PRDM9 losses in vertebrates are coupled to the loss of at least three other meiotic genes

Maria Izabel A. Cavassim[1,2,*,+,&], Zachary Baker[2,*,^,&], Carla Hoge[2], Mikkel H. Schierup[1], Molly Schumer[4], and Molly Przeworski[2,3]

Author affiliations:

[1]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[2]Department of Biological Sciences, Columbia University, New York City, United States
[3]Department of Systems Biology, Columbia University, New York City, United States
[4]Department of Biology, Stanford University, Stanford, United States
[*] Joint first authors
[+] Present address: Department of Ecology and Evolutionary Biology, University of California, Los Angeles, United States
[^] Present address: Department of Genetics, University of Cambridge, Cambridge, United Kingdom
[&] to whom correspondence should be addressed: izabelcavassim@gmail.com, zb67@cam.ac.uk

Abstract

In most mammals and likely throughout vertebrates, the gene *PRDM9* specifies the locations of meiotic double strand breaks; in mice and humans at least, it also aids in their repair. For both roles, many of the molecular partners remain unknown. Here, we take a phylogenetic approach to identify genes that may be interacting with PRDM9, by leveraging the fact that *PRDM9* arose before the origin of vertebrates, but was lost many times, either partially or entirely––and with it, its role in recombination. As a first step, we characterize PRDM9 domain composition across 379 vertebrate species, inferring at least eight independent losses. We then use the interdigitation of *PRDM9* orthologs across vertebrates to test whether it co-evolved with any of 241 candidate genes co-expressed with PRDM9 in mice or associated with recombination phenotypes in mammals. Accounting for the phylogenetic relationship among species, we identify three genes whose presence and absence is unexpectedly coincident with that of *PRDM9*: *ZCWPW1*, which was recently shown to be recruited to sites of PRDM9 binding and to facilitate double strand break repair; *TEX15*, which has also been suggested to play a role in repair; and *ZCWPW2*, the paralog of *ZCWPW1*. *ZCWPW2* is expected to be recruited to sites of PRDM9 binding as well; its tight coevolution with *PRDM9* across vertebrates suggests that it is a key interactor, with a role either in recruiting the recombination machinery or in double strand break repair.

Author Summary

Our understanding of meiotic recombination in mammals has seen great progress over the past 15 years, spurred in part by the convergence of lines of evidence from molecular biology, statistical genetics and evolutionary biology. We now know that in most mammals and likely in many vertebrates, the gene *PRDM9* specifies the location of meiotic double strand breaks and that in mice and humans at least, it also aids in their repair. For both roles, however, many of the molecular partners remain unknown. To search for these, we take a phylogenetic approach, leveraging the fact that the complete *PRDM9* has been lost at least eight times in vertebrates and thus that its presence is interdigitated across species. By this approach, we identify three genes whose presence or absence across vertebrates is significantly coupled to the presence or absence of *PRDM9*: *TEX15*, *ZCWPW1* and the paralog of *ZCWPW1*, *ZCWPW2*. ZCWPW1 was recently shown to be recruited to sites of PRDM9 binding and to aid in the repair of double strand breaks. TEX15 has also been suggested to be involved in the repair process. ZCWPW2 is likely recruited to sites of PRDM9 binding; its tight coevolution with *PRDM9* across vertebrates suggests that it plays an important role either in double strand break formation, potentially as the missing link that recruits the recombination machinery to sites of PRDM9 binding, or in double strand break repair.

Introduction

Meiotic recombination is initiated by the deliberate infliction of numerous double strand breaks (DSBs) in the genome, the repair of which yields crossover and non-crossover resolutions (reviewed in (Keeney and Neale 2006)). In mice and humans, and probably in most mammals, the localization of almost all DSBs is specified through the binding of PRDM9 (F. Baudat et al. 2010; Myers et al. 2010; Parvanov, Petkov, and Paigen 2010). Yet the presence of a PRDM9 binding site is far from sufficient for a DSB to be made: a number of additional factors modulate whether PRDM9 binds or act downstream of PRDM9 binding (Altemose et al. 2017; Spence and Song 2019; Jin, Fudenberg, and Pollard, n.d.).

The mechanism by which PRDM9 directs recombination to the genome is partially understood: it binds DNA through a C2H2 zinc finger (ZF) array and contains a SET domain that tri-methylates histones H3K4 and H3K36 (Eram et al. 2014; Powers et al. 2016). These epigenetic marks together recruit the DSB machinery, notably SPO11 (which makes the DSBs), through intermediates that remain unknown (Diagouraga et al. 2018). In addition to the zinc finger binding

array and SET domain, most mammalian *PRDM9* genes also have two other domains, KRAB and SSXRD, whose functions are unclear.

The complete PRDM9 protein, with all four domains, originated before the diversification of vertebrates, so has been conserved for hundreds of millions of years (Baker et al. 2017; Oliver et al. 2009). Yet the entire gene has also been lost numerous times, including in birds and canids (Axelsson et al. 2012; Muñoz-Fuentes, Di Rienzo, and Vilà 2011; Singhal et al. 2015). In these species, recombination occurs preferentially around promoter-like features, notably CpG islands (Ponting 2011; Auton et al. 2013; Baker et al. 2017; Singhal et al. 2015). A possible explanation is that in the absence of the histone marks laid down by PRDM9, the recombination machinery defaults to those residual H3K4me3 marks found in the genome, often associated with sites of transcription initiation, or perhaps simply to wherever DNA is accessible (Brick et al. 2012; Singhal et al. 2015). The same concentration of DSBs around promoter-like features is seen in *Prdm9*$^{-/-}$ mice (Brick et al. 2012) and in a woman who carries two loss of function copies of *PRDM9* identical by descent (Narasimhan et al. 2016). These findings suggest that mammals that carry an intact *PRDM9* retain the mechanism to direct recombination employed by species lacking *PRDM9*, but it is normally outcompeted by PRDM9 binding.

In addition to complete losses of *PRDM9*, multiple partial losses have occurred independently (e.g., in platypus and various fish lineages), usually involving the truncation of the N-terminal KRAB and SSXRD domains (Baker et al. 2017). Although these partial *PRDM9* orthologs evolve under selective constraint and thus must have some conserved function (Baker et al. 2017), several lines of evidence indicate that they do not direct recombination. For one, only in species with a complete *PRDM9* is the ZF unusually rapidly evolving in its binding affinity (Baker et al. 2017). Since the rapid evolution of the ZF is thought to arise from the role of PRDM9 in recombination (Úbeda and Wilkins 2011; Myers et al. 2010; Latrille, Duret, and Lartillot 2017), this evolutionary pattern suggests that all four domains are required for DSB localization. Empirical data support this notion: in swordtail fish carrying one *PRDM9* ortholog that lacks KRAB and SSXRD domains as well as in a mouse model in which only the KRAB domain is knocked out, recombination events are concentrated at promoter-like features, as in species lacking *PRDM9* altogether (Thibault-Sennett et al. 2018; Baker et al. 2017). Therefore, the KRAB domain at least appears to be necessary for PRDM9 to direct recombination, likely by mediating interactions with other proteins (Imai et al. 2017; Thibault-Sennett et al. 2018).

Conversely, the presence of a complete *PRDM9* with a rapidly evolving ZF outside of mammals (Baker et al. 2017) suggests that PRDM9 also directs recombination to the genome in these species, as has been reported for rattlesnakes (Schield et al. 2020). Thus, at least two

mechanisms for directing meiotic recombination are interdigitated within mammals as well as seemingly throughout the vertebrate phylogeny.

In addition to specifying the locations of DSBs, PRDM9 has recently been discovered to play a second role, in the downstream repair of DSBs (Huang et al. 2020; Mahgoub et al. 2020; Wells et al. 2020). In mice and humans, DSBs at which PRDM9 is bound on both homologs are more likely to be efficiently repaired and to result in a crossover; in contrast, DSBs at which PRDM9 is only bound on one of the two homologs are delayed in their repair (Hinch et al. 2019; Wells et al. 2020). If these "asymmetric" DSBs are overwhelming in number—as is the case in certain hybrid crosses in mice—this delay can lead to asynapsis and infertility (Davies et al. 2016; Gregorova et al. 2018).

While this second role of PRDM9 is still poorly understood, recent papers report that it is facilitated by ZCWPW1, which binds H3K4me3 and H3K36me3 (Huang et al. 2020; Mahgoub et al. 2020; Wells et al. 2020) and is co-expressed with PRDM9 in single cell data from mouse testes (Jung et al. 2019). One line of evidence that enabled the discovery of *ZCWPW1* is that although it too has been lost numerous times in vertebrates, it is found in seven clades that carry an intact *PRDM9* (Mahgoub et al. 2020; Wells et al. 2020).

The important hint provided by the phylogenetic distribution of *ZCWPW1* points to the potential power of co-evolutionary tests to identify additional molecular partners of PRDM9. Here, we took this approach more systematically: we considered a set of 241 candidate genes that are either known to be involved in recombination in model organisms (Frédéric Baudat, Imai, and de Massy 2013), associated with recombination phenotypes in a human genome-wide association study (Halldorsson et al. 2019), or co-expressed with PRDM9 in single cell data from mouse testes (Jung et al. 2019) and tested for their co-occurrence with *PRDM9* across 189 vertebrate species. After verifying our initial gene status calls in whole genome data and using RNA-seq data for a subset of species, we identified two genes in addition to *ZCWPW1* whose losses are unexpectedly coincident with that of an intact *PRDM9*: *TEX15* and the paralog of *ZCWPW1*, *ZCWPW2*.

Results

A revised phylogeny of PRDM9

We previously reported that the complete *PRDM9* gene, including KRAB, SSXRD and SET domains, arose before the origin of vertebrates and was lost independently a number of times, both in its entirety and partially (through the loss of its N-terminal domains; (Baker et al. 2017)). Here, we leverage the independent losses of *PRDM9* in order to identify genes that are

co-evolving with *PRDM9*—specifically, that tend to be present in the same species as *PRDM9* and lost (partially or entirely) when *PRDM9* is no longer complete.

As a first step, we characterized the phylogenetic distribution of *PRDM9* in light of new genome sequences published since our initial analysis (Baker et al. 2017). To this end, we created a curated dataset of 684 vertebrate *PRDM9* sequences by analyzing publicly available protein sequences from Refseq (O'Leary et al. 2016), whole genome sequences, and RNA-seq data from testes samples, as well as four RNA-seq datasets from testis that we generated (see Methods, **Figure S1**, **Tables S1-3**). For this analysis, we defined *PRDM9* orthologs as complete if they contain both KRAB and SET domains; we did not consider the SSXRD domain, because its short length makes its detection at a given e-value threshold unreliable, or the ZF array, because its repetitive structure makes it difficult to sequence and assemble reliably.

