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], Chris Holland[4,6], Gil G. Rosenthal[4,6], and M. 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 Cientificas de las Huastecas "Aguazarca"
[5] Dept. of Evolution, Ecology and Environmental Biology, Columbia University
[6] 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 a handful of species reveal 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 identified PRDM9 from genome and transcriptome data in 225 species, finding the complete PRDM9 ortholog across distantly related vertebrates. Yet, despite its broad conservation, we inferred 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 its domains are necessary for directing recombination. Indeed, as we show, swordtail fish carrying a partial 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 (hundreds to thousands of base pairs; Lange et al. 2016) intervals 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 appear to be at least two mechanisms regulating their location. In mammalian species, including apes, mice and more indirectly cattle, the locations of recombination hotspots are specified through binding of the PRDM9 protein (Baudat et al. 2010; Myers et al. 2010; Parvanov et al. 2010; Sandor et al. 2012). In these species, PRDM9 has four major functional domains: a KRAB, SSXRD and PR/SET domain, followed by a C2H2 zinc finger (ZF) array (**Figure 1**). The ZF array of PRDM9 plays a key role in directing the locations of recombination hotspots. During meiosis, PRDM9 binds sites across the genome specified by its ZF array (reviewed in Segurel et al. 2011). Subsequently, the PR/SET domain of PRDM9 creates H3K4me3 marks which, possibly in conjunction with other genomic features, recruit SPO11 and initiate DSBs (reviewed in Lichten and de Massy 2011). The KRAB and

SSXRD domains may play a role in this recruitment, but their specific functions remain unclear. One consequence of using PRDM9 to specify recombination hotspots is that recombination tends to be directed away from genes (Myers et al. 2005; Coop et al. 2008), with only a small proportion of hotspots occurring at transcription start sites (Brick et al. 2012).

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 transcriptional start sites and CpG islands, 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 marks (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. In species in which recombination is driven by PRDM9, hotspot locations are not shared between closely related ape species or between mouse subspecies, and differ even among individuals (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). This asymmetry in binding presumably leads to hotter alleles more often experiencing a DSB. 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 colder allele in hot/cold heterozygotes acts analogously to selection for the colder allele (Nagylaki and Petes 1982) and is expected to lead to the rapid loss of hotspots from the population (“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; 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. PRDM9 has the fastest evolving C2H2 ZF array in the human genome (Oliver et al. 2009; Myers et al. 2010) and mammalian PRDM9 genes show strong evidence of positive selection at known DNA-binding sites of ZFs (Oliver et al. 2009). The mechanism driving this rapid evolution is unclear, but it has been hypothesized that once the under-transmission of hotter alleles has led to the erosion of a sufficient number of hotspots that the proper alignment or segregation of homologs during meiosis is jeopardized, new PRDM9 ZF alleles may be strongly favored (Coop and Myers 2007; Myers et al. 2010; Ubeda and Wilkins 2011). 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.

In contrast, in species that do not use PRDM9 to direct meiotic recombination events, this rapid evolution is not seen. In birds that lack an ortholog of PRDM9, the locations of recombination hotspots are conserved over long evolutionary time scales (Singhal et al. 2015). Similarly, both the location and heats of recombination hotspots are conserved across highly diverged yeast species, in which H3K4me3 is performed by a single gene without a DNA binding domain (Lam and Keeney 2015). In these taxa, recombination is concentrated at or near functional genomic elements such as transcriptional start sites and CpG islands, which are highly conserved (Brick et al. 2012; Auton et al. 2013; Choi et al. 2013; Lam and Keeney 2015; Singhal et al. 2015). Whether this apparent targeting is facilitated by specific binding motifs or simply by greater accessibility to the genome remains unknown. However, even if there were specific motifs that increase rates of recombination near functional genomic elements, they likely have important, pleiotropic functions (Nicolas et al. 1989). Thus, in these species, 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 genes 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 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 identified 227 orthologs of PRDM9 from 149 of 225 species of vertebrates investigated, using a combination of genome sequences and RNAseq data.

Results

Initial identification of PRDM9 orthologs

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 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 PR/SET domains to distinguish *bona fide* PRDM9 orthologs from members of paralagous gene families and to characterize the distribution of PRDM9 duplication events (**Supplementary Figure S1**; **Supplementary Figure S2)**. Since the placement of the major clades used in our analysis is not controversial, in tracing the evolution of PRDM9 orthologs, we assumed that the true phylogenetic relationships between clades are those reported by several recent papers (synthesized by the TimeTree project, Hedges et al. 2015).

This approach identified 227 PRDM9 orthologs (**Supplementary Table S1; Supplementary Table S2**), 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 Table S2**). 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 taxon (**Figure 1**). We used both phylogenetic methods and analysis of the ZF structure to distinguish these copies (see **Supplementary Figure S3, 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 clades, suggesting at minimum three losses of PRDM9α type orthologs within teleost fish (**Figure 2, Supplementary Table S1**). 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 Table S1**).

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 **Supplementary Figure S4**). We selected representatives of each major vertebrate group with publically available testis expression or testis RNA-seq datasets (**Supplementary Table S3**); 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 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 Table S4; Supplementary Table S5**).

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 (**Supplementary Figure S5, Supplementary Table S4; Supplementary Table S6**). 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 PR/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, Supplementary Figure S3**); and the absence of the KRAB domain in monotremata (*Ornithorhynchus anatinus*) and marsupial mammals (*Sarcophilus harrisii*, **Figure 1**; **Supplementary Table S1**).

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 clades of teleost fish (*Gadus morhua*, *Astyanax mexicanus*, *Ictalurus punctatus*, *Esox lucius*; **Supplementary Table S5**), as well as losses of the KRAB domain in two PRDM9 orthologs identified in monotremata (both in *O. anatinus*, **Supplementary Table S5**).

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; see **Methods** for discussion of duplicates of PRDM9 that lack evidence for functional constraint). 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 of evolution (**Figure 1, Supplementary Table S7, Supplementary Figure S3**). 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, it seems plausible that when PRDM9 directs recombination, there

is diversifying selection on the ZF, to compensate for the loss of hotspots due to under-transmission of hot alleles (see Introduction). If so, we might expect species with PRDM9-directed recombination to show evidence for rapidly-evolving PRDM9 ZF arrays.

To investigate this possibility, 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; see **Methods**). 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-mediated 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 have been associated with interspecific variation in recombination phenotypes (**Supplementary Table S8**; Sandor et al. 2012; Ma et al. 2015), and both are seen to be rapidly evolving (**Table 1, Supplementary Table S8**).

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 (**Supplementary Figure S6**). 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 Table S8**). 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 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**). We were unable to confirm the loss of the missing domain in this case (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 ($p < 4e-25$, hypergeometric test). Assuming that rapid evolution of the ZF is indicative of PRDM9-directed recombination, these observations strongly suggest that KRAB and SSXRD domains are required for this role.

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

Together, these findings suggest that although partial PRDM9 orthologs retain a function, they are unlikely to direct recombination, as the complete ortholog does in placental mammals. To test this hypothesis, we examined patterns of recombination using publicly available and new data from naturally-occurring swordtail fish hybrids (*X. birchmanni* x *X. malinche*; see Methods). Like other percomorphs (Setiamarga et al. 2008), swordtail fish have a PRDM9β type gene lacking the KRAB and SSXRD domains, and a slowly evolving ZF array with testis-specific expression (**Figure 3**).

Briefly, we used ancestry switchpoints in hybrids to identify crossover events (see **Methods**). We then asked whether the frequency of recombination events is correlated with distance to promoter-like features, as is observed in species that do not use PRDM9 to specify recombination hotspots, but not in apes or mice (see Introduction). By this approach, there is a clear peak in recombination rates near transcriptional start sites and CpG islands (**Figure 3**), similar to what is observed in species with PRDM9-independent recombination (Auton et al. 2013; Lam and Keeney 2015; Singhal et al. 2015). We also used the computationally predicted PRDM9 binding motif to predict where PRDM9 binds in the swordtail genome. In contrast to what is observed in species with PRDM9-mediated recombination (**Supplementary Figure S7**), there is no elevation in recombination near predicted PRDM9 binding sites (**Figure 3C**). Thus, fine-scale recombination patterns in swordtails, whose PRDM9 ortholog lacks KRAB and SSXRD domains, resemble those in species that lack a PRDM9 ortholog altogether (**Figure 3**; (Auton et al. 2013; Lam and Keeney 2015; Singhal et al. 2015).

