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

Zachary Baker[1,+,*], Molly Schumer[2,3,4+], Yuki Haba[5], Lisa Bashkirova[6], Chris Holland[4,7], Gil G. Rosenthal[4,7], and Molly Przeworski[1, 2,*]

[1] Dept. of Systems Biology, Columbia University
[2] Dept. of Biological Sciences, Columbia University
[3] Harvard Society of Fellows, Harvard University
[4] Centro de Investigaciones Científicas de las Huastecas "Aguazarca"
[5] Dept. of Evolution, Ecology and Environmental Biology, Columbia University
[6] Dept. of Biochemistry and Molecular Biophysics, Columbia University
[7] Dept. of Biology, Texas A&M University
[+] Contributed equally
[*] To whom correspondence should be addressed: ztb2002@columbia.edu and mp3284@columbia.edu

Abstract

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

Introduction

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

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

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

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

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

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

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

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

simply by greater accessibility of the recombination machinery to these genomic regions (Brick et al. 2012; Auton et al. 2013; Choi et al. 2013; Lam and Keeney 2015; Singhal et al. 2015b). Even if there are specific motifs that increase rates of recombination near functional genomic elements in species lacking PRDM9, they are likely to have important, pleiotropic consequences on gene regulation (Nicolas et al. 1989). Thus, there may be a strong countervailing force to the loss of hotspots by under-transmission of hotter alleles, leading to the evolutionary stability of hotspots.

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

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

Results

Initial identification of PRDM9 orthologs in vertebrates

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

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

We further inferred an ancient duplication of PRDM9 in the common ancestor of teleost fish, apparently coincident with the whole genome duplication that occurred in this group (**Figure 1, Figure 2**). We used both phylogenetic methods and analysis of the ZF structure to distinguish these copies (see **Figure 2- Figure Supplement 1, Methods**) and refer to them as PRDM9α and PRDM9β in what follows. While PRDM9β orthologs were identified in each species of teleost fish examined, we were unable to identify

PRDM9α type orthologs within three major teleost taxa, suggesting at minimum three losses of PRDM9α type orthologs within teleost fish (**Figure 2, Supplementary File 1A**). Several additional duplication events appear to have occurred more recently in other vertebrate groups, including in jawless fish, cartilaginous fish, bony fish, and mammals (**Supplementary File 1A**).

Expression of PRDM9 in the germline of major vertebrate groups

Since a necessary condition for PRDM9 to play a role in meiotic recombination is for it to be expressed in the germline, we looked for PRDM9 in expression data from testis tissues in order to confirm its presence. We focused on testis expression rather than ovaries because although both obviously contain germline cells, preliminary analyses suggested that meiotic gene expression is more reliably detected in the testes (see **Methods**). We selected 23 representative species, spanning each major vertebrate group, with publically available testis expression or testis RNA-seq (**Supplementary File 2A**); we also generated testis RNA-seq data for two species of bony fish (see **Methods**). In teleost fish with both PRDM9α and PRDM9β genes, we were able to detect either the expression of both orthologs or only expression of PRDM9α orthologs. In species of teleost fish with only PRDM9β genes, we consistently identified expression of PRDM9β genes. More generally, we were able to identify PRDM9 expression in nearly all RNA-seq datasets from species in which the genome carried a putative ortholog, the elephant shark (*Callorhinchus milii*) being the sole exception (**Supplementary File 2B; Supplementary File 2C**).

Confirmation of PRDM9 loss events

Concerned that absences of PRDM9 observed in some species could reflect lower quality genome assemblies rather than true loss events, we also used testis RNAseq data to investigate putative losses of PRDM9 in amphibians and fish (PRDM9α). To this end, we relied on the fact that when PRDM9 is present, it is detectable in RNAseq data from the whole testis of vertebrates (see above). Our approach was to analyze testis transcriptome data from species lacking PRDM9 sequences in their genome assemblies, using an analysis that is not biased by the genome assembly (see **Methods**). For each

species, we confirmed that the dataset captured the appropriate cell populations and provided sufficient power to detect transcripts that are expressed during meiosis at levels comparable to PRDM9 in mammals (**Figure 1- Figure Supplement 3, Supplementary File 2B; Supplementary File 2D**). With this approach, we were able to find support for the loss of PRDM9 in salamanders (*Cynops pyrrhogaster, Ambystoma mexicanum*) and frogs (*Xenopus tropicalis*). Because of the paucity of amphibian genomes, however, it is not clear whether or not these examples represent a widespread loss of PRDM9 within amphibians or more recent, independent losses. Within bony fish, we were able to confirm the three independent losses of PRDM9α type orthologs in one species each of percomorph (*Xiphophorus birchmanni*), cypriniform (*Danio rerio*) and osteoglossomorph fish (*Osteoglossum bicirrhosum*). Thus, in all cases with sufficient power to detect expression of PRDM9 in testes data, our findings were consistent with inferences based on genome sequence data.

Inferences of PRDM9 domain architecture

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

Despite its widespread taxonomic distribution, however, the complete domain structure was not found in several of the 149 sampled lineages with PRDM9 orthologs (**Figure 1**; in addition to the complete losses of the gene described above). Instances include the absence of the SSXRD domain in some cartilaginous fish (see **Methods)**; absence of both KRAB and SSXRD domains in PRDM9β orthologs (**Figure 1**) and in PRDM9α orthologs found distributed throughout the teleost fish phylogeny (**Figure 2, Figure 2- Figure Supplement 1**); and the absence of the KRAB domain in monotremata (*Ornithorhynchus anatinus*) and marsupial mammals (*Sarcophilus harrisii*, **Figure 1**; **Supplementary File 1A**).

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

transcriptomes for evidence of these missing domains (see **Methods**). We required not only that missing domains homologous to PRDM9 be absent from the genome in a whole genome search, but also that the missing domain not be present in the transcriptome, when other domains of PRDM9 were. This approach necessarily limits our ability to verify putative losses when there are no suitable transcriptome data, but nonetheless allowed us to confirm the losses of the KRAB and SSXRD domains in a PRDM9 ortholog from holostean fish (*Lepisosteus oculatus*), in all PRDM9β orthologs from teleost fish (**Figure 1**), in PRDM9α orthologs that lost their complete domain structure in several taxa of teleost fish (*Gadus morhua*, *Astyanax mexicanus*, *Ictalurus punctatus*, *Esox lucius*; **Supplementary File 2C**), as well as losses of the KRAB domain in two PRDM9 orthologs identified in monotremata (both in *O. anatinus*, **Supplementary File 2C**), indicating a minimum of six N-terminal domain losses within vertebrates.

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

Evidence for rapid evolution of PRDM9 binding specificity

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

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

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

In addition to placental mammals, PRDM9 orthologs in jawless fish, some bony fish (Salmoniformes, Esociformes, Elopomorpha), turtles, snakes, lizards, and coelacanths show similarly elevated values of this statistic (**Figure 1- Figure Supplement 4**). In fact, PRDM9 is the most rapidly evolving ZF gene genome-wide in most species in these taxa and all PRDM9 orthologs with the complete domain structure were in the top 5% of the most rapidly evolving ZFs in their respective genomes (**Table 1, Supplementary File 3B**). In contrast, evidence of such rapid evolution is absent from other taxa of bony fish, including all PRDM9β orthologs and partial PRDM9α orthologs, as well as from the putatively partial PRDM9 orthologs found in the elephant shark, the Tasmanian devil, and in several species of placental mammals (see **Methods** for details). We only observed one instance (little brown bat, *Myotis lucifugus*) in which a partial PRDM9 ortholog was evolving unusually rapidly (**Table 1**); in this case, we were unable to confirm the loss of the missing KRAB domain (see **Methods**), so it remains possible this ortholog is in fact intact. In summary, with one possible exception, species show evidence of rapid evolution of the ZF binding affinity if and only if they carry the intact PRDM9 ortholog found in placental mammals. This concordance of rapid evolution with the complete domain structure is highly unlikely by chance (taking into account the

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

Analysis of SET domain catalytic residues

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

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

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

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

Within bony fish, we identified 47 PRDM9 orthologs with substitutions at one or more of these residues, including the partial PRDM9 ortholog from holosteans (see above) and all PRDM9β orthologs in teleosts (**Supplementary File 1A**). The distribution of substitutions at these residues within PRDM9β genes suggests that numerous independent substitution events have occurred in this gene family following the loss of KRAB and SSXRD domains (**Figure 3**). In contrast, no substitutions were observed at these residues in any PRDM9α orthologs, regardless of their domain architecture. These observations could be consistent with a lack of constraint on the ancestral methyltransferase activities of PRDM9 in PRDM9β genes after the PRDM9α/PRDM9β duplication event (or conceivably an indication that there has been convergent evolution towards a new functional role). Thus, PRDM9β genes not only lack KRAB and SSXRD domains, they likely lack some methyltransferase activity of the SET domain.

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

To more directly test the hypothesis that the partial ortholog of PRDM9 does not direct recombination, we examined patterns of crossing-over in naturally-occurring swordtail fish hybrids (*X. birchmanni* x *X. malinche*; see Methods). Like other percomorphs, swordtail fish have a PRDM9β type gene that lacks the KRAB and SSXRD domains and a slowly evolving ZF array with testis-specific expression (**Figure 4; Figure 4 – Figure 4 Supplement 1**); they further carry substitutions at two catalytic residues of the SET domain (Y341F and Y357P), as well as at residues of the SET

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

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

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

Recombination landscapes in vertebrates with and without PRDM9

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

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

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

Discussion

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

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

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

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

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

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

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

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

Methods

Identification of putative PRDM9 orthologs from the RefSeq database

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

Predicting PRDM9 orthologs from whole genome sequences

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

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

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

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

1

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

3 To understand the evolution of PRDM9 within vertebrates, we used a
4 phylogenetic approach. We first built an alignment of the amino acid sequences of
5 putative PRDM9 and PRDM11 SET domains using Clustal Omega (Sievers et al. 2011).
6 We included genes clustering with PRDM11 because it had been reported that PRDM11
7 arose from a duplication event of PRDM9 in the common ancestor of bony fish and
8 tetrapods (Vervoort et al. 2016), and we were interested in identifying any PRDM9
9 orthologs carried by vertebrate species that may precede this duplication event. The
10 alignment coordinates were then used to generate a nucleotide alignment, which was used
11 as input into the program RAxML (v7.2.8; Stamatakis 2006). We performed 100 rapid
12 bootstraps followed by maximum likelihood estimation of the tree under the General
13 Reversible Time substitution model, with one partition for each position within a codon.
14 The resulting phylogeny contained monophyletic groups corresponding to the PRDM9
15 and PRDM11 duplication event, with 100% bootstrap support (**Figure 1- Figure**
16 **Supplement 1**). These groups were used to label each putative ortholog as PRDM9 or
17 PRDM11. Only jawless fish have PRDM9 orthologs basal to this duplication event,
18 suggesting PRDM11 arose from PRDM9 before the common ancestor of cartilaginous
19 fish and bony vertebrates. We observed at least one PRDM11 ortholog in each of the
20 other vertebrate species examined.