Across 379 species, we identified 186 species with at least one complete *PRDM9* ortholog and 193 species without a complete *PRDM9* ortholog (**Figure 1**, **Table S4**). Notably, we were able to uncover complete *PRDM9* orthologs in a number of species for which we had previously predicted partial or complete losses (Baker et al. 2017), including in marsupials (*Sarcophilus harrisii*), the atlantic cod (*Gadus morhua*), and the atlantic herring (*Clupea harengus*), as well as in a handful of placental mammals that we had previously only investigated using RefSeq (see **Table S4** for details). We also found a complete *PRDM9* ortholog in caecilians, suggesting that the previously reported loss of *PRDM9* in amphibians (Baker et al. 2017) is restricted to frogs and salamanders. We note, finally, that by the approach taken here, the *PRDM9* ortholog from the Australian ghostshark (*Callorhinchus milii*) is considered to be complete (in contrast to in Baker et al. 2017, where we also relied on the SSXRD domain; see **Table S4** for details).

Given the phylogenetic relationships among species given by the TimeTree tool (http://timetree.org/; (Kumar et al. 2017)), we inferred 23 putative complete or partial losses of *PRDM9* across the 379 vertebrates considered (**Figure 1**, **Table S4**). These putative losses include six previously reported ones (Axelsson et al. 2012; Muñoz-Fuentes, Di Rienzo, and Vilà 2011; Baker et al. 2017; Singhal et al. 2015), each observed in two or more closely related species: in percomorph fish, cypriniformes fish, characiformes and siluriformes fish, frogs, birds and crocodiles, and canids. In addition, independent work supports our finding of a partial loss of *PRDM9* in the platypus (*Ornithorhynchus anatinus*) (J. Hussin and P. Donnelly, personal communication). We also found evidence for a putative loss in lizards. To verify this loss, we collected RNA-seq data from testes in two distantly related lizard species (divergence time ~99 Mya): the green anole, *Anolis carolinensis*, for which we had failed to find *PRDM9* in the whole genome sequence, and the fence lizard, *Sceloporus undulatus*, for which neither a Refseq nor a

genome sequence were available at the time. Despite sufficient power to detect a set of six highly conserved meiotic genes in both species, we were unable to detect PRDM9 expression in either (**Figure S2-3**, **Table S3**), confirming that it is likely absent. The remaining 15 putative absences of *PRDM9* are observed in single species; we were unable to verify their absence using testis RNA-seq data, so their *PRDM9* status remains uncertain. Thus, in total, we identified at least 8 independent *PRDM9* losses in vertebrates, and possibly as many as 23 (**Figure 1**, **Table S4**).

All eight losses for which we could find strong support are relatively old (**Figure S4**). In particular, given the presence of a complete *PRDM9* ortholog in caecilians, the loss in frogs appears to have occurred after the common ancestors of amphibians 323 million years ago (Mya), but >204 Mya. By a similar reasoning and based on the presence of a complete *PRDM9* in bearded dragons (*Pogona vitticeps*), it appears that the loss of *PRDM9* in the green anole and fence lizard reflects an ancestral loss, basal to their common ancestor, which took place over 99 Mya but less than 157 Mya (**Figure S5**). The most recent loss of an intact *PRDM9* manifest in these data seems to be either the one that happened in the branch leading to platypus or the one in canids, which could be as recent as 14.2 Mya (**Figure S4**).

Identifying genes co-evolving with PRDM9

We selected candidates based on whether they were co-expressed with PRDM9 in single cell data from mouse testes (193 genes; see Methods (Jung et al. 2019)) (**Figure S6A-B**). To this set, we added any gene associated with variation in recombination phenotypes in humans (Halldorsson et al. 2019) as well as genes known to have a role in mammalian meiotic recombination from functional studies (summarized in (Frédéric Baudat, Imai, and de Massy 2013)). Together, these three sources provided a total of 241 genes to evaluate for possible co-evolution with *PRDM9* (**Table S5**, **Figure S6C**).

We evaluated the presence or absence of these 241 genes across the NCBI RefSeq database of 189 species. These 189 species were downsampled from the larger phylogenetic tree to preserve at most three species with high quality genomes below each *PRDM9* loss, thereby minimizing phylogenetic signals of genome quality (see Methods, **Figure S7**). We encoded a gene as present when it contained all the domains found in four representative vertebrates with a complete *PRDM9* and absent if it lacked one or more of those domains (see Methods). Many of the 241 genes are present in every sampled vertebrate and hence provide no information in our co-evolutionary test of presence and absence: specifically, we found apparently complete orthologs for 102 candidate genes in all 189 species used in the phylogenetic test. We therefore focused on the remaining 139 genes, each of which has been lost at least once among

vertebrate species evaluated here. This approach yielded a matrix of 189x139 gene status calls (presented in **Table S6**).

We tested for the co-evolution of *PRDM9* and each candidate gene by comparing a null model with independent rates of gains and losses of *PRDM9* and of the focal gene to an alternative in which the state transition rates of the two genes are dependent on one another, using the maximum likelihood approach within *BayestraitsV3* (Barker and Pagel 2005; Pagel 1994) (**Table S7**, **Figure S8**). By this approach, we identified five significant hits at the 5% level (uncorrected for multiple tests): in order of increasing p-values, *MEI1*, *ZCWPW1*, *TEX15*, *ZCWPW2*, and *ANKRD31* (**Table 1, Figure 2A**).

We sought to verify the phylogenetic distribution of the top genes by developing curated datasets of high confidence orthologs, as we had for *PRDM9* (see Methods; **Figure 3, Tables S1** and **S8, Figure S9**). In doing so, we were able to identify *MEI1* orthologs from the whole genome assemblies of each species missing *MEI1* in our initial dataset, resulting in the presence of *MEI1* in every species considered (**Table S1**); thus, it appears that its inferred co-evolution with *PRDM9* based on Refseq calls was artifactual (see Methods). Moreover, when we reran the phylogenetic test on the curated ortholog sets for the remaining four genes, *ANKDR31* was no longer significant at the 5% level (p=0.376), possibly because the curation uncovered an intact ortholog in birds and in anoles.

The other three genes *TEX15*, *ZCWPW1* and *ZCWPW2* remained significant, *TEX15* (p=0.006) and *ZCWPW2* ($p<10^{-4}$) more highly so than with our initial calls. These genes are also significant if we redefine a complete PRDM9 ortholog as containing not only the KRAB and SET domain but also the SSXRD domain (**Table S9**). Like *PRDM9*, all three genes were inferred to be present in the common ancestor of vertebrates, and all three are co-expressed with PRDM9 in single cell mouse testis data (**Figure 2B**; **Figure S6**).

ZCWPW1 and PRDM9 co-evolution

Our finding that *ZCWPW1* is co-evolving with *PRDM9* (p=0.034 in the curated set; **Table 1**) is in line with previous reports that report an association in vertebrates between the presence and absence of *ZCWPW1* and *PRDM9* orthologs (Mahgoub et al. 2020; Wells et al. 2020). Here, we find an even tighter coupling of *PRDM9* and *ZCWPW1* than previously documented. Specifically, we infer 13 losses of *ZCWPW1* among 189 species used in our phylogenetic test, distributed across 17 species that lack *ZCWPW1* entirely and three species carrying partial *ZCWPW1* genes (with the PWWP domain but not the zf-CW domain; **Table S9**).

Five of the *ZCWPW1* losses occur among the eight well supported losses of *PRDM9*: in cypriniformes fish, percomorph fish (*Euacanthomorphacea*), and birds, clades in which all three representative species analyzed lack *ZCWPW1*, as well as siluriformes fish and frogs, in which one of three species lacks *ZCWPW1*. An additional three *ZCWPW1* losses occur in species for which the status of *PRDM9* is uncertain: the denticle herring (*Denticeps clupeoides*), the Asian arowana (*Scleropages formusus*), and the reedfish (*Erpetoichthys calabaricus*). The remaining five losses of *ZCWPW1* seem to break the pattern, in that they occur in lineages containing a complete *PRDM9* gene. However, four are observed only in a single species and therefore may be spurious. Thus, across the tree, there is only one well supported case of a taxon with an intact *PRDM9* that has nonetheless lost *ZCWPW1*, supported by two closely related species, the tiger snake (*Notechis scutalus*) and the eastern brown snake (*Pseudonaja textilis*) (see **Table S9** for details).

In mice as well as human cell lines, ZCWPW1 binds two marks laid down by PRDM9: the zf-CW domain binds H3K4me3 and the PWWP domain H3K36me3 (Huang et al. 2020; Mahgoub et al. 2020; Wells et al. 2020). Thus, the co-evolution across vertebrates likely reflects a conserved molecular interaction between ZCWPW1 and PRDM9 as reader and writer of these dual histone modifications.

ZCWPW2 also co-evolves with PRDM9

Intriguingly, the strongest association with the presence or absence of *PRDM9* is that of the paralog of *ZCWPW1*, *ZCWPW2*. Among the 189 species, there are 11 independent losses, distributed across 21 species that appear to lack *ZCWPW2* altogether and three that contain partial *ZCWPW2* genes (two with the PWWP domain but not the zf-CW domain, and one with the reverse; **Table S9**).

Four of the *ZCWPW2* losses occur among the eight well supported losses of an intact PRDM9: in percomorph fish, frogs, birds, and the green anole. *ZCWPW2* is additionally absent in a clade encompassing cypriniformes fish and siluriformes fish, as well as the electric eel (*Electrophorus electricus*), which has an intact *PRDM9*. This phylogenetic distribution suggests that the loss of *ZCWPW2* may have occurred before the losses of *PRDM9* in both cypriniformes fish and siluriformes fish. Three additional absences of *ZCWPW2* are observed among the 15 species for which the status of *PRDM9* is uncertain: the denticle herring (*Denticeps clupeoides*), the reedfish (*Erpetoichthys calabaricus*), and Asian arowana (*Scleropages formusus*). In this last case, *ZCWPW2* is also absent from the closest evolutionary relative in the tree, the elephantfish (*Paramormyrops kingsleyae*), which carries *PRDM9*, again suggesting that the loss of *ZCWPW2*

may have preceded the loss of *PRDM9*. The remaining three cases of *ZCWPW2* loss are each observed in a single species carrying an intact PRDM9, without supporting lines of evidence. In summary, in the few cases with *PRDM9* but not *ZCWPW2*, we cannot verify the loss of *ZCWPW2*; conversely, most of the species without *PRDM9* also lack *ZCWPW2* (see **Table S9** for details).

Like its paralog, ZCWPW2 contains zf-CW and PWWP domains, predicted to bind H3K4me3 and H3K36me3, respectively (**Figure 4A, Figure S10**). As in ZCWPW1 ((Wells et al. 2020) and (Mahgoub et al. 2020)), these domains are highly conserved, especially at residues with predicted binding properties (**Figure 4B-C**), suggesting that ZCWPW2 is also recruited to sites of PRDM9 binding.

