 in amphibians and fish (PRDM9 α). To this end,
211 we relied on the fact that when PRDM9 is present, it is detectable in RNAseq data from
212 the whole testis of vertebrates (see above). Our approach was to analyze testis
213 transcriptome data from species lacking PRDM9 sequences in their genome assemblies,
214 using an analysis that is not biased by the genome assembly (see **Methods**). For each

215 species, we confirmed that the dataset captured the appropriate cell populations and
216 provided sufficient power to detect transcripts that are expressed during meiosis at levels
217 comparable to PRDM9 in mammals (**Figure 1- Figure Supplement 3, Supplementary**
218 **File 2B; Supplementary File 2D**). With this approach, we were able to find support for
219 the loss of PRDM9 in salamanders (*Cynops pyrrhogaster*, *Ambystoma mexicanum*) and
220 frogs (*Xenopus tropicalis*). Because of the paucity of amphibian genomes, however, it is
221 not clear whether or not these examples represent a widespread loss of PRDM9 within
222 amphibians or more recent, independent losses. Within bony fish, we were able to
223 confirm the three independent losses of PRDM9 α type orthologs in one species each of
224 percomorph (*Xiphophorus birchmanni*), cypriniform (*Danio rerio*) and osteoglossomorph
225 fish (*Osteoglossum bicirrhosum*). Thus, in all cases with sufficient power to detect
226 expression of PRDM9 in testes data, our findings were consistent with inferences based
227 on genome sequence data.

228

229 *Inferences of PRDM9 domain architecture*

230 PRDM9 orthologs identified in jawless fish, some bony fish, coelacanths, lizards,
231 snakes, turtles, and placental mammals have a complete domain structure, consisting of
232 KRAB, SSXRD and SET domains, as well as a C2H2 ZF array. The phylogenetic
233 relationships between these species suggest that a complete PRDM9 ortholog was present
234 in the common ancestor of vertebrates (**Figure 1**).

235 Despite its widespread taxonomic distribution, however, the complete domain
236 structure was not found in several of the 149 sampled lineages with PRDM9 orthologs
237 (**Figure 1**; in addition to the complete losses of the gene described above). Instances
238 include the absence of the SSXRD domain in some cartilaginous fish (see **Methods**);
239 absence of both KRAB and SSXRD domains in PRDM9 β orthologs (**Figure 1**) and in
240 PRDM9 α orthologs found distributed throughout the teleost fish phylogeny (**Figure 2**,
241 **Figure 2- Figure Supplement 1**); and the absence of the KRAB domain in monotremata
242 (*Ornithorhynchus anatinus*) and marsupial mammals (*Sarcophilus harrisii*, **Figure 1**;
243 **Supplementary File 1A**).

244 Because these frequent N-terminal losses could be the result of assembly or gene
245 prediction errors, we sought to confirm them by systematically searching genomes and

246 transcriptomes for evidence of these missing domains (see **Methods**). We required not
247 only that missing domains homologous to PRDM9 be absent from the genome in a whole
248 genome search, but also that the missing domain not be present in the transcriptome,
249 when other domains of PRDM9 were. This approach necessarily limits our ability to
250 verify putative losses when there are no suitable transcriptome data, but nonetheless
251 allowed us to confirm the losses of the KRAB and SSXRD domains in a PRDM9
252 ortholog from holostean fish (*Lepisosteus oculatus*), in all PRDM9 β orthologs from
253 teleost fish (**Figure 1**), in PRDM9 α orthologs that lost their complete domain structure in
254 several taxa of teleost fish (*Gadus morhua*, *Astyanax mexicanus*, *Ictalurus punctatus*,
255 *Esox lucius*; **Supplementary File 2C**), as well as losses of the KRAB domain in two
256 PRDM9 orthologs identified in monotremata (both in *O. anatinus*, **Supplementary File**
257 **2C**), indicating a minimum of six N-terminal domain losses within vertebrates.

258 For representative cases where we were able to confirm missing N-terminal
259 domains, we further investigated whether the truncated genes had become pseudogenes
260 by testing whether the ratio of nonsynonymous to synonymous substitutions in the SET
261 domain is significantly different than 1 (see **Methods**). In all cases of N-terminal
262 truncation, the partial PRDM9 shows evidence of functional constraint (i.e., $dN/dS < 1$,
263 where dN is the rate of amino-acid substitutions and dS of synonymous substitutions; see
264 **Methods** for more details). This conservation is most strikingly seen in teleost fish, in
265 which a partial PRDM9 ortholog has been evolving under constraint for hundreds of
266 millions of years (**Figure 1**, **Figure 2- Figure Supplement 1**, **Supplementary File 3A**).
267 These observations suggest that in these species, PRDM9 has an important function that
268 it performs without KRAB or SSXRD domains. Moreover, these cases provide
269 complementary observations to full PRDM9 knockouts in amphibians and archosaurs,
270 allowing the roles of specific domains to be dissected.

271

272 ***Evidence for rapid evolution of PRDM9 binding specificity***

273 Rapid evolution of the PRDM9 ZF array has been reported previously in all
274 species with evidence for PRDM9-directed recombination, including cattle, apes and
275 mice. While it is not known whether this rapid evolution is a necessary consequence of its
276 role in recombination, plausible models suggest it is likely to be (see Introduction). If so,

277 we expect species with PRDM9-directed recombination to show evidence for rapidly-
278 evolving PRDM9 ZF arrays and can use this feature to hone in on the subset of PRDM9
279 orthologs most likely to play a role in recombination.

280 To this end, we characterized the rapid evolution of the PRDM9 ZF in terms of
281 the proportion of amino acid diversity within the ZF array that occurs at DNA-binding
282 sites (using a modification of the approach proposed by Oliver et al. 2009). This
283 summary statistic is sensitive to both rapid amino acid evolution at DNA binding sites
284 and concerted evolution between the individual ZFs (see **Methods**). Using this statistic,
285 placental mammals that have PRDM9-directed recombination show exceptionally high
286 rates of evolution of the PRDM9 ZF compared to other ZFs (**Table 1**; Baudat et al. 2010;
287 Myers et al. 2010; Parvanov et al. 2010). Moreover, two of six cattle PRDM9 orthologs
288 that we identified were previously associated with interspecific variation in
289 recombination phenotypes (**Supplementary File 3B**; Sandor et al. 2012; Ma et al. 2015),
290 and both are seen to be rapidly evolving (**Table 1, Supplementary File 3B**).

291 In addition to placental mammals, PRDM9 orthologs in jawless fish, some bony
292 fish (Salmoniformes, Esociformes, Elopomorpha), turtles, snakes, lizards, and
293 coelacanths show similarly elevated values of this statistic (**Figure 1- Figure**
294 **Supplement 4**). In fact, PRDM9 is the most rapidly evolving ZF gene genome-wide in
295 most species in these taxa and all PRDM9 orthologs with the complete domain structure
296 were in the top 5% of the most rapidly evolving ZFs in their respective genomes (**Table**
297 **1, Supplementary File 3B**). In contrast, evidence of such rapid evolution is absent from
298 other taxa of bony fish, including all PRDM9 β orthologs and partial PRDM9 α orthologs,
299 as well as from the putatively partial PRDM9 orthologs found in the elephant shark, the
300 Tasmanian devil, and in several species of placental mammals (see **Methods** for details).
301 We only observed one instance (little brown bat, *Myotis lucifugus*) in which a partial
302 PRDM9 ortholog was evolving unusually rapidly (**Table 1**); in this case, we were unable
303 to confirm the loss of the missing KRAB domain (see **Methods**), so it remains possible
304 this ortholog is in fact intact. In summary, with one possible exception, species show
305 evidence of rapid evolution of the ZF binding affinity if and only if they carry the intact
306 PRDM9 ortholog found in placental mammals. This concordance of rapid evolution with
307 the complete domain structure is highly unlikely by chance (taking into account the

308 phylogenetic relationship between orthologs, $p < 10^{-6}$; see **Methods**). Assuming that
309 rapid evolution of the ZF is indicative of PRDM9-directed recombination, these
310 observations carry two implications: KRAB and SSXRD domains are required for this
311 role and non-mammalian species such as turtles or snakes also use PRDM9 to direct
312 recombination.

313

314 *Analysis of SET domain catalytic residues*

315 While partial orthologs of PRDM9 have lost one or both of their N-terminal
316 domains, they retain the SET and ZF domains known to play a role in recombination, are
317 under purifying selection, and are expressed in testis. In principle then, these partial
318 orthologs could still play a role in directing recombination. To evaluate this possibility,
319 we started by examining whether the catalytic activities of the SET domains of partial
320 PRDM9 orthologs are conserved. We did so because the catalytic specificities of PRDM9
321 are believed to be important to its role in directing recombination: two marks made by the
322 SET domain of PRDM9, H3K4me3 and H3K36me3, are associated with hotspot activity
323 in mammals (Powers et al. 2016; Grey et al. 2017; Yamada 2017; Getun et al. 2017) and
324 the human PRDM9 is unusual in being able to add methyl groups to different lysine
325 residues of the same nucleosomes, when most other methyltransferase genes are
326 responsible for only a single mark (Eram et al. 2014; Powers et al. 2016).

327 Specifically, we focused on three tyrosine residues shown to be important for the
328 catalytic specificities of the human PRDM9 gene (Y276, Y341 and Y357; see **Methods**
329 and **Supplementary File 1A**; Wu et al. 2013) and asked if those residues were conserved
330 across vertebrates. Loss of individual residues is not necessarily evidence for loss of
331 catalytic activity, as compensatory changes may have occurred. For example, a
332 substitution at Y357 of PRDM7 has led to the loss of H3K36me3 specificity, but
333 H3K4me3 activity appears to have been retained through compensatory substitutions
334 (Blazer et al. 2016). Nonetheless, PRDM9 orthologs with substitutions at these residues
335 are unlikely to utilize the same catalytic mechanisms as human PRDM9 for any
336 methyltransferase activity that they retain.

337 We find that each of the three residues is broadly conserved across the vertebrate
338 phylogeny, with substitutions observed in only 57 of 227 PRDM9 orthologs, including 11

339 genes from placental mammals and 46 genes from bony fish. Strikingly, however, none
340 of these substitutions occur in a complete PRDM9 ortholog containing KRAB, SSXRD,
341 SET and ZF domains. Within mammals, the majority of PRDM9 orthologs that
342 experienced these substitutions are lacking the ZF array entirely, including eight PRDM7
343 genes from primates, which share a substitution at Y357, and one PRDM9 ortholog from
344 a bat (*Miniopterus natalensis*) that carries a substitution at Y276. Others are lacking the
345 KRAB domain, including PRDM9 orthologs identified from a lemur (*Galeopterus*
346 *variegatus*) and a rodent (*Octodon degus*) carry substitutions at Y276 and Y357,
347 respectively.

348 Within bony fish, we identified 47 PRDM9 orthologs with substitutions at one or
349 more of these residues, including the partial PRDM9 ortholog from holosteans (see
350 above) and all PRDM9 β orthologs in teleosts (**Supplementary File 1A**). The distribution
351 of substitutions at these residues within PRDM9 β genes suggests that numerous
352 independent substitution events have occurred in this gene family following the loss of
353 KRAB and SSXRD domains (**Figure 3**). In contrast, no substitutions were observed at
354 these residues in any PRDM9 α orthologs, regardless of their domain architecture. These
355 observations could be consistent with a lack of constraint on the ancestral
356 methyltransferase activities of PRDM9 in PRDM9 β genes after the PRDM9 α /PRDM9 β
357 duplication event (or conceivably an indication that there has been convergent evolution
358 towards a new functional role). Thus, PRDM9 β genes not only lack KRAB and SSXRD
359 domains, they likely lack some methyltransferase activity of the SET domain.

360

361 ***Fish species with a partial PRDM9 ortholog share broad patterns of recombination***
362 ***with species that lack PRDM9***

363 To more directly test the hypothesis that the partial ortholog of PRDM9 does not
364 direct recombination, we examined patterns of crossing-over in naturally-occurring
365 swordtail fish hybrids (*X. birchmanni* x *X. malinche*; see Methods). Like other
366 percomorphs, swordtail fish have a PRDM9 β type gene that lacks the KRAB and SSXRD
367 domains and a slowly evolving ZF array with testis-specific expression (**Figure 4**;
368 **Figure 4 – Figure 4 Supplement 1**); they further carry substitutions at two catalytic
369 residues of the SET domain (Y341F and Y357P), as well as at residues of the SET

370 domain implicated in H3K4me2 recognition (see **Methods**). Based on these features, we
371 predict that they should behave like a PRDM9 knockout, with no increase in
372 recombination around the PRDM9 motif.

373 To test these predictions, we collected ~1X genome coverage from 268 natural
374 hybrids and inferred crossover events from ancestry switchpoints between the two
375 parental species using a hidden Markov model (see **Methods**). By this approach, we find
376 recombination rates to be elevated near TSSs and CGIs, two promoter-like features
377 (**Figure 4**; **Figure 4 - Figure Supplement 2**). Moreover, and in contrast to what is
378 observed in species with PRDM9-mediated recombination (**Figure 4- Figure**
379 **Supplement 3**), there is no elevation in recombination rates near computationally-
380 predicted PRDM9 binding sites (**Figure 4F**). These patterns resemble those previously
381 reported for birds lacking PRDM9 (Singhal et al. 2015).

382 In addition, we performed native chromatin Chip-seq with an H3K4me3 antibody
383 in *X. birchmanni* testis and liver tissue. Consistent with a role for H3K4me3 in inducing
384 DSBs, recombination is increased around H3K4me3 peaks (testing the association with
385 distance, rho=-0.072, p=2.3e-69; **Figure 4**), an effect that remains significant after
386 correcting for distance to TSSs and CGIs (rho=-0.026, p=5.4e-10). In fact, the increase in
387 recombination rate near the TSS is almost completely explained by the joint effects of
388 proximity to H3K4me3 peaks and CGI (TSS with both: rho= -0.009, p=0.02). Windows
389 that contain testis-specific H3K4me3 peaks have significantly observed higher
390 recombination rates than those that contain liver-specific H3K4me3 peaks (**Figure 4 –**
391 **Figure 4 Supplement 4**). However, H3K4me3 peaks in the testis are not enriched for the
392 computationally predicted PRDM9 motifs (**Figure 4**), nor do they overlap with PRDM9
393 motifs in the testis more than the liver (see Methods). Conversely, sequence motifs
394 associated with testis-specific H3K4me3 peaks do not resemble the predicted PRDM9
395 motif (**Figure 4 – Figure 4 Supplement 5**). Thus, there is no evidence that PRDM9 lays
396 down the H3K4me3 marks associated with an increase in recombination.

397

398 ***Recombination landscapes in vertebrates with and without PRDM9***

399 To put the genomic patterns of recombination in swordtail fish in an explicit
400 comparative framework, we re-analyzed patterns of recombination near TSSs and CGIs

401 in previously published genetic maps based on LD data from three species without
402 functional PRDM9 genes (dog, zebra finch and long-tailed finch) and three species
403 known to use PRDM9-mediated recombination (human, gorilla and mouse), as well as
404 using a pedigree-based genetic map for one species with a complete PRDM9 ortholog,
405 but for which no direct evidence of PRDM9's role in recombination has yet been reported
406 (sheep; see **Methods** for details and references).

407 Among species with complete PRDM9 genes, recombination rates are either
408 weakly reduced near TSSs and CGIs or similar to what is seen in nearby windows
409 (**Figure 5**; see **Figure 5 – Figure Supplement 1** for results with genetic maps based on
410 pedigrees or admixture switches instead of LD data in humans and dogs). In contrast, in
411 all species lacking PRDM9 and swordtail fish, the recombination rate is notably
412 increased in windows overlapping either a TSS or CGI relative to nearby windows.
413 Quantitative comparisons are difficult because of the varying resolution of the different
414 genetic maps. Nonetheless, these results indicate that patterns of recombination near
415 TSSs and CGIs differ between species carrying complete PRDM9 orthologs and species
416 lacking PRDM9 altogether, and that swordtail fish exhibit patterns of recombination
417 similar to species that completely lack PRDM9, despite the presence of a partial ortholog.

418
419

420 **Discussion**

421

422 Based on our reconstruction of 227 PRDM9 orthologs across the vertebrate
423 phylogeny, we inferred that the ancestral domain architecture of PRDM9 consisted of
424 KRAB, SSXRD and SET domains followed by a C2H2 ZF array, and that this complete
425 architecture was likely already in place in the common ancestor of vertebrates.

426 Moreover, even though to date only the functions of the SET domain and C2H2
427 ZF array have been connected to the role of PRDM9 in directing recombination, the
428 evolutionary patterns uncovered here suggest that all four domains are important. The
429 first line of evidence is that there is no evidence of rapid evolution of the ZF domains in
430 PRDM9 orthologs from which KRAB and SSXRD domains have apparently been lost
431 (including a subset of species in which the catalytic activity of the SET domain is
432 seemingly conserved), suggesting that there has not been rapid evolution of binding
433 specificity. In contrast, we find evidence of rapid evolution of the PRDM9 ZF in all
434 species that have KRAB, SSXRD, SET, and ZF domains. Since plausible models suggest
435 that the rapid evolution of PRDM9 binding affinity is a consequence of the role of this
436 gene in directing recombination (see **Introduction**), this observation suggests that all
437 four domains are required for this role.

438 The second piece of evidence is that swordtail fish with a truncated copy of
439 PRDM9 that is missing KRAB and SSXRD domains behave like PRDM9 knockouts in
440 their fine-scale recombination patterns. It is unclear if this behavior can be attributed to
441 loss of the N terminal domains, since two key catalytic residues within the SET domain
442 were also substituted in this species. We note, however, that substitutions at catalytic
443 residues are only seen in PRDM9 genes that have lost KRAB and/or SSXRD domains or
444 have lost the ZF entirely. When the ZF is lost, PRDM9 obviously cannot induce DSBs by
445 binding DNA and its new role may not require the same methyltransferase specificities.
446 We speculate that the absence of KRAB and SSXRD domains in a PRDM9 ortholog may
447 similarly signify that PRDM9 is no longer used to direct recombination and lead to
448 reduced constraint on the catalytic activities of the SET domain. Consistent with this
449 hypothesis, a recent paper suggests that the KRAB domain may play a role in recruiting
450 the recombination machinery (Parvanov et al. 2016).

451 If the partial ortholog of PRDM9 is not used to direct recombination at all, then
452 the overall conservation of the protein points to another role of the gene. In that regard,
453 we note that partial PRDM9 orthologs share their domain architecture with other
454 members of the PRDM gene family, many of which act as transcription factors (Hayashi
455 et al. 2005; Vervoort et al. 2016).

456 Conversely, if the presence of all four domains, conservation of catalytic
457 residues, and the rapid evolution of the ZF array are sufficient indications of PRDM9-
458 directed recombination, then the role of PRDM9 in directing recombination appears to
459 have originated before the diversification of vertebrates. It would follow that many non-
460 mammalian vertebrate species, such as snakes, use the gene to determine the location of
461 recombination hotspots. One hint in that direction is provided by the high allelic diversity
462 seen in the ZF within a python species (*Python bivittatus*), reminiscent of patterns
463 observed in apes (Schwartz et al. 2014; **Figure 1- Figure Supplement 5**). Assessing the
464 role of PRDM9 in directing recombination in these species is a natural next step in
465 understanding the evolution of recombination mechanisms.

466 It further appears that the intact PRDM9 has often been duplicated, with more
467 than one copy associated with recombination rate variation in cattle (Sandor et al. 2012;
468 Ma et al. 2015). Based on the RAxML tree of the SET domain, we count 55 independent
469 cases of duplications. How commonly more than one copy of PRDM9 retains a role in
470 directing recombination remains to be investigated.

471 More generally, the distribution of PRDM9 across vertebrates raises the question
472 of why species switch repeatedly from one recombination mechanism to another.
473 Although PRDM9-directed recombination clearly confers enough of an advantage for it
474 to be widely maintained in vertebrates, at least six taxa of vertebrates carry only partial
475 PRDM9 orthologs and the gene has been lost entirely at least three times (based on 227
476 orthologs; **Figure 1, Figure 2**). Thus, PRDM9 is not essential to meiotic recombination
477 in the sense that SPO11 is, for example (Lam and Keeney 2014). Instead, the role of
478 PRDM9 is perhaps best envisaged as a classic, trans-acting recombination rate modifier
479 (Otto and Barton 1997; Otto and Lenormand 2002; Coop and Przeworski 2007), which
480 was favored enough to be adopted at some point in evolution, but not so strongly or
481 stably as to prevent frequent losses.

482 In this regard, it is worth noting that in mammalian species studied to date,
483 recombination rates are lower near promoters than in species lacking PRDM9 (Myers et
484 al. 2005; Coop et al. 2008). Because recombination hotspots have higher rates of point
485 mutations, insertions and deletions, and experience GC-biased gene conversion, there
486 may be an advantage conferred by directing recombination to non-genic regions.
487 Recombination at the TSS could have the further disadvantage of uncoupling coding and
488 regulatory variants, potentially uncovering negative epistasis, and therefore leading to
489 indirect selection for decreased recombination at the TSS. Alternatively (but non
490 mutually-exclusively), because PRDM9 binding motifs are strongly associated with
491 certain transposable element classes in mammals (Myers et al. 2008), the role of PRDM9
492 in recombination could be related to the regulation of certain families of transposable
493 elements. With a more complete picture of recombination mechanisms and their
494 consequences across the tree of life, these hypotheses can start to be tested in an
495 evolutionary context.

496

497

498

499 **Methods**

500

501 ***Identification of putative PRDM9 orthologs from the RefSeq database***

502 As a first step in understanding the distribution of PRDM9 in vertebrates, we
503 identified putative PRDM9 orthologs in the RefSeq database. We used the *blastp*
504 algorithm (Altschul et al. 1990) using the *Homo sapiens* PRDM9 sequence, minus the
505 rapidly evolving tandem ZF array, with an e-value threshold of 1e-5. We downloaded
506 GenPept files and used Batch Entrez to retrieve the corresponding GenBank files
507 (September 2016). The longest transcript for each locus and amino acid and DNA
508 sequences corresponding to the KRAB, SSXRD and SET domains of these sequences (as
509 annotated by the Conserved Domain Database; Marchler-Bauer et al. 2015), were
510 downloaded using a R script (**Supplementary Script 1**). The retrieved SET domain
511 sequences, an additional 44 retrieved from whole genome assemblies, as well as seven
512 retrieved from RNAseq datasets for five species without sequenced genomes (see
513 ***Predicting PRDM9 orthologs from whole genome sequences***) were input into
514 ClustalW2 (Larkin et al. 2007), in order to generate a neighbor-joining (NJ) guide tree
515 (see **Figure 1- Figure Supplement 2**). This approach was used to identify and remove
516 genes that cluster with known PRDM family genes from humans and that share the SET
517 domain of PRDM9 but were previously reported to have diverged from PRDM9 before
518 the common ancestor of vertebrates (Vervoort et al. 2016); see ***Phylogenetic Analysis of***
519 ***PRDM9 orthologs and related gene families***).

520

521 ***Predicting PRDM9 orthologs from whole genome sequences***

522 There were a number of groups not included in the RefSeq database or for which
523 we were unable to identify PRDM9 orthologs containing the complete domain
524 architecture. For 33 representative species from these groups, we investigated whether we
525 could find additional PRDM9 orthologs in their whole genome assemblies (see
526 **Supplementary File 1A; Supplementary File 1B**). To this end, we ran *tblastn* against
527 the whole genome assembly, using the PRDM9 ortholog from the most closely related
528 species that contained a KRAB domain, a SET domain, and at least one ZF domain
529 (**Supplementary File 1B**). The number of hits to each region was limited to ten, and

530 gene models were only predicted when a *blast* hit to the SET domain was observed with
531 an e-value threshold of 1e-10.

532 When a single contig was identified containing an alignment to the full length of
533 the query sequence, this contig was input into Genewise, along with the PRDM9 protein
534 sequence from a species with a high quality ortholog (using a closely related species
535 where possible), in order to create a new gene model. When PRDM9 domains were found
536 spread across multiple contigs, we needed to arrange them in order to generate the proper
537 sequences of the genomic regions containing PRDM9 orthologs from each species. When
538 linkage information was available and we observed the presence of PRDM9 domains on
539 linked contigs, we arranged the sequences of these contigs accordingly, with gaps padded
540 with 100 Ns, before inputting them into Genewise. In cases where linkage information
541 was not available, our approach differed depending on whether or not we identified more
542 than one hit to each region of the query sequence. In species where there appeared to be
543 only one PRDM9 ortholog, we arranged the contigs according to the expected
544 arrangements of the domains, though did not include any ZF arrays unless they were
545 found on the same contig as the complete SET domain because the repeat structure of
546 these domains makes homology difficult to infer. In species with more than one PRDM9
547 ortholog, we did not attempt to construct any gene models not supported by linkage or by
548 transcripts identified from the same species (see ***Confirming the expression or absence***
549 ***of PRDM9 in the testes of major phylogenetic groups; Supplementary File 1B*** for
550 details).

551 The positions of KRAB, SSXRD and SET domains for each gene model were
552 annotated using *CD-blast* (Domain Accessions smart00317, pfam00856, cl02566,
553 pfam09514, pfam01352, cd07765, smart00349). This approach resulted in the
554 identification of additional PRDM9 orthologs containing at minimum the SET domain, in
555 two jawless fish, two cartilaginous fish, nine bony fish, one monotreme, two marsupials,
556 one turtle, four lizards, and eight snakes (**Supplementary File 1A**). We were unable to
557 detect PRDM9 orthologs in one lizard (*Anolis carolinensis*), or in any of three
558 amphibian species (**Supplementary File 1B**). We used RNA-seq data to investigate
559 whether these negative findings are due to genome assembly quality or reflect true losses
560 (see below).

561

562 ***Phylogenetic Analysis of PRDM9 orthologs and related gene families***

563 To understand the evolution of PRDM9 within vertebrates, we used a
564 phylogenetic approach. We first built an alignment of the amino acid sequences of
565 putative PRDM9 and PRDM11 SET domains using Clustal Omega (Sievers et al. 2011).
566 We included genes clustering with PRDM11 because it had been reported that PRDM11
567 arose from a duplication event of PRDM9 in the common ancestor of bony fish and
568 tetrapods (Vervoort et al. 2016), and we were interested in identifying any PRDM9
569 orthologs carried by vertebrate species that may precede this duplication event. The
570 alignment coordinates were then used to generate a nucleotide alignment, which was used
571 as input into the program RAxML (v7.2.8; Stamatakis 2006). We performed 100 rapid
572 bootstraps followed by maximum likelihood estimation of the tree under the General
573 Reversible Time substitution model, with one partition for each position within a codon.
574 The resulting phylogeny contained monophyletic groups corresponding to the PRDM9
575 and PRDM11 duplication event, with 100% bootstrap support (**Figure 1- Figure**
576 **Supplement 1**). These groups were used to label each putative ortholog as PRDM9 or
577 PRDM11. Only jawless fish have PRDM9 orthologs basal to this duplication event,
578 suggesting PRDM11 arose from PRDM9 before the common ancestor of cartilaginous
579 fish and bony vertebrates. We observed at least one PRDM11 ortholog in each of the
580 other vertebrate species examined.

581 Within teleost fish, we identified two groups of PRDM9 orthologs, which we
582 refer to as PRDM9 α and PRDM9 β (**Figure 2- Figure Supplement 1**). While the
583 bootstrap support for the monophyly of the two groups is only 75% for PRDM9 α and
584 54% for PRDM9 β , the potential duplication event suggested by this tree is coincident
585 with the whole genome duplication event known to have occurred in the common
586 ancestor of teleost fish (Taylor et al. 2003). Moreover, the phylogenetic grouping based
587 on the SET domain is concordant with general differences in the domain architectures
588 between the two orthologs: In contrast to PRDM9 α , PRDM9 β genes have derived ZF
589 array structures, containing multiple tandem ZF arrays spread out within the same exon
590 (**Figure 2- Figure Supplement 1**) and are always found without the KRAB and SSXRD
591 domains, whereas PRDM9 α genes generally have a single tandem array of ZFs consistent

592 with the inferred ancestral domain architecture, and occasionally have KRAB and
593 SSXRD domains (**Figure 2**).

594

595 ***Confirming the expression or absence of PRDM9 in the testes of major phylogenetic***
596 ***groups***

597 A necessary condition for PRDM9 to be involved in recombination is its
598 expression in meiotic cells. For groups of taxa in which we detected a PRDM9 ortholog,
599 we evaluated whether this ortholog was expressed in the testes, using a combination of
600 publically available RNAseq data and RNAseq data that we generated. Additionally, in
601 groups of species where PRDM9 appeared to be absent from the genome, we used
602 publically available RNAseq data to confirm the absence of expression of PRDM9. In
603 both cases, we used a stringent set of criteria to try to ensure that the absence of
604 expression was not due to data quality issues (see details below).