21 Within teleost fish, we identified two groups of PRDM9 orthologs, which we
22 refer to as PRDM9α and PRDM9β (**Figure 2- Figure Supplement 1**). While the
23 bootstrap support for the monophyly of the two groups is only 75% for PRDM9α and
24 54% for PRDM9β, the potential duplication event suggested by this tree is coincident
25 with the whole genome duplication event known to have occurred in the common
26 ancestor of teleost fish (Taylor et al. 2003). Moreover, the phylogenetic grouping based
27 on the SET domain is concordant with general differences in the domain architectures
28 between the two orthologs: In contrast to PRDM9α, PRDM9β genes have derived ZF
29 array structures, containing multiple tandem ZF arrays spread out within the same exon
30 (**Figure 2- Figure Supplement 1**) and are always found without the KRAB and SSXRD
31 domains, whereas PRDM9α genes generally have a single tandem array of ZFs consistent

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

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

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

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

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

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

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

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

Three *Xiphophorus birchmanni* and three *X. malinche* were collected from the eastern Sierra Madre Oriental in the state of Hidalgo, Mexico. Fish were caught using baited minnow traps and were immediately euthanized by decapitation (Texas A&M AUP# - IACUC 2013-0168). Testis, ovaries, and liver were dissected and stored at 4°C in RNAlater. Total RNA was extracted from testis, ovary and liver tissue using the Qiagen RNeasy kit (Valencia, CA, USA) following the manufacturer's protocol. RNA was quantified and assessed for quality on a Nanodrop 1000 (Nanodrop technologies,

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

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

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

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

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

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

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

Confirmation of PRDM9 domain loss and investigation of loss of function

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

To ask whether the missing PRDM9 domains were truly absent from the genome assembly, we first used a targeted genome-wide search. To this end, we performed a *tblastn* search of the genome against the human PRDM9 ortholog with an e-value of 1e-10. For all *blast* hits, we extracted the region and 2 Mb flanking in either direction, translated them in all six frames (http://cgpdb.ucdavis.edu/DNA_SixFrames_Translation/), and performed an *rpsblast* search of these regions against the CDD (database downloaded from NCBI September 2016) with an e-value of 100 to identify any conserved domains, even with weakly supported homology. We extracted all *rpsblast* hits to the missing functional domain (SET CDD id: smart00317, pfam00856, cl02566; SSXRD CDD id: pfam09514; KRAB domains pfam01352, cd07765, smart00349) and used them as query sequences in a

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

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

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

to confirm that these transcripts were homologous to the PRDM9 ortholog identified using phylogenetic methods in these taxa. Results of this analysis can be found in **Supplementary File 2C**. In summary, there were two cases where the transcriptomes supported additional domain structures not found in the whole genome sequence (**Supplementary File 2C**): a PRDM9 ortholog from the spotted gar (*Lepisosteus oculatus*) that was observed to have a KRAB domain not identified in the genome sequence, and a PRDM9α ortholog from the Atlantic salmon (*Salmo salar*) that was observed to have both KRAB and SSXRD domains not identified in the genome search. In all other cases, we confirmed the losses of either the KRAB or SSXRD domains, including: (i) PRDM9β orthologs missing KRAB and SSXRD domains in all species of teleost fish expressing these orthologs (**Supplementary File 2B, Supplementary File 2C**) (ii) PRDM9α orthologs missing KRAB and SSXRD domains identified from *Astyanax mexicanus, Esox lucius, Gadus morhua,* and *Ictalurus punctatus*, and (iii) loss of the KRAB domain from one PRDM9 ortholog in monotremata (*O.anatinus*) and both KRAB and SSXRD domains from the other ortholog in this species.

For all groups in which we confirm that there is only a partial PRDM9 ortholog based on the above analyses, we asked whether the PRDM9 gene in question has likely become a pseudogene (as it has, for example, in canids; Oliver et al. 2009; Munoz-Fuentes et al. 2011), in which case the species can be considered a PRDM9 knockout. Though such events would be consistent with our observation of many losses of PRDM9, they would not be informative about the role of particular PRDM9 domains in recombination function. For this analysis, we aligned the SET domain of the PRDM9 coding nucleotide sequence to a high-quality PRDM9 sequence with complete domain structure from the same taxon using Clustal Omega (see **Supplementary File 3A**), except for the case of PRDM9β in bony fish and the PRDM9 ortholog from cartilaginous fish, where such a sequence was not available. In the case of PRDM9β, we compared the sequence between *X. maculatus* and *A. mexicanus*, sequences that are >200 million years diverged (Hedges et al. 2015). In the case of cartilaginous fish, we used the sequence from *R. typus* and *C. milii*, which are an estimated 400 million years diverged (Hedges et al. 2015).

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

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

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

Evolutionary patterns in the SSXRD domain

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

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

PCR and Sanger sequencing of python PRDM9

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

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

Analysis of PRDM9 ZF array evolution

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

To this end, for each species with a PRDM9 ortholog, we downloaded the nucleotide and protein sequences for all available RefSeq genes with a C2H2 ZF motif annotated in Conserved Domain Database (pfam id# PF00096). To simplify alignment generation, we only used tandem ZF arrays with four or more ZFs matching the 28 amino acid long C2H2 motif (X2-CXXC-X12-HXXXH-X5 where X is any amino acid). In all of our analyses, if a gene had multiple tandem ZF arrays that were spatially separated, only the first array of four or more adjacent ZFs was used for the following analysis (**Supplementary File 3B**). However, an alternative analysis using all ZFs or different subsets of ZFs led to qualitatively similar results for the PRDM9β orthologs from bony fish, where ZFs are commonly found in multiple tandem arrays separated by short linker regions in the predicted amino acid sequence (**Figure 2 Supplement 1; Figure 2-Figure Supplement 2**). For species with PRDM9 orthologs with fewer than five ZFs, we implemented *blastn* against the whole genome sequence using the available gene model as a query sequence, in order to determine whether or not there was a predicted gap

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

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

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

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

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 between complete domain structure and rapid evolution of the PRDM9 ZF array, taking into account phylogenetic relationships between PRDM9 orthologs. We coded these variables using a binary approach with '00' for incomplete domain structure and no evidence of rapid evolution and '11' for complete domain structure and evidence of rapid evolution. To describe the phylogenetic relationships between orthologs, we used the RAxML tree that we constructed from the SET domain for all PRDM9 orthologs. Species with missing ZF information, including species where PRDM9 has been lost, were excluded from this analysis using the drop.tip function of the ape package (Paradis et al. 2004), resulting in a tree with 91 tips. We used the phyloglm command to perform a

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

Analysis of the SET domain catalytic residues

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

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

Characterizing patterns of recombination in hybrid swordtail fish

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

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

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

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

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

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

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

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

We used the closest-feature command from the program bedops v2.4.19 (Neph et al. 2012) to determine the minimum distance between each 10 kb window and the functional feature of interest. For the TSS, we used the Ensembl annotation of the *Xiphophorus maculatus* genome with coordinates lifted over to v.4.4.2 of the linkage group assembly (Amores et al. 2014; Schumer et al. 2016) http://genome.uoregon.edu/xma/index_v1.0.php). For CGIs, we used the annotations 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=). To identify putative PRDM9 binding sites, we used the ZF prediction software available at zf.princeton.edu with the polynomial SVM settings to generate a position weight matrix for the *X.*

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

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

Because our analysis revealed a slight depression in recombination frequency near predicted PRMD9 motifs in swordtails, we evaluated whether this pattern was expected simply from base composition of the motif by shuffling the position weight matrix five times and repeating the analysis. As we had expected, we found a similar pattern in the simulations (**Figure 4- Figure Supplement 6**).

Most work in humans and mice has focused on the empirical PRDM9 binding motif rather than the computationally predicted motif. For comparison purposes, we therefore repeated the analysis described above for the computational prediction obtained for the human PRDM9A allele, using recombination rates in 10 kb windows estimated from the CEU LD map (Frazer et al. 2007; downloaded from: http://www.well.ox.ac.uk/~anjali/AAmap/). We also repeated this analysis for the gor-1 PRDM9 allele in *Gorilla gorilla*, using recombination rates in 10 kb windows estimated from a recent LD map (Schwartz et al. 2014; Stevison et al. 2016; downloaded from: https://github.com/lstevison/great-ape-recombination).

Comparisons of recombination landscapes with and without PRDM9

To investigate whether patterns of recombination rates near the TSS and CGI systematically distinguish between species that do and do not use PRDM9-directed recombination, we compared available data across species. We downloaded previously published recombination maps for three species without PRDM9 genes (dog, Auton et al. 2013, zebra finch and long-tailed finch, Singhal et al. 2015) and four species with complete PRDM9 orthologs (human, Frazer et al. 2007; Hinch et al. 2011; gorilla, Stevison et al. 2016; sheep, Johnston et al. 2016; and mouse, Brunschwig et al. 2012).

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

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

A caveat is that other than for swordtail, we relied on LD based genetics maps, which estimate population recombination rates $4N_er$, where N_e is the effective population size and r the recombination rate per meiosis. Because estimates of N_e decrease near genes as a consequence of diversity-reducing linked selection (e.g., Wright and Andolfatto 2008; Hernandez et al. 2011), a decrease in estimated population recombination rates near genes may not reflect a reduction in the recombination rate r. To

explore the potential importance of this caveat, we considered two species where both LD maps and pedigree or admixture maps were available: dogs and humans. In both cases, the qualitative results were the same as for the LD-based maps (**Figure 5 - Figure Supplement 1**). Since diversity-reducing linked selection should give rise, if anything, to a trough in diversity levels, it cannot explain the observed peaks at these features in species lacking PRDM9 or swordtail fish; in fact, since these species also experience this form of selection (e.g., Singhal et al. 2015), the true peaks in recombination rates near promoter-like features are likely somewhat more pronounced.