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 PR/SET domains followed by a C2H2 ZF array, and that this complete architecture was likely already in place by the origin of vertebrates.

Moreover, even though to date only the functions of the PR/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, 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, PR/SET, and ZF domains. Since, under plausible assumptions, rapid evolution of the ZF is expected when the gene directs recombination, 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. This observation again points to the importance of KRAB and/or SSXRD domains, but also raises a puzzle. The SET and ZF domains of PRDM9 are retained and the gene is upregulated in the germline. Since together these domains are responsible for the DNA-binding and H3K4me3 activity of PRDM9, it is not obvious why recombination patterns in this species would resemble those of a PRDM9 knockout. One possibility is that H3K4me3 marks alone are not sufficient to drive recombination in some species, but instead must interact with other proteins or functional elements. For example, H3K4me3 marks laid down by PRDM9 may be outcompeted by H3K4me3 marks occurring at the TSS when KRAB and SSXRD domains are not present to help recruit the recombination machinery. Consistent with this hypothesis, a recent paper suggests that the KRAB domain may play an indirect 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 conservation of the protein points to another role of the gene. In that regard, we note that PRDM9 was originally annotated in mice as a meiotic transcription factor, because it was shown to regulate genes expressed during meiosis (Hayashi et al. 2005), and it may play that role without KRAB and SSXRD domains, at least in a subset of species.

Conversely, if the presence of all four domains and the rapid evolution of the ZF array are sufficient indications of PRDM9-directed recombination, then this regulatory mechanism 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; **Supplementary Figure S8**). Assessing the role of PRDM9 in directing recombination in these species is a natural next step in understanding the evolution of recombination mechanisms.

The distribution of PRDM9 in 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 clades 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, PRDM9 binding tends to direct recombination away from the TSS and genes (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.

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, as well as an additional 44 retrieved from whole genome assemblies, and seven retrieved from RNAseq datasets from five species without sequenced genomes (see ***Predicting PRDM9 orthologs from whole genome sequences***), were input into ClustalW2 (Larkin et al. 2007) to generate a neighbor-joining (NJ) guide tree (see **Supplementary Figure S2)**. This approach was used to identify and remove genes clustering with known PRDM family orthologs from humans that 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

For 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 observed in mammalian PRDM9 genes, we investigated whether or not these species had additional PRDM9 orthologs in their whole genome assemblies (**see Supplementary Table S1; Supplementary Table S2**). We ran *tblastn* against the whole genome assembly of 33 species of interest 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 Table S2**). 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 or less.

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. 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 Table S2** 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 monotremata, two marsupials, one turtle, four lizards, and eight snakes (**Supplementary Table S1**). We were unable to detect PRDM9 orthologs in one lizard (*Anolis carolinenesis*), or in any of three amphibian species (**Supplementary Table S2**). We used RNA-seq data to investigate whether these negative findings are due to genome assembly quality or reflect true losses (see below).

Phylogenetic Analysis of PRDM9 orthologs and related gene families

To understand the evolution of PRDM9 within vertebrates, we used a phylogenetic approach. We first built an alignment of the amino acid sequences of putative PRDM9 and PRDM11 SET domains using Clustal Omega (Sievers et al. 2011). Genes clustering with PRDM11 were included because it has been previously reported that PRDM11 arose from a duplication event of PRDM9 in the common ancestor of bony fish and tetrapods (Vervoort et al. 2016), and we were interested in identifying any PRDM9 orthologs from vertebrate species that may precede this duplication event. The alignment coordinates were then used to generate a nucleotide alignment, which was used as input into the program RAxML (v7.2.8; Stamatakis 2006). We performed 100 rapid bootstraps followed by maximum likelihood estimation of the tree under the General Reversible Time substitution model, with one partition for each position within a codon. The resulting phylogeny contained monophyletic groups corresponding to the PRDM9 and PRDM11 duplication event, with 100% bootstrap support (**Supplementary Figure S1**). These groups were used to label each putative ortholog as PRDM9 or PRDM11. Only jawless fish have PRDM9 orthologs basal to this duplication event, suggesting PRDM11 arose from PRDM9 before the common ancestor of cartilaginous and bony fish.

Within teleost fish, we identified two groups of PRDM9 orthologs, which we refer to as PRDM9α and PRDM9β. These groups are distinguished at only 58% bootstrap support (**Supplementary Figure S3**). However, this potential duplication event is coincident with the whole genome duplication event known to have occurred in the common ancestor of teleost fish (Taylor et al. 2003). Moreover, the phylogenetic grouping is concordant with inferred differences in the domain architectures between the two orthologs: PRDM9β genes can be distinguished from PRDM9α in that they all share a unique and derived ZF array structure (**Supplementary Figure S3**) and are always found without the KRAB and SSXRD domains, whereas PRDM9α genes generally have the ancestral arrangement of ZFs and occasionally have these N-terminal 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 Table S3**; **Supplementary Table S5; Supplementary Table S6**). 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 (Leinonen et al. 2011; v2.5.7) 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 (Zerbino and Birney 2008; v1.2.1) 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 Table S3.**

In order to identify potential PRDM9 transcripts in each assembled transcriptome, 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 Table S5**). For five species lacking genome assemblies, we extracted PRDM9 orthologs with best *blast* hits to human PRDM9/7 and

included these in our analyses (see ***Phylogenetic Analysis of PRDM9 orthologs and related gene families*****).**

Failing 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 address this concern, 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 Table S3**; **Supplementary Table S6**). 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 clades (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

Xiphophorus birchmanni and *X. malinche* were collected from the eastern Sierra Madre Oriental in Hidalgo state of 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. 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 Accession #XXXXXX). 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 Table S3**. 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 clades (*Xiphophorus malinche*, *Pogona vitticeps*, and *Homo sapiens*).

To quantify expression in *X. malinche*, 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 (**Supplementary Figure S5**). 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 in 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 further investigated the potential loss of these domains). This approach allowed us to rule out genome-wide losses of PRDM9 domains in nine out of 14 species of mammals where

our initial approach 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 Table S1**). 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 may be due to assembly errors or gaps.

For species with available RNAseq data from clades 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 Table S6**). 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 Table S5**). 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 Table S3**), 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 clades. Results of this analysis can be found in **Supplementary Table S5**. In summary, there were two cases where the transcriptomes supported additional domain structures not found in the whole genome sequence (**Supplementary Table S5**): 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 Table S4, Supplementary Table S5**) (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 e.g., 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 clade using Clustal Omega (see **Supplementary Table S7**), 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. In the case of cartilaginous fish, we used the sequence from *R. typus* and *C. milii*, which are an estimated 400 million years diverged.

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 Table S7**). In all cases of N-terminal truncation analyzed, dN/dS is significantly less than one (**Supplementary Table S7**). While it is possible that some of these cases represent recently pseudogenized genes, the repeated appearance of partial PRDM9 orthologs suggests that this is not

generally the case. Instead, the widespread evidence for purifying selection on the protein 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 Table S9)**. PRDM7 is thought to have arisen from a primate specific duplication event and have undergone subsequent losses of the C2H2 ZF array and the 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 (there are fewer than four tandem ZFs in 24 of 28 orthologs), in sharp contrast to other clades 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 Table S2, Supplementary Table S5**). 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 to collect additional data on within species diversity of the ZF array (**Supplementary Figure S8**). Primers were designed based on the *Python bivattatus* genome (Castoe et al. 2013) to amplify the ZF containing exon of PRDM9 and to amplify 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). Instead, 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, 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) for each species with a PRDM9 ortholog. To simplify alignment generation, we only used tandem ZF arrays and focused on 28 amino acid long C2H2 motif arrays (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 five or more adjacent ZFs was used for the following analysis (**Supplementary Table S8**). 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 1; Supplementary Figure S9**). 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 Table S1**). Alignments with fewer than four ZFs were excluded from further analysis.