605 We downloaded data for jawless fish, cartilaginous fish, bony fish, coelacanth,
606 reptile, marsupial and monotreme species for which Illumina RNAseq data were
607 available (**Supplementary File 2A; Supplementary File 2C; Supplementary File 2D**).
608 We additionally generated RNAseq data for two percomorph fish species, *Xiphophorus*
609 *birchmanni* and *X. malinche* (see below). Downloaded reads were converted to fastq
610 format using the sratoolkit (v2.5.7; Leinonen et al. 2011) and trimmed for adapters and
611 low quality bases (Phred <20) using the program cutadapt (v1.9.1;
612 <https://cutadapt.readthedocs.io/en/stable/>). Reads shorter than 31 bp post-quality
613 trimming were discarded. The program `interleave_fastq.py` was used to combine mate
614 pairs in cases where sequence data were paired-end (Crawford 2014;
615 <https://gist.github.com/ngcrawford/2232505>). De-novo transcriptome assemblies were
616 constructed using the program `velvet` (v1.2.1; Zerbino and Birney 2008) with a kmer of
617 31; `oases` (Schulz et al. 2012; v0.2.8) was used to construct transcript isoforms.
618 Summaries of these assemblies are available in **Supplementary File 2A**.

619 In order to identify potential PRDM9 transcripts in each of 24 assembled
620 transcriptomes, we implemented `tblastn` using the human PRMD9 sequence, minus the
621 ZF domain, as the query sequence, with an e-value threshold of 1e-5. The identified
622 transcripts were extracted with a custom script and blasted to our dataset of all PRDM

623 genes (**Supplementary Script 2**). If the best blast hit was a PRDM9 ortholog, we
624 considered PRDM9 expression in the testis to be confirmed (see results in
625 **Supplementary File 2C**). For five species lacking genome assemblies, we extracted
626 PRDM9 orthologs with best *blast* hits to human PRDM9/7 and included these in our
627 phylogenetic analyses (see ***Phylogenetic Analysis of PRDM9 orthologs and related gene***
628 ***families***).

629 Failure to detect PRDM9 could mean that PRDM9 is not expressed in that tissue,
630 or that data quality and sequencing depth are too low to detect its expression. To
631 distinguish between these possibilities, we used other recombination-related genes as
632 positive controls, reasoning that if expression of several other conserved recombination-
633 related genes were detected, the absence of PRDM9 would be more strongly suggestive
634 of true lack of expression. Eight recombination-related genes are known to be conserved
635 between yeast and mice (Lam and Keeney 2014). We used the subset of seven that could
636 be reliably detected in whole genome sequences, and we asked which transcriptomes had
637 reciprocal best *tblastn* (e-value < 1e-5) hits to all of these proteins, using query sequences
638 from humans (**Supplementary File 2A; Supplementary File 2D**). In addition, in order
639 to assess whether PRDM9 expression might simply be lower than that of other meiotic
640 genes, we quantified absolute expression of PRDM9 and the seven conserved
641 recombination-related proteins in whole testes, using data from three major taxa (bony
642 fish, mammals, and reptiles); see ***Analysis of PRDM9 expression levels and expression***
643 ***levels of other conserved recombination-related genes*** for more details. Together, these
644 results suggest that not detecting PRDM9 in whole testes transcriptomes provides support
645 for its absence.

646

647 ***RNA extraction and sequencing of liver and gonad tissue from swordtail fish***

648 Three *Xiphophorus birchmanni* and three *X. malinche* were collected from the
649 eastern Sierra Madre Oriental in the state of Hidalgo, Mexico. Fish were caught using
650 baited minnow traps and were immediately euthanized by decapitation (Texas A&M
651 AUP# - IACUC 2013-0168). Testis, ovaries, and liver were dissected and stored at 4°C in
652 RNAlater. Total RNA was extracted from testis, ovary and liver tissue using the Qiagen
653 RNeasy kit (Valencia, CA, USA) following the manufacturer's protocol. RNA was

654 quantified and assessed for quality on a Nanodrop 1000 (Nanodrop technologies,
655 Willmington, DE, USA) and approximately 1 µg of total RNA was used input to the
656 Illumina TruSeq mRNA sample prep kit. Samples were prepared following the
657 manufacturer's protocol with minor modifications. Briefly, mRNA was purified using
658 manufacturer's beads and chemically fragmented. First and second strand cDNA was
659 synthesized and end repaired. Following A-tailing, each sample was individually
660 barcoded with an Illumina index and amplified for 12 cycles. The six libraries were
661 sequenced on the HiSeq 2500 at the Lewis Sigler Institute at Princeton University to
662 collect single-end 150 bp reads, while single-end 100 bp data was collected on the HiSeq
663 4000 at Weill Cornell Medical College for all other samples (SRA Accessions:
664 SRX2436594 and SRX2436597). Reads were processed and a de novo transcriptome
665 assembled for the highest coverage testis library following the approach described above
666 for publicly available samples. Details on assembly quality are available in
667 **Supplementary File 2A**. Other individuals were used in analysis of gene expression
668 levels (see next section).

669

670 ***Analysis of PRDM9 expression levels and expression levels of other conserved
671 recombination-related genes***

672 To determine whether some of the genes in our conserved recombination-related
673 gene set were expressed at similar levels to PRDM9, implying similar detection power,
674 we examined expression levels of these genes in three species representing the bony fish,
675 reptilian, and mammalian taxa (*Xiphophorus malinche*, *Pogona vitticeps*, and *Homo
676 sapiens*).

677 To quantify expression in *X. malinche*, we mapped trimmed reads from testes
678 RNAseq libraries that we generated from three individuals to the *X. maculatus* reference
679 genome (v4.4.2; Schartl et al. 2013; Amores et al. 2014) using bwa (v0.7.10; Li and
680 Durbin 2009). The number of trimmed reads per individual ranged from 9.9-27.5 million.
681 We used the program eXpress (v1.5.1; Roberts et al. 2011) to quantify fragments per
682 kilobase of transcript per million mapped reads (FPKM) for each gene, and extracted the
683 genes of interest from the results file based on their ensembl gene id. eXpress also gives
684 confidence intervals on its estimates of FPKM.

685 For the bearded lizard *Pogona vitticeps*, we only had access to one publically
686 available testis-derived RNAseq library. We followed the same steps used in analysis of
687 swordtail FPKM except that we mapped to the transcriptome generated from the data (see
688 main text) and identified transcripts belonging to recombination-related gene sets using
689 the reciprocal best *blast* hit approach described above.

690 Several publically available databases already exist for tissue specific expression
691 in humans. We downloaded the “RNA gene dataset” from the Human Protein Atlas (v15,
692 <http://www.proteinatlas.org/about/download>). This dataset reports average FPKM by
693 tissue from 122 individuals. We extracted genes of interest from this data file based on
694 their Ensembl gene id.

695 Examination of these results demonstrated that other meiotic genes (2-5) in each
696 species had expression levels comparable to PRDM9 (**Figure 1- Figure Supplement 3**).
697 This finding suggests that these genes are appropriate positive controls, in that detecting
698 their expression but not that of PRDM9 provides evidence against expression of PRDM9
699 in testes.

700

701 ***Confirmation of PRDM9 domain loss and investigation of loss of function***

702 In addition to complete losses of PRDM9, we were unable to identify one or more
703 functional domains of PRDM9 in orthologs identified from the platypus, Tasmanian
704 devil, elephant shark, all bony fish and several placental mammals.

705 To ask whether the missing PRDM9 domains were truly absent from the genome
706 assembly, we first used a targeted genome-wide search. To this end, we performed a
707 *tblastn* search of the genome against the human PRDM9 ortholog with an e-value of 1e-
708 10. For all *blast* hits, we extracted the region and 2 Mb flanking in either direction,
709 translated them in all six frames

710 (http://cgpdb.ucdavis.edu/DNA_SixFrames_Translation/), and performed an *rpsblast*
711 search of these regions against the CDD (database downloaded from NCBI September
712 2016) with an e-value of 100 to identify any conserved domains, even with weakly
713 supported homology. We extracted all *rpsblast* hits to the missing functional domain
714 (SET CDD id: smart00317, pfam00856, cl02566; SSXRD CDD id: pfam09514; KRAB
715 domains pfam01352, cd07765, smart00349) and used them as query sequences in a

716 *blastp* search against all KRAB, SSXRD and SET containing proteins in the human
717 genome. If PRDM9 or PRMD7 was the top *blast* hit in this search, we considered that the
718 missing domain could be a result of assembly or gene model prediction error (if not, we
719 investigated the potential loss of these domains further). This approach allowed us to rule
720 out genome-wide losses of PRDM9 domains in nine out of 14 species of mammals for
721 which our initial approach had failed to identify complete PRDM9 orthologs. In each
722 case, we checked whether or not the identified domains were found adjacent to any of our
723 predicted gene models and adjusted the domain architecture listed for these RefSeq genes
724 accordingly in our dataset (see **Supplementary File 1A**). In five species of mammals
725 (Tasmanian devil, three bat species, and the aardvark), we only identified a partial
726 PRDM9 ortholog, but we were unable to confirm the loss of domains using RNAseq data
727 (see next section). Within bats, each partial gene model starts within 500 bp of an
728 upstream gap in the assembly. Moreover, we were able to identify a KRAB domain
729 corresponding to PRDM9 from a closely related species of bat (*Myotis brandtii*). Thus,
730 we believe that in the case of bats, these apparent domain losses are due to assembly
731 errors or gaps.

732 For species with available RNAseq data from taxa in which we predicted PRDM9
733 N-terminal truncation based on our initial analyses, we sought to confirm the domain
734 structure observed in the genome with de novo transcriptome assemblies from testis
735 RNAseq (described above). As before, we only considered transcriptomes that passed our
736 basic quality control test (**Supplementary File 2D**). Because RNAseq data are not
737 available for all species with genome assemblies, we were only able to perform this
738 stringent confirmation in a subset of species (**Supplementary File 2C**). As a result, we
739 consider cases where N-terminal losses are confirmed in the genome as possible losses
740 but are most confident about cases where N-terminal losses are observed both in the
741 genome and transcriptome.

742 To examine the transcripts of PRDM9 orthologs from the transcriptome
743 assemblies (**Supplementary File 2A**), for each domain structure, we translated each
744 transcript with a *blast* hit to the human PRDM9 in all six frames and used *rpsblast*
745 against all of these translated transcripts, with an e-value cutoff of 100 (as described
746 above). Finally, we performed a reciprocal nucleotide *blast* (*blastn*; e-value cutoff 1e-20)

747 to confirm that these transcripts were homologous to the PRDM9 ortholog identified
748 using phylogenetic methods in these taxa. Results of this analysis can be found in
749 **Supplementary File 2C**. In summary, there were two cases where the transcriptomes
750 supported additional domain structures not found in the whole genome sequence
751 (**Supplementary File 2C**): a PRDM9 ortholog from the spotted gar (*Lepisosteus*
752 *oculatus*) that was observed to have a KRAB domain not identified in the genome
753 sequence, and a PRDM9 α ortholog from the Atlantic salmon (*Salmo* *salar*) that was
754 observed to have both KRAB and SSXRD domains not identified in the genome search.
755 In all other cases, we confirmed the losses of either the KRAB or SSXRD domains,
756 including: (i) PRDM9 β orthologs missing KRAB and SSXRD domains in all species of
757 teleost fish expressing these orthologs (**Supplementary File 2B, Supplementary File**
758 **2C**) (ii) PRDM9 α orthologs missing KRAB and SSXRD domains identified from
759 *Astyanax mexicanus*, *Esox lucius*, *Gadus morhua*, and *Ictalurus punctatus*, and (iii) loss
760 of the KRAB domain from one PRDM9 ortholog in monotremata (*O. anatinus*) and both
761 KRAB and SSXRD domains from the other ortholog in this species.
762 For all groups in which we confirm that there is only a partial PRDM9 ortholog based on
763 the above analyses, we asked whether the PRDM9 gene in question has likely become a
764 pseudogene (as it has, for example, in canids; Oliver et al. 2009; Munoz-Fuentes et al.
765 2011), in which case the species can be considered a PRDM9 knockout. Though such
766 events would be consistent with our observation of many losses of PRDM9, they would
767 not be informative about the role of particular PRDM9 domains in recombination
768 function. For this analysis, we aligned the SET domain of the PRDM9 coding nucleotide
769 sequence to a high-quality PRDM9 sequence with complete domain structure from the
770 same taxon using Clustal Omega (see **Supplementary File 3A**), except for the case of
771 PRDM9 β in bony fish and the PRDM9 ortholog from cartilaginous fish, where such a
772 sequence was not available. In the case of PRDM9 β , we compared the sequence between
773 *X. maculatus* and *A. mexicanus*, sequences that are >200 million years diverged (Hedges
774 et al. 2015). In the case of cartilaginous fish, we used the sequence from *R. typus* and *C.*
775 *mili*, which are an estimated 400 million years diverged (Hedges et al. 2015).

776 We analyzed these alignments with codeml, comparing the likelihood of two
777 models, one with a fixed omega of 1 and an alternate model without a fixed omega, and

778 performed a likelihood ratio test. A significant result for the likelihood ratio test provides
779 evidence that a gene is not neutrally-evolving (**Supplementary File 3A**). In all cases of
780 N-terminal truncation analyzed, dN/dS is significantly less than one (**Supplementary**
781 **File 3A**). While it is possible that some of these cases represent very recently
782 pseudogenized genes, the widespread evidence for purifying selection on the SET domain
783 strongly suggests that these PRDM9 orthologs are functionally important.

784 We also investigated constraint in all mammalian Ref-seq orthologs that appear to
785 lack only an annotated KRAB or SSXRD domain; for this larger number of genes, we did
786 not confirm all domain losses, due to the large number of genome searches that would be
787 required and lack of RNAseq data for most species. We found evidence of purifying
788 selection in all cases except for five PRDM7 orthologs from primates, for which we had
789 been unable to identify a KRAB domain (**Supplementary File 3C**). PRDM7 is thought
790 to have arisen from a primate specific duplication event and to have undergone
791 subsequent losses of the C2H2 ZF array and of some catalytic specificity of its SET
792 domain (Blazer et al. 2016). Thus, PRDM7 orthologs are unlikely to function in directing
793 recombination. Our findings further suggest they are evolving under very little constraint,
794 and may even be non-functional. More generally, within placental mammals, the majority
795 of partial PRDM9 orthologs that we identified lack the ZF array completely or have
796 truncated arrays (notably, there are fewer than four tandem ZFs in 24 of 28 orthologs), in
797 sharp contrast to other taxa in which partial orthologs to PRDM9 lack the N terminal
798 domains, yet have conserved ZF arrays and are constrained. Moreover, the paralogs
799 lacking a long ZF tend to be found in species that already carry a complete PRDM9
800 ortholog (21 of 24). Thus, some of these cases may represent recent duplication events in
801 which one copy of PRDM9 is under highly relaxed selection, similar to PRDM7 in
802 primates.

803

804 ***Evolutionary patterns in the SSXRD domain***

805 The SSXRD domain is the shortest functional domain in the PRDM9 protein. One
806 species of cartilaginous fish (*Rhincodon typus*), and several species of bony fish
807 (*Anguilla anguilla*, *A. rostrata*, *A. japonica*, *Salmo trutta*, *S. salar*) have weakly predicted
808 SSXRD domains (e-values > 10, see **Supplementary File 1B**, **Supplementary File 2C**).

809 This observation is potentially suggestive of functional divergence or loss of this domain.
810 Unfortunately, because the domain is so short, there is little power to reject $dN/dS = 1$:
811 though the estimate of dN/dS was 0.10 and 0.11 between cartilaginous fish and eel and
812 salmon orthologous regions, respectively, the difference between models was not
813 significant in either case. Based on these findings, we tentatively treat the weakly
814 predicted SSXRD domain in *Rhincodon typus* and in the above species of bony fish as
815 evidence that this domain is present in these species, but note that we were unable to
816 identify a similar region in predicted gene models from another species of cartilaginous
817 fish (*Callorhinchus milii*).

818

819 ***PCR and Sanger sequencing of python PRDM9***

820 We performed Sanger sequencing of *Python bivittatus* PRDM9 from a single
821 individual to collect additional data on within species diversity of the ZF array (**Figure 1-**
822 **Figure Supplement 5**). Primers were designed based on the *Python bivittatus* genome
823 (Castoe et al. 2013) to amplify the ZF containing exon of PRDM9 and through a gap in
824 the assembly. Primers were assessed for specificity and quality using NCBI Primer Blast
825 (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) against the nr reference database and
826 were synthesized by IDT (Coralville, IA, USA).

827 DNA was extracted from approximately 20 mg of tissue using the Zymo Quick-
828 DNA kit (Irvine, CA, USA) following the manufacturer's protocol. PCR was performed
829 using the NEB Phusion High-Fidelity PCR kit (Ipswich, MA, USA). Reactions were
830 performed following manufacturer's instructions with 60 ng of DNA and 10 μ M each of
831 the forward (ZF: 5' TTTGCCATCAGTGTCCCAGT'3; gap: 5'
832 GCTTCCAGCATTGCCAGTT'3) and reverse (ZF: 5'
833 TTGATTCACTTGTGAGTGGACAT'3; gap: 5' GAGCTTGCTGAAATCGGGT'3)
834 primers. Products were inspected for non-specific amplification on a 1% agarose gel with
835 ethidium bromide, purified using a Qiagen PCR purification kit (Valencia, CA, USA) and
836 sequenced by GeneWiz (South Plainfield, NJ, USA).

837

838 ***Analysis of PRDM9 ZF array evolution***

839 In species in which PRDM9 is known to play a role in recombination, the level of
840 sequence similarity between the individual ZFs of the tandem array is remarkably high,
841 reflective of high rates of ZF turnover due to paralogous gene conversion and duplication
842 events (Oliver et al. 2009; Myers et al. 2010; Jeffreys et al. 2013). It has further been
843 observed that DNA-binding residues show high levels of amino acid diversity, suggestive
844 of positive selection acting specifically at DNA-binding sites, i.e., on binding affinity
845 (e.g. Oliver et al. 2009; Schwartz et al. 2014). These signals have been previously studied
846 by comparing site specific rates of synonymous versus non-synonymous substitutions
847 (dN/dS) between paralogous ZFs in PRDM9's tandem ZF array (Oliver et al. 2009).
848 Assessing statistical significance using this approach is problematic, however, because
849 the occurrence of paralogous gene conversion across copies means that there is no single
850 tree relating the different ZFs, in violation of model assumptions (Schierup and Hein
851 2000; Wilson and McVean 2006). Here, we used a statistic sensitive to both rapid
852 evolution at DNA-binding sites and high rates of gene conversion: the total proportion of
853 amino acid diversity observed at DNA-binding sites within the ZF array. We then
854 assessed significance empirically by comparing the value of this statistic to other C2H2
855 ZF genes from the same species (where possible).

856 To this end, for each species with a PRDM9 ortholog, we downloaded the
857 nucleotide and protein sequences for all available RefSeq genes with a C2H2 ZF motif
858 annotated in Conserved Domain Database (pfam id# PF00096). To simplify alignment
859 generation, we only used tandem ZF arrays with four or more ZFs matching the 28 amino
860 acid long C2H2 motif (X2-CXXC-X12-HXXXH-X5 where X is any amino acid). In all
861 of our analyses, if a gene had multiple tandem ZF arrays that were spatially separated,
862 only the first array of four or more adjacent ZFs was used for the following analysis
863 (**Supplementary File 3B**). However, an alternative analysis using all ZFs or different
864 subsets of ZFs led to qualitatively similar results for the PRDM9 β orthologs from bony
865 fish, where ZFs are commonly found in multiple tandem arrays separated by short linker
866 regions in the predicted amino acid sequence (**Figure 2 Supplement 1; Figure 2- Figure**
867 **Supplement 2**). For species with PRDM9 orthologs with fewer than five ZFs, we
868 implemented *blastn* against the whole genome sequence using the available gene model
869 as a query sequence, in order to determine whether or not there was a predicted gap

870 within the ZF array, and, if there was, to identify any additional ZFs found in the
871 expected orientation at the beginning of the adjacent contig. This approach was able to
872 successfully identify additional ZF sequences on contigs adjacent to PRDM9 in the
873 genome assembly for two species (*Latimeria chalumnae* and *Probobothrops*
874 *mucrosquamatus*). These ZFs were included in subsequent analysis (**Supplementary File**
875 **1A**).

876 Using the alignments generated above, we determined the amino acid diversity
877 along the ZF domains of PRDM9 genes and all other C2H2 ZFs from the same species
878 (**Table 1, Supplementary File 3B**), and calculated the proportion of the total amino acid
879 diversity at canonical DNA-binding residues of the ZF array. Specifically, we calculated
880 the heterozygosity x_k at position k across the aligned ZFs from a single tandem array as:

$$x_k = 1 - \sum_{i=1}^m f_i^2$$

881 where m is the number of unique amino acids found at position k across the fingers,
882 and f_i is the frequency of the i^{th} unique amino acid across the fingers. The total
883 proportion P of amino acid diversity assigned to DNA-binding residues is the sum of
884 x_k at DNA-binding sites over the sum of x_k at all sites in the ZF array. To compare
885 results to those for other genes, we ranked PRDM9 by the value P compared to all other
886 C2H2 ZF genes from the same species (**Table 1, Supplementary File 3B**).

887 We used the R package phylotools (Zhang et al. 2012; <https://cran.r-project.org/web/packages/phylotools/index.html>) to calculate a p-value for the correlation
888 between complete domain structure and rapid evolution of the PRDM9 ZF array, taking
889 into account phylogenetic relationships between PRDM9 orthologs. We coded these
890 variables using a binary approach with '00' for incomplete domain structure and no
891 evidence of rapid evolution and '11' for complete domain structure and evidence of rapid
892 evolution. To describe the phylogenetic relationships between orthologs, we used the
893 RAxML tree that we constructed from the SET domain for all PRDM9 orthologs. Species
894 with missing ZF information, including species where PRDM9 has been lost, were
895 excluded from this analysis using the drop.tip function of the ape package (Paradis et al.
896 2004), resulting in a tree with 91 tips. We used the phyloglm command to perform a

898 logistic regression evaluating the relationship between domain structure and the odds of
899 rapid evolution of the ZF array.

900

901 ***Analysis of the SET domain catalytic residues***

902 In order to investigate whether the catalytic function of the SET domain is
903 conserved in the PRDM9 orthologs identified above, we asked whether any PRDM9
904 orthologs in our dataset carried substitutions at three catalytic residues shown to mediate
905 the methyltransferase activity of human PRDM9 (Wu et al. 2013). To this end, we used
906 Clustal Omega to create an amino acid alignment of the SET domain with 15 amino acids
907 of flanking sequence for each PRDM9 ortholog in our dataset and asked whether the gene
908 had substitutions to tyrosine residues at positions aligning to Y276, Y341 and Y357 in
909 human PRDM9 (**Supplementary File 1A**).

910 In total, 57 genes were identified as having substitutions in at least one of these
911 residues, including 11 from placental mammals and 46 from bony fish (**Supplementary**
912 **File 1A**). To visualize the distribution of these substitution events within bony fish, we
913 mapped these substitutions onto the phylogeny of PRDM9 orthologs generated above
914 (**Figure 3**).

915

916 ***Characterizing patterns of recombination in hybrid swordtail fish***

917 Percomorph fish have a partial ortholog of PRDM9 that lacks the KRAB and
918 SSXRD domains found in mammalian PRDM9. As a result, we hypothesized that they
919 would behave like PRDM9 knockouts, in that the predicted PRDM9 binding motif would
920 not co-localize with recombination events, and functional genomic elements such as the
921 TSS and CGIs would be enriched for recombination events.

922 To build a hybrid recombination map, we generated low coverage sequence data
923 for 268 individuals from a natural hybrid population (“Totonicapa”) formed between the
924 percomorph species *X. birchmanni* and *X. malinche* (RRID:SCR_008340) and sampled in
925 2013-2015. The two parental species are closely related, with pairwise sequence
926 divergence <0.5% (Schumer et al. 2014). Interestingly, in sharp contrast to what is seen
927 in placental mammals, the ZF is slowly evolving between *X. birchmanni* and *X. malinche*
928 (dN/dS=0.09; **Figure 4A**).

929 DNA was extracted from fin clips for the 268 individuals and libraries were
930 prepared following Stern (2015). Briefly, three to ten nanograms of DNA was mixed with
931 Tn5 transposase enzyme pre-charged with custom adapters and incubated at 55 C for 15
932 minutes. The reaction was stopped by adding 0.2% SDS and incubating at 55 C for an
933 additional seven minutes. One of 96 custom indices were added to each sample in a plate
934 with an individual PCR reaction including 1 ul of the tagmented DNA; between 13-16
935 PCR cycles were used. After amplification, 5 ul of each reaction was pool and purified
936 using Agencourt AMPpure XP beads. Library size distribution and quality was visualized
937 on the Bioanalyzer 1000 and size selected by the Princeton Lewis Sigler Core Facility to
938 be between 350-750 basepairs. Libraries were sequenced on the Illumina HiSeq 4000 at
939 Weill Cornell Medical Center across three lanes to collect paired-end 100 bp reads.

940 Ancestry assignment in hybrids was performed using the Multiplexed Shotgun
941 Genotyping (“MSG”) pipeline (Andolfatto et al. 2011). This approach has been
942 previously validated for genome-wide ancestry determination in late generation *X.*
943 *birchmanni* x *X. malinche* hybrids (Schumer et al. 2014; Schumer et al. 2015). Briefly,
944 raw data was parsed by barcode and trimmed to remove low-quality basepairs (Phred
945 quality score <20). Reads with fewer than 30 bp after trimming were discarded. Because
946 of prohibitively long computational times, reads from individuals with more than 16
947 million reads were subsampled to 16 million before running the MSG pipeline. The
948 minimum number of reads for an individual to be included was set to 300,000, since
949 ancestry inference with fewer reads is predicted to have lower accuracy based on
950 simulations (Schumer et al. 2015). This procedure resulted in 239 individuals for our
951 final analysis, with an average coverage of 8.3 million reads, or ~1X genome-wide
952 coverage.

953 The parameters used in the MSG run were based on previous work on this hybrid
954 population (Schumer et al. in review). The expected number of recombination events per
955 chromosome (recRate) was set to 8, based on a prior expectation of approximately 30
956 generations of admixture and assuming initial admixture proportions of 75% of the
957 genome derived from *X. birchmanni* and 25% derived from *X. malinche*. Similarly, priors
958 for each ancestry state were set based on these mixture proportions (par1=0.5625,

959 par1par2=0.375 and par2=0.0625). The recombination rate scaling factor was set to the
960 default value of 1.

961 Ancestry transitions were identified as the interval over which the posterior
962 probability changed from ≥ 0.95 in support of one ancestry state to ≥ 0.95 for a different
963 ancestry state. Breakpoint intervals that occurred within 10 kb of a contig edge were
964 excluded from the analysis due to concerns that false breakpoints may occur more
965 frequently near the edges of contigs. The identified recombination intervals varied
966 significantly in their lengths, i.e., in the resolution of the crossover event. The median
967 resolution was 13 kb, with 75% of breakpoints resolved within 35 kb or less.

968 To evaluate the relationship between recombination frequency and genomic
969 elements such as the TSS, CGIs, and computationally predicted PRDM9 binding sites,
970 we needed to convert the observed recombination events into an estimate of
971 recombination frequency throughout the genome. To this end, we considered the
972 proportion of events observed in a particular 10 kb window; we note that this rate is not
973 equivalent to a rate per meiosis. We filtered the data to remove windows within 10 kb of
974 a contig boundary. Because the majority of events span multiple 10 kb windows, we
975 randomly placed events that spanned multiple windows into one of the windows that they
976 spanned.

977 We used the closest-feature command from the program bedops v2.4.19 (Neph et
978 al. 2012) to determine the minimum distance between each 10 kb window and the
979 functional feature of interest. For the TSS, we used the Ensembl annotation of the
980 *Xiphophorus maculatus* genome with coordinates lifted over to v.4.4.2 of the linkage
981 group assembly (Amores et al. 2014; Schumer et al. 2016)
982 http://genome.uoregon.edu/xma/index_v1.0.php). For CGIs, we used the annotations
983 available from the UCSC genome browser beta site ([http://genome-test.cse.ucsc.edu/cgi-bin/hgTables?hgsid=391260460_COev5GTglYu74K2t24uaU4UcaTvP&clade=vertebrate&org=Southern+platyfish&db=xipMac1&hgta_group=allTracks&hgta_track=cpgIslandExt&hgta_table=0&hgta_regionType=genome&position=JH556661%3A3162916-4744374&hgta_outputType=primaryTable&hgta_outFileName="](http://genome-test.cse.ucsc.edu/cgi-bin/hgTables?hgsid=391260460_COev5GTglYu74K2t24uaU4UcaTvP&clade=vertebrate&org=Southern+platyfish&db=xipMac1&hgta_group=allTracks&hgta_track=cpgIslandExt&hgta_table=0&hgta_regionType=genome&position=JH556661%3A3162916-4744374&hgta_outputType=primaryTable&hgta_outFileName=)). To identify putative
984 PRDM9 binding sites, we used the ZF prediction software available at zf.princeton.edu
985 with the polynomial SVM settings to generate a position weight matrix for the *X*.

990 *malinche* and *X. birchmanni* PRDM9 orthologs (Persikov and Singh 2014). This
991 approach yielded identical predicted binding motifs in the two species (**Figure 4A**). We
992 used this position weight matrix to search the *X. malinche* genome (Schumer et al. 2014)
993 for putative PRDM9 binding sites with the meme-suite program FIMO (v4.11.1; Grant et
994 al. 2011). We selected all regions with a predicted PRDM9 binding score of ≥ 5 . Since the
995 individuals surveyed are interspecific hybrids, and the two species may differ in the
996 locations of predicted PRDM9 binding sites, we repeated the FIMO search against the *X.*
997 *birchmanni* genome, obtaining qualitatively identical results.