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

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

Whole testis and liver were dissected from two *X. birchmanni* adults and stored in HypoThermosol FRS (BioLife Solutions, Bothell, WA) buffer on ice until processing. Native chromatin ChIP was performed as described previously (Markenscoff-Papadimitriou et al. 2014). Briefly, tissue was homogenized and lysed; the lysate was spun through a sucrose cushion (to pellet nuclei). Nuclei were resuspended in 500ul MNase digestion buffer and digested with 1 unit of micrococcal nuclease (MNase, Sigma N5386, St. Louis, MO) for 2 minutes at 37 C, then inactivated with 20ul 0.5M EDTA and chilled on ice. The first soluble chromatin fraction was recovered by spinning for 10 min at 10,000 rcf at 4C and collecting the supernatant. To isolate the second soluble chromatin fraction, the pellet was resuspended in 500 μl dialysis buffer, rotated overnight at 4 C, then centrifuged for 10 min at 10,000 rcf at 4 C to pellet insoluble material. The digestion quality of each fraction was evaluated on an agarose gel. The two soluble fractions were combined for chromatin immunoprecipitation with 1 μg of H3K4me3 antibody (Millipore 04-745, Billerica, MA); 1/10 volume was retained as an input control. Antibody was bound to the remaining chromatin overnight while rotating at 4 C. The next day blocked Protein A and Protein G beads were added, and rotated for 3 hours. The bound beads were then washed a total of 7 times with chilled wash buffers and immunoprecipitated chromatin was eluted in elution buffer for 30 minutes at 37 C and

cleaned up with ChIP DNA Clean and Concentrator kit (Zymo Research, Irvine, CA). Libraries were prepared for sequencing using the NuGEN ultralow library prep kit (NuGEN, San Carlos, CA) following manufacturer's instructions and sequenced on an Illumina HiSeq 2500 at Hudson Alpha to collect 10.3-10.5 and 12.2-14.5 paired-end 50 bp reads for pull-down and input samples respectively.

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

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

As a secondary approach, we compared H3K4me3 peaks that are specific to the testis to H3K4me3 peaks that are specific to the liver, defined as peaks in the testis where there is no overlapping peak in the liver. Using a Chi-squared test, we asked whether

H3K4me3 peaks found only in the testis are more likely to overlap a PRDM9 binding motif than those that are liver specific (where the definition is analogous) (**Figure 4**). Because the size of H3K4me3 peaks will impact the expected overlap with PRDM9 binding motifs, we also constrained the size of the H3K4me3 peaks in the liver analysis to be the same as that inferred from the testis using the –size flag in homer (229 bp). Results were not qualitatively different with the original analysis, using liver H3K4me3 peaks that were inferred to be 238 bp. Counterintuitively liver-specific H3K4me3 peaks appear to overlap predicted PRDM9 motifs more often than testes-specific peaks (χ=14.8; p=1.2e-4). However, performing this same analysis with the 500 null motifs (generated as described above), we found that liver-specific peaks were significantly enriched in shuffled motifs in 85% of simulations (at the 0.05 level). This analysis suggests that base composition differences between liver and testes-specific H3K4me3 peaks explain the difference in overlap results.

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

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

The above analyses strongly suggest that in swordtail fish, PRDM9 does not make H3K4me3 marks but they do not indicate whether H3K4me3 peaks are associated with recombination events in swordtails. We therefore verified that recombination rates in 10 kb windows are significantly correlated with the distance of that window to the nearest H3K4me3 peak (rho=-0.072, p=2.3e-69; **Figure 4**). This relationship weakens but remains significant when accounting for distance both to TSSs and CGIs by a partial correlation analysis (rho=-0.026, p=5.4e-10). Furthermore, windows that contain a testis-specific H3K4me3 peak have a higher recombination rate than windows that contain a

liver-specific peak (**Figure 4 – Figure Supplement 4**). Finally, there is a significant positive correlation between the number of bp in a 10 kb window overlapping an H3K4me3 peak and the number of recombination events observed in that window in the testis but not in the liver (testis: rho=0.044, p=2.8e-29; liver: rho=0.002, p=0.66). Together, these analyses suggest that a relationship between H3K4me3 peaks and recombination exists in swordtails, but not one mediated through PRDM9 binding.

Acknowledgements

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

Competing Interests Statement

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

Figures

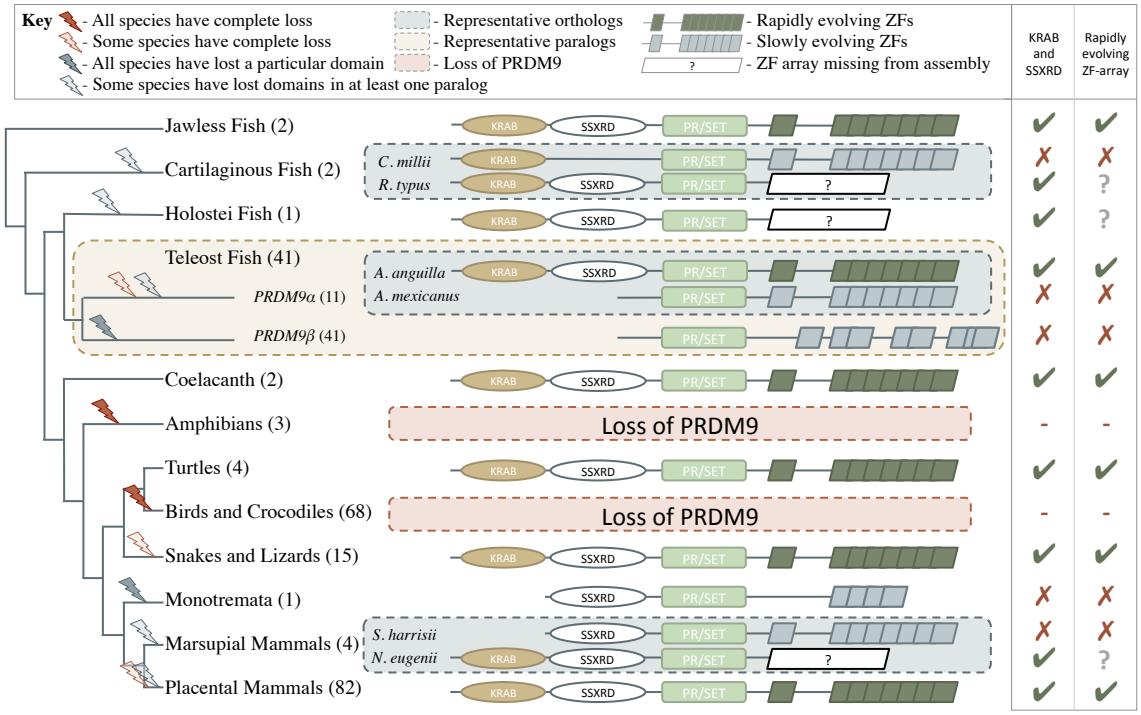

Figure 1. Phylogenetic distribution and evolution of PRDM9 orthologs in vertebrates. Shown are the four domains: KRAB domain (in tan), SSXRD (in white), PR/SET (in green) and ZF (in gray/dark green; the approximate structure of identified ZFs is also shown). The number of unique species included from 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. For select taxa, we present the most complete PRDM9 gene found in two examplar species. Within teleost fish, we additionally show a PRDM9 paralog that likely arose before the common ancestor of this taxon; in this case, the number of species observed to have each paralog is in paranthesis. Although the monotremata ZF is shaded gray, it was not included in our analysis of rapid evolution because of its small number of ZFs.

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

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

Figure 1- Figure Supplement 3. Expression levels of genes with a known role in meiotic recombination in testes tissues of three exemplar species.

Figure 1- Figure Supplement 4. Patterns of amino acid diversity as a function of amino acid position in the ZF alignment.

Figure 1- Figure Supplement 5. Examples of differences in computationally predicted PRDM9 binding motifs for species from three taxa.

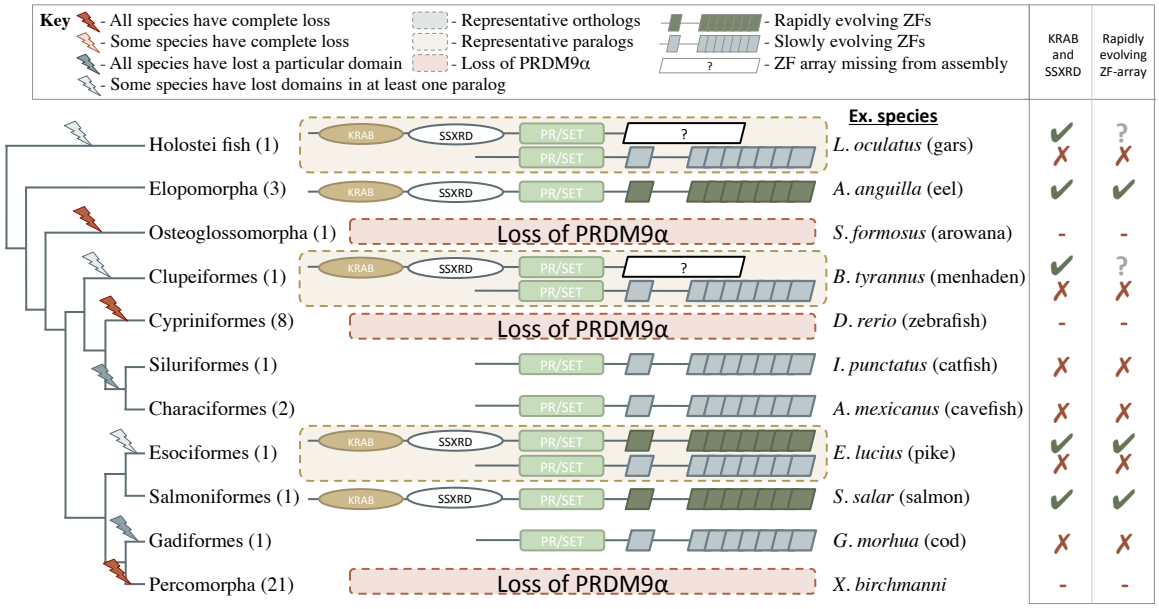

Figure 2. Phylogenetic distribution and functional domains of PRDM9α orthologs in teleost fish and in holostean fish that are outgroups to the PRDM9α/PRDM9β duplication event. Shown are the four domains: KRAB domain (in tan), SSXRD (in white), PR/SET (in green) and ZF (in gray/dark green; the approximate 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.