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 Table S8**), and calculated the proportion of the total amino acid diversity at canonical DNA-binding residues of the ZF array. To compare results to those for other genes, we ranked each PRDM9 gene by this value against all other C2H2 ZF genes from the same species (**Table 1, Supplementary Table S8**).

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 CpG islands would be enriched for recombination events.

To build a hybrid recombination map, we took advantage of data from a natural hybrid population formed between the percomorph species *X. birchmanni* and *X. malinche*. These 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 3D**). We used a combination of publicly available and new data from swordtail fish hybrids to build a hybrid recombination map and ask what genomic features are associated with higher rates of recombination. Specifically, we used previously collected reduced-representation sequencing data from 170 hybrids (Schumer et al. 2014) and from an additional 98 hybrids collected from the Tlatemaco hybrid zone in the Sierra Madre Oriental of Mexico. These hybrids derive approximately 25% of their genome from *X. birchmanni* and 75% from *X. malinche* (Schumer et al. 2014). Previous work suggested that this hybrid population formed within the last 56 generations (Schumer et al. 2014).

To identify crossover events in hybrids, we used ancestry switchpoints. Specifically, we applied the Multiplexed Shotgun Genotyping approach (Andolfatto et al. 2011; "MSG") to assign posterior probabilities for three ancestry states across the genome (homozygous parent 1, heterozygous, homozygous parent 2). Low quality

basepairs were trimmed from reads (Phred quality score <20) and reads with fewer than 30 bp of contiguous high quality sequence were removed. The maximum number of reads per individual including in the analysis was 2 million for computational speed, and the minimum was 300,000, based on power simulations suggesting that this minimal number of reads is required for accuracy (Schumer et al. 2015). Parameters were specified based on available information from these hybrid populations. The expected number of recombination events (recRate) was set to 400 based on a prior expectation of 1 recombination event per meiosis, 55 generations of admixture (Schumer et al. 2014) and detection probability of ~30%, given admixture proportions (Gravel 2012). The recombination tuning parameter (rfac) was set to the default value of 1. The ancestry priors were set to par1=0.0625, par1par2=0.375 and par2=0.5625 based on previously reported genome-wide admixture proportions (Schumer et al. 2014).

We defined a recombination event interval as the interval over which the posterior probability changed from ≥0.95 one ancestry state to ≥0.95 another ancestry state. This resulted in 131,282 inferred crossovers across the 24 linkage groups. This number is approximately concordant with expectations based on simulations, given this number of generations of admixture (Cui et al. 2016). The median resolution of these events was 44 kb, with 17% of events within 20 kb.

To evaluate the relationship between recombination frequency and genomic elements such as the TSS, CpG islands, and 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 number 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 it 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 transcriptional start site, 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 CpG islands, we used the annotations available from the UCSC genome browser beta site. 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 3D**). 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). 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 3**). 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 CpG islands 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. Indeed, we found a similar pattern in simulations (**Supplementary Figure S10**).

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 predicted obtained for the human PRDM9A allele, using recombination rates estimated from the CEU LD map in 10 kb windows (Frazer et al. 2007; downloaded from: http://www.well.ox.ac.uk/~anjali/AAmap/). We repeated this analysis for *Gorilla gorilla*

for the gor-1 PRDM9 allele, using recombination rates estimated from a recent LD map in 10 kb windows (Schwartz et al. 2014; Stevison et al. 2016; downloaded from: https://github.com/lstevison/great-ape-recombination).

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, 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

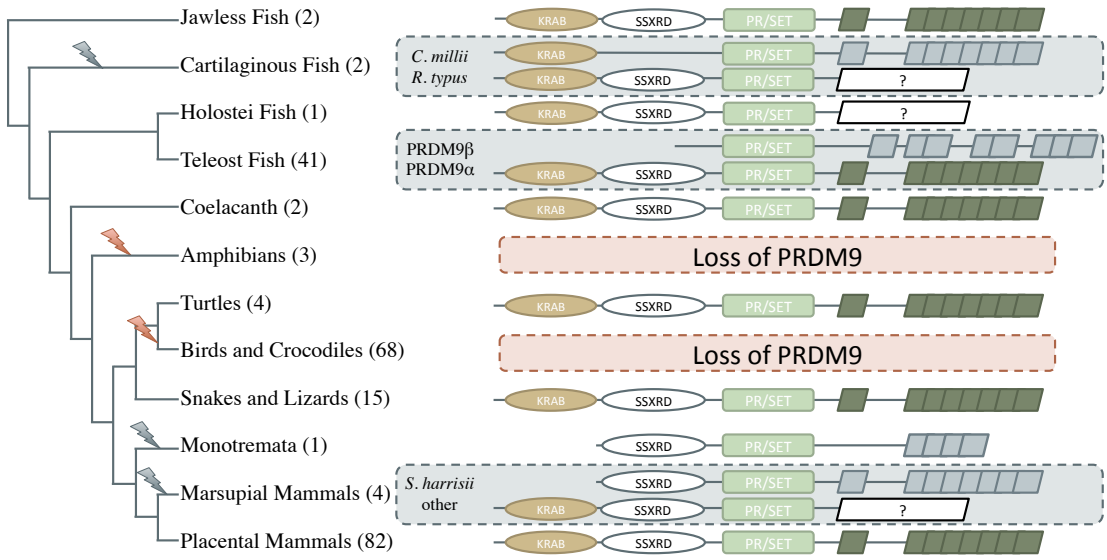


Figure 1. Phylogenetic distribution and functional domains of PRDM9 orthologs in vertebrates. Shown are the four domains of the PRDM9 gene: KRAB (in tan), SSXRD (in white), PR/SET (in green) and ZF (in gray/dark green; the approximate structure of identified ZFs is also shown). Complete losses are indicated on the phylogeny by red lightning bolts and partial losses by gray lightning bolts. ZF arrays shown in dark green are those taxa in which species show evidence of rapid evolution. White rectangles indicate cases where we could not determine whether the ZF was present. In clades where the most complete ortholog of PRDM9 varied across species both forms are shown. Note: though the monotremata ZF is shaded gray it was not included in analysis of rapid evolution due to a small number of ZFs.

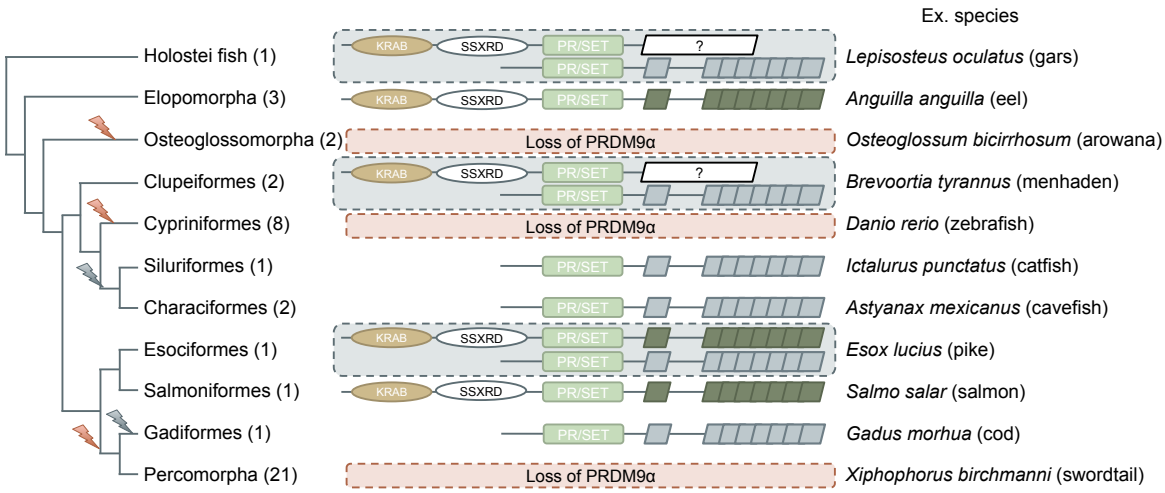


Figure 2. Phylogenetic distribution and functional domains of PRDM9α orthologs in teleost fish and in holostean fish that are outgroups to the PRDM9α/PRDM9β duplication event. The number of unique species included from each clade in our analysis of RefSeq, whole genome, and transcriptome datasets is shown in parenthesis. Domains are: KRAB (in tan), SSXRD (in white), PR/SET (in green) and ZF (in gray or dark green). Complete losses are indicated on the phylogeny by red lightning bolts and partial losses in clades (without other complete PRDM9 orthologs) by gray lightning bolts. White rectangles indicate cases where we could not determine whether the ZF was present. While many clades 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 PRDM9 orthologs 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 PRDM9α ortholog.