998 After determining the minimum distance between each 10 kb window and the
999 features of interest, we calculated the average recombination frequency in hybrids as a
1000 function of distance from the feature of interest in 10 kb windows (**Figure 4**; **Figure 4 -**
1001 **Figure Supplement 2**). To estimate the uncertainty associated with rates at a given
1002 distance from a feature, we repeated this analysis 500 times for each feature,
1003 bootstrapping windows with replacement. Because we found a positive correlation
1004 between distance from the TSS and CGIs in 10 kb windows with recombination
1005 frequency, we checked that power (i.e., the proportion of ancestry informative sites) was
1006 not higher near these features.

1007 Most work in humans and mice has focused on the empirical PRDM9 binding
1008 motif rather than the computationally predicted motif. Since we expect the
1009 computationally predicted motif to be a poorer predictor of PRDM9 binding, we checked
1010 how its use would affect the analyses, by repeating the analysis described above for the
1011 computational prediction obtained for the human PRDM9A allele, using recombination
1012 rates in 10 kb windows estimated from the CEU LD map (Frazer et al. 2007; downloaded
1013 from: <http://www.well.ox.ac.uk/~anjali/Amap/>). We also repeated this analysis for the
1014 gor-1 PRDM9 allele in *Gorilla gorilla*, using recombination rates in 10 kb windows
1015 estimated from a recent LD map (Schwartz et al. 2014; Stevison et al. 2016; downloaded
1016 from Stevison 2016: <https://github.com/lstevison/great-ape-recombination>).

1017
1018 ***Comparisons of recombination landscapes with and without PRDM9***

1019 To investigate whether patterns of recombination rates near the TSS and CGI
1020 systematically distinguish between species that do and do not use PRDM9-directed

1021 recombination, we compared available data across species. We downloaded previously
1022 published recombination maps for three species without PRDM9 genes (dog, Auton et al.
1023 2013, zebra finch and long-tailed finch, Singhal et al. 2015) and four species with
1024 complete PRDM9 orthologs (human, Frazer et al. 2007; Hinch et al. 2011; gorilla,
1025 Stevison et al. 2016; sheep, Johnston et al. 2016; and mouse, Brunschwig et al. 2012).

1026 For each species, we binned recombination rate into 10kb windows along the
1027 genome, excluding the sex chromosomes and windows overlapping with assembly gaps
1028 from all analyses. For each species, we downloaded annotations of assembly gaps, TSSs
1029 and CGIs from the UCSC genome browser website. For CGI positions in the gorilla
1030 genome, we used the LiftOver tool (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>) to
1031 convert the available coordinates for the GorGor4 genome assembly to the GorGor3
1032 assembly. For zebra finch and long-tailed finches, we used the coordinates of CGIs and
1033 TSSs as annotated for the TaeGut3.2 genome assembly, noting that these coordinates are
1034 consistent with the TaeGut3.1 assembly for all chromosomes for which genetic distances
1035 were inferred in (Singhal et al. 2015).

1036 For each map, we calculated the distance to the nearest TSS and to the nearest
1037 CGI by from the midpoint of each 10 kb window. To visualize these patterns, we fit a
1038 Gaussian loess curve using the distance to nearest TSS or CGI and recombination rate for
1039 each species, using only windows within 100kb of a representative element. For visual
1040 comparison, we scaled the resulting curves by setting the y-value (recombination rate) of
1041 the last point to one.

1042 A caveat is that other than for swordtail and sheep, we relied on LD based
1043 genetics maps, which estimate population recombination rates $4N_e r$, where N_e is the
1044 effective population size and r the recombination rate per meiosis. Because estimates of
1045 N_e decrease near genes as a consequence of diversity-reducing linked selection (e.g.,
1046 Wright and Andolfatto 2008; Hernandez et al. 2011), a decrease in estimated population
1047 recombination rates near genes may not reflect a reduction in the recombination rate r . To
1048 explore the potential importance of this caveat, we considered two species where both LD
1049 maps and pedigree or admixture maps were available: dogs and humans. In both cases,
1050 the qualitative results were the same as for the LD-based maps (**Figure 5 - Figure**
1051 **Supplement 1**). Since diversity-reducing linked selection should give rise, if anything, to

1052 a trough in diversity levels, it cannot explain the observed peaks at these features in
1053 species lacking PRDM9 or swordtail fish; in fact, since these species also experience this
1054 form of selection (e.g., Singhal et al. 2015), the true peaks in recombination rates near
1055 promoter-like features are likely somewhat more pronounced.

1056 We note further that although the peak in recombination rate at these features in
1057 swordtail fish appears to be less prominent than in dog or birds, quantitative comparisons
1058 of different species are difficult because these maps differ in their resolution.

1059

1060 **Native chip-seq of *X. birchmanni* testis and liver tissue**

1061 Whole testis and liver were dissected from two *X. birchmanni* adults and stored in
1062 HypoThermosol FRS (BioLife Solutions, Bothell, WA) buffer on ice until processing.
1063 Native chromatin ChIP was performed as described previously (Markenscoff-
1064 Papadimitriou et al. 2014). Briefly, tissue was homogenized and lysed; the lysate was
1065 spun through a sucrose cushion (to pellet nuclei). Nuclei were resuspended in 500ul
1066 MNase digestion buffer and digested with 1 unit of micrococcal nuclease (MNase, Sigma
1067 N5386, St. Louis, MO) for 2 minutes at 37 C, then inactivated with 20ul 0.5M EDTA and
1068 chilled on ice. The first soluble chromatin fraction was recovered by spinning for 10 min
1069 at 10,000 rcf at 4C and collecting the supernatant. To isolate the second soluble
1070 chromatin fraction, the pellet was resuspended in 500 μ l dialysis buffer, rotated overnight
1071 at 4 C, then centrifuged for 10 min at 10,000 rcf at 4 C to pellet insoluble material. The
1072 digestion quality of each fraction was evaluated on an agarose gel. The two soluble
1073 fractions were combined for chromatin immunoprecipitation with 1 μ g of H3K4me3
1074 antibody (Millipore 04-745, Billerica, MA); 1/10 volume was retained as an input
1075 control. Antibody was bound to the remaining chromatin overnight while rotating at 4 C.
1076 The next day blocked Protein A and Protein G beads were added, and rotated for 3 hours.
1077 The bound beads were then washed a total of 7 times with chilled wash buffers and
1078 immunoprecipitated chromatin was eluted in elution buffer for 30 minutes at 37 C and
1079 cleaned up with ChIP DNA Clean and Concentrator kit (Zymo Research, Irvine, CA).
1080 Libraries were prepared for sequencing using the NuGEN ultralow library prep kit
1081 (NuGEN, San Carlos, CA) following manufacturer's instructions and sequenced on an

1082 Illumina HiSeq 2500 at Hudson Alpha to collect 10.3-10.5 and 12.2-14.5 paired-end 50
1083 bp reads for pull-down and input samples respectively.

1084 Raw reads were trimmed to remove adapter sequences and reads with fewer than
1085 18 bp after adapter trimming using the program cutadapt. These trimmed reads were then
1086 mapped to the *X. maculatus* reference genome with bowtie2 (Langmead and Salzberg
1087 2012) and the resulting bam file was sorted with samtools (Li et al. 2009). Homer (Heinz
1088 et al. 2010) was used to generate bigWig files and call peaks using the option style –
1089 factor. We also performed the analysis using the option style –histone and found that the
1090 results were qualitatively similar. Peak files were converted to bed files and bedtools2
1091 (Quinlan and Hall 2010) was used to analyze overlap between the locations of H3K4me3
1092 peaks and predicted PRDM9 binding motifs in the swordtail genome (see above). Based
1093 on Homer analysis, which identified 20,662 peaks in the testis and 15,050 in the liver, the
1094 IP efficiency was estimated to be 38% for the testis sample and 40% for the liver sample;
1095 the peak width was estimated to be 229 bp for the testis sample and 238 for the liver
1096 sample.

1097 Having identified H3K4me3 peaks in testis and liver tissue, we next asked about
1098 the relationship between these peaks and predicted PRDM9 binding sites (see above). If
1099 PRDM9 is making H3K4me3 peaks during meiosis, we expect to see an association
1100 between predicted PRDM9 binding motifs in the swordtail genome and H3K4me3 peaks.
1101 To test for such an association, we generated 500 null motifs by randomly shuffling
1102 without replacement the position weight matrix of the *X. birchmanni* PRDM9 and re-
1103 running FIMO as described above. We then asked how frequently randomly shuffled
1104 PRDM9 motifs overlap H3K4me3 peaks compared to the real motif. We found that no
1105 evidence that the real motif overlapped H3K4me3 peaks more frequently than the
1106 shuffled versions of the motif (**Figure 4C**).

1107 As a secondary approach, we compared H3K4me3 peaks that are specific to the
1108 testis to H3K4me3 peaks that are specific to the liver, defined as peaks in the testis where
1109 there is no overlapping peak in the liver. Using a Chi-squared test, we asked whether
1110 H3K4me3 peaks found only in the testis are more likely to overlap a PRDM9 binding
1111 motif than those that are liver specific (where the definition is analogous) (**Figure 4**).
1112 Because the size of H3K4me3 peaks will impact the expected overlap with PRDM9

1113 binding motifs, we also constrained the size of the H3K4me3 peaks in the liver analysis
1114 to be the same as that inferred from the testis using the –size flag in homer (229 bp).
1115 Results were not qualitatively different with the original analysis, using liver H3K4me3
1116 peaks that were inferred to be 238 bp. Counterintuitively liver-specific H3K4me3 peaks
1117 appear to overlap predicted PRDM9 motifs more often than testes-specific peaks
1118 ($\chi=14.8$; $p=1.2e-4$). However, performing this same analysis with the 500 null motifs
1119 (generated as described above), we found that liver-specific peaks were significantly
1120 enriched in shuffled motifs in 85% of simulations (at the 0.05 level). This analysis
1121 suggests that base composition differences between liver and testes-specific H3K4me3
1122 peaks explain the difference in overlap results.

1123 We also repeated the above analysis for clusters of three ZFs in the swordtail
1124 PRDM9 ZF array, using a smaller number of shuffled sequences (n=20). We observed
1125 the same qualitative patterns for each of the ZF clusters as reported above.

1126 Finally, we used a third approach to ask about the association of H3K4me3 peaks
1127 and PRDM9 binding sites. We generated five replicate datasets of H3K4me3 sequences
1128 and their flanking 250 bp regions from both the testis and the liver. We ran the program
1129 MEME to predict motifs enriched in the testis-specific H3K4me3 peaks using the liver as
1130 a background sequence set on these five replicate datasets. We then examined the top ten
1131 predicted motifs to ask whether any of these motifs resembled the computationally
1132 predicted PRDM9 binding motifs (**Figure 4 – Figure Supplement 5**).

1133 The above analyses suggest that in swordtail fish, PRDM9 does not make
1134 H3K4me3 marks but they do not indicate whether H3K4me3 peaks are associated with
1135 recombination events in swordtails. We therefore verified that recombination rates in 10
1136 kb windows are significantly correlated with the distance of that window to the nearest
1137 H3K4me3 peak ($\rho=-0.072$, $p=2.3e-69$; **Figure 4**). This relationship weakens but
1138 remains significant when accounting for distance both to TSSs and CGIs by a partial
1139 correlation analysis ($\rho=-0.026$, $p=5.4e-10$). Furthermore, windows that contain a testis-
1140 specific H3K4me3 peak have a higher recombination rate than windows that contain a
1141 liver-specific peak (**Figure 4 – Figure Supplement 4**). Finally, there is a significant
1142 positive correlation between the number of bp in a 10 kb window overlapping an
1143 H3K4me3 peak and the number of recombination events observed in that window in the

1144 testis but not in the liver (testis: rho=0.044, p=2.8e-29; liver: rho=0.002, p=0.66).
1145 Together, these analyses suggest that a relationship between H3K4me3 peaks and
1146 recombination exists in swordtails, but not one mediated through PRDM9 binding.

1147

1148

1149 **Acknowledgements**

1150

1151 We thank the federal government of Mexico for permission to collect fish under a
1152 scientific collecting permit to Guillermina Alcaraz (PPF/DGOPA-173/14). We are
1153 grateful to Dana Pe'er for generous use of lab space, Joe Derisi for sending us python
1154 tissue, Ammon Corl and Rasmus Nielsen for access to additional lizard transcriptomes,
1155 and Nick Altemose, Scott Keeney, Simon Myers, Laure Segurel, Guy Sella, Sonal
1156 Singhal and members of the Pickrell, Przeworski and Sella labs for helpful discussions.
1157 This project was supported by R01 GM83098 grant to MP and NSF DDIG DEB-1405232
1158 to MS.

1159

1160 **Competing Interests Statement**

1161

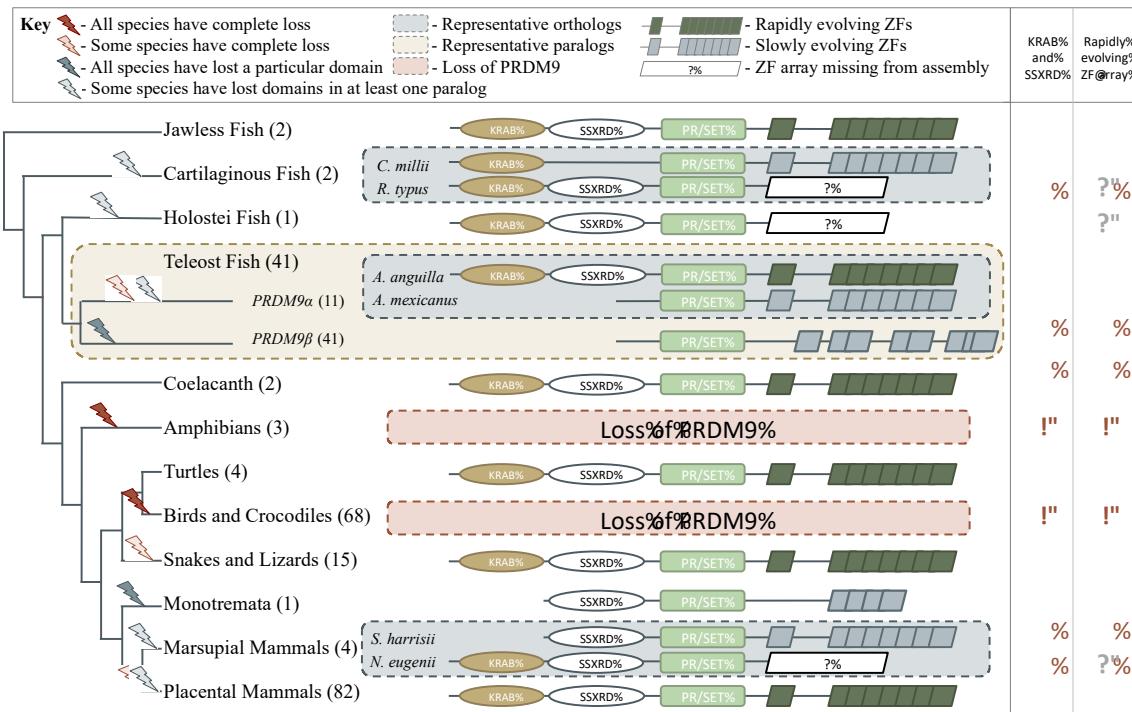
1162 The authors declare no competing financial or non-financial interests.

1163

1164

1165 **Figures**

1166



1167
1168

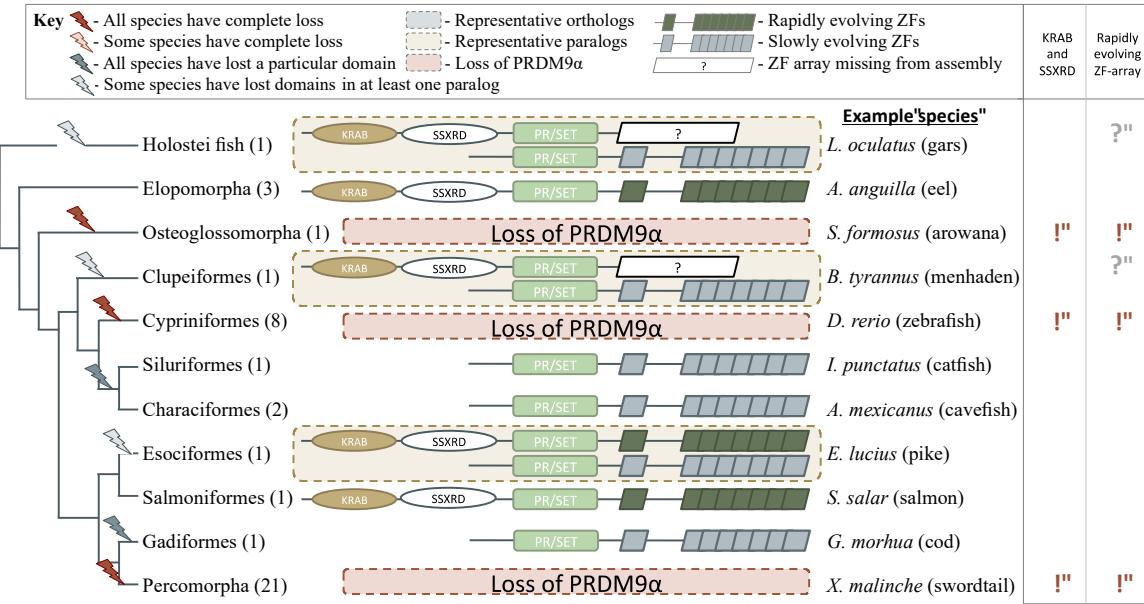
1169 **Figure 1. Phylogenetic distribution and evolution of PRDM9 orthologs in**
1170 **vertebrates.** Shown are the four domains: KRAB domain (in tan), SSXRD (in white),
1171 PR/SET (in green) and ZF (in gray/dark green; the approximate structure of identified
1172 ZFs is also shown). The number of unique species included from each taxon is shown in
1173 parenthesis. Complete losses are indicated on the phylogeny by red lightning bolts and
1174 partial losses by gray lightning bolts. Lightning bolts are shaded dark when all species in
1175 the indicated lineage have experienced the entire loss or same partial loss. Lightning bolts
1176 are shaded light when it is only true of a subset of species in the taxon. ZF arrays in dark
1177 green denote those taxa in which species show evidence of rapid evolution. White
1178 rectangles indicate cases where we could not determine whether the ZF was present,
1179 because of the genome assembly quality. For select taxa, we present the most complete
1180 PRDM9 gene found in two exemplar species. Within teleost fish, we additionally show a
1181 PRDM9 paralog that likely arose before the common ancestor of this taxon; in this case,
1182 the number of species observed to have each paralog is in parenthesis. Although the
1183 monotremata ZF is shaded gray, it was not included in our analysis of rapid evolution
1184 because of its small number of ZFs.

1185
1186

1187 **Figure 1- Figure Supplement 1.** Phylogenetic approach to identifying PRDM9
1188 orthologs and related gene families.

1189 **Figure 1- Figure Supplement 2.** Neighbor joining guide tree based on the SET domain.

1190 **Figure 1- Figure Supplement 3.** Expression levels of genes with a known role in
1191 meiotic recombination in testes tissues of three exemplar species.
1192 **Figure 1- Figure Supplement 4.** Patterns of amino acid diversity as a function of amino
1193 acid position in the ZF alignment.
1194 **Figure 1- Figure Supplement 5.** Examples of differences in computationally predicted
1195 PRDM9 binding motifs for species from three taxa.
1196
1197



1198

1199

Figure 2. Phylogenetic distribution and functional domains of PRDM9α orthologs in teleost fish and in holostean fish that are outgroups to the PRDM9α/PRDM9β duplication event. Shown are the four domains: KRAB domain (in tan), SSXRD (in white), PR/SET (in green) and ZF (in gray/dark green; the approximate structure of identified ZFs is also shown). The number of unique species included from each taxon is shown in parenthesis. Complete losses are indicated on the phylogeny by red lightning bolts and partial losses by gray lightning bolts. Lightning bolts are shaded dark when all species in the indicated lineage have experienced the entire loss or same partial loss. Lightning bolts are shaded light when it is only true of a subset of species in the taxon. ZF arrays in dark green denote those taxa in which species show evidence of rapid evolution. White rectangles indicate cases where we could not determine whether the ZF was present, because of the genome assembly quality. While many taxa shown have more than one PRDM9α ortholog, the genes identified from each species generally have similar domain architectures. Exceptions include Clupeiformes, Esociformes, and Holostean fish, for which two alternative forms of PRDM9α paralogs are shown. ZF arrays shown in dark green are those in which we detected evidence of rapid evolution. Based on this distribution, we infer that the common ancestor of ray-finned fish likely had a rapidly evolving and complete PRDM9α ortholog.

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

Figure 2- Figure Supplement 1. Subset of maximum-likelihood phylogeny of the SET domain showing bony fish PRDM9 orthologs α and β.

Figure 2- Figure Supplement 2. Analysis of ZF evolution in PRDM9β.

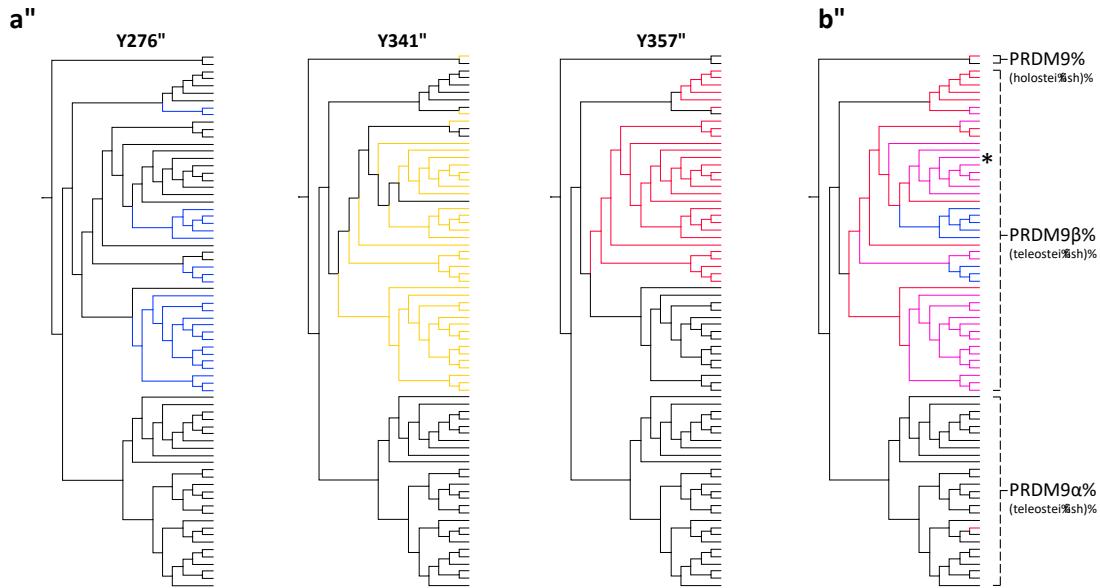


Figure 3. Substitutions at SET domain catalytic residues in bony fish PRDM9 genes.
 (a) Lineages within bony fish carrying substitutions at each of three tyrosine residues involved in H3K4me3 catalysis in human PRDM9 are shown in blue, yellow and red. (b) Lineages carrying substitutions at one, two or three of these residues are shown in red, pink and blue respectively. All PRDM9 β genes as well as a partial PRDM9 ortholog from holostei fish carry one or more substitutions at these residues. The PRDM9 β gene from *Xiphophorus* is indicated by the presence of asterisk.

1224
 1225
 1226
 1227
 1228
 1229
 1230
 1231
 1232
 1233

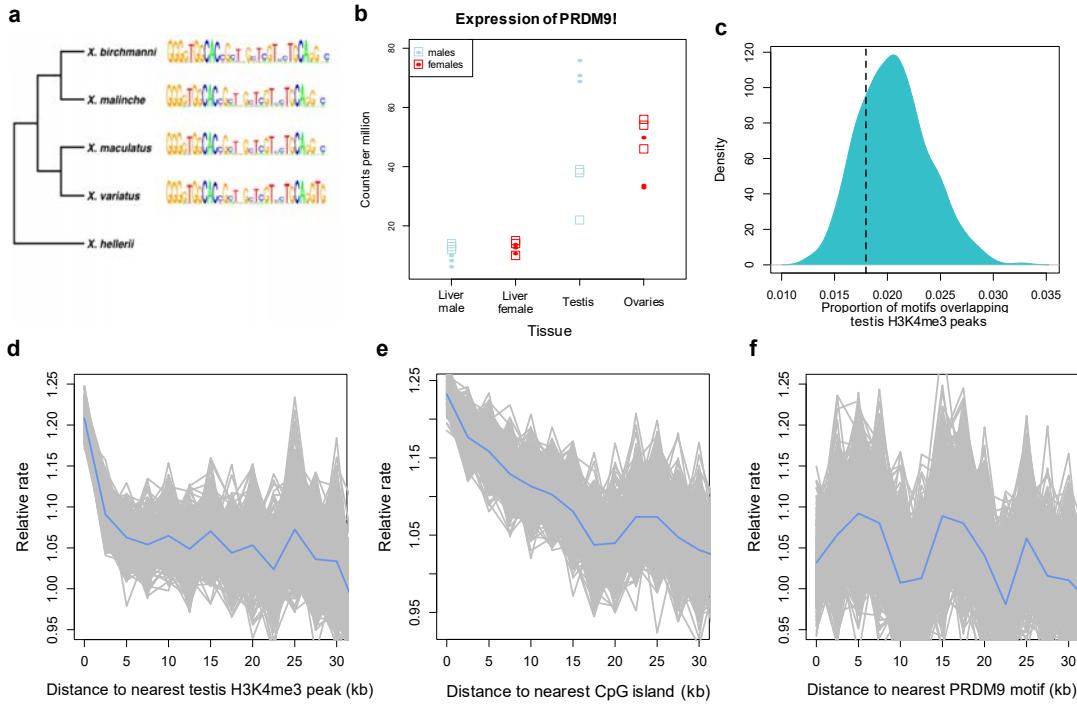


Figure 4. Patterns of recombination and PRDM9 evolution in swordtail fish. (a) The ZF array of PRDM9 appears to be evolving slowly in *Xiphophorus*, with few changes over over 1 million years of divergence (Cui et al. 2013; Jones et al. 2013). (b) PRDM9 is upregulated in the germline relative to the liver in *Xiphophorus birchmanni* (circles) and *X. malinche* (squares; panel shows three biological replicates for each species). (c) The computationally-predicted PRDM9 binding sites is not unusually associated with H3K4me3 peaks in testes (d) Crossover rates increase near H3K4me3 peaks in testis (e) Crossover rates increase near CGIs (f) Crossover rates do not increase near computationally-predicted PRDM9 binding sites (see **Figure 4- Figure Supplement 3** for comparison). Crossover rates were estimated from ancestry switchpoints between naturally occurring hybrids of two species (see **Methods**).

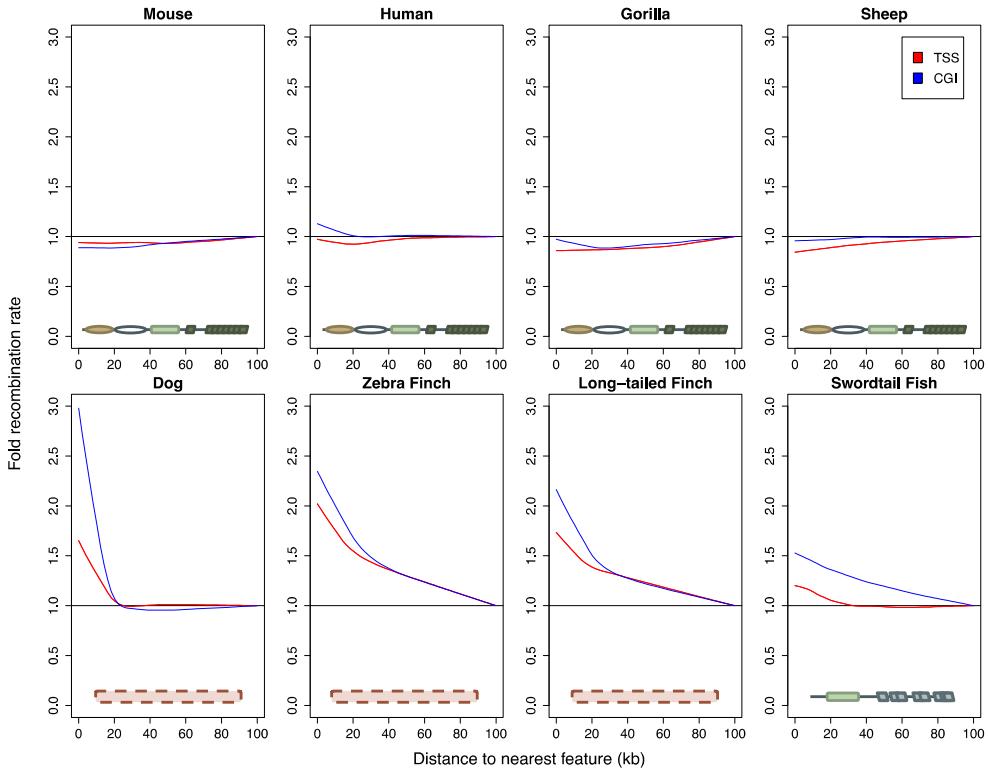
Figure 4 - Figure Supplement 1. Expression levels of meiosis-related genes in swordtail fish tissues.