The distribution of TEX15 orthologs

We identify a third gene that is co-evolving with the presence of *PRDM9* across vertebrates, TEX15, which is expressed at multiple stages of spermatogenesis (F. Yang et al. 2008), and is co-expressed with PRDM9 in two components inferred from single cell data from mice, active during pre-leptotene and zygotene (**Figure S6**). Across the 189 species, *TEX15* is missing in three taxa lacking *PRDM9*: we found no ortholog of *TEX15* in birds and percomorphs and only a partial *TEX15* ortholog in the *Anolis carolinensis*. Interestingly, *TEX15* is also absent in the Atlantic cod (*Gadus morhua*), which is the immediate outgroup to percomorph fish (*Euacanthomorphacea*) in our tree but carries an intact *PRDM9*, raising the possibility that the loss of *TEX15* that led to its absence in percomorpha occurred before that of *PRDM9*. All other species have an intact *TEX15* (see **Table S9** for details).

Discussion

Roles of PRDM9 across vertebrates

By extending the reconstruction of *PRDM9* to 379 vertebrate species, we identified eight losses that are supported by more than one species or by independent evidence, and possibly as many as 23. Focusing on a subset of 189 species that capture all the state transitions of *PRDM9*, we tested whether *PRDM9* transitions coincide with those of 139 candidate genes lost at least once across vertebrates. After carefully vetting the ortholog calls for our top five signals, we identified three genes that are co-evolving in their presence and absence with *PRDM9*: *ZCWPW1*, its paralog *ZCWPW2* and *TEX15*.

We had previously hypothesized that PRDM9 plays a role in directing recombination not only in mammals but across vertebrates, based on the presence of an intact ortholog across vertebrates with a rapidly-evolving zinc finger (Baker et al. 2017). Consistent with our prediction,

there is tentative evidence for the influence of PRDM9 binding on recombination in rattlesnakes (Schield et al. 2020). That two genes with a known role in recombination, *ZCWPW1* and *TEX15*, co-evolve with *PRDM9* across vertebrates lends further support to this hypothesis.

The precise nature of the molecular interactions between PRDM9 and ZCWPW1 or TEX15 remains unknown, but recent evidence suggests that ZCWPW1 interacts with PRDM9 to facilitate the repair of PRDM9-dependent DSBs: notably, *Zcwpw1*-/- male mice and older female mice are sterile (Li et al. 2019; Wells et al. 2020) and exhibit defects in their ability to repair DSBs (Mahgoub et al. 2020; Wells et al. 2020; Huang et al. 2020). In turn, the genomic locations of DSBs are not altered in *Zcwpw1*-/- mice, indicating that the gene does not play a role in DSB positioning (Mahgoub et al. 2020; Wells et al. 2020; Huang et al. 2020). Similarly, male mouse knockouts of *Tex15* are sterile (F. Yang et al. 2008); DSBs are formed, but not repaired, and in particular the loading of DNA repair proteins RAD51 and DMC1 is compromised (F. Yang et al. 2008). In light of these experimental results, the co-evolution of *PRDM9* with *ZCWPW1* and *TEX15* across vertebrates indicates that PRDM9 likely plays a role in the efficient repair of DSBs not only in mice and humans (Wells et al. 2020; Mahgoub et al. 2020; Huang et al. 2020; F. Yang et al. 2008), but across the vertebrate phylogeny.

Roles of genes that interact with PRDM9

If a gene interacts with PRDM9 by reading its histone modifications, as is the case for ZCWPW1 (Wells et al. 2020; Huang et al. 2020; Mahgoub et al. 2020) and likely ZCWPW2 (**Figure 4**), and has no other roles, we would expect that gene to be dispensable in species that no longer have an active SET domain. Previous papers reported that *ZCWPW1* is more likely to be missing from ray-finned fish with substitutions in catalytic tyrosine residues of the SET domain, in addition to clades lacking the entire *PRDM9* gene (Mahgoub et al., 2020; Wells et al., 2020). In our analysis, we find that both *ZCWPW1* and *ZCWPW2* are more likely to be absent from species carrying only *PRDM9* orthologs with substitutions in at least one catalytic tyrosine residue, as well as those lacking *PRDM9* altogether (**Figure 3**).

While this suggests a dependence of ZCWPW1 and ZCWPW2 on the intact catalytic activity of PRDM9, the interpretation is complicated by the fact that nearly all species with substitutions at the tyrosine residues in all *PRDM9* copies are also carrying only partial *PRDM9* orthologs lacking KRAB and SSXRD domains. Among species investigated here, exceptions to this pattern include six placental mammals for which we had labeled *PRDM9* status as uncertain, which we suspect carry complete *PRDM9* orthologs that we were unable to detect. In addition, the platypus and siluriformes fish carry *PRDM9* orthologs putatively missing the KRAB domain

but have intact tyrosine residues. These exceptions raise the possibility that the dependence is on the KRAB domain of *PRDM9*, rather than the catalytic activity of its SET domain: *ZCWPW2* is absent from all three siluriformes fish species while *ZCWPW1* is absent in one of three species apparently lacking a KRAB domain (*Tachysurus fulvidraco*), and the other two carry a partial KRAB domain at higher e-values than our default cutoff (*Pangasianodon hypophthalmus* and *Ictalurus punctatus*; e-values 6.5 and 450 respectively, **Table S1** and **S9**). Thus, the presence of *ZCWPW1* and *ZCWPW2* may depend on that of the KRAB domain rather than, or in addition to, the tyrosine residues remaining intact.

Similar considerations suggest that in the rare lineages where *ZCWPW1*, *ZCWPW2* and *TEX15* are present in the absence of *PRDM9*, we should expect the genes to be under relaxed selective constraint. To examine this prediction, we tested whether $\omega = dn/ds$ was higher in lineages without a complete *PRDM9* (see Methods). For *ZCWPW1*, there was no evidence for a relaxation of selection (p>0.13; **Table S10**)**.** The intriguing exception is in platypus (p=0.038), a rare species that has an intact SET domain but is lacking the KRAB domain. This observation lends further support to the notion that *ZCWPW1* may be dispensable once the KRAB domain is lost (p=0.038; in all other cases, p>0.13; **Table S10**).

For *ZCWPW2*, the same test revealed that a model in which all species have the same $\omega$is significantly less likely than one in which $\omega$is elevated in lineages lacking a complete *PRDM9*; in particular, in canids and platypus (p=0.0003; **Table S10**). In fact, in canid lineages (fox and dogs), for which the loss of *PRDM9* is ancestral, *ZCWPW2* is no longer under any discernible selective constraint (testing a null model of ω=1, p=0.307; **Table S10**). For *TEX15*, ω is also higher in lineages where *PRDM9* is absent or incomplete (p=0.0036 for fish and p=0.015 for mammals), but remains significantly below 1 (**Table S10**). Together, these evolutionary analyses suggest that whereas *TEX15* may play other roles in a subset of lineages, *ZCWPW2* is dispensable in the absence of a complete *PRDM9* ortholog.

The precise molecular function of ZCWPW2 is to our knowledge completely unknown. Like its paralog, it could be involved in the processing or repair of DSBs. If so, the observation that *Zcwpw1*-/- mice show defective DSB processing and repair (Huang et al. 2020; Mahgoub et al. 2020; Wells et al. 2020) suggests that the role of ZCWPW2 cannot be completely redundant with that of its paralog. Alternatively, by reading the dual marks laid down by PRDM9, ZCWPW2 might help to recruit the recombination machinery (in particular SPO11) and thus play an earlier role in the positioning of DSBs. While in yeast, the link between histone modifications (specifically, H3K4me3) and the recruitment of Spo11 is made by Spp1 (Acquaviva et al. 2013), in mammals, the ortholog of Spp1, CXXC1, is not essential for meiosis (Tian, Billings, and Petkov 2018), and

the gene that plays the analogous role has not yet been identified. Our analysis highlights ZCWPW2 as a potential candidate for this role, to be tested experimentally.

Which loss came first?

While PRDM9 has two distinct roles—in specifying the location of DSBs and in facilitating their repair—the three candidate genes that we have identified may only be involved in one of these two roles. If so, the dependencies between the presence of PRDM9 and of these genes may be asymmetric. For instance, if ZCWPW1 and TEX15 play roles in repair, but not in DSB localization, we might predict that their loss occurs after that of PRDM9; if ZCWPW2 is involved in DSB localization, but not repair, we might predict it to be lost before PRDM9. The phylogenetic data considered here do not allow us to distinguish between these scenarios: there is statistical evidence for a dependence of state transitions of *TEX15*, *ZCWPW1* and *ZCWPW2* on *PRDM9* as well as vice versa (maximum $p<0.00154$, testing the null model of no dependence against either dependence as an alternative model; **Table S11**, also see **Table S9B**). These scenarios could potentially be distinguished by collecting more fine-grained phylogenetic information to pinpoint the specific lineages in which the first loss occurred, as well as in light of further experimental data.

Phylogenetic analyses

Our phylogenetic analysis allowed us to identify novel putative interactors of PRDM9 that are promising candidates for functional studies. For this analysis, the power comes from the repeated losses of *PRDM9*—in our case, eight transitions from presence to absence. Confounding these kinds of analyses, however, are issues of data quality and in particular absences of complete *PRDM9* orthologs that reflect poor genome quality rather than true losses. To address this issue, we validated any absence in Refseq with whole genome searches and where possible, *de novo* assemblies from RNA-seq data, leading us to realize that in two cases, the apparent co-evolution with *PRDM9* was in fact spurious. A more subtle but related issue stems from a phylogenetic signal of genome quality, which can lead to apparent clustering of losses. To minimize this issue, we restricted our analysis to genomes that included most “core” eukaryotic genes (**Fig. S7**) and downsampled our tree to include only the three highest quality genomes below every inferred *PRDM9* loss. As genome qualities improve and as their assemblies become more uniform (eg., (Rhie et al. 2021)), these issues should be alleviated.

Beyond the application to PRDM9 and meiotic recombination, our analysis illustrates how long-standing phylogenetic approaches can now be applied to comparative genomic data to identify novel molecular interactions (Smith et al. 2020). Such analyses need not be restricted to

measurements of presence or absence of whole genes, as we have done here, but could focus exclusively on specific domains, indicative of specific subfunctions, or consider how rates of evolution in specific domains depend on the presence or absence of other genes. With the explosion of high quality and more representative sets of genomes now coming on line (e.g., (Rhie et al. 2021); (Koepfli et al. 2015)), and the development of statistical methods that consider both binary and continuous character evolution jointly, we expect this type of approach to become increasingly widespread.

Material and Methods

1. Identification of PRDM9 orthologs

As a first step towards characterizing the distribution of *PRDM9* in vertebrates, we identified putative *PRDM9* orthologs in the RefSeq database with a *blastp* search [30], using the N-terminal portion of the *Homo sapiens* PRDM9 protein sequence containing KRAB, SSXRD and SET domains as the query sequence (RefSeq accession: NP_001297143; amino acid residues 1-364). We downloaded the corresponding GenBank file for 5,000 hits (3,384 unique genes from 378 species) and characterized the presence or absence of KRAB, SSXRD and SET domains for each record using the Conserved Domain Database (CDD; (A. Marchler-Bauer 2004); accessions cl02581, cl09744 and cl40432), annotating each domain as present if that domain had an e-value less than 1. We then removed alternative transcripts from the dataset by preferentially keeping, for each unique gene, the transcript with the maximal number of annotated domains. When there were multiple transcripts with the same maximal number of domains, we kept the longest one.