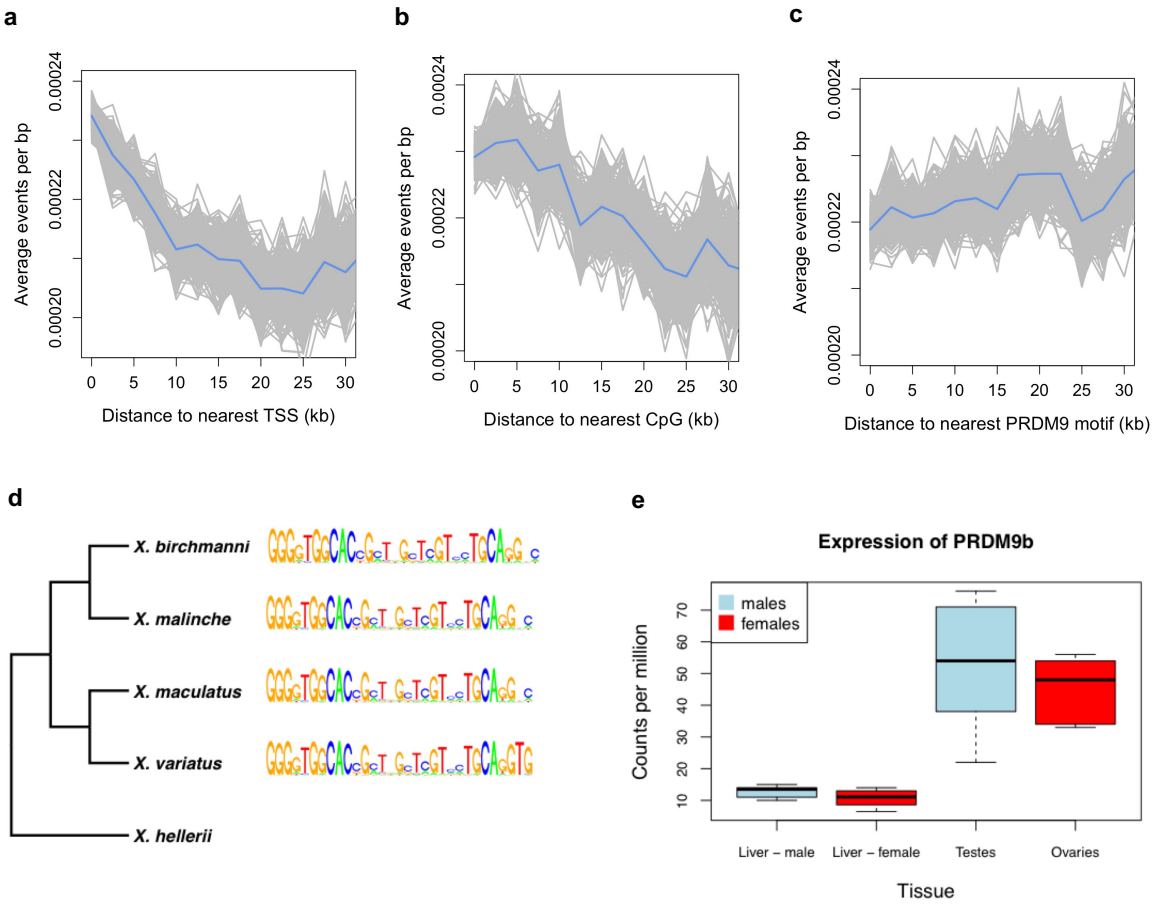


Figure 3. **Patterns of recombination and PRDM9 evolution in swordtail fish**. Recombination rates were estimated from ancestry switchpoints between naturally occurring hybrids of two species (see **Methods**). Recombination events are concentrated at the TSS and CpG islands (panels A and B) but not at computationally predicted PRDM9 binding sites (panel C; see **Supplementary Figure S7** for comparison). Simulations show that the slight dip in recombination frequency near PRDM9 binding sites could occur by chance, due to base composition differences (**Supplementary Figure S10**, **Methods**). The ZF array of PRDM9 appears to be evolving slowly in *Xiphophorus* (panel D), with few changes over >1 million years of divergence (Cui et al. 2013; Jones et al. 2013). Despite its apparent lack of involvement in specifying recombination hotspots, PRDM9 is upregulated in the germline relative to the liver (panel E).

Tables

Table 1. Evolution of the ZF for PRDM9 orthologs with different domain architectures. PRDM9 orthologs for which an empirical comparison dataset is available are ordered by one of three domain structures: KRAB-SSXRD-SET, SSXRD-SET, or only SET. A row is shaded 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. 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 meiotic recombination phenotypes.

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	bony fish	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	bony fish	KRAB-SSXRD-SET	0.538	9	4	510
Salmo salar	bony fish	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	bony fish	SET	0.258	60	2	158
Astyanax mexicanus	bony fish	SET	0.167	152	2	158
Clupea harengus	bony fish	SET	0.279	6	4	118
Clupea harengus	bony fish	SET	0.278	7	4	118
Clupea harengus	bony fish	SET	0.274	10	4	118
Clupea harengus	bony fish	SET	0.158	114	4	118
Cynoglossus semilaevis	bony fish	SET	0.182	80	1	107
Danio rerio	bony fish	SET	0.179	345	1	367
Esox lucius	bony fish	SET	0.295	32	4	234
Esox lucius	bony fish	SET	0.192	176	4	234
Esox lucius	bony fish	SET	0.192	177	4	234
Fundulus heteroclitus	bony fish	SET	0.189	158	1	206
Haplochromis burtoni	bony fish	SET	0.180	148	1	168
Ictalurus punctatus	bony fish	SET	0.320	14	8	140
Ictalurus punctatus	bony fish	SET	0.319	15	8	140
Ictalurus punctatus	bony fish	SET	0.306	24	8	140
Ictalurus punctatus	bony fish	SET	0.303	25	8	140
Ictalurus punctatus	bony fish	SET	0.286	33	8	140
Ictalurus punctatus	bony fish	SET	0.276	39	8	140
Ictalurus punctatus	bony fish	SET	0.253	55	8	140
Ictalurus punctatus	bony fish	SET	0.179	127	8	140
Larimichthys crocea	bony fish	SET	0.192	70	1	115
Lepisosteus oculatus	bony fish	SET	0.223	48	1	106
Maylandia zebra	bony fish	SET	0.173	161	1	176
Neolamprologus brichardi	bony fish	SET	0.173	141	1	152
Nothobranchius furzeri	bony fish	SET	0.180	245	1	266
Notothenia coriiceps	bony fish	SET	0.167	83	1	87
Oreochromis niloticus	bony fish	SET	0.173	173	1	190
Oryzias latipes	bony fish	SET	0.213	104	1	191
Otolemur garnettii	placental	SET	0.266	121	1	285
Poecilia formosa	bony fish	SET	0.191	184	1	242
Poecilia latipinna	bony fish	SET	0.191	175	1	235
Poecilia mexicana	bony fish	SET	0.191	187	1	244
Poecilia reticulata	bony fish	SET	0.191	162	1	212
Pundamilia nyererei	bony fish	SET	0.173	134	1	147
Pygocentrus nattereri	bony fish	SET	0.331	12	2	142
Pygocentrus nattereri	bony fish	SET	0.179	124	2	142
Salmo salar	bony fish	SET	0.188	411	4	510
Salmo salar	bony fish	SET	0.180	454	4	510
Sinocyclocheilus anshuiensis	bony fish	SET	0.185	224	2	284
Sinocyclocheilus anshuiensis	bony fish	SET	0.185	225	2	284
Sinocyclocheilus grahami	bony fish	SET	0.185	211	1	271
Sinocyclocheilus rhinocerous	bony fish	SET	0.185	208	2	269
Sinocyclocheilus rhinocerous	bony fish	SET	0.185	209	2	269
Takifugu rubripes	bony fish	SET	0.188	66	1	98
Xiphophorus maculatus	bony fish	SET	0.191	117	1	158