Figure 4 - Figure Supplement 2. Recombination frequency in swordtails as a function of distance to the TSS.

Figure 4 - Figure Supplement 3. Recombination rates show a peak near the computationally predicted PRDM9A binding motif in humans and gor-1 allele in gorillas.

Figure 4 - Figure Supplement 4. Higher observed recombination rate at testis-specific H3K4me3 peaks than liver-specific H3K4me3 peaks.

Figure 4 - Figure Supplement 5. MEME prediction of sequences enriched in testis-H3K4me3 peaks relative to liver-specific H3K4me3 peaks.



1259
1260 **Figure 5. Patterns of recombination near TSSs and CGIs in species with and**
1261 **without complete PRDM9 orthologs.** For each species, recombination rates estimated
1262 from patterns of LD (or in swordtail fish, from admixture switches) were binned in 10kb
1263 windows along the genome; curves were fit using gaussian loess smoothing. The fold
1264 change in recombination rates shown on the y-axis is relative to recombination rates at
1265 the last point shown. Species shown in the top row have complete PRDM9 orthologs
1266 (mouse, human, gorilla and sheep), whereas species in the bottom row have no PRDM9
1267 ortholog (dog, zebra finch, long-tailed finch), or a partial PRDM9 ortholog (swordtail
1268 fish).

1269
1270 **Figure 5 – Figure supplement 1.** Relationships of recombination rates to the nearest
1271 TSS and CGI using maps inferred from pedigrees in dog and admixture maps in human,
1272 compared to those estimated from LD data.

1273 **Tables**

1274

1275 **Table 1. Evolution of the ZF in PRDM9 orthologs with different domain**
1276 **architectures.** PRDM9 orthologs for which an empirical comparison dataset is available
1277 are ordered by their domain structures: from the top, we present cases of complete
1278 PRDM9 orthologs with KRAB-SSXRD-SET domains; partial orthologs putatively
1279 lacking KRAB or SSXRD domains or partial orthologs lacking both; then those
1280 containing only the SET domain. A row is shaded green if the ZF is in the top 5% most
1281 rapidly evolving C2H2 ZF in the species, as summarized by the proportion of amino-acid
1282 diversity at DNA-binding sites, and is dark green if it is ranked first. A complete PRDM9
1283 ortholog from dolphins (*Balaenoptera acufoorostrata scammoni*) is shaded in gray
1284 because there is no amino acid diversity between ZFs of the tandem array. The empirical
1285 rank is also shown, as are the number of PRDM9 orthologs identified in the species.
1286 Asterisks indicate PRDM9 orthologs known to play a role in directing recombination. For
1287 PRDM9 genes from teleost fish, under major group, we additionally indicate whether or
1288 not the gene is a PRDM9 α or PRDM9 β gene.
1289

Organism	Major group	PRDM9 structure	Proportion AA diversity at DNA-binding sites	Rank	Number of PRDM9 genes from species	Number of ZF genes evaluated from species
<i>Balaenoptera acutorostrata scammoni</i>	placental	KRAB-SSXRD-SET	NA	NA	1	272
<i>Bison bison bison</i>	placental	KRAB-SSXRD-SET	0.667	1	1	285
<i>Bos taurus* (chr1)</i>	placental	KRAB-SSXRD-SET	0.684	1	3	313
<i>Bos taurus (chrX)</i>	placental	KRAB-SSXRD-SET	0.414	6	3	313
<i>Bos taurus* (chrX)</i>	placental	KRAB-SSXRD-SET	0.414	7	3	313
<i>Bubalus bubalis</i>	placental	KRAB-SSXRD-SET	0.667	1	1	268
<i>Chelonia mydas</i>	turtle	KRAB-SSXRD-SET	0.414	11	1	235
<i>Chlorocebus sabaeus</i>	placental	KRAB-SSXRD-SET	0.500	1	1	344
<i>Chrysemys picta bellii</i>	turtle	KRAB-SSXRD-SET	0.478	1	1	308
<i>Cricetus griseus</i>	placental	KRAB-SSXRD-SET	0.781	3	1	259
<i>Dasyprocta novercinctus</i>	placental	KRAB-SSXRD-SET	0.614	1	1	289
<i>Dipodomys ordii</i>	placental	KRAB-SSXRD-SET	0.567	1	1	194
<i>Esox lucius</i>	teleost fish (α)	KRAB-SSXRD-SET	0.455	1	4	234
<i>Fukomys damarensis</i>	placental	KRAB-SSXRD-SET	0.430	3	1	227
<i>Homo sapiens*</i>	placental	KRAB-SSXRD-SET	0.687	1	1	357
<i>Latimeria chalumnae</i>	coelacanth	KRAB-SSXRD-SET	0.545	2	1	227
<i>Loxodonta africana</i>	placental	KRAB-SSXRD-SET	0.617	1	1	381
<i>Macaca fascicularis</i>	placental	KRAB-SSXRD-SET	0.680	1	1	364
<i>Macaca mulatta</i>	placental	KRAB-SSXRD-SET	0.645	1	1	366
<i>Marmota marmota marmota</i>	placental	KRAB-SSXRD-SET	0.483	1	1	277
<i>Microcebus murinus</i>	placental	KRAB-SSXRD-SET	1.000	1	1	326
<i>Mus musculus*</i>	placental	KRAB-SSXRD-SET	0.910	1	1	224
<i>Nannopaleax galii</i>	placental	KRAB-SSXRD-SET	1.000	1	1	307
<i>Octodon degus</i>	placental	KRAB-SSXRD-SET	0.333	5	3	227
<i>Octodon degus</i>	placental	KRAB-SSXRD-SET	0.331	6	3	227
<i>Ovis aries</i>	placental	KRAB-SSXRD-SET	0.615	1	2	252
<i>Ovis aries</i>	placental	KRAB-SSXRD-SET	0.398	4	2	252
<i>Ovis aries musimon</i>	placental	KRAB-SSXRD-SET	0.353	12	1	285
<i>Papio anubis</i>	placental	KRAB-SSXRD-SET	0.565	1	1	404
<i>Pelodiscus sinensis</i>	turtle	KRAB-SSXRD-SET	0.692	1	1	221
<i>Peromyscus maniculatus bairdii</i>	placental	KRAB-SSXRD-SET	1.000	1	1	243
<i>Protobothrops mucrosquamatus</i>	squamata	KRAB-SSXRD-SET	0.462	5	1	195
<i>Python bivittatus</i>	squamata	KRAB-SSXRD-SET	0.571	1	1	206
<i>Rattus norvegicus</i>	placental	KRAB-SSXRD-SET	0.570	1	1	255
<i>Rousettus aegyptiacus</i>	placental	KRAB-SSXRD-SET	0.742	1	1	258
<i>Salmo salar</i>	teleost fish (α)	KRAB-SSXRD-SET	0.538	9	4	510
<i>Salmo salar</i>	teleost fish (α)	KRAB-SSXRD-SET	0.500	11	4	510
<i>Sus scrofa</i>	placental	KRAB-SSXRD-SET	0.542	1	1	248
<i>Thamnophis sirtalis</i>	squamata	KRAB-SSXRD-SET	0.459	3	1	179
<i>Tupai chinenensis</i>	placental	KRAB-SSXRD-SET	1.000	1	1	249
<i>Tursiops truncatus</i>	placental	KRAB-SSXRD-SET	0.939	1	1	233
<i>Myotis lucifugus</i>	placental	SSXRD-SET	0.524	1	2	308
<i>Myotis lucifugus</i>	placental	SSXRD-SET	0.310	68	2	308
<i>Octodon degus</i>	placental	SSXRD-SET	0.282	46	3	227
<i>Sarcophilus harisi</i>	marsupial	SSXRD-SET	0.224	277	2	344
<i>Callorhinus milii</i>	cartilaginous fish	KRAB-SET	0.314	6	1	63
<i>Astyanax mexicanus</i>	teleost fish (α)	SET	0.258	60	2	158
<i>Astyanax mexicanus</i>	teleost fish (β)	SET	0.167	152	2	158
<i>Clupea harengus</i>	teleost fish (α)	SET	0.279	6	4	118
<i>Clupea harengus</i>	teleost fish (α)	SET	0.278	7	4	118
<i>Clupea harengus</i>	teleost fish (α)	SET	0.274	10	4	118
<i>Clupea harengus</i>	teleost fish (β)	SET	0.158	114	4	118
<i>Cynoglossus semilaevis</i>	teleost fish (β)	SET	0.182	80	1	107
<i>Danio rerio</i>	teleost fish (β)	SET	0.179	345	1	367
<i>Esox lucius</i>	teleost fish (α)	SET	0.295	32	4	234
<i>Esox lucius</i>	teleost fish (β)	SET	0.192	176	4	234
<i>Esox lucius</i>	teleost fish (β)	SET	0.192	177	4	234
<i>Fundulus heteroclitus</i>	teleost fish (β)	SET	0.189	158	1	206
<i>Haplochromis burtoni</i>	teleost fish (β)	SET	0.180	148	1	168
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.320	14	8	140
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.319	15	8	140
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.306	24	8	140
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.303	25	8	140
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.286	33	8	140
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.276	39	8	140
<i>Ictalurus punctatus</i>	teleost fish (α)	SET	0.253	55	8	140
<i>Ictalurus punctatus</i>	teleost fish (β)	SET	0.179	127	8	140
<i>Larimichthys crocea</i>	teleost fish (β)	SET	0.192	70	1	115
<i>Lepisosteus osseus</i>	holoste fish	SET	0.223	48	1	106
<i>Maylandia zebra</i>	teleost fish (β)	SET	0.173	161	1	176
<i>Neolamprologus brichardi</i>	teleost fish (β)	SET	0.173	141	1	152
<i>Nothobranchius furzeri</i>	teleost fish (β)	SET	0.180	245	1	266
<i>Nothobranchius furzeri</i>	teleost fish (β)	SET	0.167	83	1	87
<i>Oreochromis niloticus</i>	teleost fish (β)	SET	0.173	173	1	190
<i>Oryzias latipes</i>	teleost fish (β)	SET	0.213	104	1	191
<i>Otolemur garnettii</i>	placental	SET	0.266	121	1	285
<i>Poecilia formosa</i>	teleost fish (β)	SET	0.191	184	1	242
<i>Poecilia latipinna</i>	teleost fish (β)	SET	0.191	175	1	235
<i>Poecilia mexicana</i>	teleost fish (β)	SET	0.191	187	1	244
<i>Poecilia reticulata</i>	teleost fish (β)	SET	0.191	162	1	212
<i>Pundamilia nyererei</i>	teleost fish (β)	SET	0.173	134	1	147
<i>Pygocentrus nattereri</i>	teleost fish (α)	SET	0.331	12	2	142
<i>Pygocentrus nattereri</i>	teleost fish (β)	SET	0.179	124	2	142
<i>Salmo salar</i>	teleost fish (β)	SET	0.188	411	4	510
<i>Salmo salar</i>	teleost fish (β)	SET	0.180	454	4	510
<i>Sinocyclocheilus anshuiensis</i>	teleost fish (β)	SET	0.185	224	2	284
<i>Sinocyclocheilus anshuiensis</i>	teleost fish (β)	SET	0.185	225	2	284
<i>Sinocyclocheilus grahami</i>	teleost fish (β)	SET	0.185	211	1	271
<i>Sinocyclocheilus rhinocerous</i>	teleost fish (β)	SET	0.185	208	2	269
<i>Sinocyclocheilus rhinocerous</i>	teleost fish (β)	SET	0.185	209	2	269
<i>Takifugu rubripes</i>	teleost fish (β)	SET	0.188	66	1	98
<i>Xiphophorus maculatus</i>	teleost fish (β)	SET	0.191	117	1	158

1291 **Supplementary Files**
1292

1293 Supplementary Files 1, 2 and 3 are provided as excel documents. Alignments of KRAB
1294 and SET domains of PRDM9 genes included in this study are available online at the
1295 Dryad Digital Repository (doi: XXX).

1296
1297 **Supplementary File 1A.** PRDM9 orthologs identified in RefSeq and whole genome
1298 databases. Includes which amino acids are found aligning to each of three catalytic
1299 tyrosine residues of the human PRDM9 SET domain for each PRDM9 ortholog.

1300
1301 **Supplementary File 1B.** Genomes targeted for the PRDM9 search. Major groups or
1302 individual species lacking PRDM9 in RefSeq were targeted for further analysis of their
1303 whole genome sequences, with the exception of previously reported bird and crocodilian
1304 losses. Species included and results of this search are reported here.

1305
1306 **Supplementary File 2A.** Accession numbers and assembly descriptions of publicly
1307 available testes RNAseq samples used for de novo assembly and assessment of PRDM9
1308 expression. N50 describes the shortest contig length in which 50% of the assembled
1309 transcriptome is contained.

1310
1311 **Supplementary File 2B.** Summary of expression results of PRDM9 in the testis in
1312 representative species from major taxa. Only species that passed the core recombination
1313 protein quality test (see Methods, **Supplementary File 2C**) are included in this table,
1314 with the exception of cases, indicated with asterisks, in which PRDM9 was detected but
1315 one or more conserved recombination proteins were not.

1316
1317 **Supplementary File 2C.** Results of a *rpsblast* search of assembled transcriptomes and a
1318 reciprocal best *blast* test to PRMD9. Domain structures found in transcripts that blasted
1319 to PRDM9 for each species are also listed.

1320
1321 **Supplementary File 2D.** Results of the core recombination protein test for each species
1322 for which a transcriptome was assembled. Blue shading indicates that a reciprocal best
1323 *blast* test did not identify the gene in the transcriptome.

1324
1325 **Supplementary File 3A.** Rates of amino acid evolution in SET domains of
1326 representative PRDM9 orthologs lacking other functional domains. To determine whether
1327 PRDM9 orthologs lacking functional domains are non-functional, we compared rates of
1328 evolution between each PRDM9 ortholog missing a domain and another sequence (listed
1329 here) with the complete domain structure. The number of aligned bases and the results of
1330 a likelihood ratio test of non-neutral versus neutral evolution are also shown. See
1331 Methods for details.

1332
1333 **Supplementary File 3B.** Amino acid diversity levels of PRDM9 ZF arrays and the
1334 proportion localized to known DNA-binding residues. Columns labeled V1-V28 indicate
1335 the amount of amino acid diversity observed at each amino acid in the ZF array. For each
1336 gene, we also report the ranking of this proportion relative to all other C2H2 ZF genes

1337 from the same species, when such a ranking was feasible. This table additionally includes
1338 the average percent DNA identity between ZFs used in our analysis of rapid evolution.
1339

1340 **Supplementary File 3C.** Results of the likelihood ratio test of neutral versus not non-
1341 neutral evolution along the SET domain of mammalian PRDM9 orthologs lacking a
1342 KRAB or SSXRD domain, as annotated in RefSeq (see **Methods**). We also indicate
1343 whether another annotated ortholog exists with a KRAB domain.

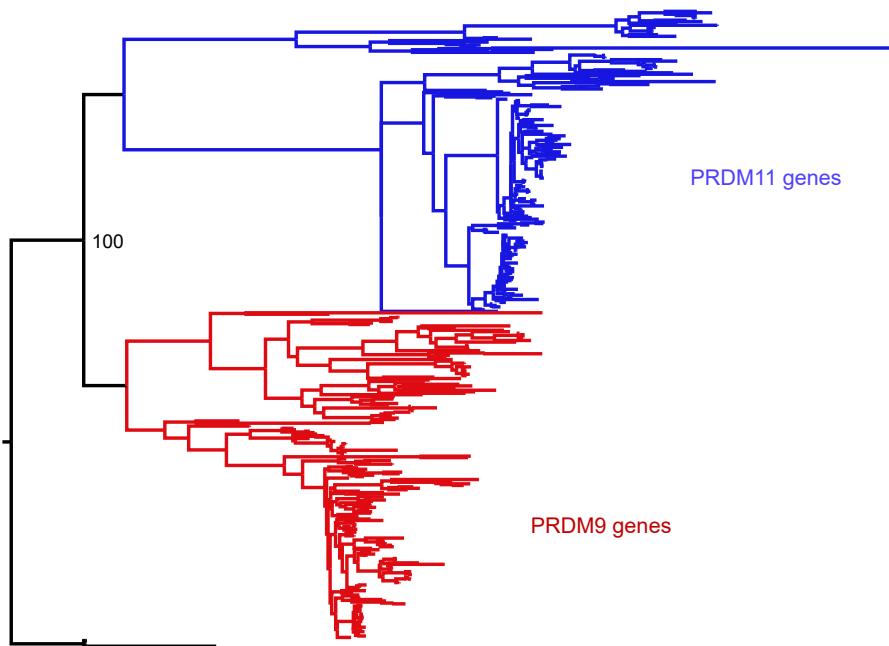
1344
1345 **Supplementary Script 1.** R script to convert GenPept/GenBank files for RefSeq genes
1346 into table format.
1347

1348 **Supplementary Script 2.** Shell script to perform reciprocal best blast search of
1349 transcripts from de novo assembly of testis transcriptome.
1350

1351 **Figure Supplements**

1352

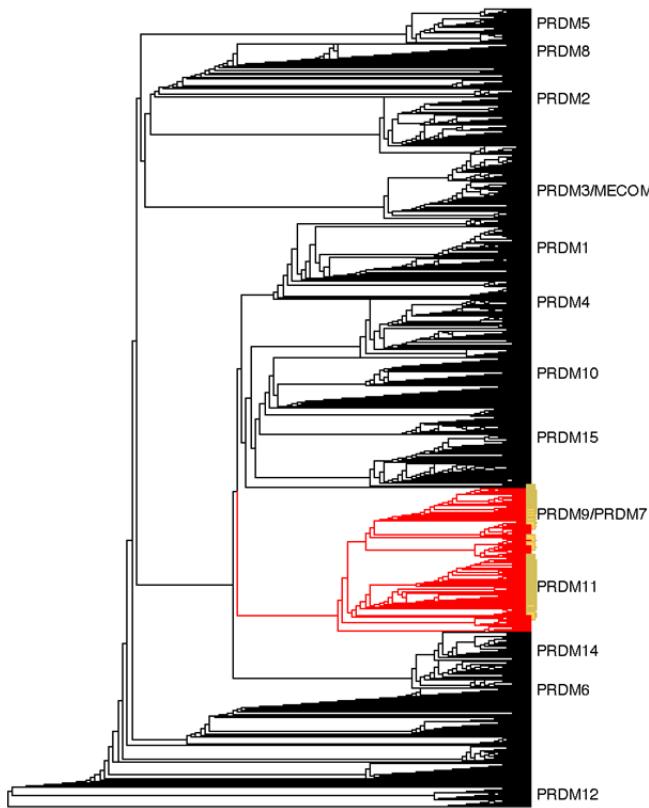
1353



1354

1355 **Figure 1- Figure Supplement 1.** Phylogenetic approach to identifying PRDM9
1356 orthologs and related gene families. A maximum likelihood phylogeny built with
1357 RAxML, using an alignment of SET domains, distinguishes between genes that cluster
1358 with mammalian PRDM9 and PRDM11 with 100% bootstrap support. Genes shown in
1359 black, which are orthologous to both PRDM9 and PRDM11, are only found in jawless
1360 fish.
1361

1362



1363

1364

1365 **Figure 1- Figure Supplement 2.** Neighbor-joining (NJ) guide tree based on the SET
1366 domain. A NJ guide tree analysis on SET domains identified in our RefSeq, whole
1367 genome assembly, and transcriptome datasets was used as an initial step to identify
1368 sequences clustering with human PRDM9/7 or PRDM11. These sequences (in red) were
1369 selected for phylogenetic analysis with RAxML; they included all RefSeq genes in our
1370 dataset that have been previously annotated as PRDM9/7 or PRDM11 (in yellow). Genes
1371 more closely related to known PRDM genes other than PRDM9 or PRDM11 (in black)
1372 were excluded from further analysis.

1373

1374

1375

1376

1377

1378

1379

1380

1381

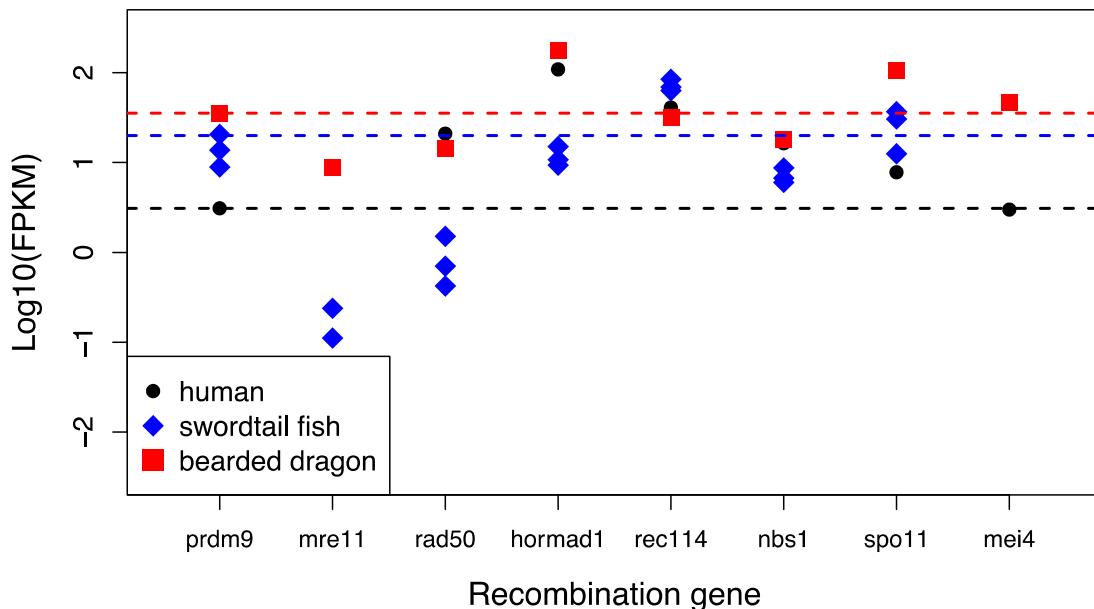
1382

1383

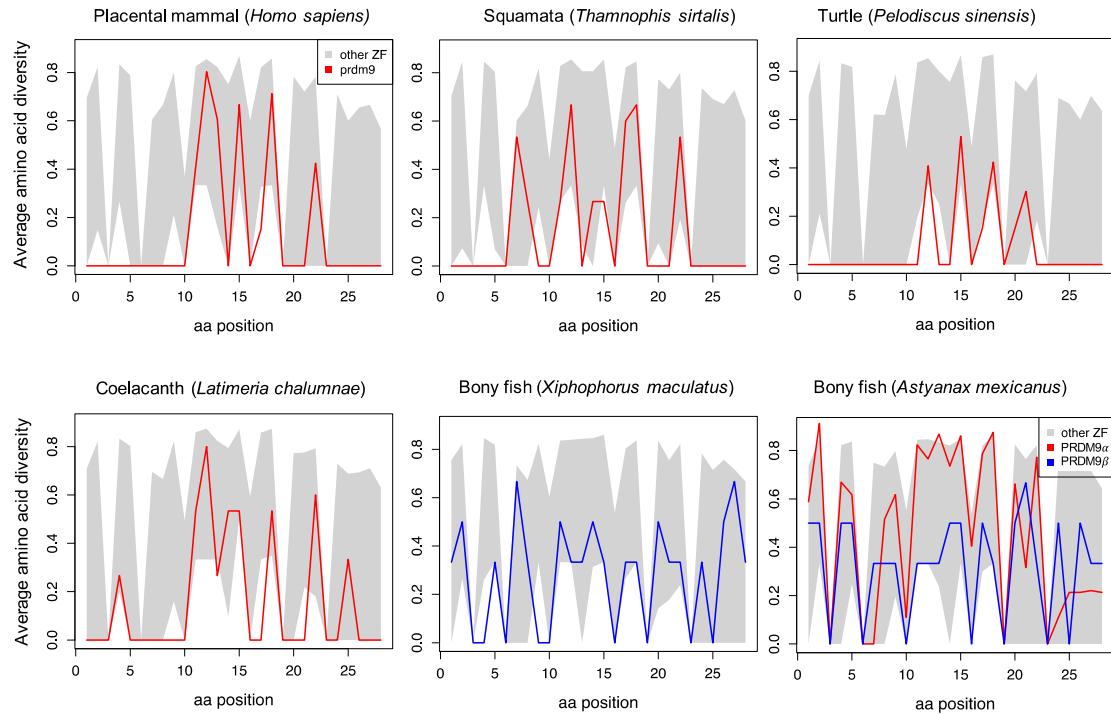
1384

1385

1386

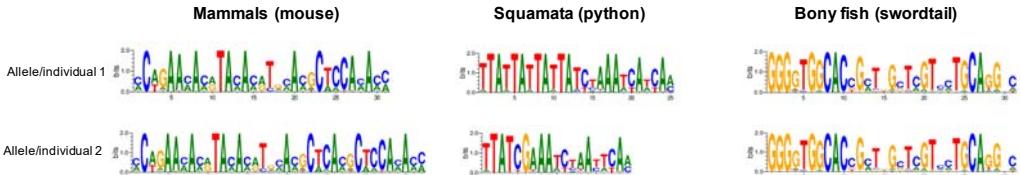


1387
1388
1389 **Figure 1- Figure Supplement 3.** Expression levels of genes with a known role in
1390 meiotic recombination in testes of three exemplar species: human, swordtail fish and
1391 bearded dragon (a lizard). For three swordtails (*X. malinche*) and one bearded dragon,
1392 the FPKM per individual is plotted for each transcript. For humans, the point represents the
1393 average expression of 122 individuals from the gene expression atlas (see **Methods**). For
1394 bearded dragons, PRDM9 and RAD50 were represented by multiple transcripts (two and
1395 three respectively), and the average expression level is shown. Dashed lines show the
1396 point estimate or average expression level of PRDM9 to highlight that several genes in
1397 each species have expression levels comparable to or lower than PRDM9 in testes.
1398



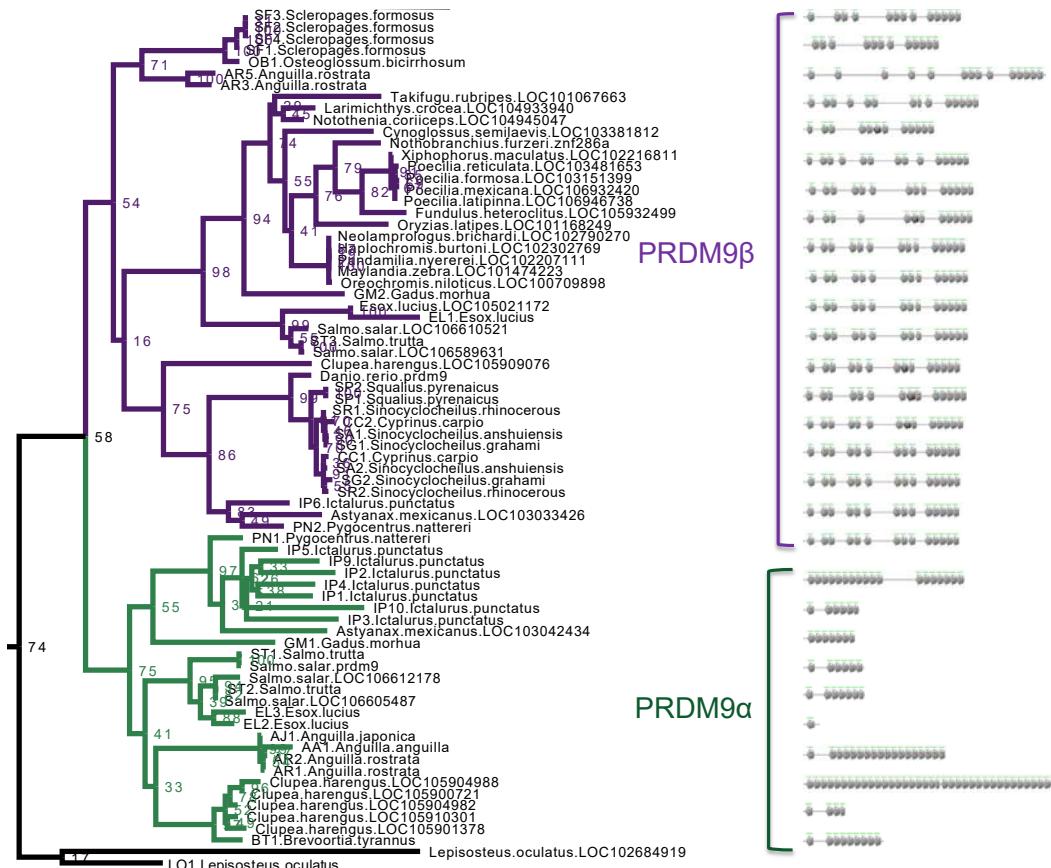
1399

1400 **Figure 1- Figure Supplement 4.** Amino acid diversity as a function of amino acid
 1401 position in the ZF alignment for six exemplar species. Each plot shows the 95% range of
 1402 diversity levels at that site for all C2H2 ZFs from a species of that taxon (gray); the
 1403 values at PRDM9 are show in red or blue. Turtles, snakes and coelacanth show a pattern
 1404 of diversity that is similar to those in mammalian species with a complete PRDM9
 1405 ortholog, with higher diversity at DNA-binding sites (residues 11, 12, 15 and 18) and
 1406 reduced diversity at most other sites. In bony fish, this pattern is not observed in
 1407 PRDM9 β genes (blue) or in partial PRDM9 α genes (shown for *A. mexicanus*), where
 1408 PRDM9 ZF diversity is more typical of other C2H2 ZFs.
 1409
 1410