Because *PRDM9* shares its SET domain with other PRDM family genes and its N-terminal domains with members of the KRAB-ZF and SSX gene families, many of these hits are potential PRDM9 paralogs. To identify bona fide *PRDM9* orthologs from this initial set of genes, we sought to build phylogenetic trees specific to the KRAB, SSXRD, and SET domains and remove homologs that cluster with genes annotated as distantly related paralogs of *PRDM9*. To this end, we extracted the amino acid sequences for complete KRAB, SSXRD, and SET domains, and for each domain, constructed neighbor-joining trees using Clustal Omega (Sievers and Higgins 2014). Utilizing the KRAB and SSXRD domain-based trees, we identified and removed 113 genes that visually cluster with members of the SSX gene family (**Fig. S1A-B**). Analyzing the SET domain-based tree, we identified and removed 2,646 genes that group with other members of the *PRDM* gene family (**Fig. S1C**; see figure legend for details). We ultimately retained 625 genes, each of which cluster with *PRDM9* in one or more of these trees.

By this approach, in the 378 species considered, we identified 188 *PRDM9* orthologs containing KRAB, SSXRD and SET domains from 141 species, as well as 13 *PRDM9* orthologs containing KRAB and SET domains from an additional 13 species for which we were unable to detect an SSXRD domain with an e-value less than 1. For the remaining 224 species, we were unable to identify a *PRDM9* ortholog spanning KRAB and SET domains from the RefSeq database. Of these, 175 are in clades for which the complete or partial loss of *PRDM9* has been documented previously (Baker et al. 2017; Axelsson et al. 2012; Singhal et al. 2015) i.e., in birds and crocodilians (86), percomorph fish (71), cypriniformes fish (6), characiforme fish and siluriformes fish (6), frogs (3), canids (2), or the platypus (1). In turn, we sought to verify that *PRDM9* was truly absent in the remaining 49 species using a number of approaches.

First, we performed a series of *tblastn* searches of the whole genome of each focal species using the N-terminal portion of the *Homo sapiens* PRDM9 protein as a query. When we were unable to retrieve any promising hits with the human protein sequence, we re-performed the *tblastn* search using the N-terminal portion of a *PRDM9* ortholog from a species closely related to the focal species. In order to identify which of the identified contigs corresponded to genuine *PRDM9* orthologs (as opposed to paralogs such as *PRDM11*), we performed blastp searches against the *Homo sapiens* RefSeq database using the aligned protein sequences as query sequences. Contigs containing the relevant alignments spanning KRAB and/or SET domains were then downloaded and the aligned region including 10,000 of flanking sequence was extracted and input into *Genewise (Birney, Clamp, and Durbin 2004)*, using the PRDM9 protein sequence from *Homo sapiens* or a closely related species as a guide sequence (see **Table S2** for details). In genomes from nine species, we identified separate contigs containing the KRAB domain and the SET domain. In these cases, the contigs were concatenated before use as input in *Genewise*. These approaches enabled us to identify an additional 57 *PRDM9* orthologs from 35 species, including 24 *PRDM9* orthologs containing KRAB, SSXRD and SET domains from 22 species, and 24 *PRDM9* orthologs containing KRAB and SET domains but for which we were unable to identify the SSXRD domain from 8 species.

These analyses left 19 species for which we were unable to identify a *PRDM9* ortholog with both KRAB and SET domains and which do not fall within the clades previously reported to have a complete or partial loss of *PRDM9*. For these species, we additionally searched testis RNA-seq datasets when possible, including those generated for this study (see below; **Table S3**); this approach enabled us to identify two additional *PRDM9* orthologs containing KRAB and SET domains from two species of fish. Additionally, we were able to confirm one absence, in *Anolis carolinensis*, as well as establish one in *Scleropages formusus*, a closely related species that had

not been included in our initial RefSeq based search. Of the 16 species remaining, one is outgroup to a previously reported loss (the pinecone soldierfish, *Myripristis murdian*, an outgroup to percomorph fish), whereas the other 15 putative losses would lead to the inference of 15 independent losses across the vertebrate phylogeny. Given that we do not have confirmatory evidence of absence, we treat these 15 species as having an uncertain *PRDM9* status.

Altogether, this pipeline resulted in the identification of 163 species in which we find a complete *PRDM9* ortholog containing KRAB, SSXRD and SET domains, 23 species for which we identify *PRDM9* orthologs containing KRAB and SET domains but not SSXRD domains, 178 species for which we have evidence for the absence of a complete *PRDM9* gene, and 15 species for which we were unable to make a confident determination (see **Tables S1**-**S4, Figure 1**).

For each of the *PRDM9* orthologs that we identified, we characterized the conservation of three key tyrosine residues that have been shown to underlie the catalytic function of the human SET domain in vitro (i.e., Y276, Y341, and Y357; (Wu et al. 2013)) and for Y357, in vivo in mouse (Diagouraga et al. 2018). To this end, we constructed an alignment of the SET domain using Clustal Omega (Sievers and Higgins 2014) and extracted the residues aligning to the human tyrosine residues from each of 678 SET domains (**Table S1**).

2. Verification of genomic calls using RNA-seq data

For four species in which we identified no *PRDM9* ortholog or only a partial ortholog, we investigated whether a complete *PRDM9* ortholog may nonetheless be present using RNA-seq data. We therefore sought to verify its absence from *Anolis carolinensis,* a species in which we had been unable to find a *PRDM9* ortholog in the genome assembly or Refseq, as well as a second reptile species, *Sceloporus undulatus,* for which Refseq data and a genome sequence were not available. To this end, we built a *de novo* RNA transcriptome assembly and tested for the expression of PRDM9 in testis and other tissue samples (see below).

Similarly, in two species in which we had originally identified only a partial ortholog of *PRDM9* (*Astyanax mexicanus* and *Clupea harengus*), we wanted to verify the incomplete domain structure inferred from the genome sequence by conducting a *de novo* transcriptome assembly (in *Clupea harengus*, this analysis turned out to be unnecessary, as an updated reference genome, GCA_000966335.1, contains a complete *PRDM9*). To this end, we analyzed RNA-seq data from male gonad surface from *Astyanax mexicanus* and liver and testis from *Clupea harengus*.

Dissected tissue samples preserved in RNAlater were kindly provided to us by Arild Folkvord and Leif Andersson (*Clupea harengus*), Cliff Tabin (*Astyanax mexicanus*), Tonia

Schwartz and Tracy Langkilde (*Sceloporus undulatus*), and Athanasia Tzika (*Anolis carolinensis*). These samples were stored at -20°C until extraction and library preparation. Total RNA was extracted using the Qiagen RNeasy kit (Valencia, CA, USA) following the manufacturer's protocol. RNA was quantified and assessed for quality on a Qubit fluorometer and approximately 1 µg of total RNA was input for library preparation using the Kapa RNA-seq kit. Samples were prepared following the manufacturer's protocol, except that half reactions were used. 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. In order to evaluate the library quality and size distribution, libraries were evaluated on an Agilent Tapestation. The libraries were then sequenced over two runs on the NextSeq 550 at Columbia University to collect paired-end 150 bp reads.

Illumina sequencing reads (248,820,547 2x150 base pair (bp) paired-end reads) were demultiplexed into individual sample fastq files with the software bcl2fastq2 (v2.20.0, Illumina). The FastQC software (Schmieder and Edwards 2011) was used for visual inspection of read quality. Adapters and low-quality reads were trimmed with the Trimmomatic software, which is bundled as a plugin within the Trinity *de novo* assembler (Grabherr et al. 2011) (v2.8.5) and was enabled using the *--trimmomatic* flag. The default trimming settings (phredscore>=5; slidingwindow:4:5; leading:5, trailing:5; minlen:25) were used following (Macmanes 2014) recommendations. The pair-end reads were trimmed and *de novo* transcriptomes assembled with Trinity (v2.8.5) using the following parameters: --seqType fq --SS_lib_type FR --max_memory 100G --min_kmer_cov 1 --trimmomatic --CPU 32. Details on assembly quality are shown in **Fig. S2**. Gene expression data for all four species (*Anolis carolinensis, Sceloporus undulatus, Clupea harengus, Sceloporus undulatus*) are available from the NCBI sequence read archive (Bioproject PRJNA605699, SRA accessions: SRR11050679-SRR11050687).

To evaluate whether PRDM9 was present in the transcriptome data, we conducted a *tblastn* search (e-value ≤ 1e-5) against each *de novo* assembly using the human PRDM9 protein sequence (without its rapidly evolving zinc finger array) as a query, and we classified the domain presence of up to five top hits using CDD blast (A. Marchler-Bauer 2004). For a given species, if the KRAB and SET domains were not identified in any transcript, *PRDM9* was considered incomplete. The inability to identify PRDM9 could indicate either that the gene is not expressed or that we lack the appropriate cell types or sequence coverage to detect it. To assess our power to detect PRDM9 from the testis RNA-seq data, we followed methods outlined in (Baker et al.

2017). Specifically, for each transcriptome, we evaluated whether we could identify transcripts from six genes with highly conserved roles in meiotic recombination (Lam and Keeney 2014) (*HORMAD1*, *MEI4*, *MRE11A*, *RAD50*, *REC114*, and *SPO11*). To identify the transcripts orthologous to each of these genes, we performed a *tblastn* search (e-value ≤ 1e-5) of the *Homo sapiens* reference protein sequence against each *de novo* transcriptome. We considered PRDM9 to be absent if we detected expression of all six genes but not a complete *PRDM9*; by these criteria, we found PRDM9 to be missing from *A. carolinensis*, *S. undulatus*, and *A. mexicanus*.

Using the same approach to *de novo* assembly and gene detection, we also analyzed publicly available RNA-seq datasets from testis for 25 additional species (**Table S3**), either to verify the absence of PRDM9 (see above) or of one of the candidate genes (see below).

To estimate the expression levels of the Trinity-reconstructed transcripts, we used RSEM (Applied Research Applied Research Press 2015) (v1.3.1) implemented through Trinity (v2.8.5). We first aligned the RNA-seq reads from each sample to the newly generated *de novo* assembled transcriptome (see above) using the alignment method bowtie (Langmead and Salzberg 2012) (v1.2.2). We then extracted quantification information for each gene of interest from the RSEM output (in fragments per kilobase of transcript per million mapped reads or FPKM) (**Fig. S3**).