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.

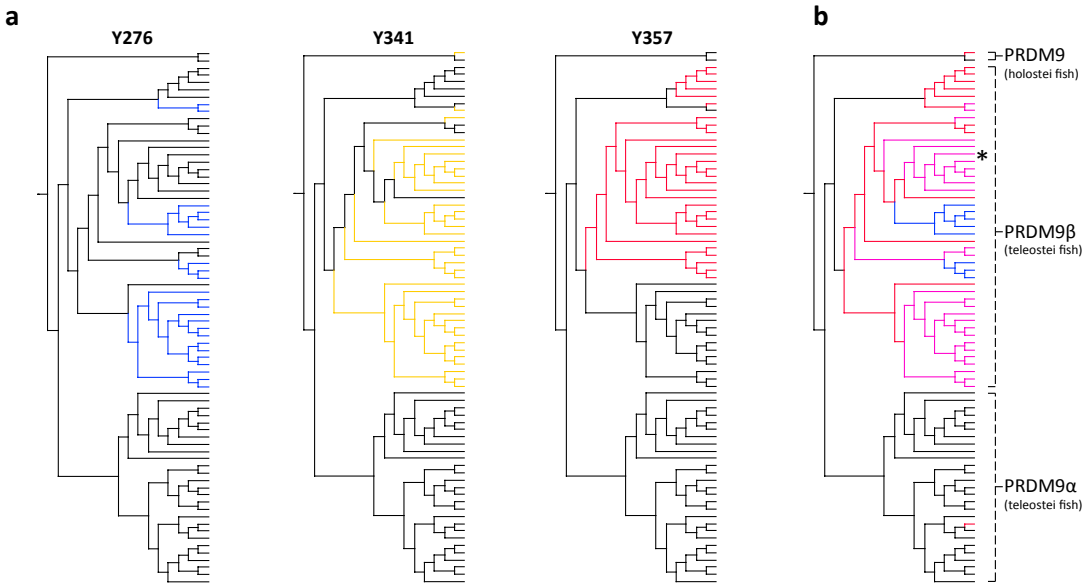

Figure 3. Substitutions at SET domain catalytic residues in bony fish PRDM9 genes. **(a)** Lineages within bony fish carrying substitutions at each of three tyrosine residues involved in H3K4me3 catalysis in human PRDM9 are shown in blue, yellow and red. **(b)** Lineages carrying substitutions at one, two or three of these residues are shown in red, pink and blue respectively. All PRDM9β genes as well as a partial PRDM9 ortholog from holostei fish carry one or more substitutions at these residues. The PRDM9β gene from *Xiphophorus* is indicated by the presence of asterisk.

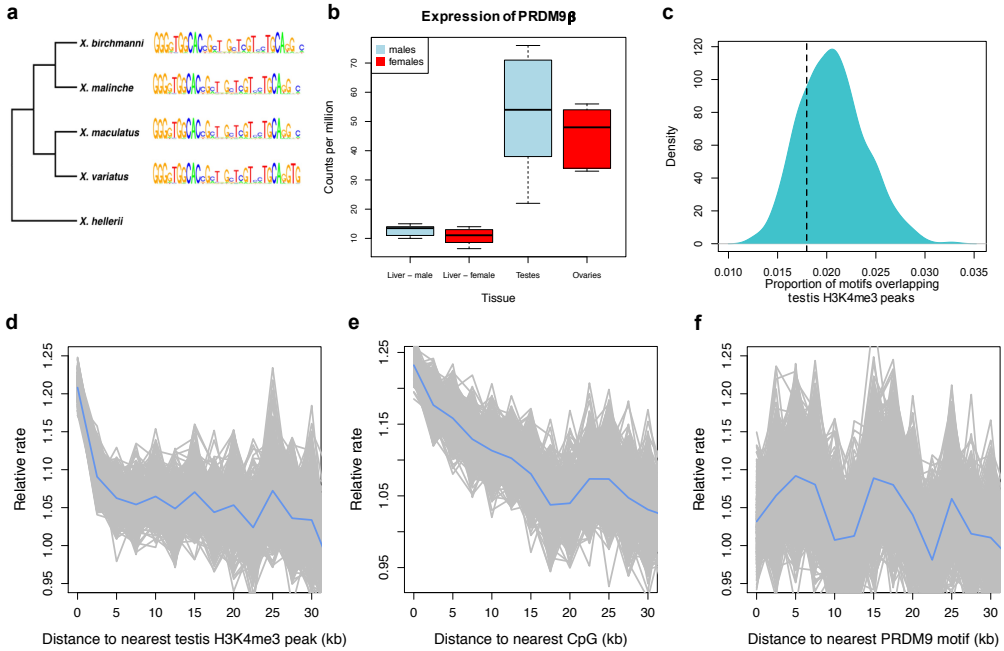

Figure 4. **Patterns of recombination and PRDM9 evolution in swordtail fish**. (a) The ZF array of PRDM9 appears to be evolving slowly in *Xiphophorus*, with few changes over over 1 million years of divergence (Cui et al. 2013; Jones et al. 2013). (b) PRDM9 is upregulated in the germline relative to the liver in *Xiphophorus birchmanni* and X. *malinche* (boxplot shows three biological replicates for each of the two species, and there is no difference between 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**). Simulations show that the slight dip in recombination frequency near PRDM9 binding sites could occur by chance, due to base composition differences (**Figure 4- Figure Supplement 6**, **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.

Figure 4 - Figure Supplement 6. Recombination frequency in swordtails as a function of distance to three shuffled versions of the swordtail PRDM9 recognition motif.

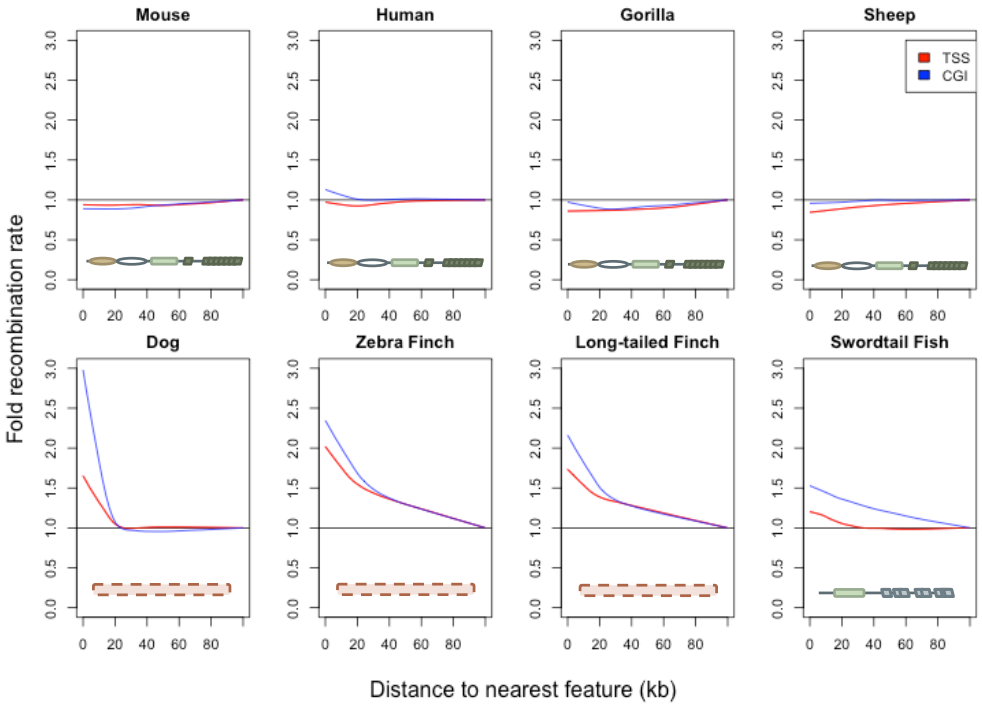

Figure 5. Patterns of recombination near TSSs and CGIs in species with and without complete PRDM9 orthologs. For each species, recombination rates estimated from patterns of LD (or in swordtail fish, from admixture switches) were binned in 10kb windows along the genome; curves were fit using gaussian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

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

Tables

Table 1. Evolution of the ZF in PRDM9 orthologs with different domain architectures. PRDM9 orthologs for which an empirical comparison dataset is available are ordered by their domain structures: from the top, we present cases of complete PRDM9 orthologs with KRAB-SSXRD-SET domains; partial orthologs putatively lacking KRAB or SSXRD domains or partial orthologs lacking both; then those containing only the SET domain. A row is shaded green if the ZF is in the top 5% most rapidly evolving C2H2 ZF in the species, as summarized by the proportion of amino-acid diversity at DNA-binding sites, and is dark green if it is ranked first. A complete PRDM9 ortholog from dolphins (*Balaenoptera acuforostrata scammoni*) is shaded in gray because there is no amino acid diversity between ZFs of the tandem array. The empirical rank is also shown, as are the number of PRDM9 orthologs identified in the species. Asterisks indicate PRDM9 orthologs known to play a role in directing recombination.