Supplementary Tables

Supplementary Tables 1, 2, 5, and 7-9 are provided as excel documents.

Supplementary Table S1. PRDM9 orthologs identified in RefSeq and whole genome databases.

Supplementary Table S2. 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 statistics are reported here.

Supplementary Table S5. 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 Table S6. 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 Table S8. Amino acid diversity levels of PRDM9 ZF arrays and the proportion localized to known DNA-binding residues. Columns labeled V1-V28 indicate the proportion 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 included in our analysis.

Supplementary Table S9. Likelihood ratio of neutral versus not non-neutral evolution along the SET domain of mammalian PRDM9 orthologs lacking a KRAB or SSXRD domain annotated in RefSeq, and whether another annotated ortholog exists with a KRAB domain.

1 **Supplementary Table S3.** Accession numbers and assembly descriptions of publicly
2 available testes RNAseq samples used for de novo assembly and assessment of PRDM9
3 expression.

Species (common name)	**SRA Accession**	**Number of sequences assembled**	**Assembly N50 (bp)**
Ambystoma mexicanum (Mexican salamander)	SRR2885289	65,430	1,646
Anguilla anguilla (European eel)	SRR1532765	204,457	2,305
Astyanax mexicanus (Mexican tetra)	SRR2045409	132,268	2,648
Brevoortia tyrannus (Atlantic Menhaden)	SRR1820331	208,748	1,359
Callorhinchus milii (elephant shark)	SRR513757	265,956	812
Cynops pyrrhogaster (Japanese fire belly newt)	SRR1553344	163,230	1,926
Danio rerio (zebrafish)	SRR1524249	194,569	2,354
Esox lucius (pike)	SRR1533661	137,688	2,724
Gadus morhua (cod)	SRR2045424	78,280	1,966
*Ictalurus punctatus** (Channel catfish)	SRR653303	275,791	856
Latimeria menadoensis (Indonesian coelacanth)	SRR576101	94,325	984
Lepisosteus oculatus (spotted gar)	SRR1524260	144,631	3,076
Macropus eugenii (Tammar wallaby)	DRR013676	110,948	1,592
Monodelphis domestica (gray short-tailed opossum)	SRR649379; SRR649380	178,722	699
Mus musculus (mouse)	SRR306775	49,230	1,316
Ornithorhynchus anatinus (platypus)	SRR649383; SRR649384	136,099	880
Petromyzon marinus (sea lamprey)	SRR369904	124,994	1048
Phascolarctos cinereus (Koala)	SRR1207975	481,174	1,227
Pogona vitticeps (bearded dragon)	ERR413070	147,543	1,279
*Salmo salar** *(Salmon)*	SRR1422872	237,736	1,699
Salmo trutta (brown trout)	SRR1532786	301,091	2,606
Xenopus tropicalis (western clawed frog)	SRR649399; SRR943353	116,163; 103,312	795; 2,304
Xiphophorus birchmanni (lowland swordtail)	This study (SRRXXXX)	64,607	2,423

4 * For assembly, channel catfish data and salmon data were downsampled to 15% and 20% of the
5 original number of reads, respectively.
6
7

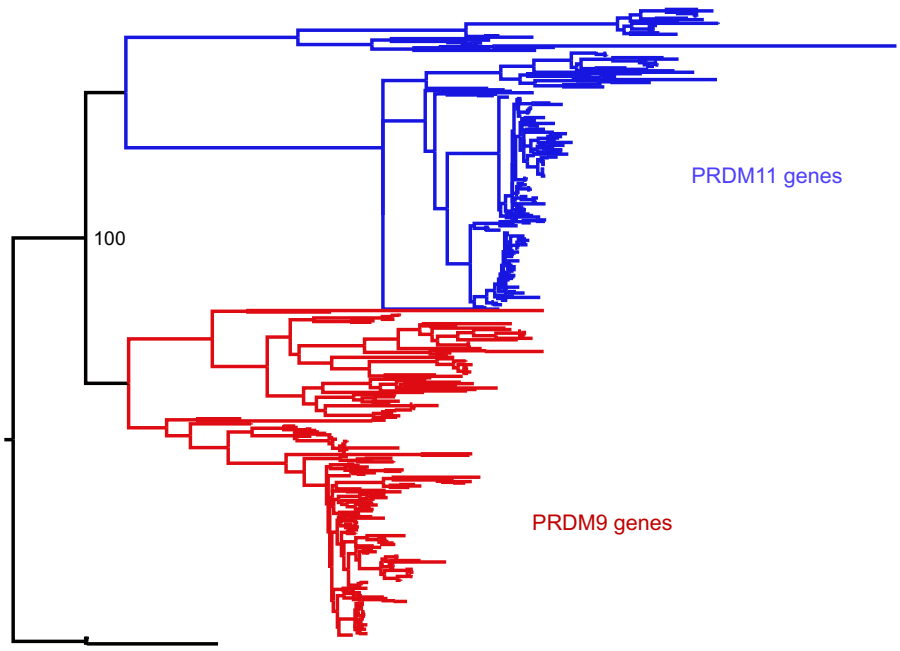
Supplementary Table S4. Summary of expression results of PRDM9 in the testis in representative species from major clades. Only species that passed the core recombination protein quality test (see Methods, **Supplementary Table S5**) are included in this table, with the exception of cases, indicated with asterisks, where PRDM9 was detected but one or more conserved recombination proteins were not.

Species	**Major group**	**PRDM9 expression detected (Y/N)**
Ambystoma mexicanum (Mexican salamander)	Amphibian	N
Anguilla anguilla (European eel)	Teleost fish – with PRDM9α and PRDM9β	Y (PRDM9α)
Astyanax mexicanus (Mexican tetra)	Teleost fish – with PRDM9α and PRDM9β	Y (both)
Brevoortia tyrannus (Atlantic menhaden)	Teleost fish – with PRDM9α and PRDM9β	Y (PRDM9α)
Callorhinchus milii (Elephant shark)	Cartilaginous fish	N
Cynops pyrrhogaster (Japanese fire belly newt)	Amphibian	N
Danio rerio (zebrafish)	Teleost fish – with PRDM9β	Y
Esox lucius (pike)*	Teleost fish – with PRDM9α and PRDM9β	Y (multiple)
Gadus morhua (cod)	Teleost fish – with PRDM9α and PRDM9β	Y (multiple)
Ictalurus punctatus (Channel catfish)	Teleost fish – with PRDM9α and PRDM9β	Y (PRDM9α)
Latimeria menadoensis (Indonesian coelacanth)	Coelacanth	Y
Lepisosteus oculatus (spotted gar)	Holostei fish	Y (multiple)
Macropus eugenii (Tamar wallaby)	Marsupial	Y
Mus musculus (mouse)	Mammal	Y
Monodelphis domestica (opossum)	Marsupial	Y
Ornithorhynchus anatinus (platypus)	Monotreme	N
Osteoglossum bicirrhosum (silver arowana)	Teleost fish – with PRDM9β	Y
Petromyzon marinus (sea lamprey)*	Jawless fish	Y
Phascolarctos cinereus (Koala)	Marsupial	Y
Pogona vitticeps (bearded dragon)	Lizard	Y
Salmo salar (Atlantic salmon)	Teleost fish – with PRDM9α and PRDM9β	Y (multiple)
Salmo trutta (brown trout)	Teleost fish – with PRDM9α and PRDM9β	Y (both)
Xenopus tropicalis (western clawed frog)	Amphibian	N
Xiphophorus birchmanni (lowland swordtail)	Teleost fish – with PRDM9β	Y