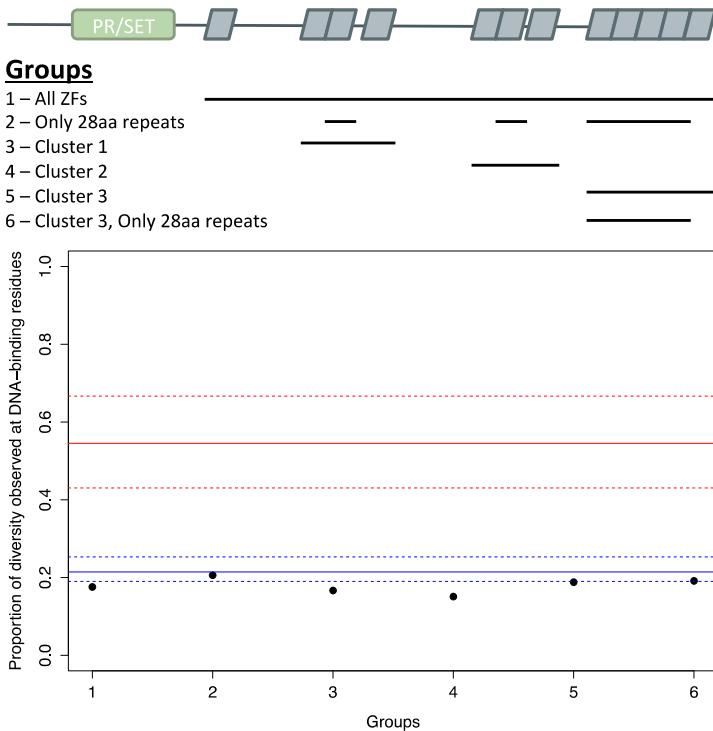


1411
1412
1413
1414
1415
1416
1417
1418
1419

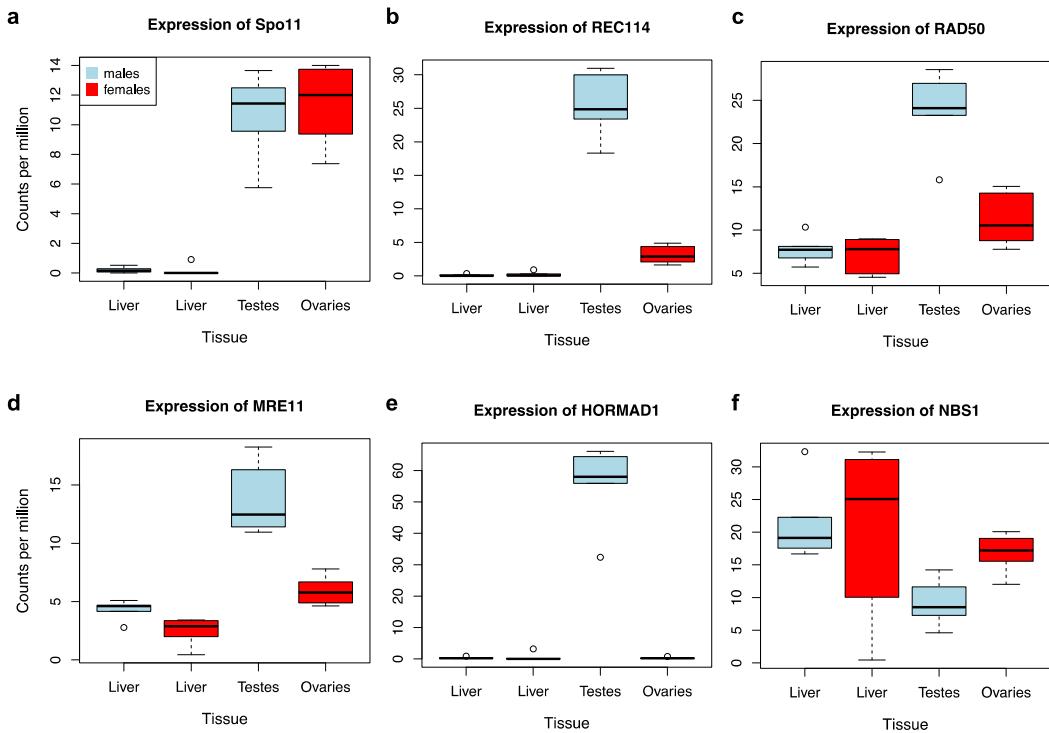
Figure 1- Figure Supplement 5. Examples of differences in computationally predicted PRDM9 binding motifs for species from three taxa. Shown are two mouse from the same species (*Mus musculus* subspecies; Genbank: AB844114.1; FJ899852.1), two pythons from the same species (*Python bivittatus*; the genome sequence and a Sanger resequenced individual; see **Methods**), and two species of swordtail fish (*X. birchmanni* and *X. malinche*; genome sequences). The position weight matrix was obtained using C2H2 prediction tools available at <http://zf.princeton.edu>.



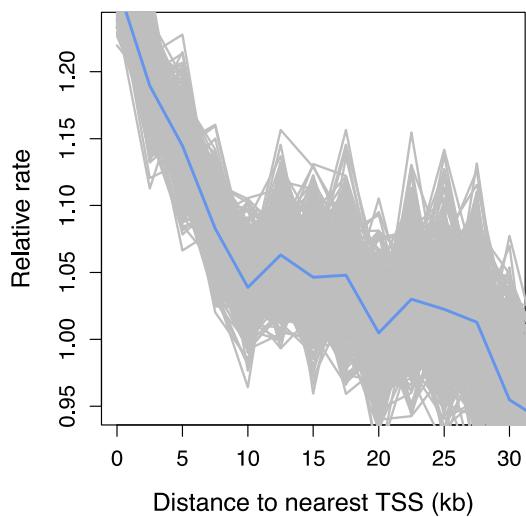
1420
 1421 **Figure 2- Figure Supplement 1.** Section of maximum-likelihood phylogeny of the SET
 1422 domain showing bony fish PRDM9 orthologs α and β . The reciprocal monophyly of
 1423 PRDM9 orthologs α and β is reasonably well supported and in particular bootstrap
 1424 support for the monophyly of PRDM9 α genes is 75%. The ZF domains for representative
 1425 PRDM9 orthologs of each type are shown to the right, with each gray pentagon
 1426 indicating the location of a ZF. In swordtail fish, the complete ZF array is found within a
 1427 single exon, and the last tandem array of six ZFs forms a minisatellite structure.
 1428



1430 **Figure 2- Figure Supplement 2.** Analysis of ZF evolution in PRDM9 β . Red lines show
1431 the median (solid) and first and third quantiles (dashed lines) for all 48 complete PRDM9
1432 orthologs identified in vertebrates that have four or more ZFs. Blue lines show the
1433 median (solid) and first and third quantiles (dashed lines) for all other C2H2 ZF genes
1434 from *X. maculatus* (157 genes). Results about the rate of ZF evolution in the PRDM9 β
1435 gene from *Xiphophorus maculatus* are qualitatively similar regardless of our choice of
1436 which cluster of individual ZFs domains to include in our analysis, indicating that our
1437 ability to detect evidence of positive selection at DNA-binding residues in these arrays, or
1438 lack thereof, is unlikely to be influenced by this choice.
1439



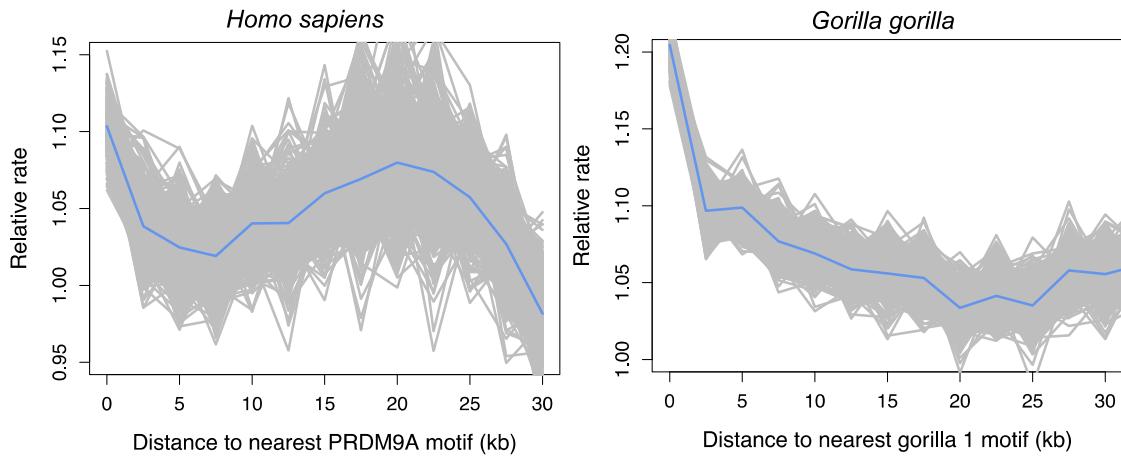
1440
 1441 **Figure 4- Figure Supplement 1.** Expression levels of meiosis-related genes in swordtail
 1442 fish tissues. In general, the seven meiosis-related genes surveyed had higher expression in
 1443 tissues containing germline cells than liver tissue, but this pattern was much more
 1444 pronounced in testis tissue (compared to ovary tissue). As a result, we focused our
 1445 analysis of meiosis related genes on RNAseq data generated from testis. Results shown
 1446 are based on analysis of three male and female biological replicates from each swordtail
 1447 species (*X. birchmanni* and *X. malinche*).
 1448



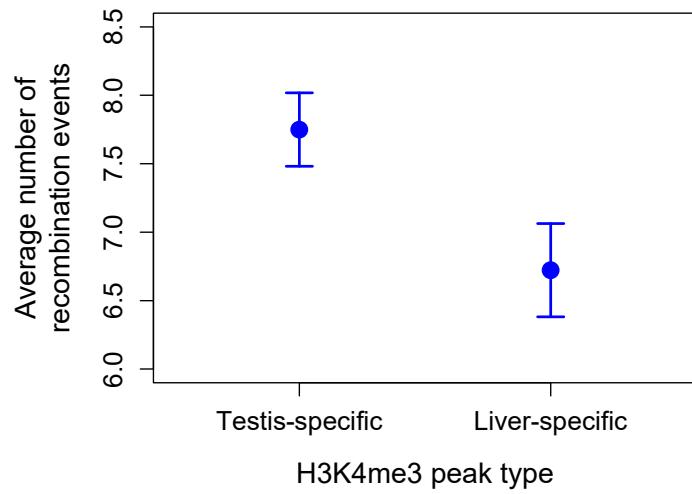
1449
1450
1451
1452
1453

Figure 4 – Figure Supplement 2. Recombination frequency in swordtails as a function of distance to the TSS. Partial correlation analyses suggest that the association between the TSS and recombination rate in swordtails is explained by H3K4me3 peaks and CGIs.

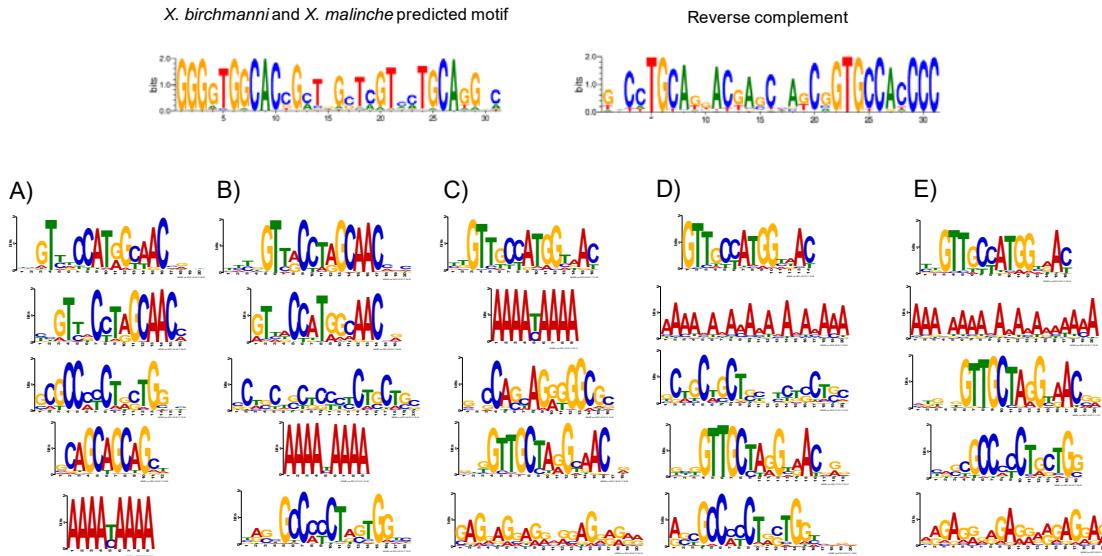
1454
1455



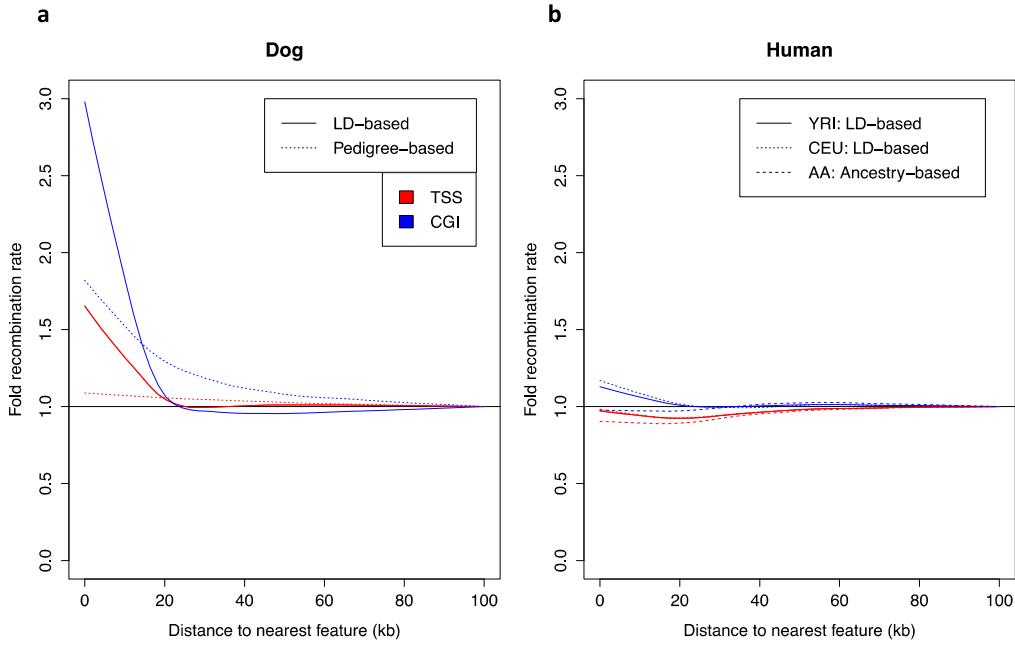
1456
1457
1458
1459
1460
1461
1462 we present for swordtails in **Figure 4F**, we therefore determined recombination rate in
1463 humans (using the map based on LD patterns in the CEU; Frazer et al. 2007) as a
1464 function of distance to computationally predicted binding sites for the PRDM9A motif in
1465 humans and as a function of distance to computationally predicted binding sites for the
1466 gor-1 PRDM9 allele (Schwartz et al. 2014) in gorillas (using the LD-based map from
1467 Stevenson et al. 2016).
1468



1470
1471 **Figure 4 - Figure Supplement 4.** Higher observed recombination rate at testis-specific
1472 H3K4me3 peaks than liver-specific H3K4me3 peaks. H3K4me3 peaks found only in the
1473 testis and not in the liver of *X. birchmanni* have higher observed recombination rates in
1474 *X. birchmanni* – *X. malinche* hybrids. This pattern supports the conclusion that H3K4me3
1475 peaks are associated with recombination in swordtails.
1476



1477
 1478 **Figure 4 - Figure Supplement 5.** MEME prediction of sequences enriched in testis-
 1479 H3K4me3 peaks relative to liver-specific H3K4me3 peaks. Results shown in A-E are
 1480 from five replicate runs of 2,000 testis-specific sequences using liver-specific sequences
 1481 as a background comparison set. The swordtail computationally-predicted PRDM9
 1482 binding motif is shown for comparison.
 1483



1484
1485 **Figure 5 – Figure supplement 1. Dependence of patterns of recombination near**
1486 **TSSs and CGIs in dog and human on the type of genetic map. (a)** Recombination
1487 rates near the TSS and CGI in dogs are shown using recombination maps inferred either
1488 from LD patterns or pedigree data. The magnitude of the peak near these features is lower
1489 in the map with lower resolution. This observation raises the possibility that a higher
1490 resolution map in swordtail fish would have a higher peak near these features. **(b)**
1491 Recombination rates near the TSS and CGI in humans are shown using recombination
1492 maps inferred either from LD patterns or ancestry switches in African-American samples.
1493 Recombination rates near the TSS and CGI in human do not seem to be strongly
1494 influenced by the choice of genetic map, though peaks at these features are slightly
1495 reduced in admixture- and pedigree-based methods.

1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507

1508 **References**

1509

1510 Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local
1511 alignment search tool. *Journal of Molecular Biology* 215:403-410.

1512 Amores, A., J. Catchen, I. Nanda, W. Warren, R. Walter, M. Schartl, and J. H.
1513 Postlethwait. 2014. A RAD-Tag Genetic Map for the Platypus (*Xiphophorus*
1514 *maculatus*) Reveals Mechanisms of Karyotype Evolution Among Teleost Fish.
1515 *Genetics* 197:625-U307.

1516 Andolfatto, P., D. Davison, D. Eryilmaz, T. T. Hu, J. Mast, T. Sunayama-Morita, and
1517 D. L. Stern. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic
1518 mapping. *Genome Res* 21:610-617.

1519 Auton, A., A. Fledel-Alon, S. Pfeifer, O. Venn, L. Segurel, T. Street, E. M. Leffler, R.
1520 Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R.
1521 D. Hernandez, C. Melton, A. Venkat, M. A. Nobrega, R. Bontrop, S. Myers, P.
1522 Donnelly, M. Przeworski, and G. McVean. 2012. A Fine-Scale Chimpanzee
1523 Genetic Map from Population Sequencing. *Science* 336:193-198.

1524 Auton, A., Y. R. Li, J. Kidd, K. Oliveira, J. Nadel, J. K. Holloway, J. J. Hayward, P. E.
1525 Cohen, J. M. Greally, J. Wang, C. D. Bustamante, and A. R. Boyko. 2013.
1526 Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs.
1527 *Plos Genetics* 9.

1528 Baker, C. L., S. Kajita, M. Walker, R. L. Saxl, N. Raghupathy, K. Choi, P. M. Petkov,
1529 and K. Paigen. 2015. PRDM9 Drives Evolutionary Erosion of Hotspots in *Mus*
1530 *musculus* through Haplotype-Specific Initiation of Meiotic Recombination. *Plos*
1531 *Genetics* 11.

1532 Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B.
1533 de Massy. 2010. PRDM9 Is a Major Determinant of Meiotic Recombination
1534 Hotspots in Humans and Mice. *Science* 327:836-840.

1535 Berg, I. L., R. Neumann, K. W. G. Lam, S. Sarbajna, L. Odenthal-Hesse, C. A. May, and
1536 A. J. Jeffreys. 2010. PRDM9 variation strongly influences recombination hot-spot
1537 activity and meiotic instability in humans. *Nature Genetics* 42:859-+.

1538 Blazer, L. L., E. Lima-Fernandes, E. Gibson, M. S. Eram, P. Loppnau, C. H. Arrowsmith,
1539 M. Schapira, and M. Vedadi. 2016. PR Domain-containing Protein 7 (PRDM7) Is
1540 a Histone 3 Lysine 4 Trimethyltransferase. *Journal of Biological Chemistry*
1541 291:13509-13519.

1542 Borde, V., N. Robine, W. Lin, S. Bonfils, V. Geli, and A. Nicolas. 2009. Histone H3
1543 lysine 4 trimethylation marks meiotic recombination initiation sites. *Embo Journal*
1544 28:99-111.

1545 Boulton, A., R. S. Myers, and R. J. Redfield. 1997. The hotspot conversion paradox and
1546 the evolution of meiotic recombination. *Proceedings of the National Academy of*
1547 *Sciences of the United States of America* 94:8058-8063.

1548 Brick, K., F. Smagulova, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova. 2012.
1549 Genetic recombination is directed away from functional genomic elements in
1550 mice. *Nature* 485:642-645.

1551 Brunschwig, H., L. Levi, E. Ben-David, R. W. Williams, B. Yakir, and S. Shifman. 2012.
1552 Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome.
1553 *Genetics* 191:757-U169.

1554 Board, J., P. Barthes, C. Grey, and B. de Massy. 2009. Distinct histone modifications
1555 define initiation and repair of meiotic recombination in the mouse. *Embo Journal*
1556 28:2616-2624.

1557 Castoe, T. A., A. P. J. de Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K. Fujita, R.
1558 P. Ruggiero, J. F. Degner, J. M. Daza, W. J. Gu, J. Reyes-Velasco, K. J. Shaney,
1559 J. M. Castoe, S. E. Fox, A. W. Poole, D. Polanco, J. Dobry, M. W. Vandewege,
1560 Q. Li, R. K. Schott, A. Kapusta, P. Minx, C. Feschotte, P. Uetz, D. A. Ray, F. G.
1561 Hoffmann, R. Bogden, E. N. Smith, B. S. W. Chang, F. J. Vonk, N. R. Casewell,
1562 C. V. Henkel, M. K. Richardson, S. P. Mackessy, A. M. Bronikowski, M. Yandell,
1563 W. C. Warren, S. M. Secor, and D. D. Pollock. 2013. The Burmese python
1564 genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings*
1565 of the National Academy of Sciences of the United States of America 110:20645-
1566 20650.

1567 Chan, A. H., P. A. Jenkins, and Y. S. Song. 2012. Genome-Wide Fine-Scale
1568 Recombination Rate Variation in *Drosophila melanogaster*. *Plos Genetics* 8.

1569 Choi, K. H., X. H. Zhao, K. A. Kelly, O. Venn, J. D. Higgins, N. E. Yelina, T. J.
1570 Hardcastle, P. A. Ziolkowski, G. P. Copenhaver, F. C. H. Franklin, G. McVean,
1571 and I. R. Henderson. 2013. *Arabidopsis* meiotic crossover hot spots overlap with
1572 H2A.Z nucleosomes at gene promoters. *Nature Genetics* 45:1327-+.

1573 Coop, G. and S. R. Myers. 2007. Live hot, die young: Transmission distortion in
1574 recombination hotspots. *Plos Genetics* 3:377-386.

1575 Coop, G. and M. Przeworski. 2007. An evolutionary view of human recombination.
1576 *Nature Reviews Genetics* 8:23-34.

1577 Coop, G., X. Q. Wen, C. Ober, J. K. Pritchard, and M. Przeworski. 2008. High-resolution
1578 mapping of crossovers reveals extensive variation in fine-scale recombination
1579 patterns among humans. *Science* 319:1395-1398.

1580 Crawford, N. 2014. `interleave_fastq.py`. GitHubGist.
1581 <https://gist.github.com/ngcrawford/2232505>

1582 Cui, R., M. Schumer, K. Kruesi, R. Walter, P. Andolfatto, and G. Rosenthal. 2013.
1583 Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes.
1584 *Evolution* 67:2166-2179.

1585 Davies, B., E. Hatton, N. Altemose, J. G. Hussin, F. Pratto, G. Zhang, A. G. Hinch, D.
1586 Moralli, D. Biggs, R. Diaz, C. Preece, R. Li, E. Bitoun, K. Brick, C. M. Green, R.
1587 D. C. Amerini-Otero, S. R. Myers, and P. Donnelly. 2016. Re-engineering the
1588 zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* 530:171-+.

1589 de Massy, B. 2013. Initiation of Meiotic Recombination: How and Where? Conservation
1590 and Specificities Among Eukaryotes. *Annual Review of Genetics*, Vol 47 47:563-
1591 599.

1592 Duret, L. and N. Galtier. 2009. Biased Gene Conversion and the Evolution of
1593 Mammalian Genomic Landscapes. *Annual Review of Genomics and Human*
1594 *Genetics* 10:285-311.

1595 Eram, M. S., S. P. Bustos, E. Lima-Fernandes, A. Siarheyeva, G. Senisterra, T. Hajian, I.
1596 Chau, S. L. Duan, H. Wu, L. Dombrovski, M. Schapira, C. H. Arrowsmith, and
1597 M. Vedadi. 2014. Trimethylation of Histone H3 Lysine 36 by Human
1598 Methyltransferase PRDM9 Protein. *Journal of Biological Chemistry* 289:12177-
1599 12188.

1600 Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W.
1601 Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T.
1602 D. Willis, F. L. Yu, H. M. Yang, C. Q. Zeng, Y. Gao, H. R. Hu, W. T. Hu, C. H.
1603 Li, W. Lin, S. Q. Liu, H. Pan, X. L. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q.
1604 R. Zhang, H. B. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel,
1605 A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H.
1606 Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D.
1607 Altshuler, Y. Shen, Z. J. Yao, W. Huang, X. Chu, Y. G. He, L. Jin, Y. F. Liu, Y.
1608 Y. Shen, W. W. Sun, H. F. Wang, Y. Wang, X. Y. Xiong, L. Xu, M. M. Y. Waye,
1609 S. K. W. Tsui, J. T. F. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S.
1610 Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M.
1611 Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J.
1612 Hudson, P. Y. Kwok, D. M. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P.
1613 Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y.
1614 Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi,
1615 A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T.
1616 Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P.
1617 Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R.
1618 Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J.
1619 Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, L. D. Stein,
1620 L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P.
1621 E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. H. Guan, Y. Li, H.
1622 M. Munro, Z. H. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N.
1623 Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M.
1624 Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S.
1625 Weir, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol and C. Int
1626 HapMap. 2007. A second generation human haplotype map of over 3.1 million
1627 SNPs. *Nature* 449:851-U853.
1628 Fullerton, S. M., A. B. Carvalho, and A. G. Clark. 2001. Local rates of recombination are
1629 positively correlated with GC content in the human genome. *Molecular Biology*
1630 and *Evolution* 18:1139-1142.
1631 Getun, I. V., Z. Wu, M. Fallahi, S. Ouizem, Q. Liu, W. M. Li, R. Costi, W. R. Roush, J.
1632 L. Cleveland, and P. R. J. Bois. 2017. Functional Roles of Acetylated Histone
1633 Marks at Mouse Meiotic Recombination Hot Spots. *Molecular and Cellular*
1634 *Biology* 37.
1635 Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. FIMO: scanning for occurrences of a
1636 given motif. *Bioinformatics* 27:1017-1018.
1637 Grey, C., J. A. J. Clement, J. Buard, B. Leblanc, I. Gut, M. Gut, L. Duret, and B. de
1638 Massy. 2017. In vivo binding of PRDM9 reveals interactions with noncanonical
1639 genomic sites. *Genome Research* 27:580-590.
1640 Hayashi, K., K. Yoshida, and Y. Matsui. 2005. A histone H3 methyltransferase controls
1641 epigenetic events required for meiotic prophase. *Nature* 438:374-378.
1642 Hedges, S. B., J. Marin, M. Suleski, M. Paymer, and S. Kumar. 2015. Tree of Life
1643 Reveals Clock-Like Speciation and Diversification. *Molecular Biology and*
1644 *Evolution* 32:835-845.

1645 Heil, C. S. S., C. Ellison, M. Dubin, and M. A. F. Noor. 2015. Recombining without
1646 Hotspots: A Comprehensive Evolutionary Portrait of Recombination in Two
1647 Closely Related Species of *Drosophila*. *Genome Biology and Evolution* 7:2829-
1648 2842.

1649 Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre,
1650 H. Singh, and C. K. Glass. 2010. Simple Combinations of Lineage-Determining
1651 Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage
1652 and B Cell Identities. *Molecular Cell* 38:576-589.

1653 Hellsten, U., K. M. Wright, J. Jenkins, S. Q. Shu, Y. W. Yuan, S. R. Wessler, J. Schmutz,
1654 J. H. Willis, and D. S. Rokhsar. 2013. Fine-scale variation in meiotic
1655 recombination in *Mimulus* inferred from population shotgun sequencing.
1656 *Proceedings of the National Academy of Sciences of the United States of America*
1657 110:19478-19482.

1658 Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G.
1659 Sella, M. Przeworski, and P. Genomes. 2011. Classic Selective Sweeps Were
1660 Rare in Recent Human Evolution. *Science* 331:920-924.