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
Balaenoptera acutorostrata scammoni	placental	KRAB-SSXRD-SET	NA	NA	1	272
Bison bison bison	placental	KRAB-SSXRD-SET	0.667	1	1	285
*Bos taurus** (chr1)	placental	KRAB-SSXRD-SET	0.684	1	3	313
Bos taurus (chrX)	placental	KRAB-SSXRD-SET	0.414	6	3	313
*Bos taurus** (chrX)	placental	KRAB-SSXRD-SET	0.414	7	3	313
Bubalus bubalis	placental	KRAB-SSXRD-SET	0.667	1	1	268
Chelonia mydas	turtle	KRAB-SSXRD-SET	0.414	11	1	235
Chlorocebus sabaeus	placental	KRAB-SSXRD-SET	0.500	1	1	344
Chrysemys picta bellii	turtle	KRAB-SSXRD-SET	0.478	1	1	308
Cricetulus griseus	placental	KRAB-SSXRD-SET	0.781	3	1	259
Dasypus novemcinctus	placental	KRAB-SSXRD-SET	0.614	1	1	289
Dipodomys ordii	placental	KRAB-SSXRD-SET	0.567	1	1	194
Esox lucius	teleost fish (α)	KRAB-SSXRD-SET	0.455	1	4	234
Fukomys damarensis	placental	KRAB-SSXRD-SET	0.430	3	1	227
*Homo sapiens**	placental	KRAB-SSXRD-SET	0.687	1	1	357
Latimeria chalumnae	coelacanth	KRAB-SSXRD-SET	0.545	2	1	227
Loxodonta africana	placental	KRAB-SSXRD-SET	0.617	1	1	381
Macaca fascicularis	placental	KRAB-SSXRD-SET	0.680	1	1	364
Macaca mulatta	placental	KRAB-SSXRD-SET	0.645	1	1	366
Marmota marmota marmota	placental	KRAB-SSXRD-SET	0.483	1	1	277
Microcebus murinus	placental	KRAB-SSXRD-SET	1.000	1	1	326
*Mus musculus**	placental	KRAB-SSXRD-SET	0.910	1	1	224
Nannospalax galili	placental	KRAB-SSXRD-SET	1.000	1	1	307
Octodon degus	placental	KRAB-SSXRD-SET	0.333	5	3	227
Octodon degus	placental	KRAB-SSXRD-SET	0.331	6	3	227
Ovis aries	placental	KRAB-SSXRD-SET	0.615	1	2	252
Ovis aries	placental	KRAB-SSXRD-SET	0.398	4	2	252
Ovis aries musimon	placental	KRAB-SSXRD-SET	0.353	12	1	285
Papio anubis	placental	KRAB-SSXRD-SET	0.585	1	1	404
Pelodiscus sinensis	turtle	KRAB-SSXRD-SET	0.692	1	1	221
Peromyscus maniculatus bairdii	placental	KRAB-SSXRD-SET	1.000	1	1	243
Protobothrops mucrosquamatus	squamata	KRAB-SSXRD-SET	0.462	5	1	195
Python bivittatus	squamata	KRAB-SSXRD-SET	0.571	1	1	206
Rattus norvegicus	placental	KRAB-SSXRD-SET	0.570	1	1	255
Rousettus aegyptiacus	placental	KRAB-SSXRD-SET	0.742	1	1	258
Salmo salar	teleost fish (α)	KRAB-SSXRD-SET	0.538	9	4	510
Salmo salar	teleost fish (α)	KRAB-SSXRD-SET	0.500	11	4	510
Sus scrofa	placental	KRAB-SSXRD-SET	0.542	1	1	248
Thamnophis sirtalis	squamata	KRAB-SSXRD-SET	0.459	3	1	179
Tupaia chinensis	placental	KRAB-SSXRD-SET	1.000	1	1	249
Tursiops truncatus	placental	KRAB-SSXRD-SET	0.939	1	1	233
Myotis lucifugus	placental	SSXRD-SET	0.524	1	2	308
Myotis lucifugus	placental	SSXRD-SET	0.310	68	2	308
Octodon degus	placental	SSXRD-SET	0.282	46	3	227
Sarcophilus harrisii	marsupial	SSXRD-SET	0.224	277	2	344
Callorhinchus millii	cartilaginous fish	KRAB-SET	0.314	6	1	63
Astyanax mexicanus	teleost fish (α)	SET	0.258	60	2	158
Astyanax mexicanus	teleost fish (β)	SET	0.167	152	2	158
Clupea harengus	teleost fish (α)	SET	0.279	6	4	118
Clupea harengus	teleost fish (α)	SET	0.278	7	4	118
Clupea harengus	teleost fish (α)	SET	0.274	10	4	118
Clupea harengus	teleost fish (β)	SET	0.158	114	4	118
Cynoglossus semilaevis	teleost fish (β)	SET	0.182	80	1	107
Danio rerio	teleost fish (β)	SET	0.179	345	1	367
Esox lucius	teleost fish (α)	SET	0.295	32	4	234
Esox lucius	teleost fish (β)	SET	0.192	176	4	234
Esox lucius	teleost fish (β)	SET	0.192	177	4	234
Fundulus heteroclitus	teleost fish (β)	SET	0.189	158	1	206
Haplochromis burtoni	teleost fish (β)	SET	0.180	148	1	168
Ictalurus punctatus	teleost fish (α)	SET	0.320	14	8	140
Ictalurus punctatus	teleost fish (α)	SET	0.319	15	8	140
Ictalurus punctatus	teleost fish (α)	SET	0.306	24	8	140
Ictalurus punctatus	teleost fish (α)	SET	0.303	25	8	140
Ictalurus punctatus	teleost fish (α)	SET	0.286	33	8	140
Ictalurus punctatus	teleost fish (α)	SET	0.276	39	8	140
Ictalurus punctatus	teleost fish (α)	SET	0.253	55	8	140
Ictalurus punctatus	teleost fish (β)	SET	0.179	127	8	140
Larimichthys crocea	teleost fish (β)	SET	0.192	70	1	115
Lepisosteus oculatus	holostei fish	SET	0.223	48	1	106
Maylandia zebra	teleost fish (β)	SET	0.173	161	1	176
Neolamprologus brichardi	teleost fish (β)	SET	0.173	141	1	152
Nothobranchius furzeri	teleost fish (β)	SET	0.180	245	1	266
Notothenia coriiceps	teleost fish (β)	SET	0.167	83	1	87
Oreochromis niloticus	teleost fish (β)	SET	0.173	173	1	190
Oryzias latipes	teleost fish (β)	SET	0.213	104	1	191
Otolemur garnettii	placental	SET	0.266	121	1	285
Poecilia formosa	teleost fish (β)	SET	0.191	184	1	242
Poecilia latipinna	teleost fish (β)	SET	0.191	175	1	235
Poecilia mexicana	teleost fish (β)	SET	0.191	187	1	244
Poecilia reticulata	teleost fish (β)	SET	0.191	162	1	212
Pundamilia nyererei	teleost fish (β)	SET	0.173	134	1	147
Pygocentrus nattereri	teleost fish (α)	SET	0.331	12	2	142
Pygocentrus nattereri	teleost fish (β)	SET	0.179	124	2	142
Salmo salar	teleost fish (β)	SET	0.188	411	4	510
Salmo salar	teleost fish (β)	SET	0.180	454	4	510
Sinocyclocheilus anshuiensis	teleost fish (β)	SET	0.185	224	2	284
Sinocyclocheilus anshuiensis	teleost fish (β)	SET	0.185	225	2	284
Sinocyclocheilus grahami	teleost fish (β)	SET	0.185	211	1	271
Sinocyclocheilus rhinocerous	teleost fish (β)	SET	0.185	208	2	269
Sinocyclocheilus rhinocerous	teleost fish (β)	SET	0.185	209	2	269
Takifugu rubripes	teleost fish (β)	SET	0.188	66	1	98
Xiphophorus maculatus	teleost fish (β)	SET	0.191	117	1	158

Supplementary Files

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

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

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

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

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

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

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

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

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

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

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

Figure Supplements

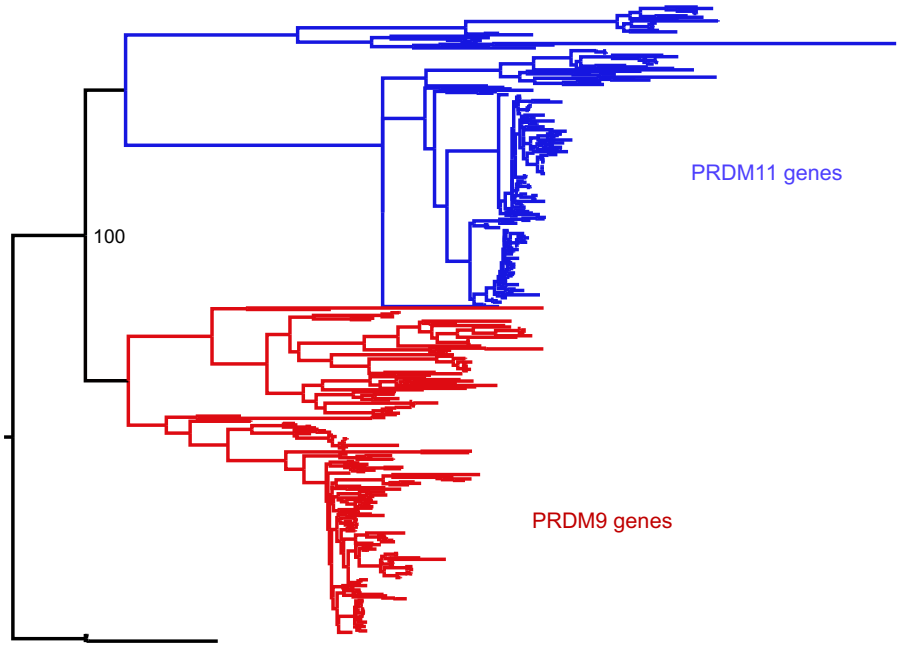

Figure 1- Figure Supplement 1. Phylogenetic approach to identifying PRDM9 orthologs and related gene families. A maximum likelihood phylogeny built with RAxML, using an alignment of SET domains, distinguishes between genes that cluster with mammalian PRDM9 and PRDM11 with 100% bootstrap support. Genes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