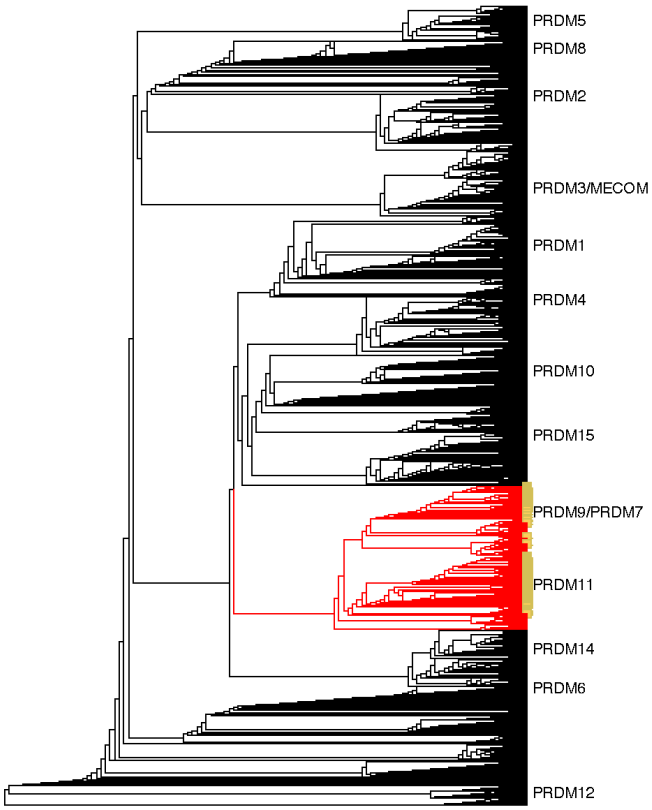
Supplementary Table S7. Rates of amino acid evolution in SET domains of representative PRDM9 orthologs lacking other functional domains. To determine whether PRDM9 orthologs lacking functional domains are non-functional, we compared rates of evolution between each PRDM9 ortholog missing a domain and 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.

Focal species (common name)	**Aligned species**	**Number of aligned base pairs**	**dN/dS**	**Likelihood difference (p-value)**
Ornithorhynchus anatinus (platypus)	Human	348	0.14	23 (2.5e-11)
Sarcophilus harrisii (Tasmanian devil)	Human	354	0.06	46 (<1e-10)
Xiphophorus maculatus PRDM9β (southern platyfish)	*Astyanax mexicanus* PRDM9β	270	0.17	15 (6.3e-8)
Astyanax mexicanus PRDM9α (Mexican tetra)	*Salmo salar* PRDM9α	348	0.0039	32 (1.1e-15)
Callorhinchus milii (Australian ghost shark)	*Rhincodon typus* (whale shark)	312	0.11	21 (1e-10)
Gadus morhua PRDM9 (Atlantic cod)	*Salmo salar* PRDM9α	327	0.15	14 (3.9e-4)