1661 Hinch, A. G., A. Tandon, N. Patterson, Y. Song, N. Rohland, C. D. Palmer, G. K. Chen,
1662 K. Wang, S. G. Buxbaum, E. L. Akylbekova, M. C. Aldrich, C. B. Ambrosone, C.
1663 Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, E.
1664 Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L. A. Cupples, S. L. Deming, W. R.
1665 Diver, J. Divers, M. Fornage, E. M. Gillanders, J. Glessner, C. C. Harris, J. J. Hu,
1666 S. A. Ingles, W. Isaacs, E. M. John, W. H. L. Kao, B. Keating, R. A. Kittles, L. N.
1667 Kolonel, E. Larkin, L. Le Marchand, L. H. McNeill, R. C. Millikan, A. Murphy,
1668 S. Musani, C. Neslund-Dudas, S. Nyante, G. J. Papanicolaou, M. F. Press, B. M.
1669 Psaty, A. P. Reiner, S. S. Rich, J. L. Rodriguez-Gil, J. I. Rotter, B. A. Rybicki, A.
1670 G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. J. Thun, M. A. Tucker,
1671 Z. Wang, J. K. Wiencke, J. S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K. A.
1672 Zanetti, W. Zheng, R. G. Ziegler, X. Zhu, S. Redline, J. N. Hirschhorn, B. E.
1673 Henderson, H. A. Taylor, Jr., A. L. Price, H. Hakonarson, S. J. Chanock, C. A.
1674 Haiman, J. G. Wilson, D. Reich, and S. R. Myers. 2011. The landscape of
1675 recombination in African Americans. *Nature* 476:170-U167.

1676 Janousek, V., P. Munclinger, L. Wang, K. C. Teeter, and P. K. Tucker. 2015. Functional
1677 Organization of the Genome May Shape the Species Boundary in the House
1678 Mouse. *Molecular Biology and Evolution* 32:1208-1220.

1679 Jeffreys, A. J., V. E. Cotton, R. Neumann, and K. W. G. Lam. 2013. Recombination
1680 regulator PRDM9 influences the instability of its own coding sequence in
1681 humans. *Proceedings of the National Academy of Sciences of the United States of*
1682 *America* 110:600-605.

1683 Johnston, S. E., C. Berenos, J. Slate, and J. M. Pemberton. 2016. Conserved Genetic
1684 Architecture Underlying Individual Recombination Rate Variation in a Wild
1685 Population of Soay Sheep (*Ovis aries*). *Genetics* 203:583-+.

1686 Jones, J. C., S. Fan, P. Franchini, M. Schartl, and A. Meyer. 2013. The evolutionary
1687 history of Xiphophorus fish and their sexually selected sword: a genome-wide
1688 approach using restriction site-associated DNA sequencing. *Molecular Ecology*
1689 22:2986-3001.

1690 Kauppi, L., M. P. H. Stumpf, and A. J. Jeffreys. 2005. Localized breakdown in linkage
1691 disequilibrium does not always predict sperm crossover hot spots in the human
1692 MHC class II region. *Genomics* 86:13-24.

1693 Lam, I. and S. Keeney. 2014. Mechanism and Regulation of Meiotic Recombination
1694 Initiation. *Cold Spring Harbor Perspectives in Biology* 7.

1695 Lam, I. and S. Keeney. 2015. Nonparadoxical evolutionary stability of the recombination
1696 initiation landscape in yeast. *Science* 350:932-937.

1697 Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2.
1698 *Nature Methods* 9:357-U354.

1699 Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H.
1700 McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J.
1701 Gibson, and D. G. Higgins. 2007. Clustal W and clustal X version 2.0.
1702 *Bioinformatics* 23:2947-2948.

1703 Leinonen, R., H. Sugawara, M. Shumway, and C. Int Nucleotide Sequence Database.
1704 2011. The Sequence Read Archive. *Nucleic Acids Research* 39:D19-D21.

1705 Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-
1706 Wheeler transform. *Bioinformatics* 25.

1707 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis,
1708 R. Durbin, and P. Genome Project Data. 2009. The Sequence Alignment/Map
1709 format and SAMtools. *Bioinformatics* 25:2078-2079.

1710 Lichten, M. and A. S. H. Goldman. 1995. Meiotic recombination hotspots. *Annual
1711 Review of Genetics* 29:423-444.

1712 Ma, L., J. R. O'Connell, P. M. VanRaden, B. T. Shen, A. Padhi, C. Y. Sun, D. M.
1713 Bickhart, J. B. Cole, D. J. Null, G. E. Liu, Y. Da, and G. R. Wiggans. 2015. Cattle
1714 Sex-Specific Recombination and Genetic Control from a Large Pedigree
1715 Analysis. *Plos Genetics* 11.

1716 Marchler-Bauer, A., M. K. Derbyshire, N. R. Gonzales, S. N. Lu, F. Chitsaz, L. Y. Geer,
1717 R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H.
1718 Marchler, J. S. Song, N. Thanki, Z. X. Wang, R. A. Yamashita, D. C. Zhang, C. J.
1719 Zheng, and S. H. Bryant. 2015. CDD: NCBI's conserved domain database.
1720 *Nucleic Acids Research* 43:D222-D226.

1721 Markenscoff-Papadimitriou, E., W. E. Allen, B. M. Colquitt, T. Goh, K. K. Murphy, K.
1722 Monahan, C. P. Mosley, N. Ahituv, and S. Lomvardas. 2014. Enhancer
1723 Interaction Networks as a Means for Singular Olfactory Receptor Expression.
1724 *Cell* 159:543-557.

1725 McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly.
1726 2004. The fine-scale structure of recombination rate variation in the human
1727 genome. *Science* 304:581-584.

1728 Munoz-Fuentes, V., A. Di Rienzo, and C. Vila. 2011. Prdm9, a Major Determinant of
1729 Meiotic Recombination Hotspots, Is Not Functional in Dogs and Their Wild
1730 Relatives, Wolves and Coyotes. *Plos One* 6.

1731 Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map
1732 of recombination rates and hotspots across the human genome. *Science* 310:321-
1733 324.

1734 Myers, S., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G.
1735 McVean, and P. Donnelly. 2010. Drive Against Hotspot Motifs in Primates
1736 Implicates the PRDM9 Gene in Meiotic Recombination. *Science* 327:876-879.

1737 Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean. 2008. A common
1738 sequence motif associated with recombination hot spots and genome instability in
1739 humans. *Nature Genetics* 40:1124-1129.

1740 Nagylaki, T. and T. D. Petes. 1982. INTRACHROMOSOMAL GENE CONVERSION
1741 AND THE MAINTENANCE OF SEQUENCE HOMOGENEITY AMONG
1742 REPEATED GENES. *Genetics* 100:315-337.

1743 Narasimhan, V. M., K. A. Hunt, D. Mason, C. L. Baker, K. J. Karczewski, M. E. R.
1744 Barnes, A. H. Barnett, C. Bates, S. Bellary, N. A. Bockett, K. Giorda, C. J.
1745 Griffiths, H. Hemingway, Z. L. Jia, M. A. Kelly, H. A. Khawaja, M. Lek, S.
1746 McCarthy, R. McEachan, A. O'Donnell-Luria, K. Paigen, C. A. Parisinos, E.
1747 Sheridan, L. Southgate, L. Tee, M. Thomas, Y. L. Xue, M. Schnall-Levin, P. M.
1748 Petkov, C. T. Smith, E. R. Maher, R. C. Trembath, D. G. MacArthur, J. Wright,
1749 R. Durbin, and D. A. Heel. 2016. Health and population effects of rare gene
1750 knockouts in adult humans with related parents. *Science* 352:474-477.

1751 Neph, S., M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E.
1752 Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, and J.
1753 A. Stamatoyannopoulos. 2012. BEDOPS: high-performance genomic feature
1754 operations. *Bioinformatics* 28:1919-1920.

1755 Nicolas, A., D. Treco, N. P. Schultes, and J. W. Szostak. 1989. AN INITIATION SITE
1756 FOR MEIOTIC GENE CONVERSION IN THE YEAST SACCHAROMYCES-
1757 CEREVISIAE. *Nature* 338:35-39.

1758 Oliver, P. L., L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, N. Phadnis, S. A. Beatson,
1759 G. Lunter, H. S. Malik, and C. P. Ponting. 2009. Accelerated Evolution of the
1760 Prdm9 Speciation Gene across Diverse Metazoan Taxa. *Plos Genetics* 5.

1761 Otto, S. P. and N. H. Barton. 1997. The evolution of recombination: Removing the limits
1762 to natural selection. *Genetics* 147:879-906.

1763 Otto, S. P. and T. Lenormand. 2002. Resolving the paradox of sex and recombination.
1764 *Nature Reviews Genetics* 3:252-261.

1765 Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and
1766 Evolution in R language. *Bioinformatics* 20:289-290.

1767 Parvanov, E. D., P. M. Petkov, and K. Paigen. 2010. Prdm9 Controls Activation of
1768 Mammalian Recombination Hotspots. *Science* 327:835-835.

1769 Parvanov, E. D., H. Tian, T. Billings, R. L. Saxl, R. Aithal, L. Krejci, K. Paigen, and P.
1770 M. Petkov. 2016. PRDM9 forms a multiprotein complex tethering recombination
1771 hotspots to the chromosomal axis. *biorxiv*.

1772 Parvanov, E. D., H. Tian, T. Billings, R. L. Saxl, C. Spruce, R. Aithal, L. Krejci, K.
1773 Paigen, and P. M. Petkov. 2017. PRDM9 interactions with other proteins provide
1774 a link between recombination hotspots and the chromosomal axis in meiosis.
1775 *Molecular Biology of the Cell* 28:488-499.

1776 Persikov, A. V. and M. Singh. 2014. De novo prediction of DNA-binding specificities for
1777 Cys(2)His(2) zinc finger proteins. *Nucleic Acids Research* 42:97-108.

1778 Pineda-Krch, M. and R. J. Redfield. 2005. Persistence and loss of meiotic recombination
1779 hotspots. *Genetics* 169:2319-2333.

1780 Powers, N. R., E. D. Parvanov, C. L. Baker, M. Walker, P. M. Petkov, and K. Paigen.
1781 2016. The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36
1782 and H3K4 at Recombination Hotspots In Vivo. *Plos Genetics* 12.

1783 Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski,
1784 K. A. Frazer, and S. Paabo. 2005. Fine-scale recombination patterns differ
1785 between chimpanzees and humans. *Nature Genetics* 37:429-434.

1786 Ptak, S. E., A. D. Roeder, M. Stephens, Y. Gilad, S. Paabo, and M. Przeworski. 2004.
1787 Absence of the TAP2 human recombination hotspot in chimpanzees. *Plos Biology*
1788 2:849-855.

1789 Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing
1790 genomic features. *Bioinformatics* 26:841-842.

1791 Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. 2011. Improving RNA-
1792 Seq expression estimates by correcting for fragment bias. *Genome Biology* 12.

1793 Rockman, M. V. and L. Kruglyak. 2009. Recombinational Landscape and Population
1794 Genomics of *Caenorhabditis elegans*. *Plos Genetics* 5.

1795 Sandor, C., W. B. Li, W. Coppelters, T. Druet, C. Charlier, and M. Georges. 2012.
1796 Genetic Variants in REC8, RNF212, and PRDM9 Influence Male Recombination
1797 in Cattle. *Plos Genetics* 8.

1798 Schartl, M., R. B. Walter, Y. Shen, T. Garcia, J. Catchen, A. Amores, I. Braasch, D.
1799 Chalopin, J.-N. Volff, K.-P. Lesch, A. Bisazza, P. Minx, L. Hillier, R. K. Wilson,
1800 S. Fuerstenberg, J. Boore, S. Searle, J. H. Postlethwait, and W. C. Warren. 2013.
1801 The genome of the platyfish, *Xiphophorus maculatus*, provides insights into
1802 evolutionary adaptation and several complex traits. *Nat Genet* 45:567-U150.

1803 Schierup, M. H. and J. Hein. 2000. Recombination and the molecular clock. *Molecular
1804 Biology and Evolution* 17:1578-1579.

1805 Schulz, M. H., D. R. Zerbino, M. Vingron, and E. Birney. 2012. Oases: robust de novo
1806 RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*
1807 28:1086-1092.

1808 Schumer, M., R. Cui, D. Powell, R. Dresner, G. Rosenthal, and P. Andolfatto. 2014.
1809 High-resolution Mapping Reveals Hundreds of Genetic Incompatibilities in
1810 Hybridizing Fish Species. *eLife*.

1811 Schumer, M., R. F. Cui, D. L. Powell, G. G. Rosenthal, and P. Andolfatto. 2016. Ancient
1812 hybridization and genomic stabilization in a swordtail fish. *Molecular Ecology*
1813 25:2661-2679.

1814 Schwartz, J. J., D. J. Roach, J. H. Thomas, and J. Shendure. 2014. Primate evolution of
1815 the recombination regulator PRDM9. *Nature Communications* 5.

1816 Segurel, L., E. M. Leffler, and M. Przeworski. 2011. The Case of the Fickle Fingers:
1817 How the PRDM9 Zinc Finger Protein Specifies Meiotic Recombination Hotspots
1818 in Humans. *Plos Biology* 9.

1819 Setiamarga, D. H. E., M. Miya, Y. Yamanoue, K. Mabuchi, T. P. Satoh, J. G. Inoue, and
1820 M. Nishida. 2008. Interrelationships of Atherinomorpha (medakas, flyingfishes,
1821 killifishes, silversides, and their relatives): The first evidence based on whole
1822 mitogenome sequences. *Molecular Phylogenetics and Evolution* 49:598-605.

1823 Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Z. Li, R. Lopez, H.
1824 McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. 2011.

1825 Fast, scalable generation of high-quality protein multiple sequence alignments
1826 using Clustal Omega. *Molecular Systems Biology* 7.

1827 Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. Hooper, A. Strand, Q.
1828 Li, B. Raney, C. Balakrishnan, S. Griffith, G. McVean, and M. Przeworski. 2015.
1829 Stable recombination hotspots in birds. *Science* 350:928-932.

1830 Smagulova, F., K. Brick, Y. M. Pu, R. D. Camerini-Otero, and G. V. Petukhova. 2016.
1831 The evolutionary turnover of recombination hot spots contributes to speciation in
1832 mice. *Genes & Development* 30:266-280.

1833 Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic
1834 analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.

1835 Steviston, L. S., A. E. Woerner, J. M. Kidd, J. L. Kelley, K. R. Veeramah, K. F.
1836 McManus, C. D. Bustamante, M. F. Hammer, J. D. Wall, and P. Great Ape
1837 Genome. 2016. The Time Scale of Recombination Rate Evolution in Great Apes.
1838 *Molecular Biology and Evolution* 33:928-945.

1839 Steviston, L. 2016. great-ape-recombination. GitHub. <https://github.com/lsteviston/great-ape-recombination>

1840 Taylor, J. S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome
1841 duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Research*
1842 13:382-390.

1843 Ubeda, F. and J. F. Wilkins. 2011. The Red Queen theory of recombination hotspots.
1844 *Journal of Evolutionary Biology* 24:541-553.

1845 Vervoort, M., D. Meulemeester, J. Behague, and P. Kerner. 2016. Evolution of Prdm
1846 Genes in Animals: Insights from Comparative Genomics. *Molecular Biology and*
1847 *Evolution* 33:679-696.

1848 Wilson, D. J. and G. McVean. 2006. Estimating diversifying selection and functional
1849 constraint in the presence of recombination. *Genetics* 172:1411-1425.

1850 Wright, S. I. and P. Andolfatto. 2008. The Impact of Natural Selection on the Genome:
1851 Emerging Patterns in *Drosophila* and *Arabidopsis*. *Annual Review of Ecology*
1852 *Evolution and Systematics* 39:193-213.

1853 Wu, H., N. Mathioudakis, B. Diagouraga, A. P. Dong, L. Dombrovski, F. Baudat, S.
1854 Cusack, B. de Massy, and J. Kadlec. 2013. Molecular Basis for the Regulation of
1855 the H3K4 Methyltransferase Activity of PRDM9. *Cell Reports* 5:13-20.

1856 Yamada, S., STischfield, SE Lange, JJasin, M Keeney, S. 2017. Genomic and chromatin
1857 features shaping meiotic double-strand break formation and repair in mice.
1858 *bioRxiv*.

1859 Zerbino, D. R. and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly
1860 using de Bruijn graphs. *Genome Research* 18:821-829.