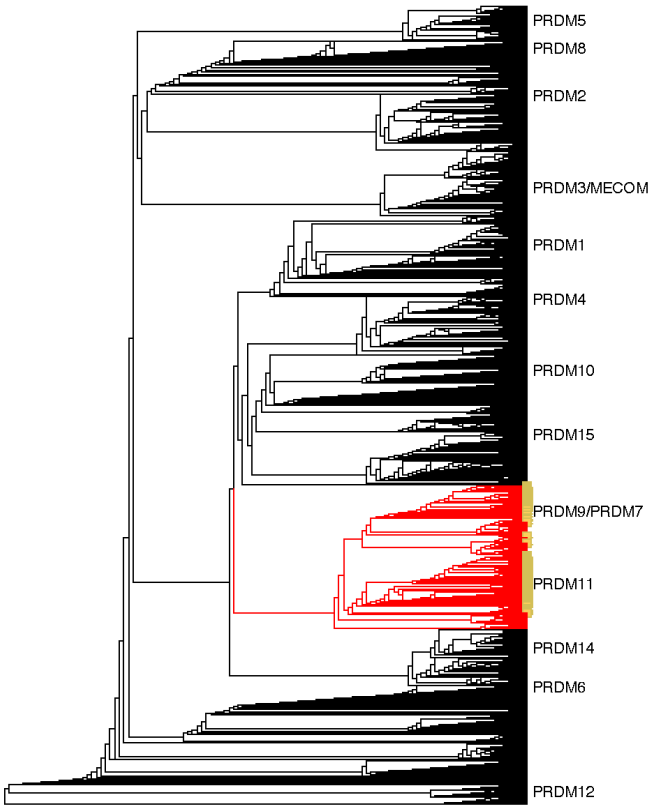

Figure 1- Figure Supplement 2. Neighbor-joining (NJ) guide tree based on the SET domain. A NJ guide tree analysis on SET domains identified in our RefSeq, whole genome assembly, and transcriptome datasets was used as an initial step to identify sequences clustering with human PRDM9/7 or PRDM11. These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

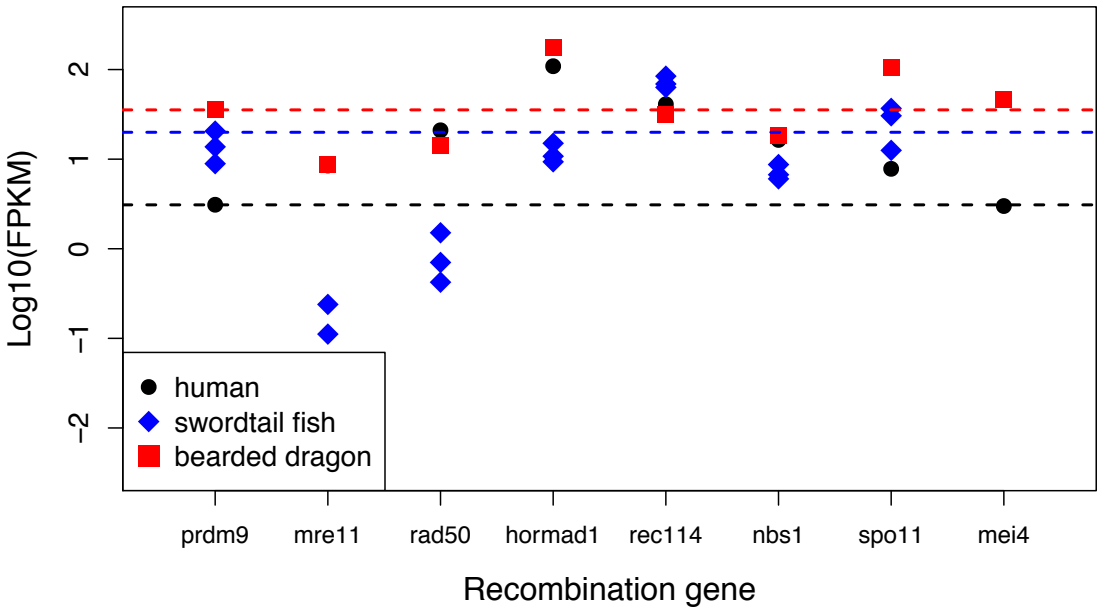

Figure 1- Figure Supplement 3. Expression levels of genes with a known role in meiotic recombination in testes of three exemplar species: human, swordtail fish and bearded dragon (a lizard). For three swordtails (*X. malinche*) and one bearded dragon, the FPKM per individual is plotted for each transcript.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

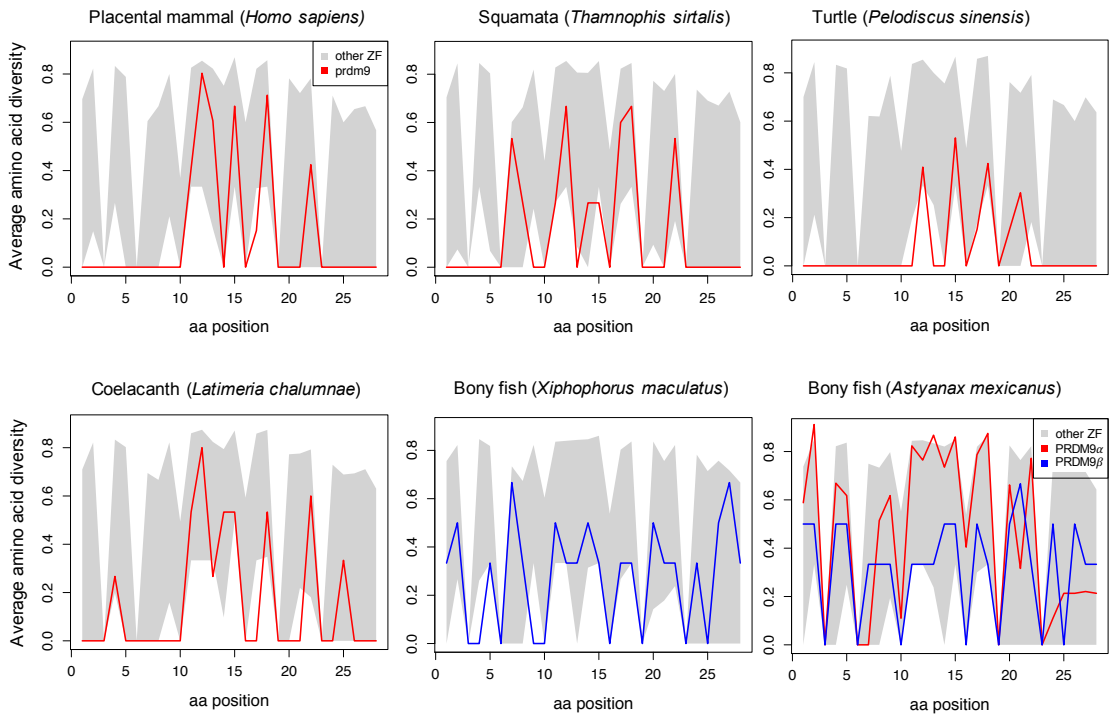

Figure 1- Figure Supplement 4. Amino acid diversity as a function of amino acid position in the ZF alignment for six examplar species. Each plot shows the 95% range of diversity levels at that site for all C2H2 ZFs from a species of that taxon (gray); the values at PRDM9 are show in red or blue. Turtles, snakes and coelacanth show a pattern of diversity that is similar to those in mammalian species with a complete PRDM9 ortholog, with higher diversity at DNA-binding sites (residues 11, 12, 15 and 18) and reduced diversity at most other sites. In bony fish, this pattern is not observed in PRDM9β genes (blue) or in partial PRDM9α genes (shown for *A. mexicanus*), where PRDM9 ZF diversity is more typical of other C2H2 ZFs.

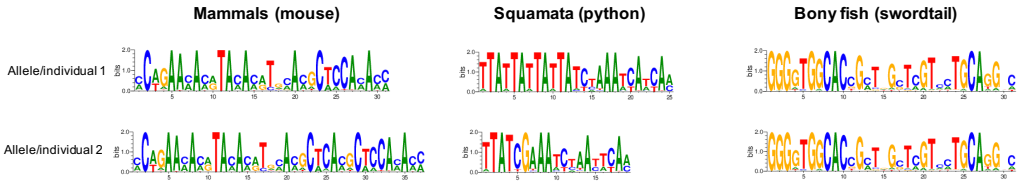

Figure 1- Figure Supplement 5. Examples of differences in computationally predicted PRDM9 binding motifs for species from three taxa. Shown are two mouse from the same species (*Mus musculus* subspecies; Genbank: AB844114.1; FJ899852.1), two pythons from the same species (*Python bivittatus*; the genome 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.

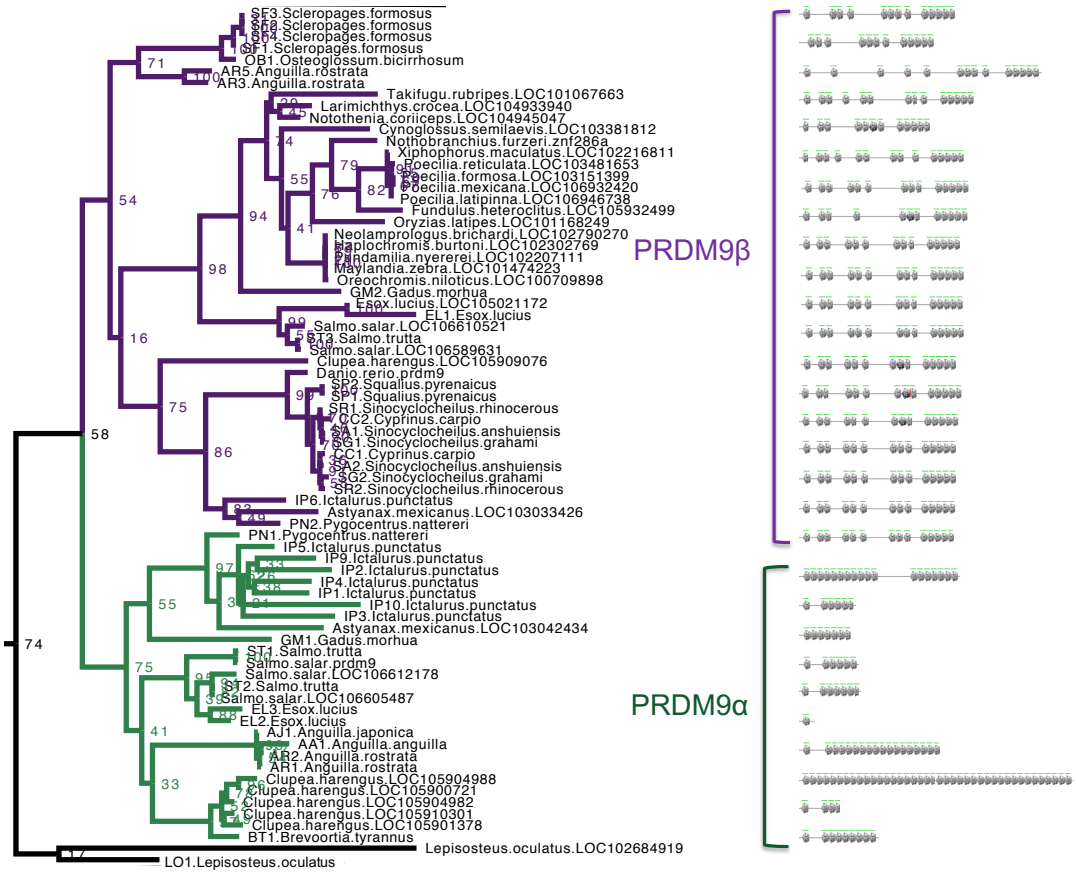

Figure 2- Figure Supplement 1. Section of maximum-likelihood phylogeny of the SET domain showing bony fish PRDM9 orthologs α and β. The reciprocal monophyly of PRDM9 orthologs α and β is reasonably well supported and in particular bootstrap support for the monophyly of PRDM9α genes is 75%. The ZF domains for representative PRDM9 orthologs of each type are shown to the right, with each gray pentagon indicating the location of a ZF.