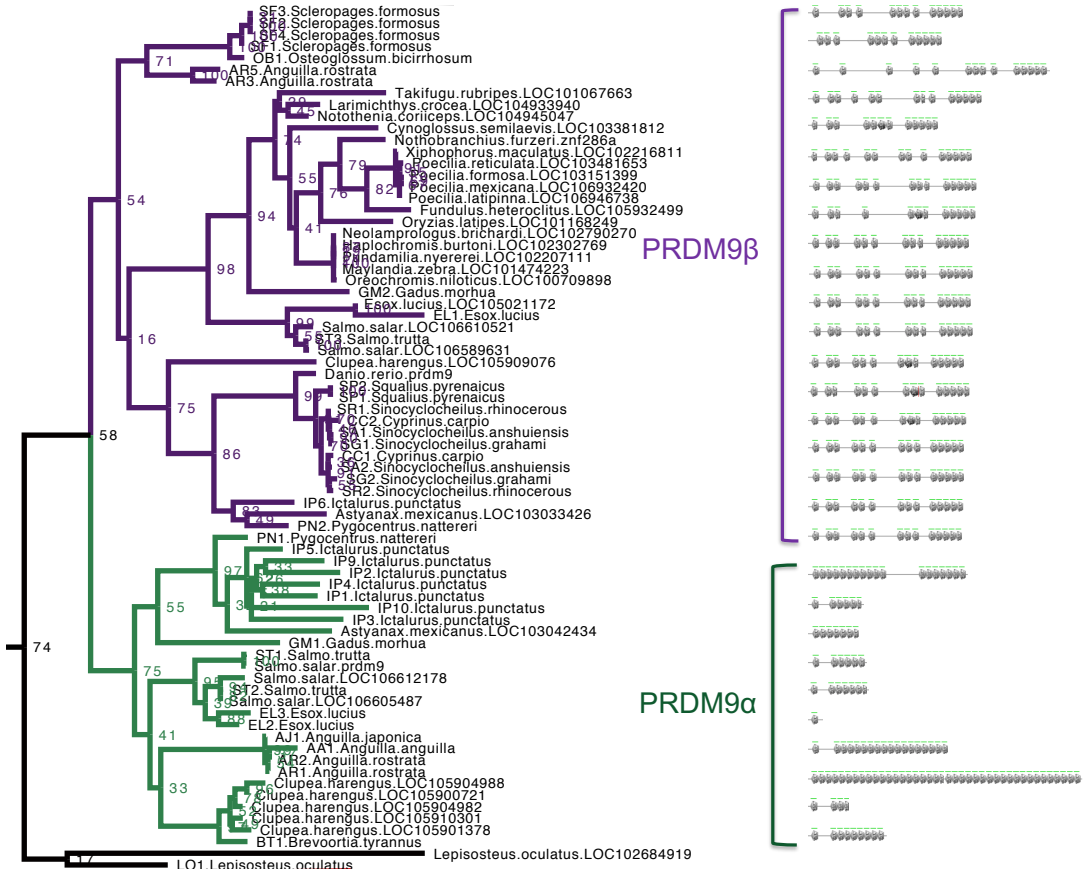
Supplementary Figures



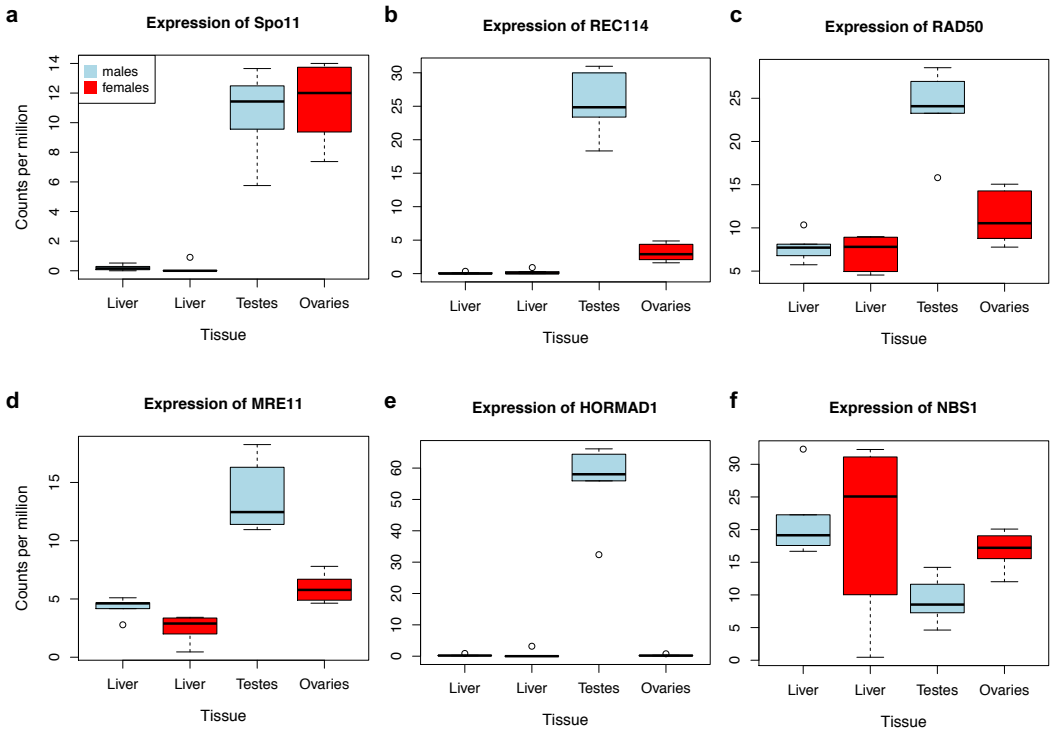
Supplementary Figure S1. Phylogenetic approach to identifying PRDM9 orthologs and related gene families. A maximum likelihood phylogeny built with RAxML using an alignment of SET domains distinguishes between genes that cluster with mammalian PRDM9 and PRDM11 with 100% bootstrap support.



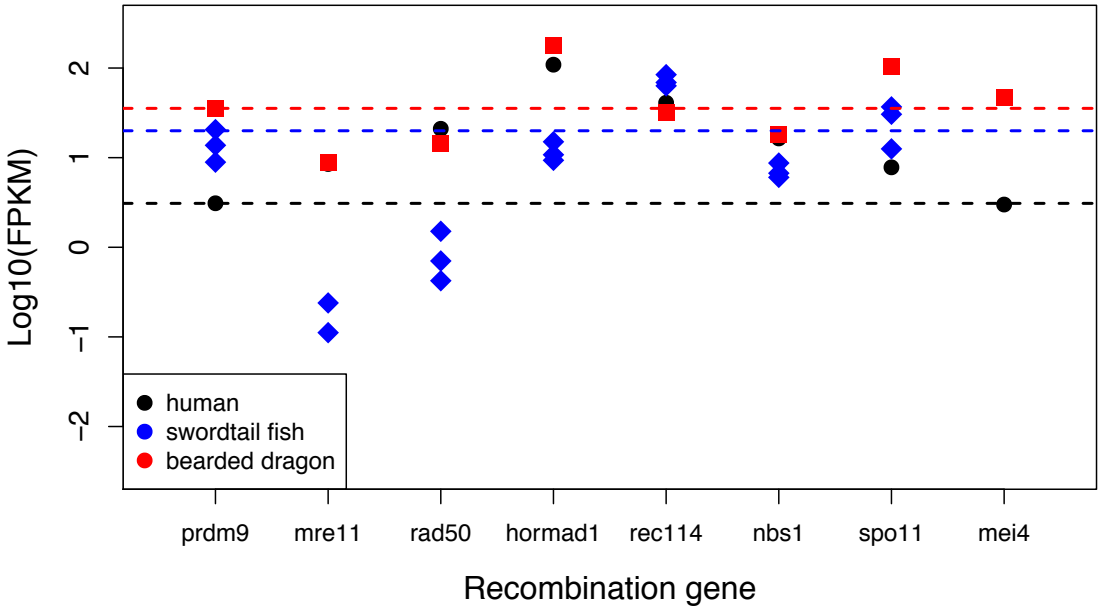
Supplementary Figure S2. Neighbor joining guide tree based on the SET domain. A NJ guide tree analysis on SET domains identified in our RefSeq, whole genome assembly, and transcriptome datasets was used as an initial step to identify sequences clustering with human PRDM9/7 or PRDM11. These sequences (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 were excluded from further analysis (in black).



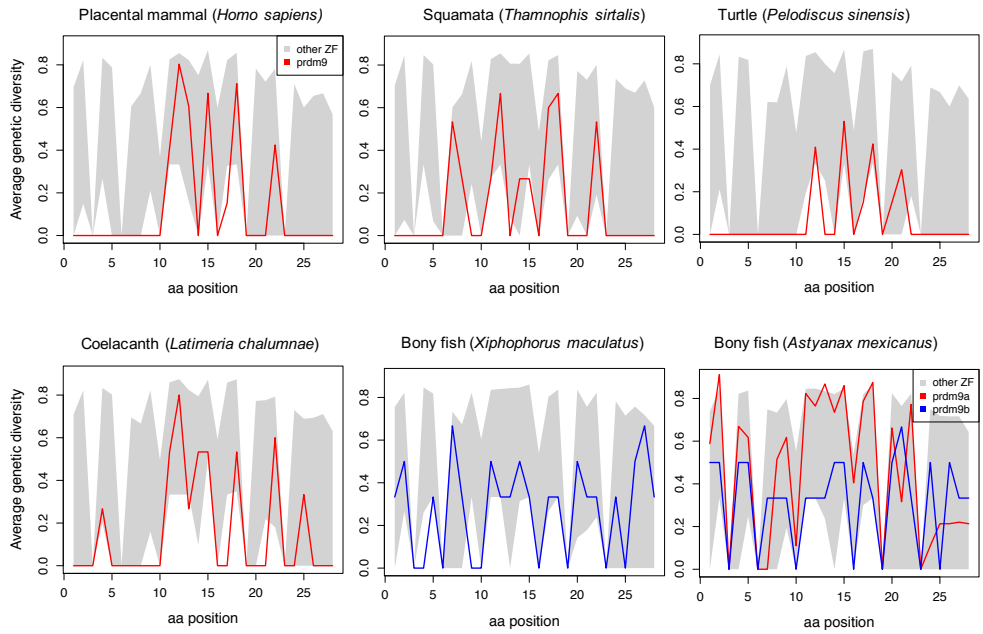
Supplementary Figure S3. Section of maximum-likelihood phylogeny of the SET domain showing bony fish PRDM9 orthologs α and β. These duplicates are distinguished with 58% bootstrap support in this phylogeny. The ZF domains for representative PRDM9 orthologs of each type are shown to the right, with each gray pentagon indicating the location of a ZF.



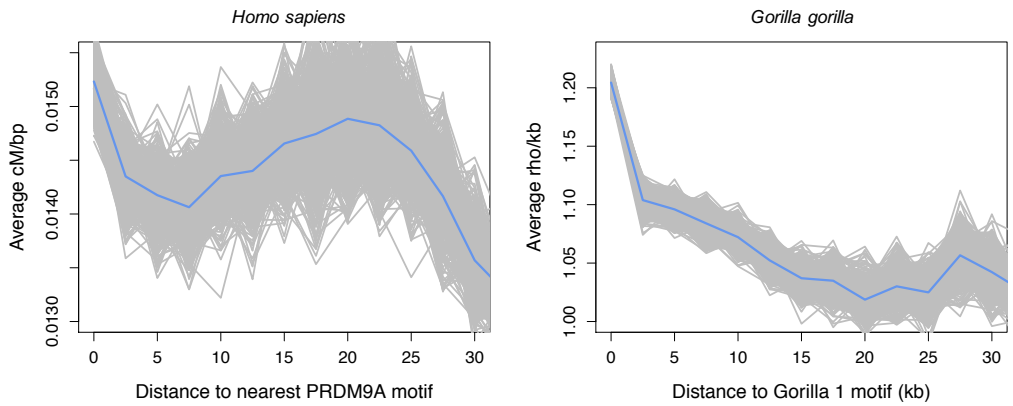
Supplementary Figure S4. Expression levels of meiosis-related genes in swordtail fish tissues. In general, the seven meiosis-related genes surveyed had higher expression in tissues containing germline cells than liver tissue, but this pattern was much more pronounced in testis tissue (compared to ovary tissue). As a result, we focused our analysis of meiosis related genes on RNAseq data generated from testis.