3. Choice of candidate genes and orthology assignments

To identify a set of genes that may co-evolve with *PRDM9*, we relied on three publicly available datasets, namely: (i) 38 genes associated with variation in recombination phenotypes in a genome-wide association study in humans (Halldorsson et al. 2019). Of the variants reported to be associated with recombination phenotypes, six were found in intergenic regions; we included the subset of two cases in which the authors assigned these variants to nearby genes (*ZNF84* and *ZNF140*). (ii) 193 genes co-expressed with PRDM9 in single cell data from mouse testes. Specifically, we considered the top 1% of genes based on their gene expression loadings in component 5, the component in which PRDM9 is most highly expressed (Jung et al. 2019). (iii) 36 genes known to have a role in mammalian meiotic recombination based on functional studies (Frédéric Baudat, Imai, and de Massy 2013).

Genes co-expressed with PRDM9 in mouse spermatogenesis were converted to human gene symbols using the package biomaRt in R (Durinck et al. 2009). Fifteen of these genes did not have an orthologous human gene symbol (*Gm7972*, *H2-K1*, *Gm4349*, *Ddx43*, *Atad2*, *Xlr4c*, *Gm364*, *Tex16*, *4933427D06Rik*, *AI481877*, *H2-D1*, *Trap1a*, *Xlr4a*, *2310035C23Rik*, and *Tmem5*) and eight other genes mapped to more than one human gene symbol (*Msh5*, *Cbwd1*, *Nxf2, Cbwd1*, *Fam90a1b*, *Srgap2*, *Cdk11b*, *Gm15262*). Keeping all mapped gene symbols

yielded 185 genes; combined with the two other sources, 241 genes were tested for their co-evolution with *PRDM9* (**Figure S6**). A supplementary file describing each meiosis candidate gene is available in **Table S5**.

For the 241 genes, we characterized whether the ortholog is present in its complete form across vertebrate species. To this end, we first downloaded all the vertebrate RefSeq protein sequences available on the NCBI database (accessed on June 3, 2020), corresponding to 339 species. Of these, we filtered out 32 species that were missing 10 or more BUSCO core genes (out of a total of 255 genes) (Waterhouse et al. 2018), reasoning that their genomes were sufficiently incomplete that they may be missing orthologs by chance (see **Figure S7**). Of the remaining 307 species, we further excluded 29 species in order to remove polytomies observed in the phylogeny; specifically, we removed the minimal number of species necessary to remove each polytomy while preserving any transitions in the state of *PRDM9*. Moreover, to minimize possible phylogenetic signals generated by genome assembly quality, we thinned the tree such that for each *PRDM9* loss along the phylogeny, we kept at most three species representing that loss. In cases where a loss was ancestral to more than three species in our dataset, we picked three distantly related species with the best genome assemblies, as measured by the BUSCO score. In the end, we retained 189 species: 134 mammals, 3 birds, 6 amphibians, 18 reptiles, 2 percomorph fish, 3 cypriniformes fish, 20 other ray-finned fish, 2 cartilaginous fish, and one jawless fish.

For each candidate gene in each species, we performed a blastp search of the human ortholog against the RefSeq database of the species and kept up to five top hits obtained at an e-value threshold of 1e-5. We inferred the domain structure of each hit using the Batch CD-Search (Aron Marchler-Bauer et al. 2015), and considered a domain as present in a species if the e-value was ≤ 0.1. We considered genes to be complete orthologs if they contained the superfamily domains found in four representative species of the vertebrates phylogeny carrying a complete PRDM9 (*Esox lucius* (fish) *Geotrypetes seraphini* (caecilian), and *Pseudonaja textilis* (snake)), at an e-value threshold of 1e-4. For the 15 genes (*FANCB*, *FMR1NB*, *GPR137C*, *HAUS8*, *M1AP*, *MEI1*, *SPATA22*, *CLSPN*, *FBXO47*, *HMGA2*, *HSF2BP*, *IQCB1*, *LRRC42*, *PRAME*, *SYCE2*) in which no detectable domains were present, we annotated the presence or absence of the gene using the blastp results alone. In the end, we built a matrix of presence or absence across species and candidate genes to be used in the phylogenetic test (see **Table S6**).

4. Testing for the co-evolution of PRDM9 and candidate genes

To test for the co-evolution of *PRDM9* and each candidate gene, we need to account for the phylogenetic relationships among the species considered. To obtain these relationships and time-calibrated branch lengths, we used the TimeTree resource (http://timetree.org/; (Kumar et al. 2017), accessed on June 10, 2020). Of the 189 species included in the phylogenetic tests, 9 were not present in the TimeTree database; in those cases, we used information from a close evolutionary relative to determine their placement and branch lengths.

For this test, we consider *PRDM9* as present if it contains KRAB and SET domains or incomplete/missing if one of those domains is absent (**Tables S4** and **S9**). We do not rely on the SSXRD domain when making these calls because its short length makes its detection at a given e-value threshold unreliable. Notably, for 17 of the 23 species with *PRDM9* orthologs containing KRAB and SET domains, but not SSXRD domains with an e-value < 1, we are able to detect the SSXRD domain when using an e-value threshold of 1000 (**Tables S1** and **S4**). We additionally do not rely on the ZF array because its repetitive nature makes it difficult to sequence reliably.

We tested whether state changes of intact candidate genes were unexpectedly coincident with state changes of the intact *PRDM9* using the software *BayesTraitsV3* (Pagel, Pagel, and Meade 2006). We did so by comparing the statistical support for two models: a null model in which *PRDM9* and a given candidate gene evolve independently of one another along the phylogeny versus an alternative model in which the gain ("1") and loss ("0") of a gene is dependent on the status of *PRDM9* and vice versa. We compared the likelihoods of the two models using a likelihood ratio test with 4 degrees of freedom, and reported a p-value uncorrected for multiple tests (**Table S7**). For each gene and model, 100 maximum likelihood tries were computed and the maximum likelihood value was retained.

Given the phylogenetic distribution of *PRDM9*, it is likely that a *PRDM9* ortholog was present in the common ancestor of vertebrates (Oliver et al. 2009; Baker et al. 2017). Based on this prior knowledge, we restricted the state of *PRDM9* at the root of the phylogeny to always be present. In turn, for each candidate gene, we set a prior in which it had 50% probability of being present and 50% probability of being absent. We also used this prior for the state of *PRDM9* in the 15 species that lacked *PRDM9* but where the loss was not supported by a closely related species (i.e., for which we considered the status uncertain).

For *TEX15*, *ZCWPW1* and *ZCWPW2*, we also explored restrictions on the rates in the dependent model, such that their state transitions depend on PRDM9 (model X) or the state transitions of *PRDM9* depends on theirs (model Y), rather than both being true. For these tests, we compared the likelihoods of each dependent model against our independent null model using

a likelihood ratio test with 2 degrees of freedom. For each gene and model, 100 maximum likelihood tries were computed and the maximum likelihood value was retained.

5. Improving gene status calls of top candidate genes

For the five genes found to be significant at the 5% level (**Figure 2A**), we sought to improve our calls by building phylogenetic trees based on domains in the genes and examining the clustering patterns visually, as well as by searching for orthologs in whole genome assemblies and testis transcriptomes (following the same procedures described for *PRDM9*). These improved calls were then used to rerun the phylogenetic independent contrast tests, following the same implementation as previously; the p-values for these improved gene models are shown alongside the original ones in **Table 1**. Below we provide an overview of the steps we took for each candidate gene. For each gene we provide descriptions of identified orthologs and how they were identified in **Table S1**, specific details about orthologs identified from whole genome assemblies in **Table S2**, our improved calls per species in **Table S9A**, and a summary of loss events in **Table S9B**.

i. *MEI1*

For *MEI1*, an initial blastp search of the vertebrate RefSeq database using the human sequence as query and an e-value threshold of 10e-5 resulted in the identification of 422 *MEI* orthologs from 372 species. We note that, for *MEI1*, we did not find any domain annotations, and therefore did not perform phylogenetic analysis to support the identification of these orthologs. However, each homolog identified in our initial RefSeq analysis was annotated as either *MEI1* or *MEI1-like*. We thus labeled each species as having a complete ortholog if an ortholog was present. This approach resulted in the identification of *MEI1* ortholog for 187 of the 189 species used for our co-evolutionary test. For the remaining 2 species, we sought to identify *MEI1* orthologs from whole genome sequences following the same procedures described for *PRDM9*. This approach allowed us to identify a *MEI1* ortholog in every species, revealing that in fact, *MEI1* has not been lost among the vertebrate species examined (**Tables S1** and **S2**).

ii. *ZCWPW1* and *ZCWPW2*

Because *ZCWPW1* and *ZCWPW2* are paralogs, we performed our analyses of these genes together. To this end, we combined the datasets of genes identified in our initial RefSeq blastp search to create a dataset of 977 putative orthologs from 363 species. We then extracted amino acid sequences and built neighbor-joining trees using Clustal Omega for both the zf-CW

and PWWP domains ((Sievers and Higgins 2014); accessions cl06504 and cl02554; **Figure S9A-B**). Utilizing these trees, we removed 573 genes that visually clustered with genes annotated as distantly related paralogs, such as members of the MORC and NSD gene families. We additionally relied on these trees to more confidently label which genes were *ZCWPW1* orthologs and which were *ZCWPW2* orthologs based on where they clustered in the tree. We considered orthologs as complete if they contain both PWWP and zf-CW domains with e-values < 1. This approach resulted in the identification of 193 complete *ZCWPW1* orthologs from 188 species, and 187 complete *ZCWPW2* orthologs from 180 species.

Among the 189 species used in our co-evolutionary test, 164 had complete *ZCWPW1* orthologs and 154 had complete *ZCWPW2* orthologs on the basis of this initial search. For the 25 species missing a complete *ZCWPW1* ortholog, and for the 35 missing a complete *ZCWPW2* ortholog, we sought to identify the orthologs from whole genome sequences following the same procedures as described for *PRDM9*. This approach enabled us to identify an additional 3 complete *ZCWPW1* orthologs from 3 species, and an additional 11 complete *ZCWPW2* orthologs from 11 species (**Tables S1** and **S2**). For the remaining species, we checked the putative loss of *ZCWPW1* or *ZCWPW2* using RNA-seq data when available, which led to the identification of an additional 2 complete *ZCWPW1* orthologs, but no additional *ZCWPW2* orthologs (**Tables S1** and **S3**).

iii. *TEX15*

For *TEX15*, our initial blastp search resulted in the identification of 900 putative orthologs from 363 species. We similarly utilized a tree built using the DUF3715 domain to remove 667 genes that cluster with distantly related paralogs, in particular, *TASOR* and *TASOR2* (**Figure S9C**). When making our final calls about *TEX15* orthologs, we labeled them as complete if they contained both DUF3715 and TEX15 domains (accessions pfam12509 and pfam15326). This approach resulted in the identification of 179 complete *TEX15* orthologs from 175 species. Among the 189 species used for our co-evolutionary test, 150 had complete *TEX15* orthologs on the basis of this initial search. For the 39 species missing a complete *TEX15* ortholog, we sought to identify the orthologs from whole genome sequences, following the same procedures as described for *PRDM9*. In this way, we identified an additional 28 complete *TEX15* orthologs from 27 species (**Table S2**). For the remaining species, we checked if we could find *TEX15* using RNA-seq data, when available, and found one additional complete *TEX15* ortholog by this approach (**Tables S1** and **S3**).