1861

1862

1863

1864

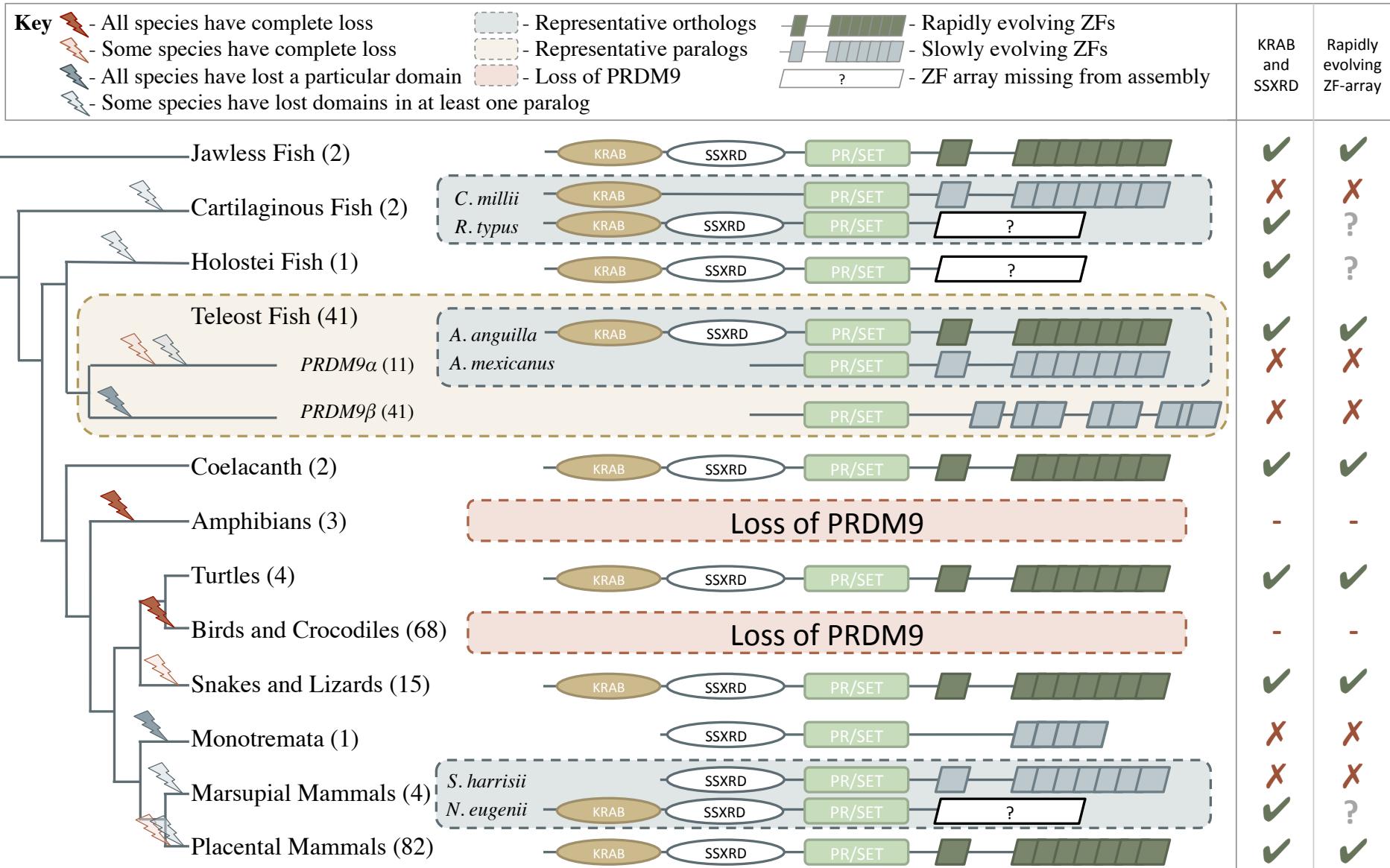
1865

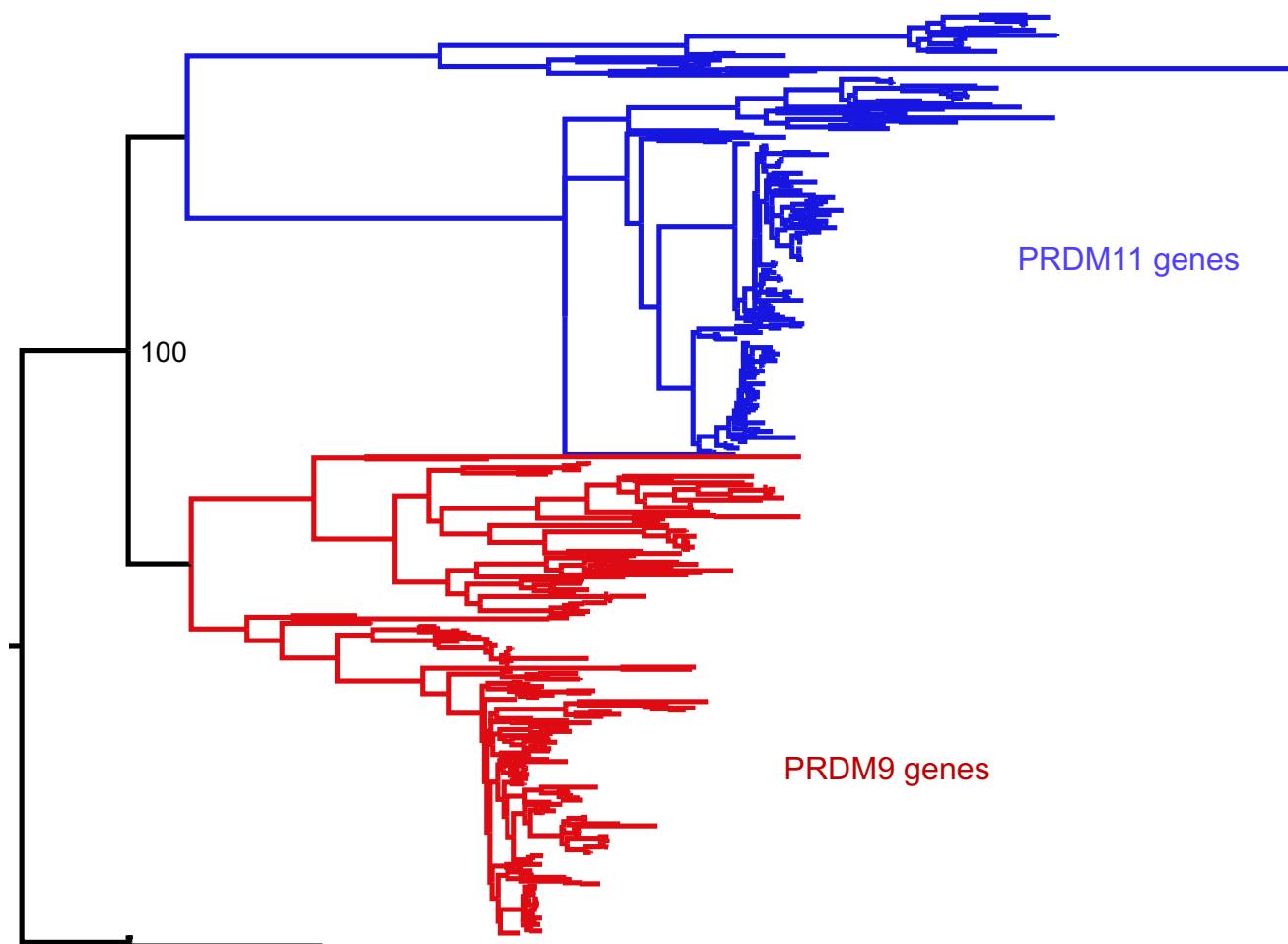
1866

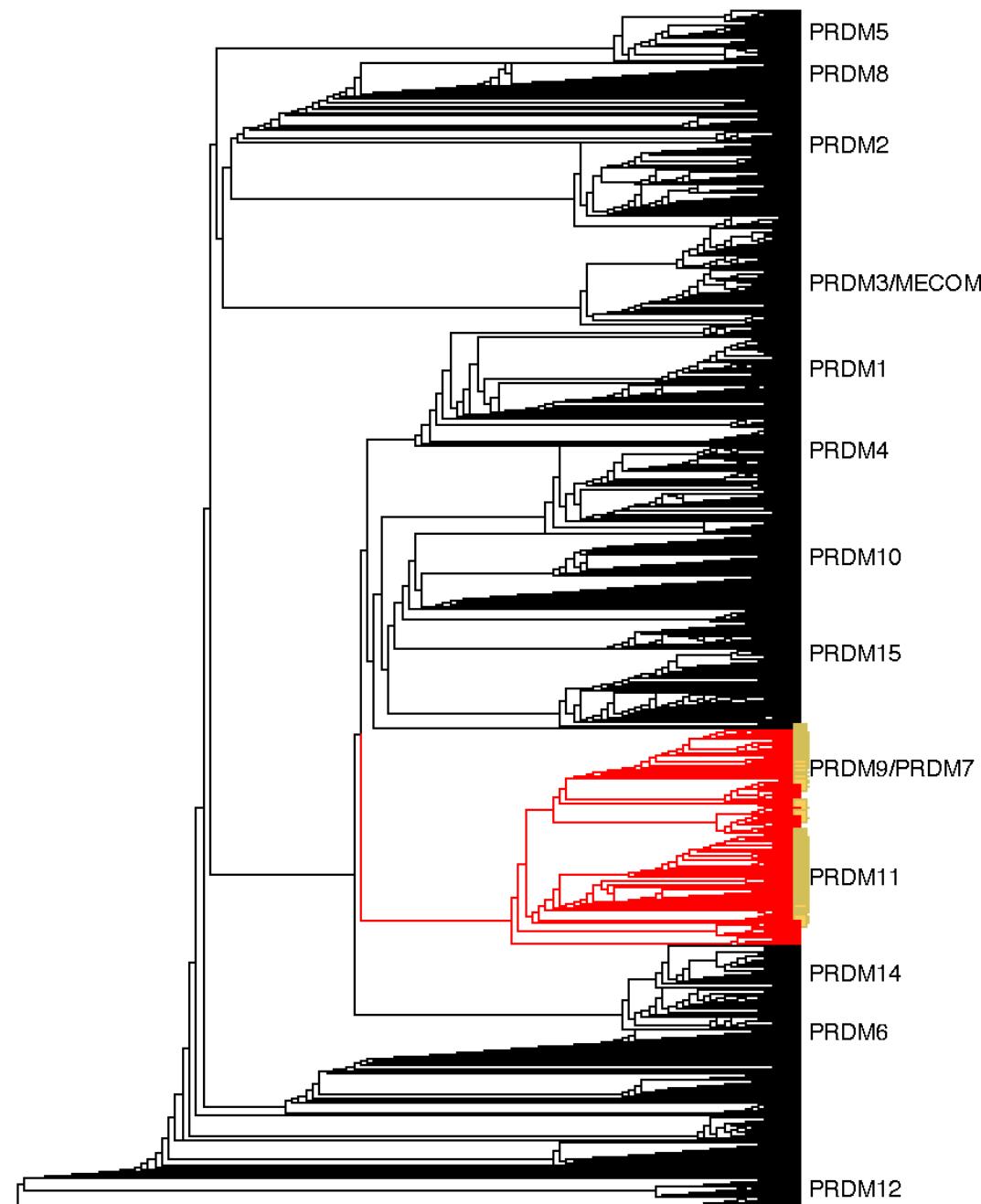
1867

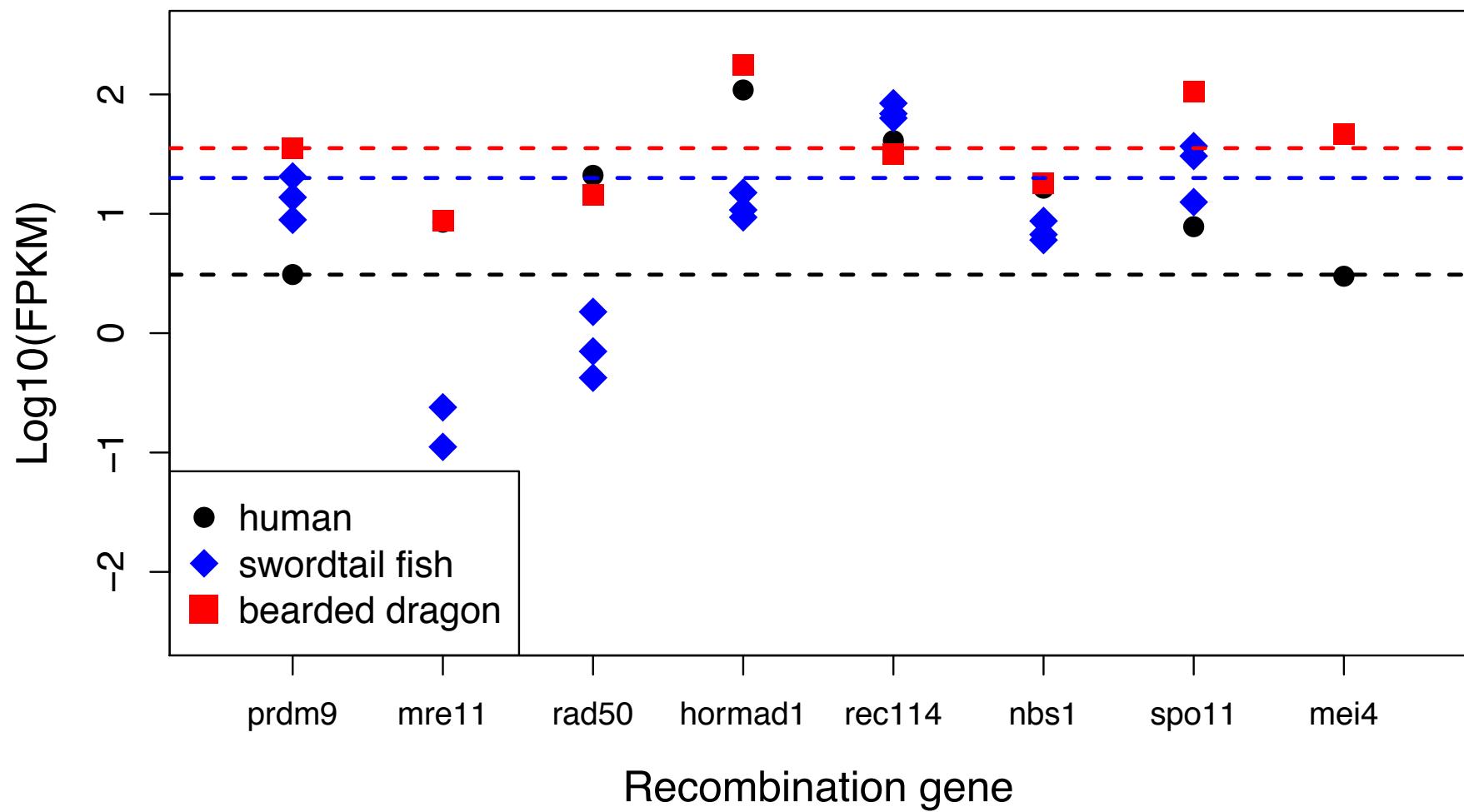
1868

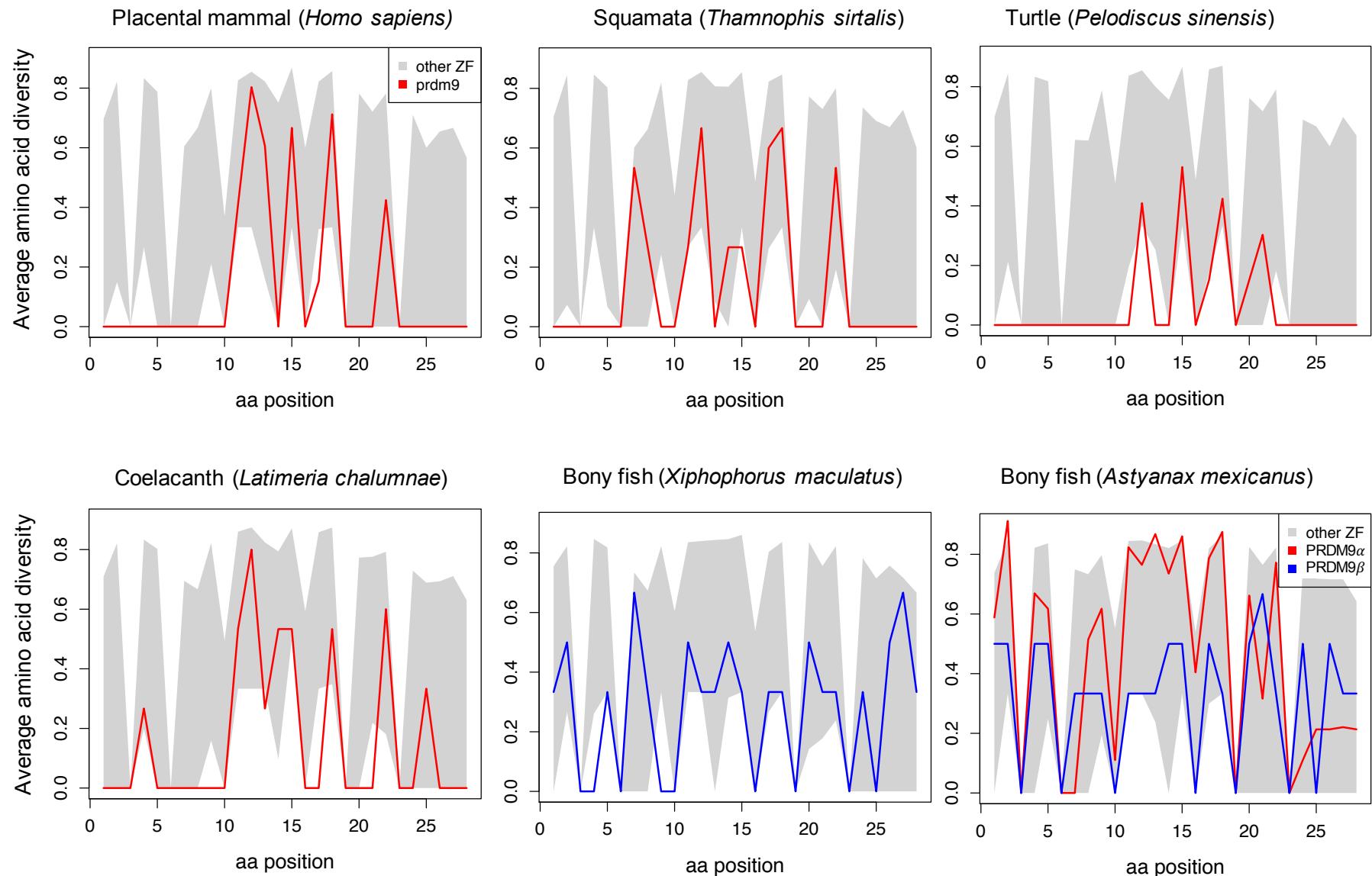
1869

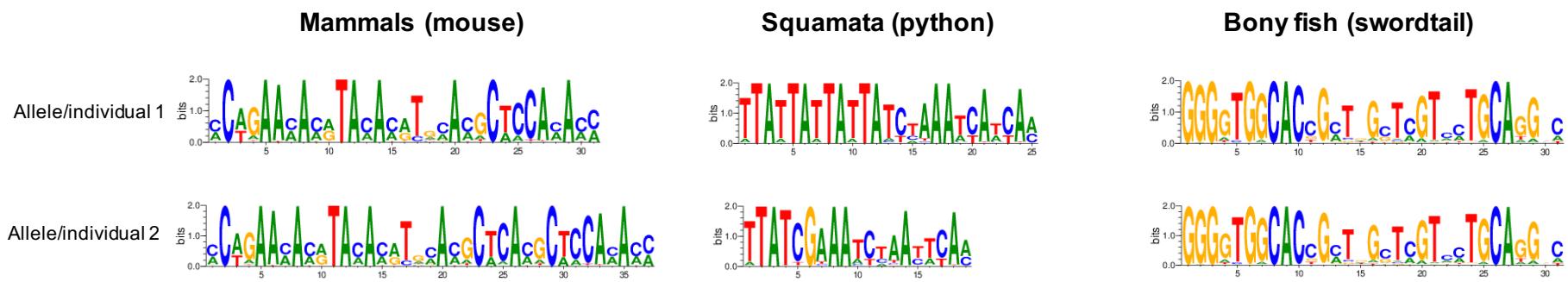


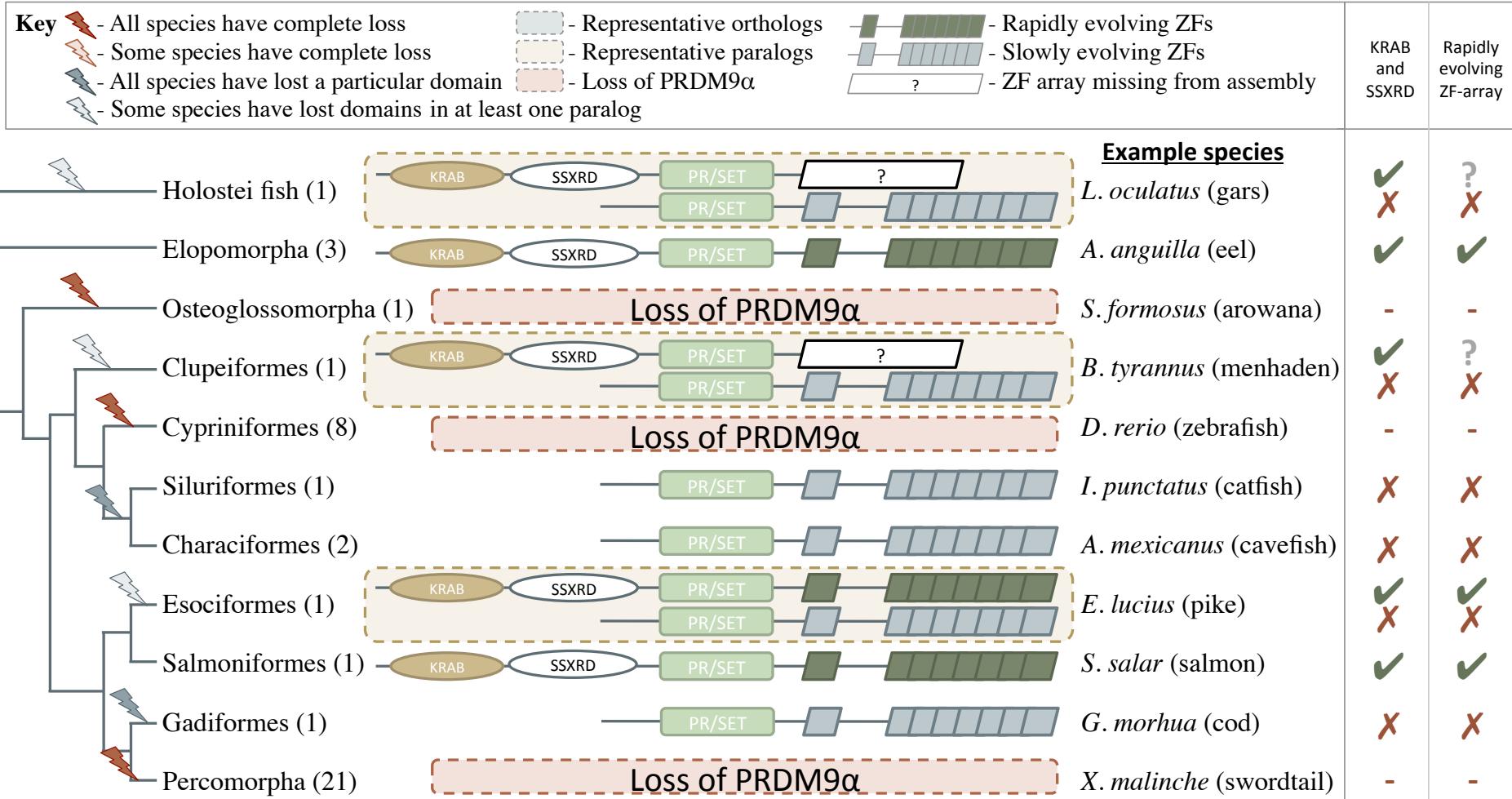


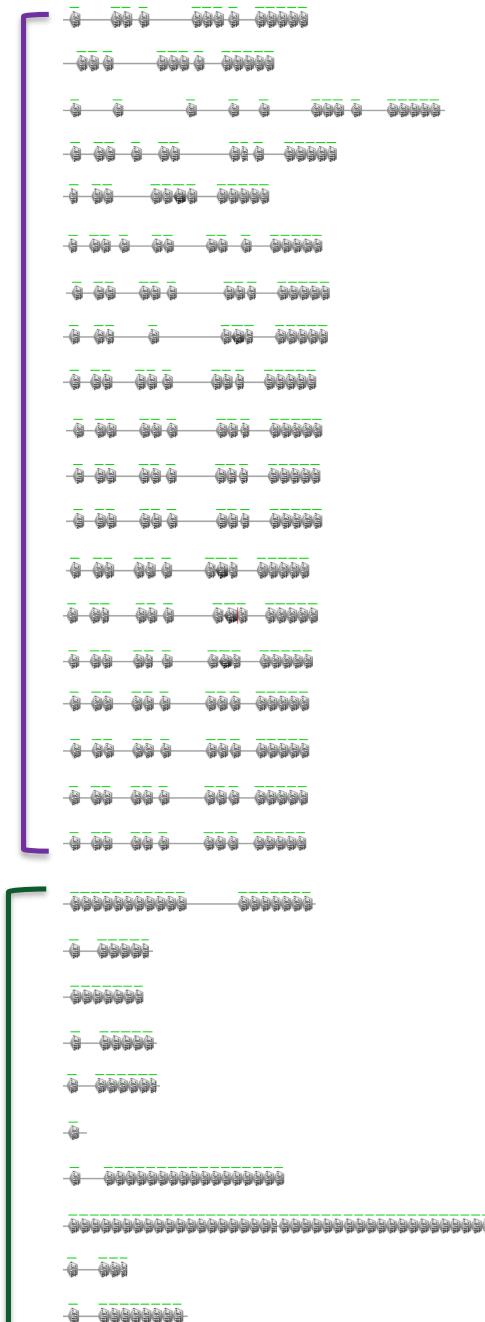
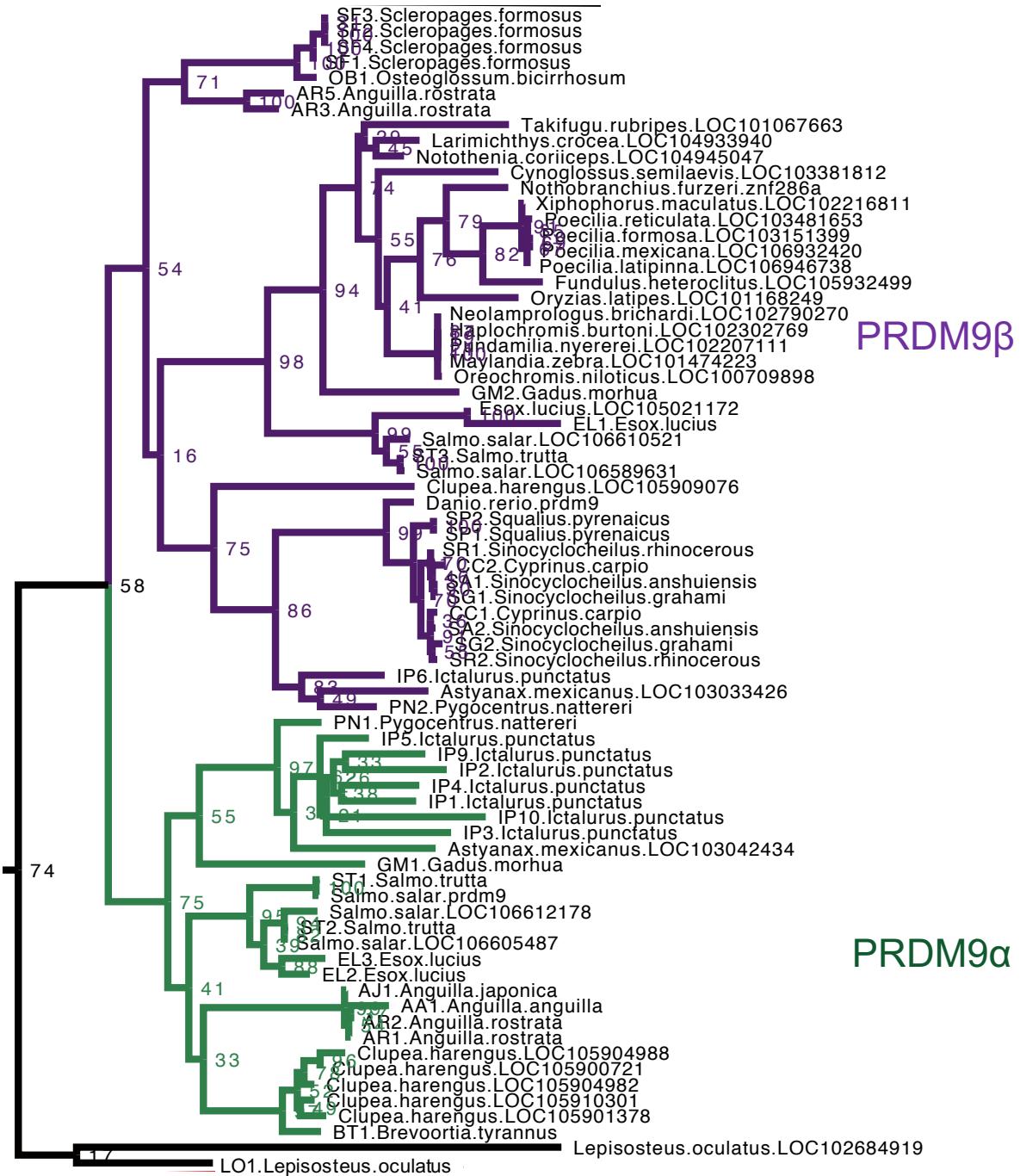








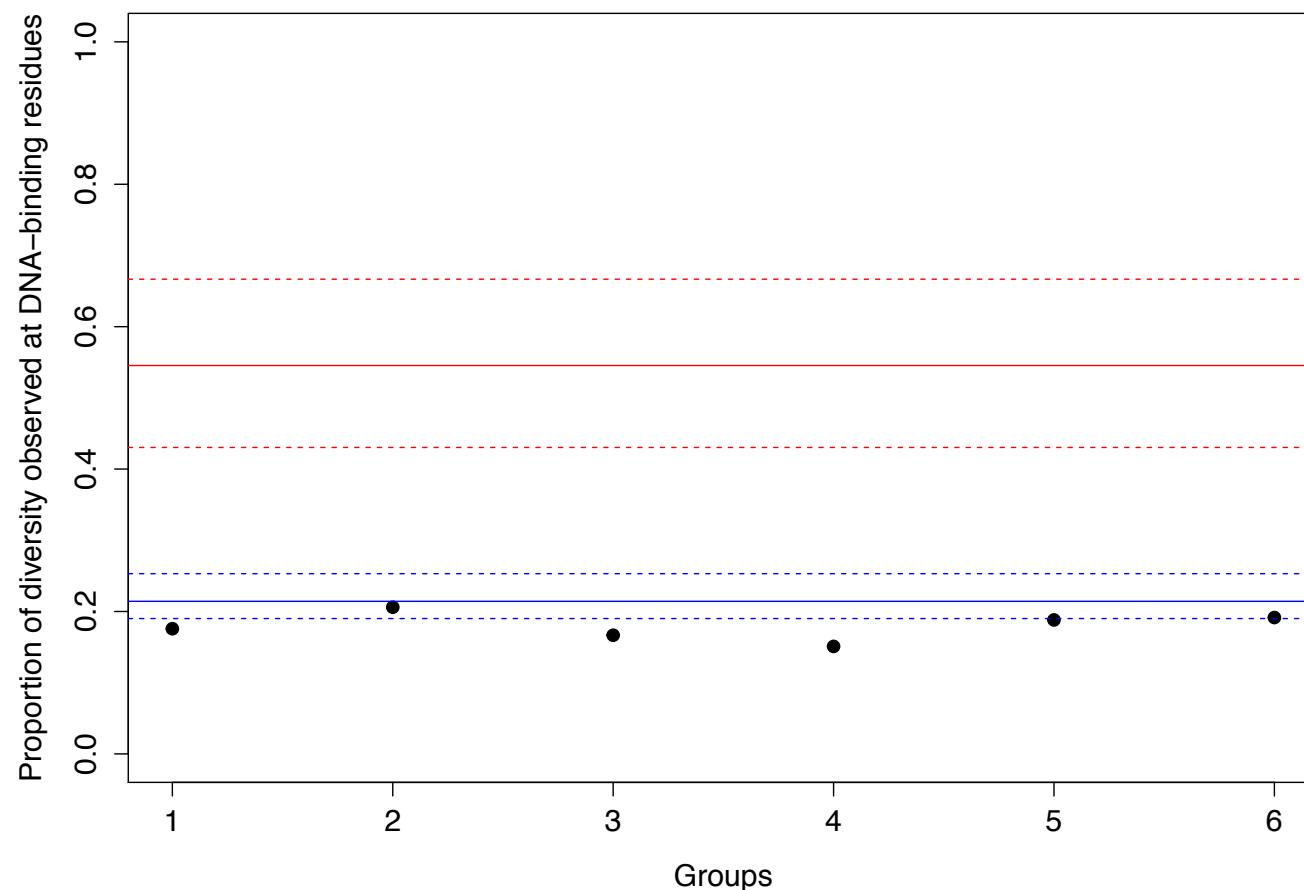


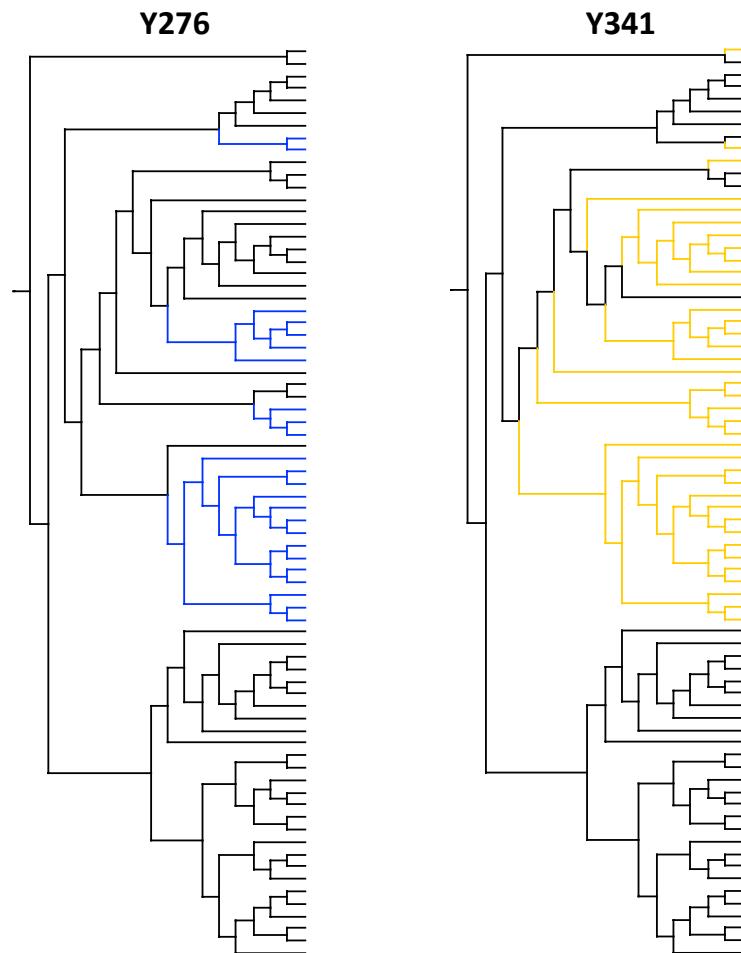
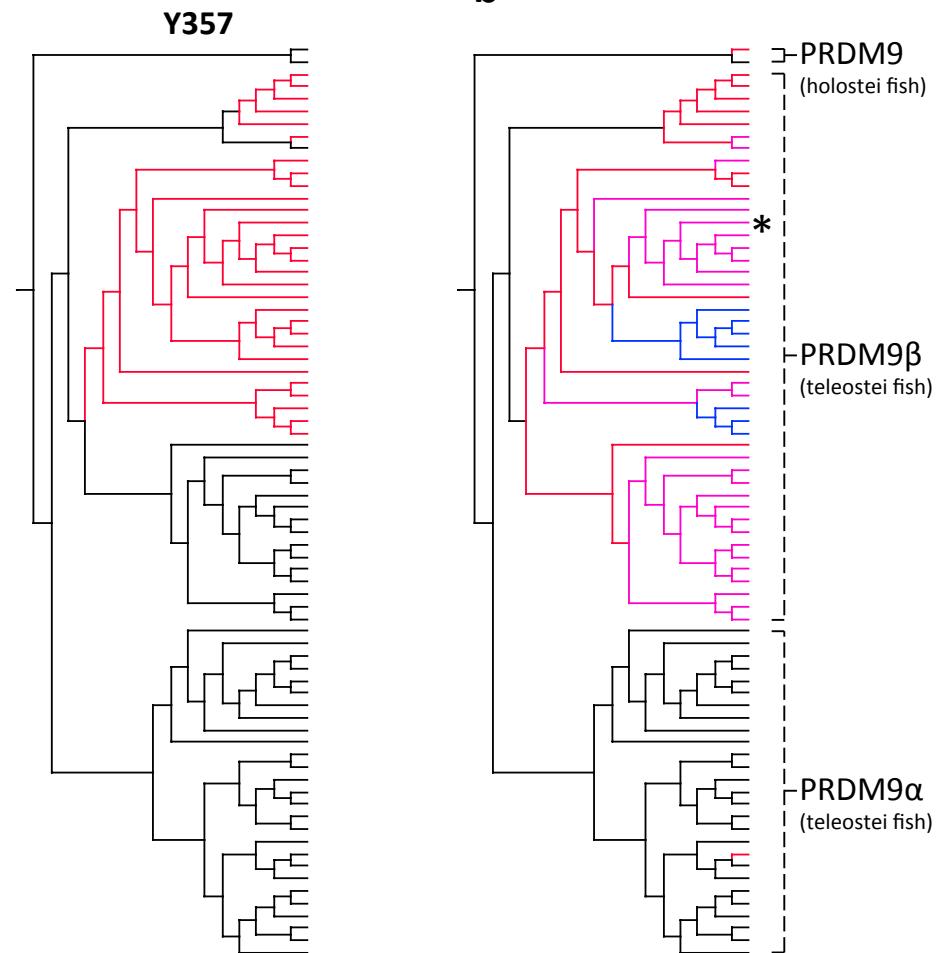