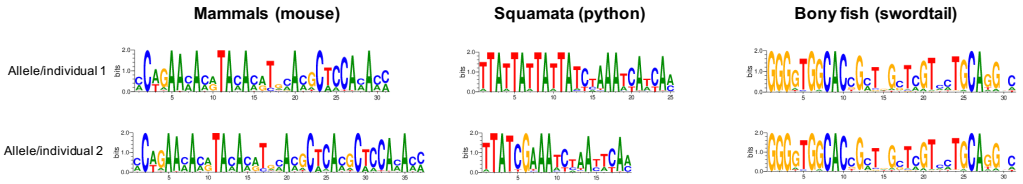
Supplementary Figure S5. Expression levels of genes with a known role in meiotic recombination in testes tissues of three exemplar species. FPKM per individual is plotted for swordtail (*X. malinche*) and bearded dragon for each transcript. For humans, the point represents the average expression of 122 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.



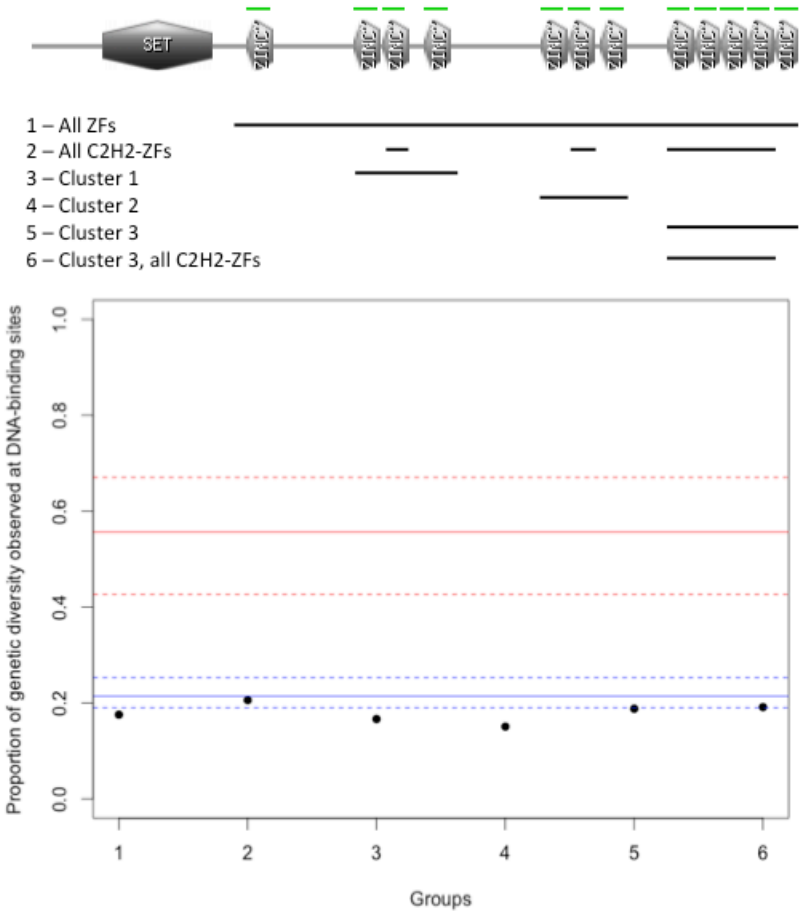
Supplementary Figure S6. Patterns of amino acid diversity as a function of amino acid position in the ZF alignment. Each plot shows the 95% range of diversity levels at that site for all C2H2 ZFs from a species of that taxon (gray); the values at PRDM9 are show in red or blue. Turtles, snakes and coelacanth 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). In bony fish, this pattern is not observed and PRDM9 ZF diversity is more typical of other C2H2 ZFs.



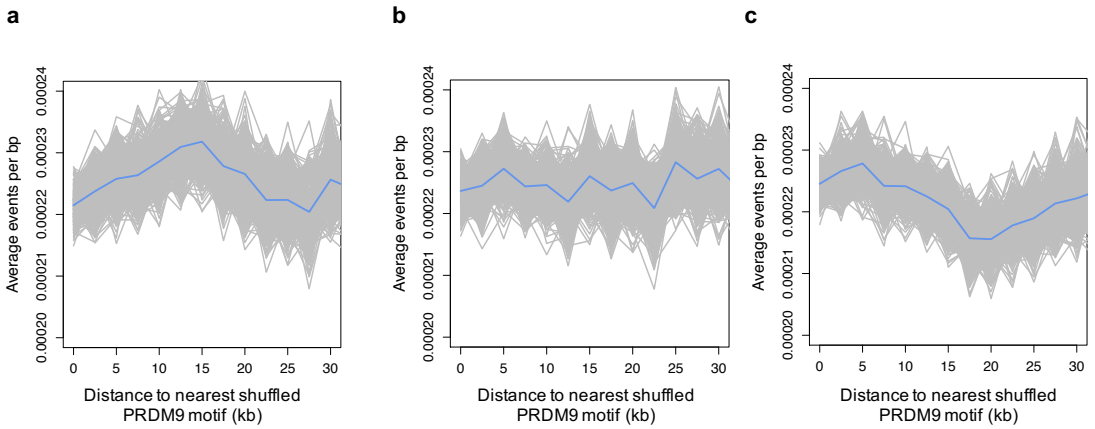
Supplementary Figure S7. Recombination rates show a peak near the computationally predicted PRDM9A binding motif in humans and gor-1 allele in gorillas. Most work investigating relationships between PRDM9 motifs and recombination rates have focused on the PRDM9 motif empirically inferred from recombination hotspots, but the empirical motif is unknown for species where PRDM9 is not associated with hotspots or hotspots are not sufficiently well resolved. To generate results comparable to those we present for swordtails in **Figure 3C**, we therefore determined recombination rate in humans (using the CEU LD map; Frazer et al. 2007) as a function of distance to computationally predicted binding sites for the PRDM9A motif and in gorillas (using the LD map from Stevison et al. 2016) as a function of distance to computationally predicted binding sites for the gor-1 PRDM9 allele (Schwartz et al. 2014).



Supplementary Figure S8. 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**), versus two species of swordtail fish (*X. birchmanni* and *X. malinche*; genome sequences). The position weight matrix was obtained by online prediction tools available at http://zf.princeton.edu.



Supplementary Figure S9. Analysis of ZF evolution in PRDM9β. Red lines show the median (solid) and first and third quantiles (dashed lines) for 48 complete PRDM9 orthologs identified across vertebrates. 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.



Supplementary Figure S10. 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 3**) is not unexpected, given the base composition of the PRDM9 binding motif in swordtails.

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 Research 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.

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.

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.

Cui, R. F., M. Schumer, and G. G. Rosenthal. 2016. Admix'em: a flexible framework for forward-time simulations of hybrid populations with selection and mate choice. Bioinformatics 32:1103-1105.

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.

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.

Gravel, S. 2012. Population Genetics Models of Local Ancestry. Genetics 191:607-619.

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.

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.

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.

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.

Lange, J., S. Yamada, S. E. Tischfield, J. Pan, S. Kim, X. Zhu, N. D. Socci, M. Jasin, and S. Keeney. 2016. The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. Cell 167:695.

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. 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.

Lichten, M. and B. de Massy. 2011. The Impressionistic Landscape of Meiotic Recombination. Cell 147:267-270.

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.

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 cervisiae*. 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.

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.

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.

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.

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. Nature Genetics 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. Cui, G. Rosenthal, and P. Andolfatto. 2015. simMSG: an experimental design tool for high-throughput genotyping of hybrids. Molecular Ecology Resources doi: 10.1111/1755-0998.12434.

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.

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