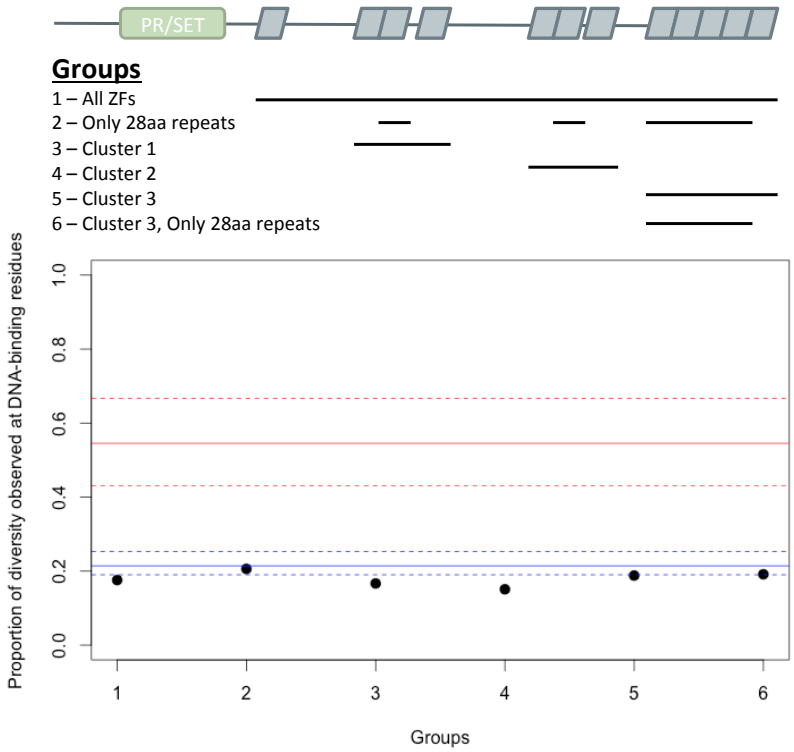

Figure 2- Figure Supplement 2. Analysis of ZF evolution in PRDM9β. Red lines show the median (solid) and first and third quantiles (dashed lines) for all 48 complete PRDM9 orthologs identified in vertebrates that have four or more ZFs. Blue lines show the median (solid) and first and third quantiles (dashed lines) for all other C2H2 ZF genes from *X. maculatus* (157 genes). Results about the rate of ZF evolution in the PRDM9β gene from *Xiphophorus maculatus* are qualitatively similar regardless of our choice of which cluster of individual ZFs domains to include in our analysis, indicating that our ability to detect evidence of positive selection at DNA-binding residues in these arrays, or lack thereof, is unlikely to be influenced by this choice.

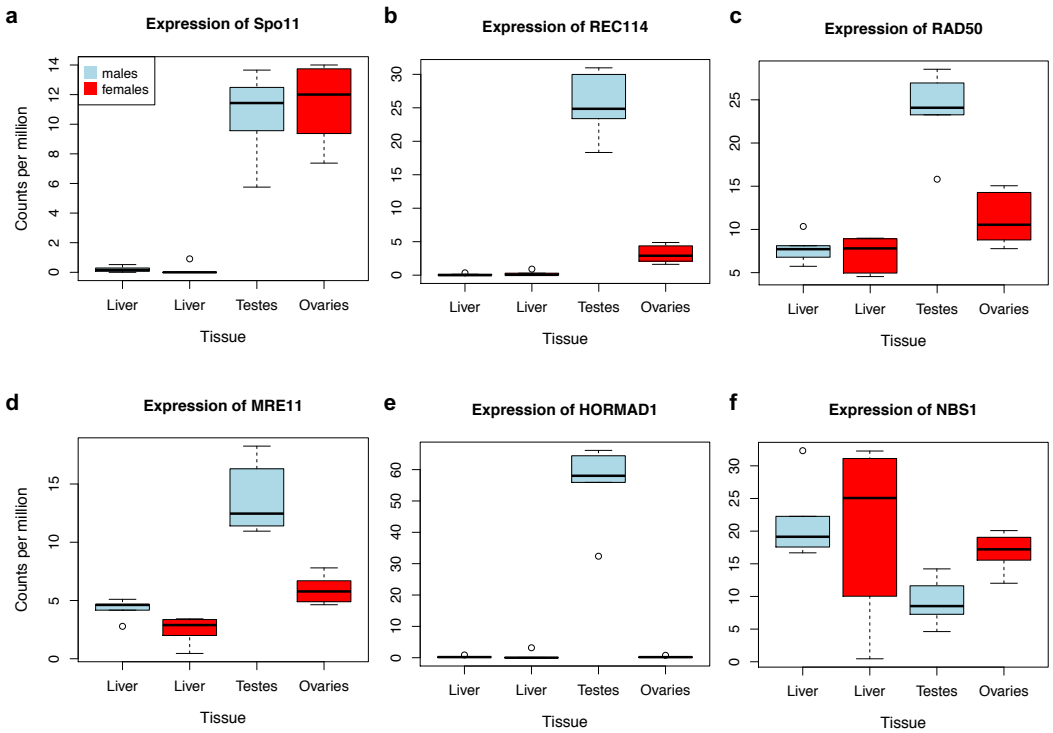

Figure 4- Figure Supplement 1. Expression levels of meiosis-related genes in swordtail fish tissues. In general, the seven meiosis-related genes surveyed had higher expression in tissues containing germline cells than liver tissue, but this pattern was much more pronounced in testis tissue (compared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

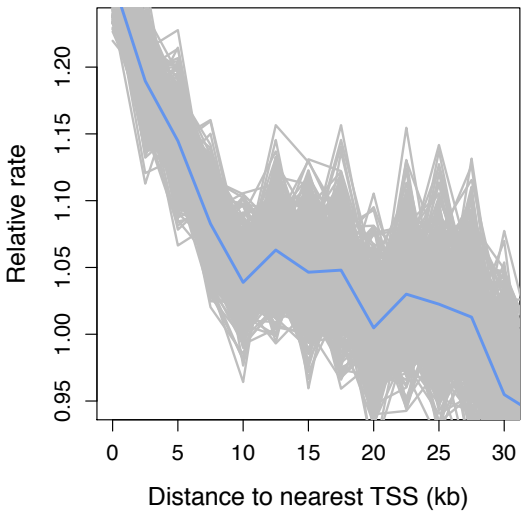

Figure 4 – Figure Supplement 2. Recombination frequency in swordtails as a function of distance to the TSS. Partial correlation analyses suggest that the association between the TSS and recombination rate in swordtails is explained by H3K4me3 peaks and CGIs.

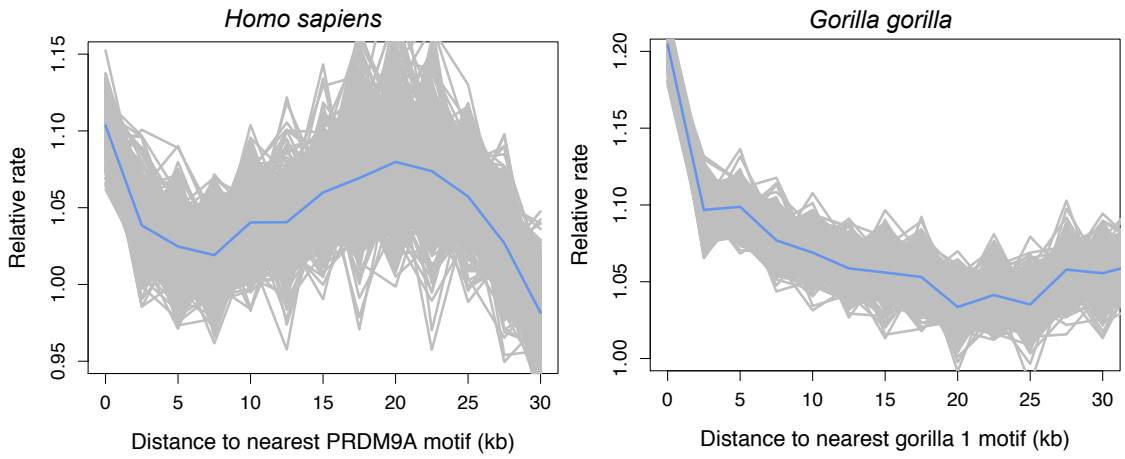

Figure 4- Figure Supplement 3. Recombination rates show a peak near the computationally predicted PRDM9A binding motif in humans and gor-1 allele in gorillas. Most work investigating relationships between PRDM9 motifs and recombination rates have focused on the PRDM9 motif empirically inferred from recombination hotspots, but the empirical motif is unknown for many species. To generate results comparable to those we present for swordtails in **Figure 4F**, we therefore determined recombination rate in humans (using the map based on LD patterns in the CEU; Frazer et al. 2007) as a function of distance to computationally predicted binding sites for the PRDM9A motif in humans and as a function of distance to computationally predicted binding sites for the gor-1 PRDM9 allele (Schwartz et al. 2014) in gorillas (using the LD-based map from Stevison et al. 2016).