iv. *ANKRD31*

For *ANKRD31*, our initial blastp search resulted in the identification of 2,563 putative orthologs from 388 species. In humans, the *ANKRD31* gene contains two sets of three Ankyrin repeats, and a RAMA domain (accession cl16759). While the RAMA domain appears to be unique to *ANKRD31* orthologs amongst our initial RefSeq results, Ankyrin repeats are found in a broad set of genes. We identified 12 distinct domain annotations corresponding to the Ankyrin repeats in the human *ANKRD31* gene, usually distinguished by the number of repeats they contain. We therefore utilized a tree built using the ANKYR domain (accession cl34000), which spans one such set of repeats in the human ortholog, to remove 2,196 genes clustering with distantly related paralogs, such as *SLF1*, *BCORL1*, *TONSL*, *PSMD10*, and other members of the ANKRD gene family, such as *ANKRD11* and *ANKRD12* (**Figure S9D**). There were 12 genes from our initial RefSeq results that were not included in this tree because they lacked ANKYR domains. Of these, one lacked any domain annotations and was removed from subsequent analysis. The remaining 11 genes each contained RAMA domains, indicating that they are genuine *ANKRD31* orthologs. When making our final calls about the presence or absence of complete *ANKRD31* orthologs, we labeled orthologs as complete if they contained Ankyrin repeats and the RAMA domain. To this end, we used the Ank_2 domain (accession cl39094) as a proxy for the presence of Ankyrin repeats; we did so having observed that each gene containing at least one of the 12 domain annotations corresponding to Ankyrin repeats in the human ortholog contained the Ank_2 domain.

This approach resulted in the identification of 285 complete *ANKRD31* orthologs from 282 species. Among the 189 species used for our co-evolutionary test, 159 had complete *ANKRD31* orthologs on the basis of this initial search. For the 30 species missing a complete *ANKRD31* ortholog, we sought to identify the orthologs from whole genome sequences, following the same procedures as described for PRDM9, thus identifying an additional 11 complete *ANKRD31* orthologs from 11 species (**Table S2**). For the remaining species, we checked RNA-seq data, when available (**Table S3**); this approach did not allow us to identify any additional complete ANKRD31 orthologs.

6. Conservation of residues in ZCWPW2

We carried out a residue conservation analysis using an approach proposed by (Capra and Singh 2007), using code *score_conservation.py* available at https://compbio.cs.princeton.edu/conservation/. This approach quantifies the Jensen–Shannon divergence between the amino acid distribution of the focal residue and a “background amino acid distribution.” The alignment of *ZCWPW2* was produced using Clustal Omega (using default

parameters) within MEGA (version 7, (Kumar et al. 2017; Kumar, Stecher, and Tamura 2016)). As recommended, the overall background amino acid distribution was drawn based on the BLOSUM62 amino acid substitution matrix provided by the software (Capra and Singh 2007). Any column of the gene sequence alignment with more than 30% gaps was ignored. A window size of 3 was used to incorporate information from sequential amino acids, as recommended by the default settings.

7. Evidence for relaxed selective constraint in the absence of PRDM9

To test for possible relaxed selection in species without a complete *PRDM9*, we used the program *codeml* within PAML (Z. Yang 2007, 1997). *Codeml* uses protein coding sequences to estimate the ratio of non-synonymous to synonymous substitution rates ($\omega = d_N/d_S$). Values of ωsignificantly less than 1 are indicative of purifying selection, i.e., of the functional importance of the gene.

To this end, we considered each major clade (fish, mammals, reptiles, amphibians) separately and extracted and aligned coding nucleotide sequences from NCBI for multiple species. We aligned those sequences in a codon-aware manner using Clustal Omega (using default parameters) within MEGA (version 7, (Kumar et al. 2017; Kumar, Stecher, and Tamura 2016)) and inspected the codon-aware alignment visually to ensure that the same isoforms were used across species. For each multi-species alignment, we tried two approaches: (i) We estimated ω under a null model assuming the same ω across all branches and an alternative model in which there are two ω allowed: one ω value in species with a complete *PRDM9* and a second ω for the branches in which *PRDM9* is absent or incomplete (including the internal branches on which PRDM9 may have been lost); (ii) We considered the same null model with the same ω across all branches; and an alternative model with one ω value in species with a complete *PRDM9*, a second ω for the branches in which *PRDM9* is absent or incomplete and additional ω values for each branch on which *PRDM9* was inferred to be lost (a different one for each independent loss, as the ω value averaged over the branch will depend on when along the branch *PRDM9* was lost). For (i), significance was assessed using a likelihood ratio test with 1 degree of freedom; for (ii), by the number of degrees of freedom corresponded to the number of distinct ω values minus 1. If ω values were found to be significantly higher in species without a complete *PRDM9*, we tested whether or not we could reject $\omega = 1$ for these species. For two cases in which we could not obtain a multi-species alignment that included the whole coding sequence (*ZCWPW1* in fish and *TEX15* in amphibians), we instead used the pairwise model (runmode: -2 within PAML) on alignments for a pair of species, and tested whether we could reject $\omega = 1$ for

species lacking *PRDM9* by comparing a model allowing ω to vary versus a null model fixing the ω value at 1, with 1 degree of freedom.

Figures

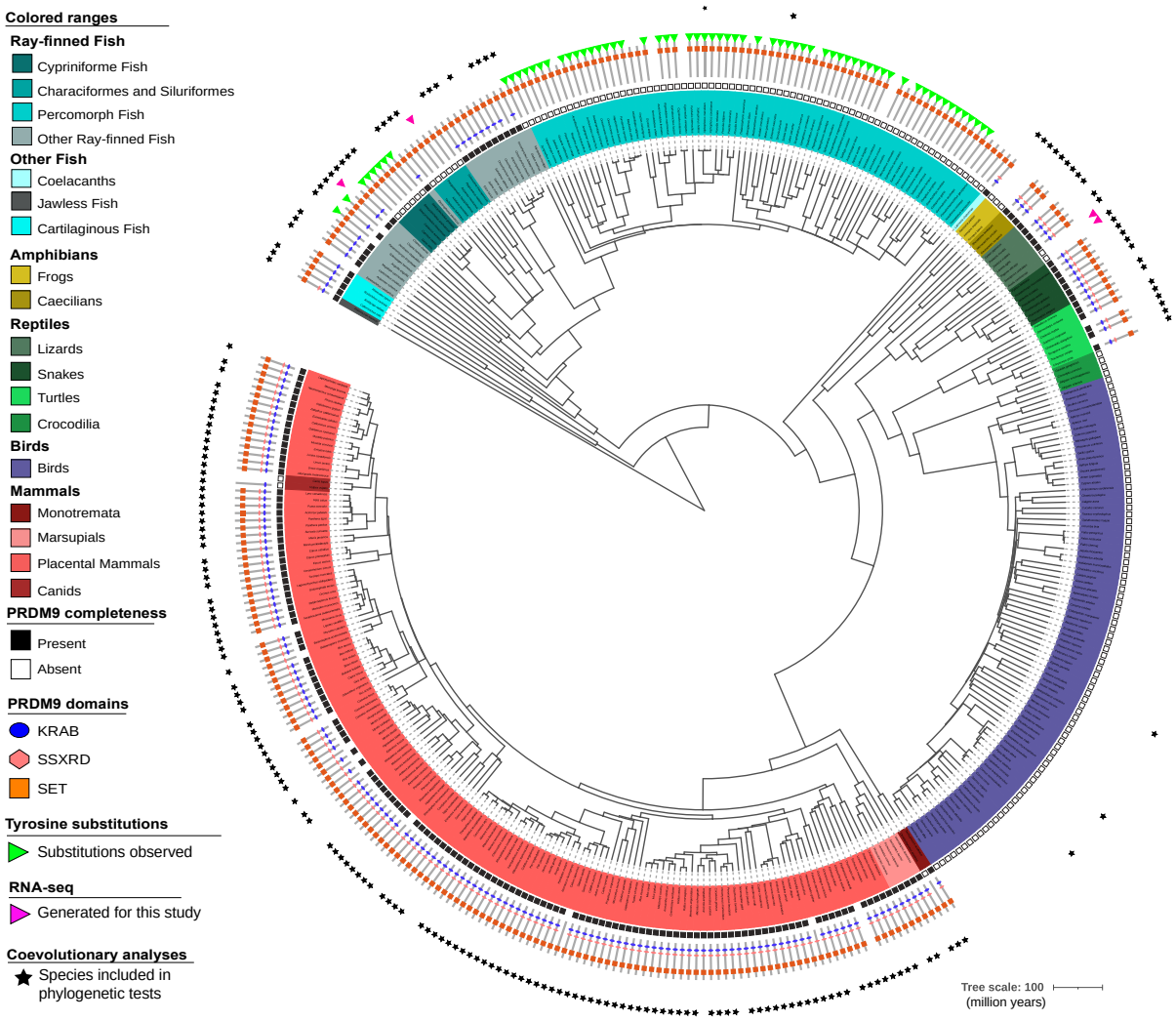

Figure 1. The phylogenetic distribution of PRDM9 and its domain architecture across vertebrates. The inferred *PRDM9* status of 379 vertebrate species. Branch lengths were computed based on the TimeTree database; for 33 species not present in the database, we used branch length information from a close evolutionary relative. Different vertebrate clades are indicated by colored segments, with salmon for mammals, cyan for fish, mustard for amphibians, green for reptiles, and purple for birds. In the inner circle, squares indicate whether *PRDM9* is complete (filled black) or incomplete/absent (empty black); for species with an uncertain *PRDM9* status, no box is shown. The PRDM9 domain architecture of each species is shown with a cartoon, where the presence of a KRAB domain is indicated in blue, of SSXRD in pink, and of the SET domain in orange. Green triangles indicate species that only carry *PRDM9* orthologs with substitutions at putatively important catalytic residues in the SET domain (see **Table S4**). The tree was drawn using itool (https://itol.embl.de/); an interactive version is available at https://itol.embl.de/shared/izabelcavassim.