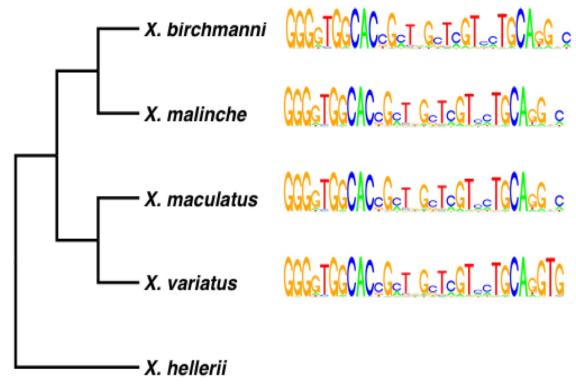
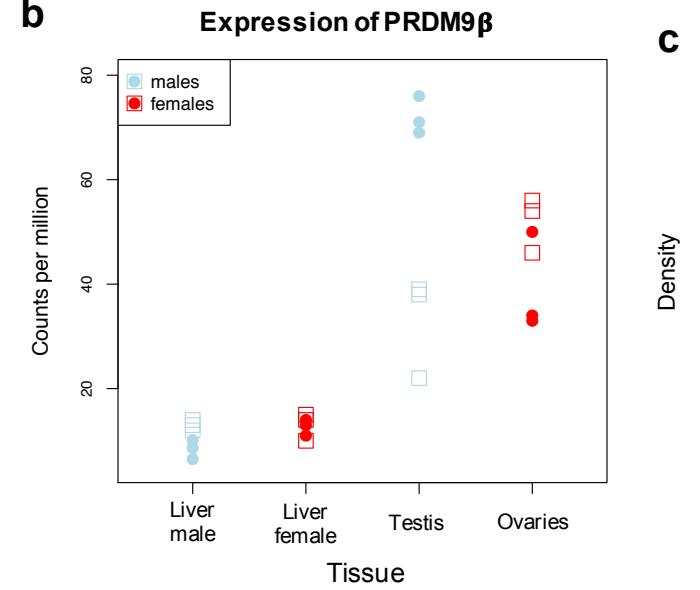
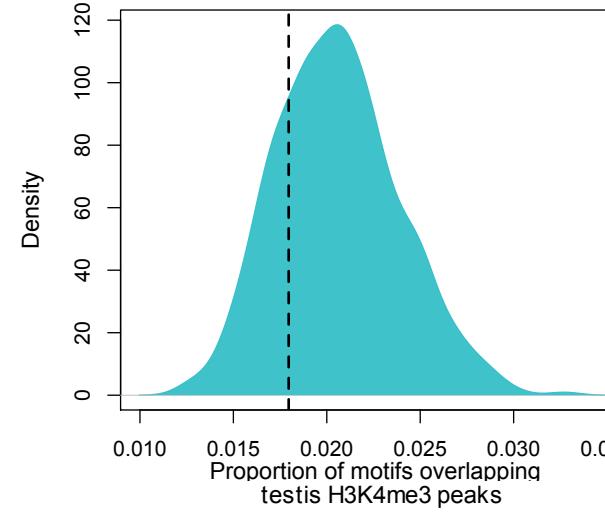
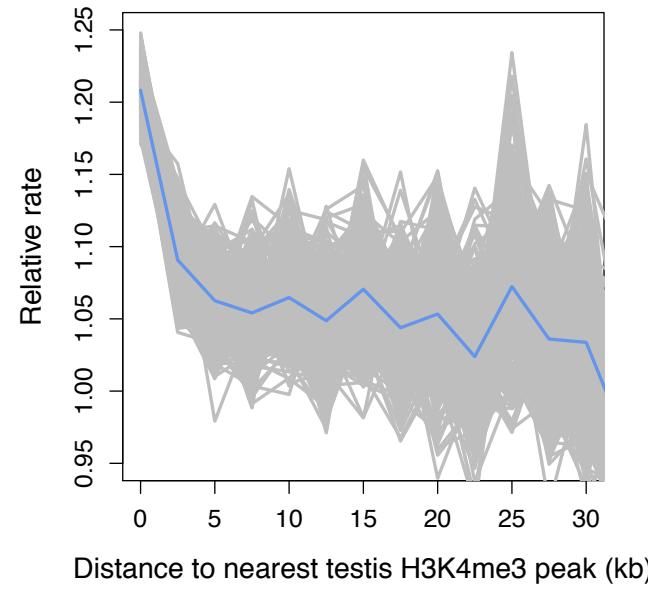
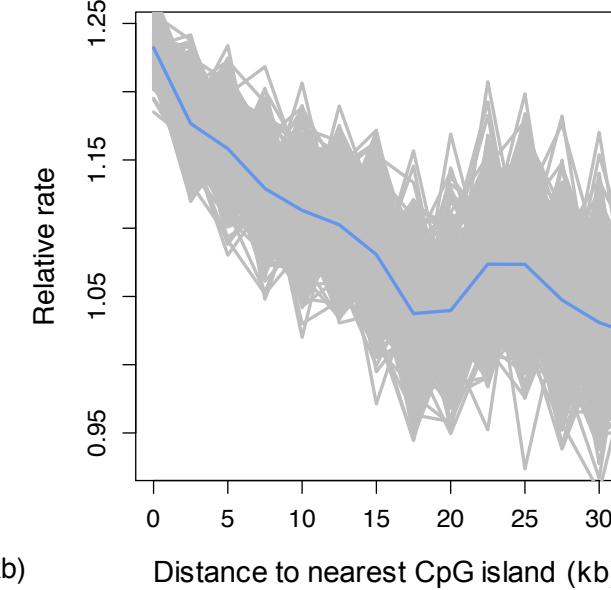
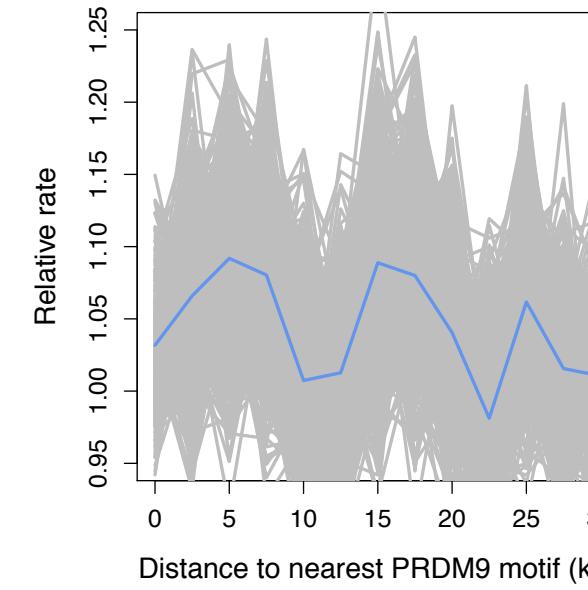


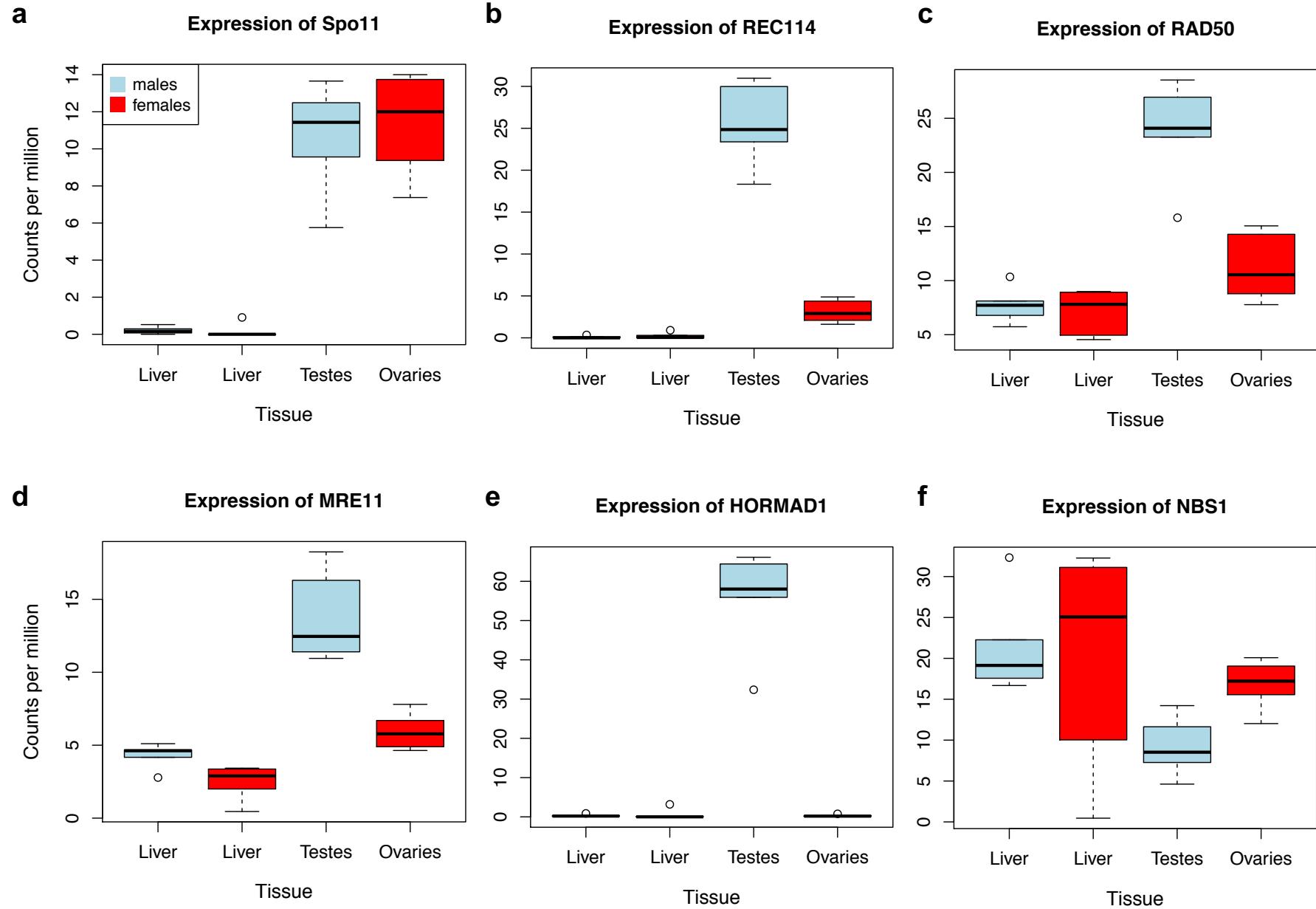
Groups

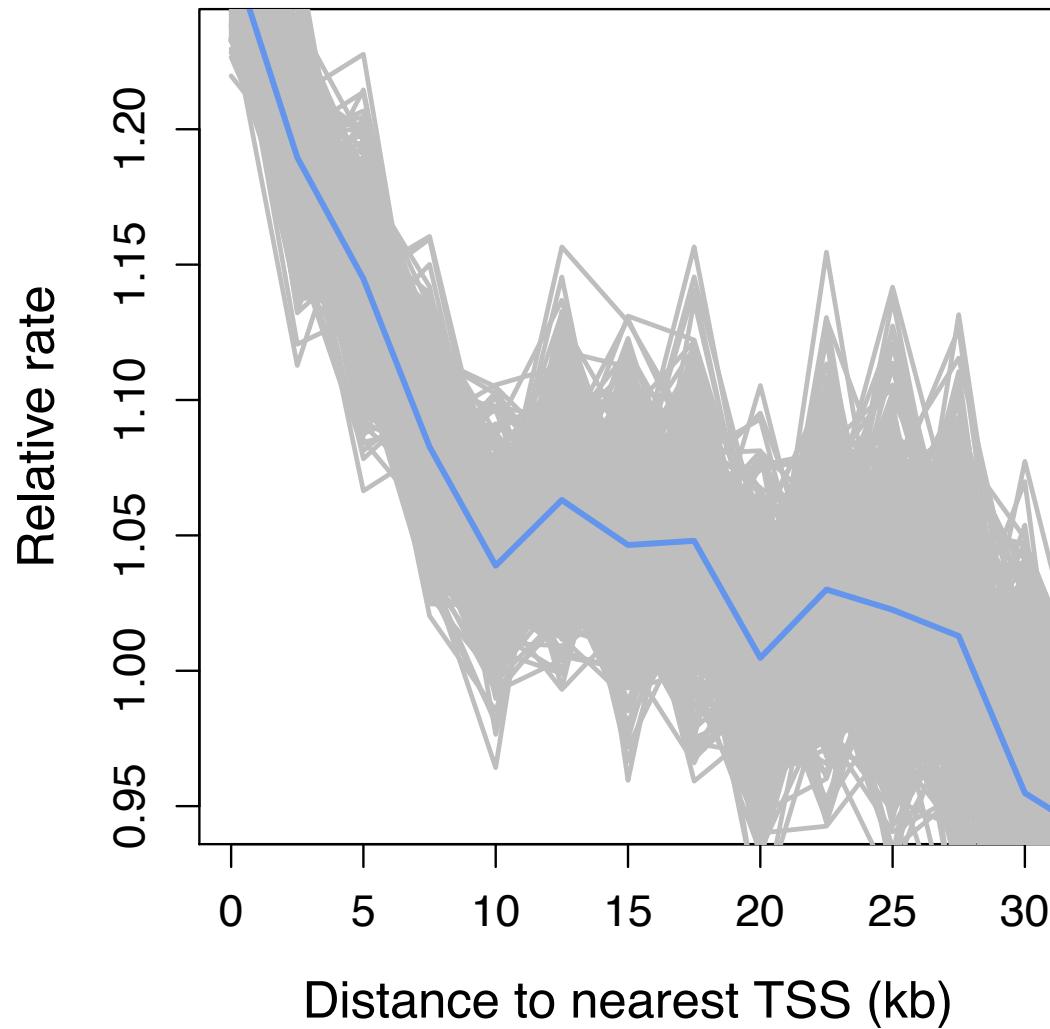
- 1 – All ZFs
- 2 – Only 28aa repeats
- 3 – Cluster 1
- 4 – Cluster 2
- 5 – Cluster 3
- 6 – Cluster 3, Only 28aa repeats

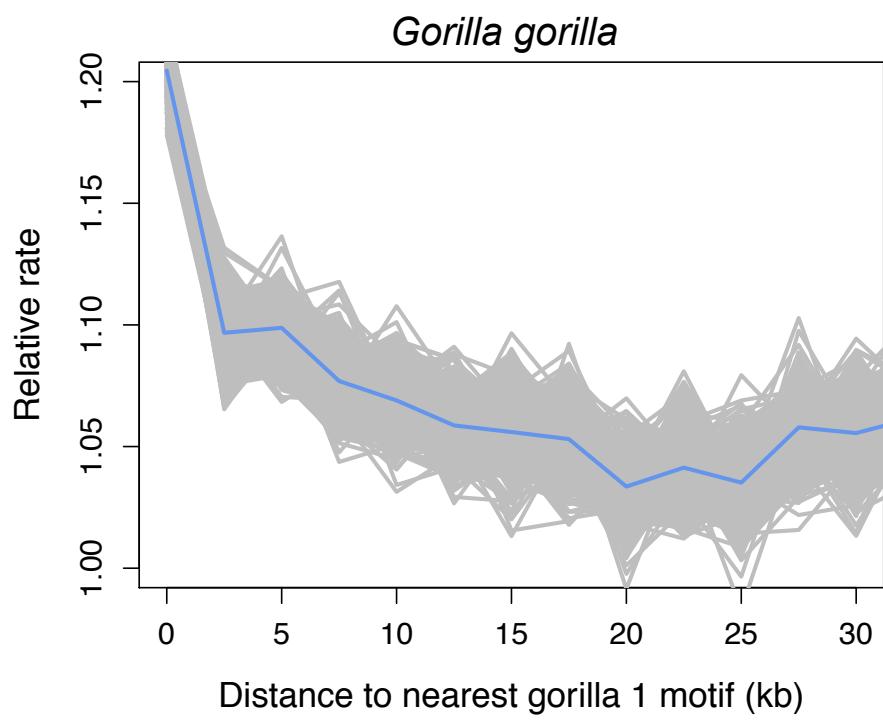
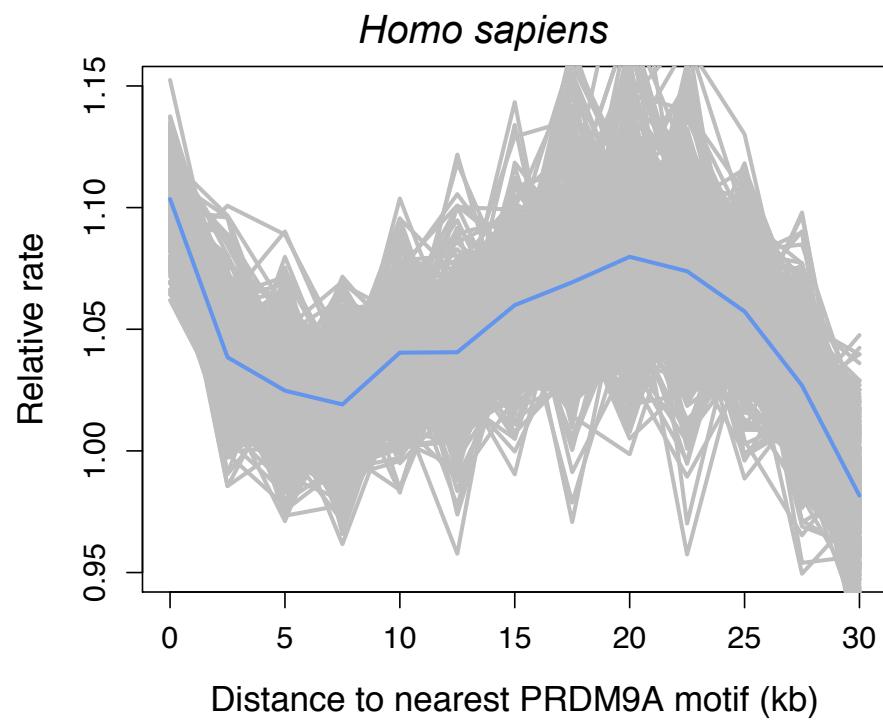


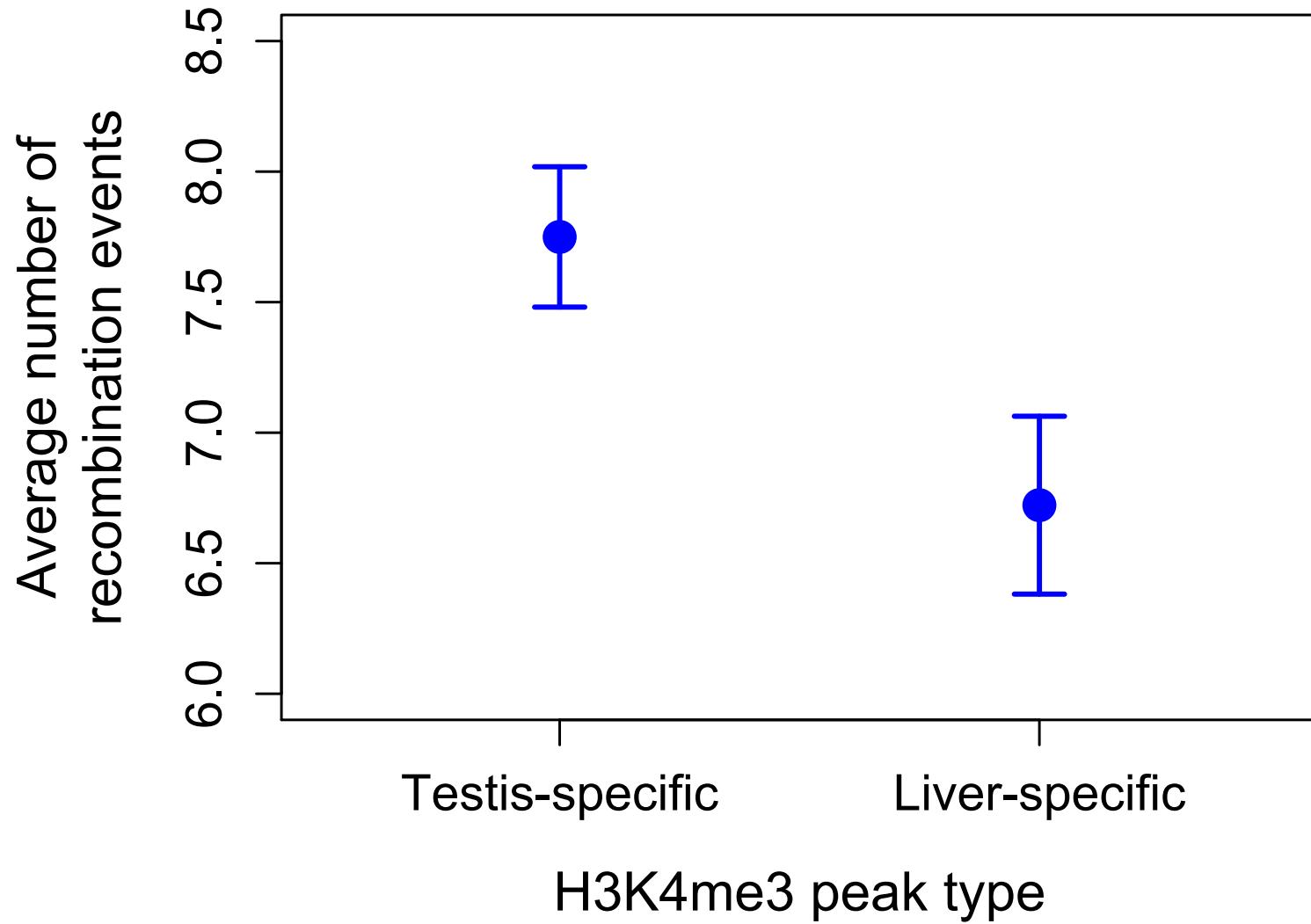
a**b**

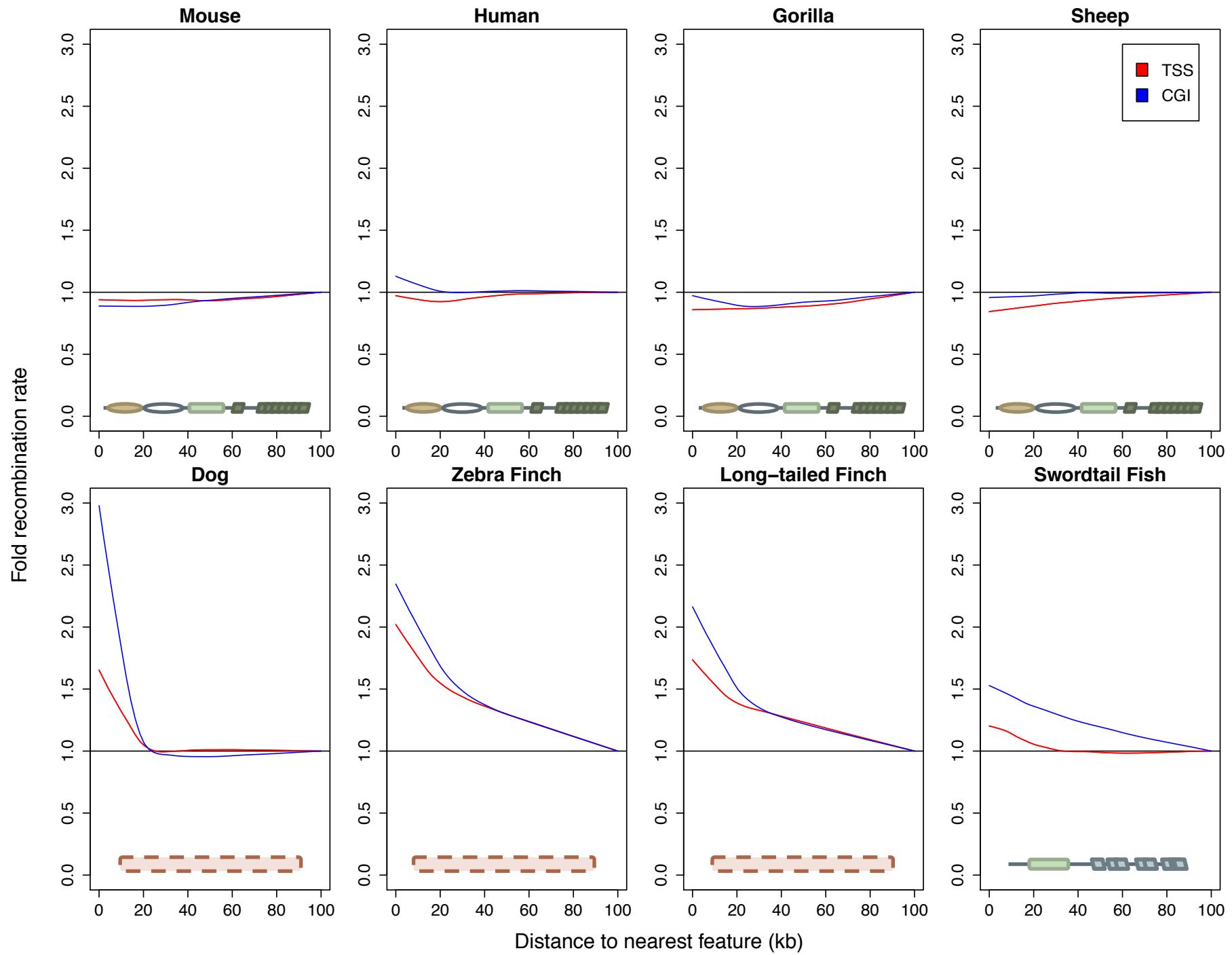
a**b****c****d****e****f**



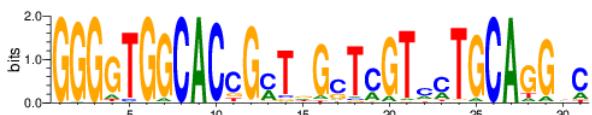




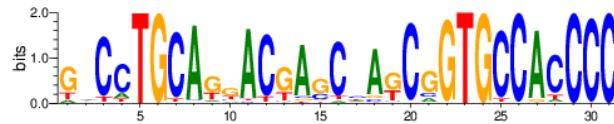




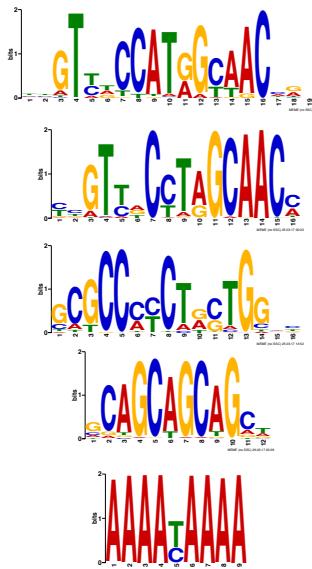
X. birchmanni and *X. malinche* predicted motif



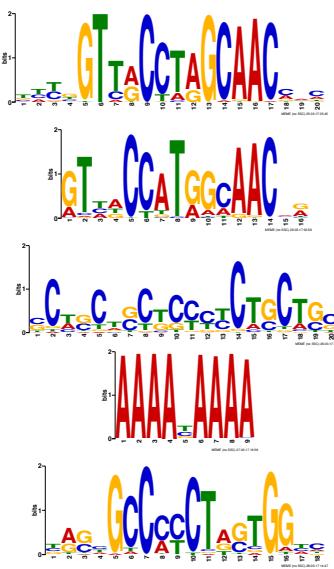
Reverse complement



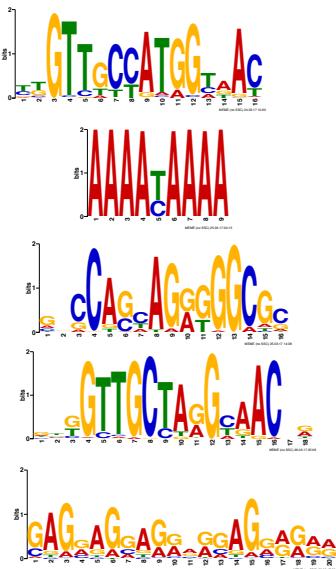
A)



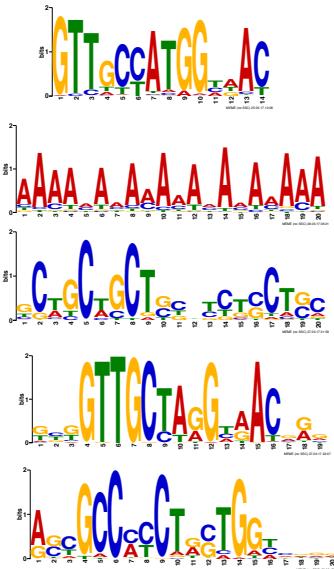
B)



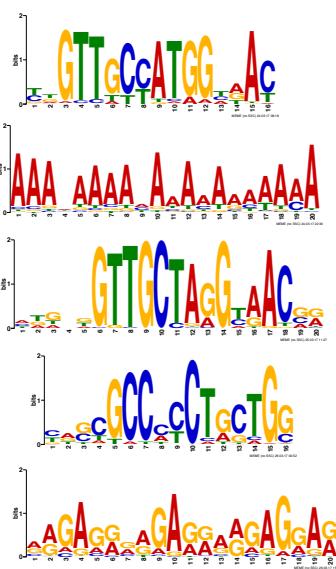
C)

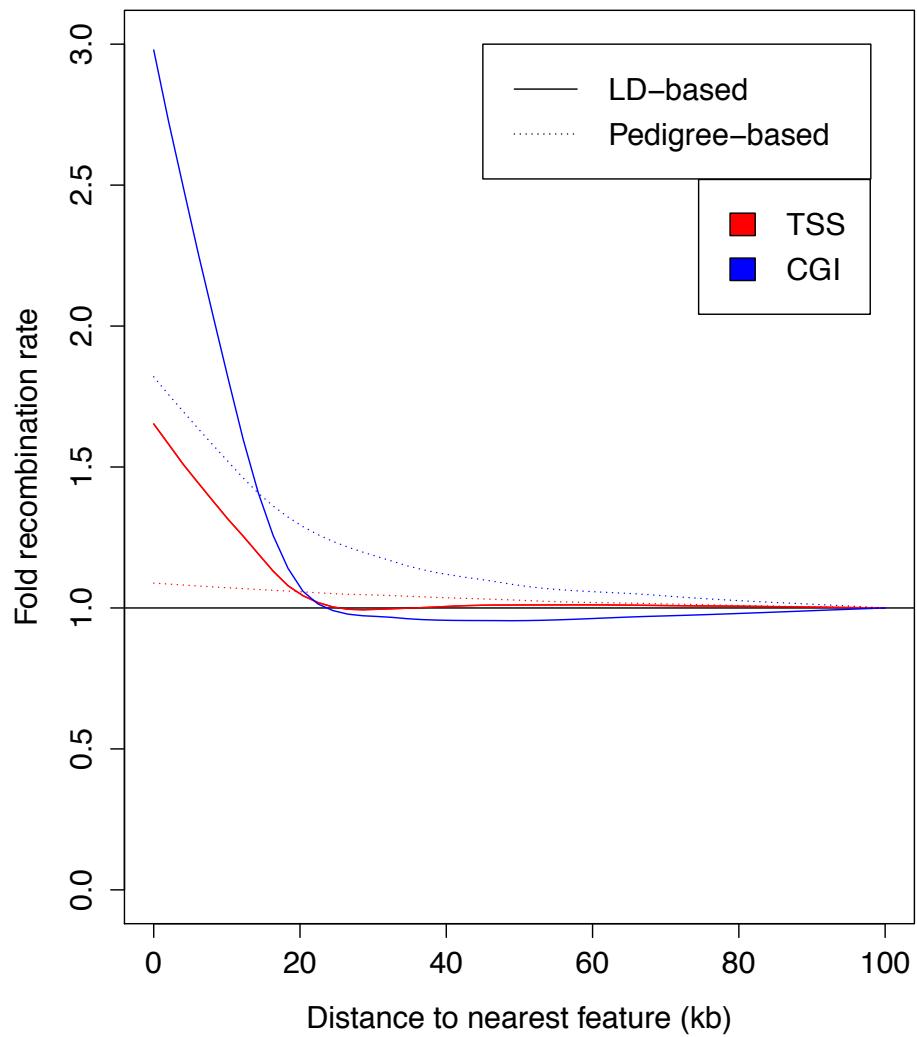


D)



E)



a**Dog****b****Human**