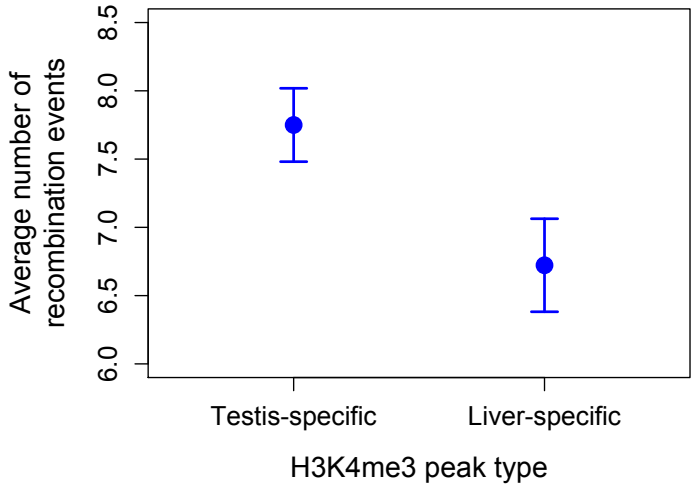

Figure 4 - Figure Supplement 4. Higher observed recombination rate at testis-specific H3K4me3 peaks than liver-specific H3K4me3 peaks. H3K4me3 peaks found only in the testis and not in the liver of *X. birchmanni* have higher observed recombination rates in *X. birchmanni – X. malinche* hybrids. This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

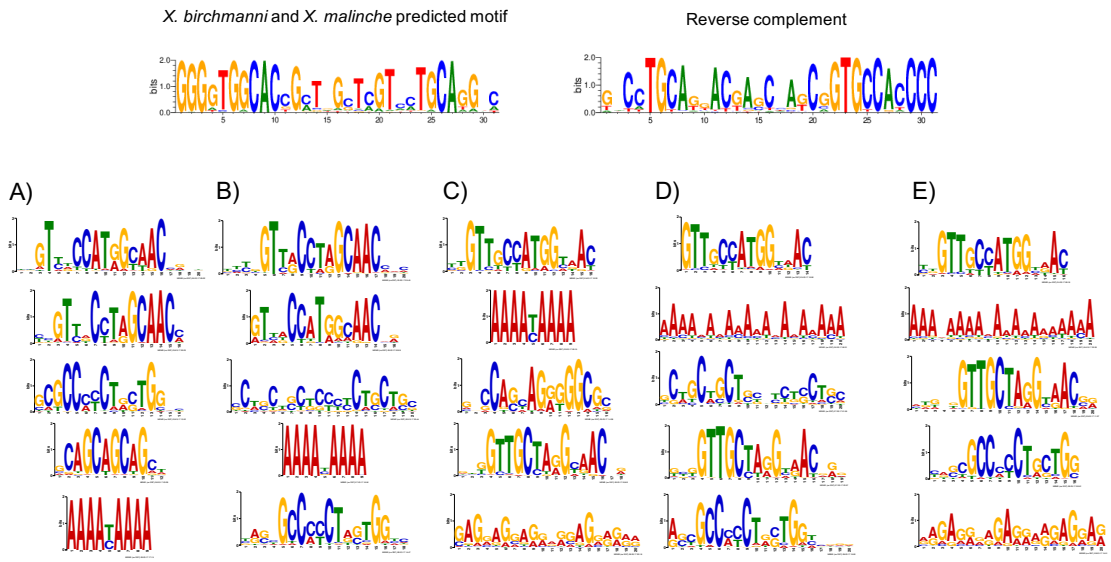

Figure 4 - Figure Supplement 5. MEME prediction of sequences enriched in testis-H3K4me3 peaks relative to liver-specific H3K4me3 peaks. Results shown in A-E are from five replicate runs of 2,000 testis-specific sequences using liver-specific sequences as a background comparison set. The swordtail computationally-predicted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

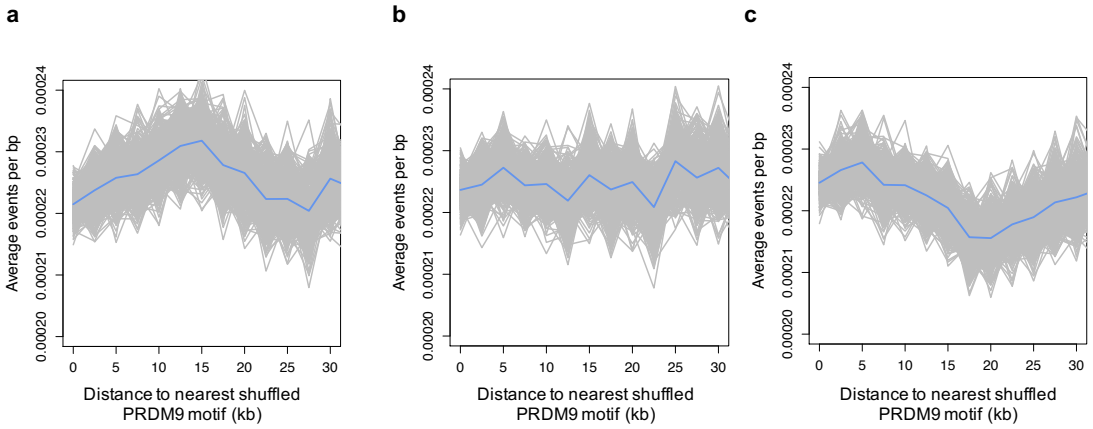

Figure 4- Figure Supplement 6. Recombination frequency in swordtails as a function of distance to three shuffled versions of the swordtail PRDM9 recognition motif. These results suggest that the pattern of depressed recombination near the PRDM9 binding motif seen in the real data (**Figure 4**) is not unexpected, given the base composition of the PRDM9 binding motif in swordtails.

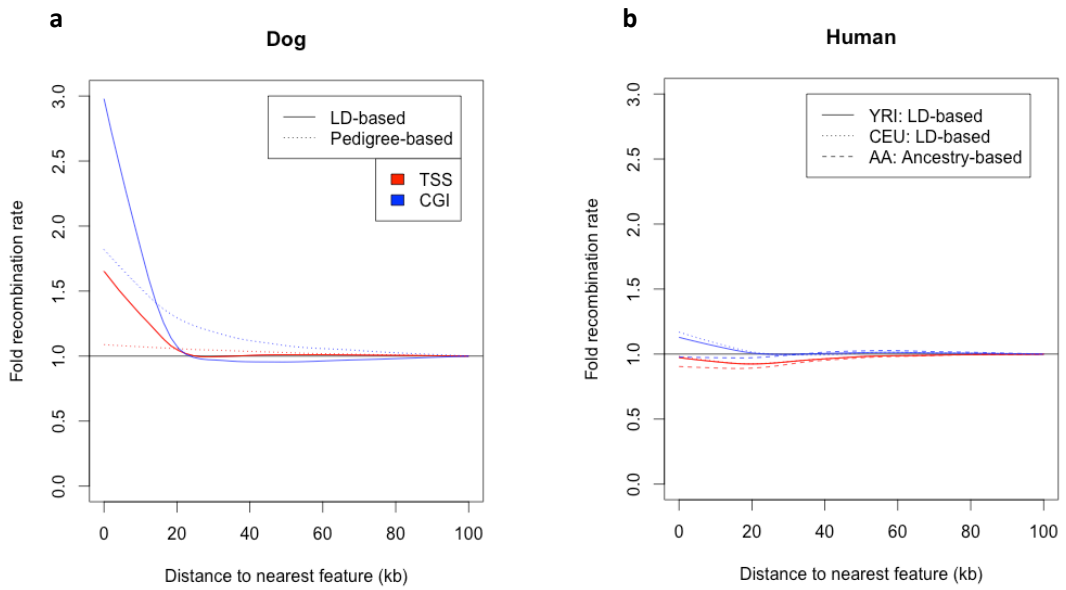

Figure 5 – Figure supplement 1. Dependence of patterns of recombination near TSSs and CGIs in dog and human on the type of genetic map. (a) Recombination rates near the TSS and CGI in dogs are shown using recombination maps inferred either from LD patterns or pedigree data. The magnitude of the peak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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T.

D. Willis, F. L. Yu, H. M. Yang, C. Q. Zeng, Y. Gao, H. R. Hu, W. T. Hu, C. H. Li, W. Lin, S. Q. Liu, H. Pan, X. L. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. R. Zhang, H. B. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. J. Yao, W. Huang, X. Chu, Y. G. He, L. Jin, Y. F. Liu, Y. Y. Shen, W. W. Sun, H. F. Wang, Y. Wang, X. Y. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, J. T. F. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. M. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. H. Guan, Y. Li, H. M. Munro, Z. H. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol and C. Int HapMap. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-U853.

Fullerton, S. M., A. B. Carvalho, and A. G. Clark. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Molecular Biology and Evolution 18:1139-1142.

Getun, I. V., Z. Wu, M. Fallahi, S. Ouizem, Q. Liu, W. M. Li, R. Costi, W. R. Roush, J. L. Cleveland, and P. R. J. Bois. 2017. Functional Roles of Acetylated Histone Marks at Mouse Meiotic Recombination Hot Spots. Molecular and Cellular Biology 37.

Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017-1018.

Grey, C., J. A. J. Clement, J. Buard, B. Leblanc, I. Gut, M. Gut, L. Duret, and B. de Massy. 2017. In vivo binding of PRDM9 reveals interactions with noncanonical genomic sites. Genome Research 27:580-590.

Hayashi, K., K. Yoshida, and Y. Matsui. 2005. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature 438:374-378.

Hedges, S. B., J. Marin, M. Suleski, M. Paymer, and S. Kumar. 2015. Tree of Life Reveals Clock-Like Speciation and Diversification. Molecular Biology and Evolution 32:835-845.

Heil, C. S. S., C. Ellison, M. Dubin, and M. A. F. Noor. 2015. Recombining without Hotspots: A Comprehensive Evolutionary Portrait of Recombination in Two

Closely Related Species of Drosophila. Genome Biology and Evolution 7:2829-2842.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Z. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7.

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

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

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

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

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

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

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

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

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

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

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

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