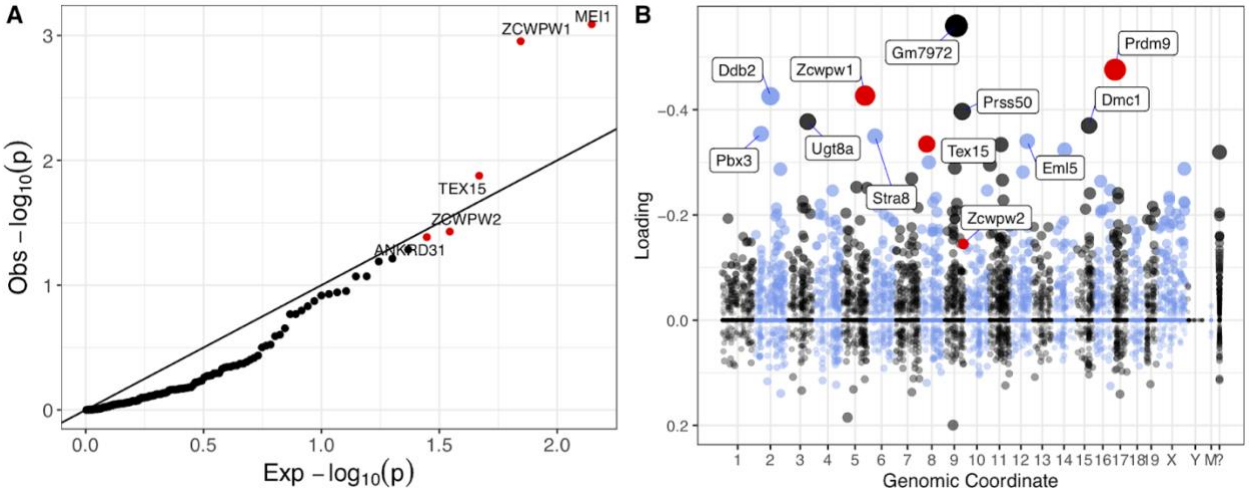

Figure 2. Phylogenetic tests and genes co-expressed with PRDM9 in single cell mouse testes data (A) Quantile-Quantile plot of the p-values obtained from the phylogenetic tests run on 139 genes that appeared to have been lost at least once in the 189 vertebrate species considered. Genes that are significant at the 5% level are shown in red. **(B)** Loadings for one of 46 components (component 5) inferred from single cell expression data in mouse testes (Jung et al. 2019), in which PRDM9 is most highly expressed. The dot sizes are proportional to the square of the absolute value of the loading, indicative of higher expression within component 5. *PRDM9* and the three genes identified in our phylogenetic tests with $p<0.05$ are shown in red. Mouse genomic coordinates are displayed. Panel B was made from summary statistics provided by (Jung et al. 2019), using SDAtools (https://github.com/marchinilab/SDAtools/).

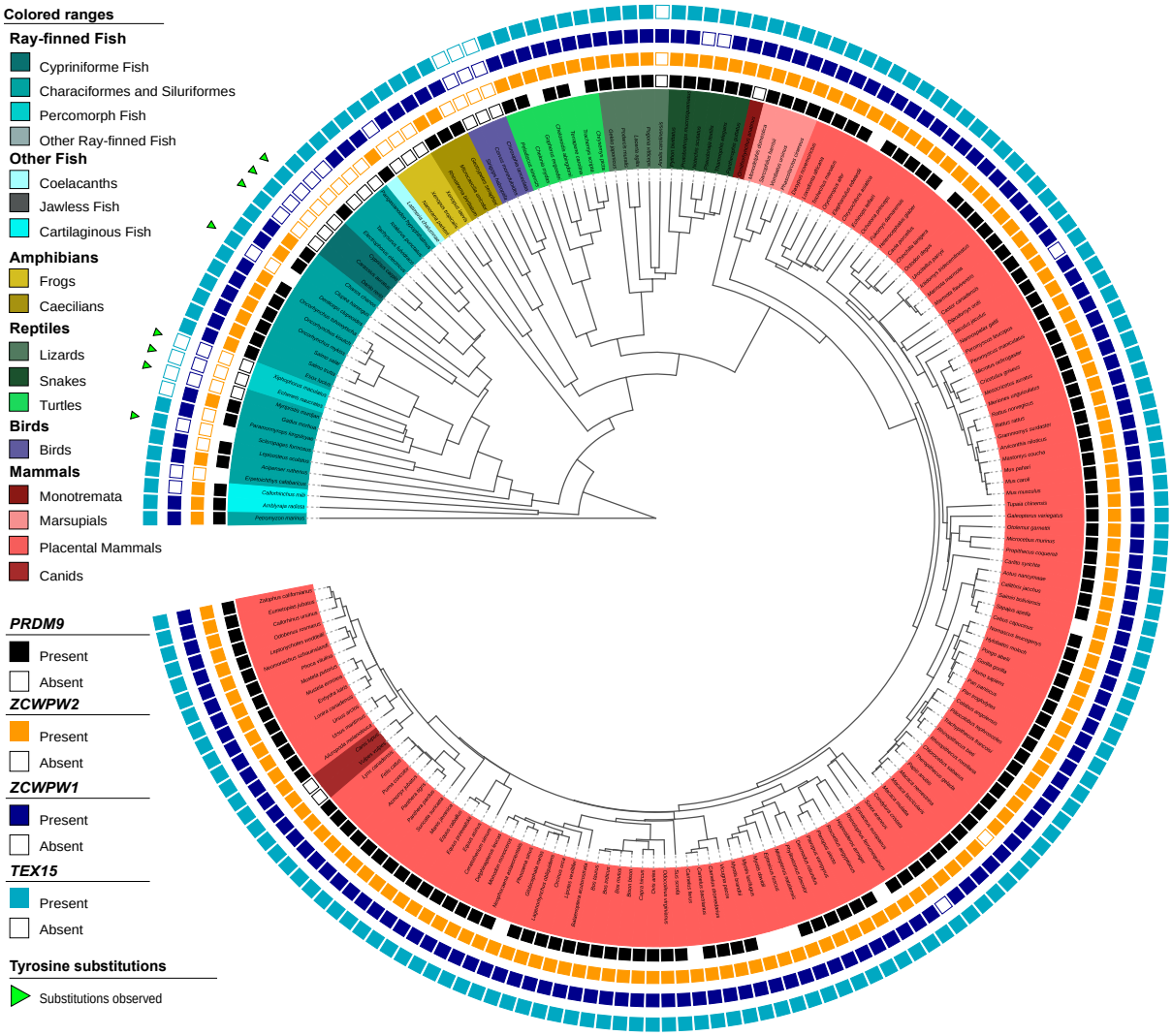

Figure 3. The phylogenetic distribution of *PRDM9* and co-evolving genes across 189 species. Filled teal and empty teal squares indicate whether *PRDM9* is present or absent, respectively (see Methods). If nothing is indicated, the status of *PRDM9* is uncertain. Likewise, filled orange and empty squares indicate whether *ZCWPW2* is present or absent/incomplete and filled and empty navy squares indicate whether *ZCWPW1* is present or absent/incomplete. Green triangles indicate species that only carry *PRDM9* orthologs with substitutions at putatively important catalytic residues in the SET domain (see **Table S4**). The status of candidate genes (for which $p<0.05$; Figure 2A) was re-evaluated based on a search of gene models within whole genome sequences (see Methods); updated p-values for the phylogenetic test are shown in Table 1. The tree was drawn using itool (https://itol.embl.de/); an interactive version is available at https://itol.embl.de/shared/izabelcavassim.

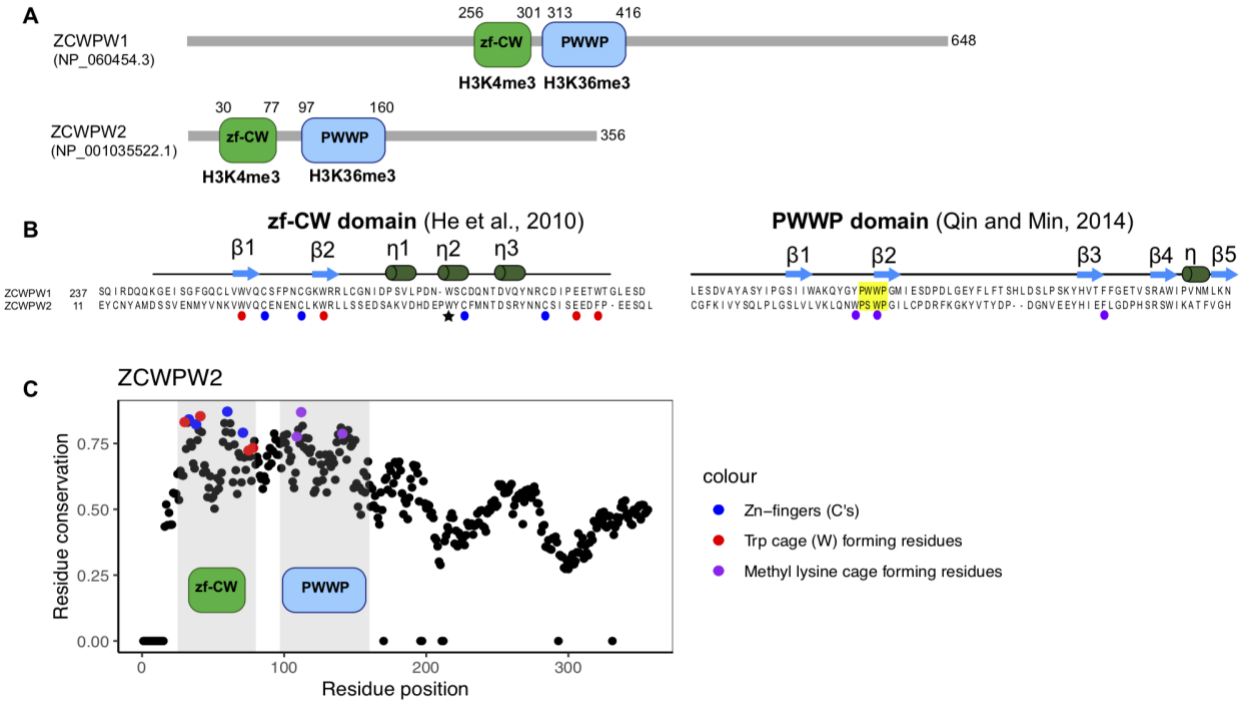

Figure 4. Domain architecture of ZCWPW1 and ZCWPW2. (A) Amino acid sequence and domain structure composition of genes ZCWPW1 and ZCWPW2 in humans. **(B)** The ZF-CW domain structure includes the fingers (residues indicated by blue circles) and an aromatic cage (red) expected to bind to H3K4me3 (He et al. 2010), and the star indicates the third Trp residue that is thought to stabilize the fold by hydrophobic interactions (He et al. 2010). The PWWP domain (yellow) is expected to bind to histone H3K36me3 through a hydrophobic cavity composed of three aromatic residues (purple) (Qin and Min 2014). The secondary structures of zf-CW and PWWP domains are represented above sequences. **(C)** Conservation of residues in ZCWPW2 across vertebrates, with those residues recognizing modifications on the histone tail colored in blue, red and purple. Positions in the ZCWPW2 alignment with › 30% of gaps were ignored and the conservation score was set to 0.

Tables

Table 1. Results of phylogenetic tests

We focused on the five genes for which the likelihood ratio test yielded p<0.05 (uncorrected for multiple tests), improved the ortholog status calls, and reran the phylogenetic tests for four of them (all but *MEI1*, which turned out to be present in all species considered; see Methods). Gene source refers to the criterion by which the gene was originally included among our lists of candidates: (1) it is among top 1% of the genes of component 5 inferred from single cell mouse testes data (Jung et al. 2019) or (2) it is associated with variation in recombination phenotypes in humans (Halldorsson et al. 2019) or (3) it was previously known to have a role in mammalian meiotic recombination from functional studies (Frédéric Baudat, Imai, and de Massy 2013).

Gene	LogLik H0	LogLik Ha	P-value	P-value for improved status calls	Position (human coordinate)	Gene source
MEI1	-41.06	-50.52	0.0008	NA	chr22: 41699503	3,1
ZCWPW1	-62.27	-53.15	0.0011	0.0341	chr7: 100400826	1
TEX15	-134.76	-128.44	0.0132	0.006	chr8: 30831544	1
ZCWPW2	-64.04	-58.94	0.0372	0.000096	chr3: 28348721	1
ANKRD31	-74.12	-69.14	0.0412	0.376	chr5: 75068275	1

Acknowledgments

The authors would like to thank Arild Folkvord and Leif Andersson (*Clupea harengus*), Cliff Tabin (*Astyanax mexicanus)*, Tonia Schwartz, and Tracy Langkilde (*Sceloporus undulatus*), and Athanasia Tzika (*Anolis carolinensis*) for providing testis, muscles, and liver samples, and Mackenzie Keegan for extraction and library preparation of samples. The authors are also grateful to Andres Bendesky for use of his lab space and sequencing facilities and to Andrew Meade for helpful discussions about the software *Bayestraits*.

Data and Code availability

The following data sets were generated

1. **RNA-seq data for two fish species and two reptile species (fastq files)** Cavassim M I A, Baker Z, Hoge C, Schierup M, Schumer M and Przeworski M (2020)

   Available from the NCBI BioProject (accession number: PRJNA605699).

2. **Code availability** Code generated for this study can be found at

   https://github.com/izabelcavassim/PRDM9_analyses

3. **Supplementary material**

   Supplementary material is available on

   https://www.dropbox.com/sh/pihq6a643fz21js/AAANGJWpALT42MCrrsJdlSUHa?dl=0

Author contributions

Conceptualization: M.I.A.C., Z.B., and M.P.; **Methodology:** M.I.A.C., Z.B., C.H., M.S., and M.P.; **Formal Analysis:** M.I.A.C., Z.B.; **Investigation:** M.I.A.C., Z.B., C.H., M.S.; **Resources:** M.H.S., M.S., M.P.; **Data curation:** M.I.A.C., Z.B., M.S., and M.P.; **Writing - Original Draft:** M.I.A.C. Z.B., M.S., and M.P. **Writing - Review and Editing:** M.I.A.C., Z.B., C.H., M.S., and M.P; **Visualization:** M.I.A.C. and Z.B. **Supervision:** M.H.S., M.S., and M.P. **Project administration:** M.P.; **Funding acquisition:** M.H.S., M.S., and M.P.

Funding information

This work was partly supported by R01 GM83098 to M.P. and NIH 1R35GM133774 to M.S..

Competing interest

The authors declare that they have no competing interests.

References

Acquaviva, Laurent, Julie Drogat, Pierre-Marie Dehé, Christophe de La Roche Saint-André, and Vincent Géli. 2013. “Spp1 at the Crossroads of H3K4me3 Regulation and Meiotic Recombination.” *Epigenetics: Official Journal of the DNA Methylation Society* 8 (4): 355–60.

Altemose, Nicolas, Nudrat Noor, Emmanuelle Bitoun, Afidalina Tumian, Michael Imbeault, J. Ross Chapman, A. Radu Aricescu, and Simon R. Myers. 2017. “A Map of Human PRDM9 Binding Provides Evidence for Novel Behaviors of PRDM9 and Other Zinc-Finger Proteins in Meiosis.” *eLife* 6 (October). https://doi.org/10.7554/eLife.28383.

Applied Research Applied Research Press. 2015. *RSEM: Accurate Transcript Quantification from RNA-Seq Data with Or Without a Reference Genome*.

Auton, Adam, Ying Rui Li, Jeffrey Kidd, Kyle Oliveira, Julie Nadel, J. Kim Holloway, Jessica J. Hayward, et al. 2013. “Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs.” *PLoS Genetics* 9 (12): e1003984.

Axelsson, Erik, Matthew T. Webster, Abhirami Ratnakumar, LUPA Consortium, Chris P. Ponting, and Kerstin Lindblad-Toh. 2012. “Death of PRDM9 Coincides with Stabilization of the Recombination Landscape in the Dog Genome.” *Genome Research* 22 (1): 51–63.

Baker, Zachary, Molly Schumer, Yuki Haba, Lisa Bashkirova, Chris Holland, Gil G. Rosenthal, and Molly Przeworski. 2017. “Repeated Losses of PRDM9-Directed Recombination despite the Conservation of PRDM9 across Vertebrates.” *eLife* 6 (June). https://doi.org/10.7554/eLife.24133.

Barker, Daniel, and Mark Pagel. 2005. “Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes.” *PLoS Computational Biology* 1 (1): e3.

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 (5967): 836–40.

Baudat, Frédéric, Yukiko Imai, and Bernard de Massy. 2013. “Meiotic Recombination in Mammals: Localization and Regulation.” *Nature Reviews. Genetics* 14 (11): 794–806.

Birney, Ewan, Michele Clamp, and Richard Durbin. 2004. “GeneWise and Genomewise.” *Genome Research* 14 (5): 988–95.

Brick, Kevin, Fatima Smagulova, Pavel Khil, R. Daniel Camerini-Otero, and Galina V. Petukhova. 2012. “Genetic Recombination Is Directed Away from Functional Genomic Elements in Mice.” *Nature* 485 (7400): 642–45.

Capra, John A., and Mona Singh. 2007. “Predicting Functionally Important Residues from Sequence Conservation.” *Bioinformatics* 23 (15): 1875–82.

Davies, Benjamin, Edouard Hatton, Nicolas Altemose, Julie G. Hussin, Florencia Pratto, Gang Zhang, Anjali Gupta Hinch, et al. 2016. “Re-Engineering the Zinc Fingers of PRDM9 Reverses Hybrid Sterility in Mice.” *Nature* 530 (7589): 171–76.

Diagouraga, Boubou, Julie A. J. Clément, Laurent Duret, Jan Kadlec, Bernard de Massy, and Frédéric Baudat. 2018. “PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites.” *Molecular Cell* 69 (5): 853–65.e6.

Durinck, Steffen, James Bullard, Paul T. Spellman, and Sandrine Dudoit. 2009. “GenomeGraphs: Integrated Genomic Data Visualization with R.” *BMC Bioinformatics* 10 (January): 2.

Eram, Mohammad S., Susan P. Bustos, Evelyne Lima-Fernandes, Alena Siarheyeva, Guillermo Senisterra, Taraneh Hajian, Irene Chau, et al. 2014. “Trimethylation of Histone H3 Lysine 36 by Human Methyltransferase PRDM9 Protein.” *The Journal of Biological Chemistry* 289 (17): 12177–88.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data

without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52.
Gregorova, Sona, Vaclav Gergelits, Irena Chvatalova, Tanmoy Bhattacharyya, Barbora Valiskova, Vladana Fotopulosova, Petr Jansa, Diana Wiatrowska, and Jiri Forejt. 2018. "Modulation of Controlled Meiotic Chromosome Asynapsis Overrides Hybrid Sterility in Mice." *eLife* 7 (March). https://doi.org/10.7554/eLife.34282.
Halldorsson, Bjarni V., Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, et al. 2019. "Characterizing Mutagenic Effects of Recombination through a Sequence-Level Genetic Map." *Science* 363 (6425). https://doi.org/10.1126/science.aau1043.
He, Fahu, Takashi Umehara, Kohei Saito, Takushi Harada, Satoru Watanabe, Takashi Yabuki, Takanori Kigawa, et al. 2010. "Structural Insight into the Zinc Finger CW Domain as a Histone Modification Reader." *Structure* 18 (9): 1127–39.
Hinch, Anjali Gupta, Gang Zhang, Philipp W. Becker, Daniela Moralli, Robert Hinch, Benjamin Davies, Rory Bowden, and Peter Donnelly. 2019. "Factors Influencing Meiotic Recombination Revealed by Whole-Genome Sequencing of Single Sperm." *Science* 363 (6433). https://doi.org/10.1126/science.aau8861.
Huang, Tao, Shenli Yuan, Lei Gao, Mengjing Li, Xiaochen Yu, Jianhong Zhang, Yingying Yin, et al. 2020. "The Histone Modification Reader ZCWPW1 Links Histone Methylation to PRDM9-Induced Double-Strand Break Repair." *eLife* 9 (May). https://doi.org/10.7554/eLife.53459.
Imai, Yukiko, Frédéric Baudat, Miguel Taillepierre, Marcello Stanzione, Attila Toth, and Bernard de Massy. 2017. "The PRDM9 KRAB Domain Is Required for Meiosis and Involved in Protein Interactions." *Chromosoma* 126 (6): 681–95.
Jin, Xiaofan, Geoff Fudenberg, and Katherine S. Pollard. n.d. "Genome-Wide Variability in Recombination Activity Is Associated with Meiotic Chromatin Organization." https://doi.org/10.1101/2021.01.06.425599.
Jung, Min, Daniel Wells, Jannette Rusch, Suhaira Ahmad, Jonathan Marchini, Simon R. Myers, and Donald F. Conrad. 2019. "Unified Single-Cell Analysis of Testis Gene Regulation and Pathology in Five Mouse Strains." *eLife* 8 (June). https://doi.org/10.7554/eLife.43966.
Keeney, S., and M. J. Neale. 2006. "Initiation of Meiotic Recombination by Formation of DNA Double-Strand Breaks: Mechanism and Regulation." *Biochemical Society Transactions* 34 (Pt 4): 523–25.
Koepfli, Klaus-Peter, Benedict Paten, Genome 10K Community of Scientists, and Stephen J. O'Brien. 2015. "The Genome 10K Project: A Way Forward." *Annual Review of Animal Biosciences* 3: 57–111.
Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times." *Molecular Biology and Evolution* 34 (7): 1812–19.
Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." *Molecular Biology and Evolution* 33 (7): 1870–74.
Lam, Isabel, and Scott Keeney. 2014. "Mechanism and Regulation of Meiotic Recombination Initiation." *Cold Spring Harbor Perspectives in Biology* 7 (1): a016634.
Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
Latrille, Thibault, Laurent Duret, and Nicolas Lartillot. 2017. "The Red Queen Model of Recombination Hot-Spot Evolution: A Theoretical Investigation." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372 (1736). https://doi.org/10.1098/rstb.2016.0463.
Li, Miao, Tao Huang, Meng-Jing Li, Chuan-Xin Zhang, Xiao-Chen Yu, Ying-Ying Yin, Chao Liu, et al. 2019. "The Histone Modification Reader ZCWPW1 Is Required for Meiosis Prophase I

in Male but Not in Female Mice." *Science Advances* 5 (8): eaax1101.
Macmanes, Matthew D. 2014. "On the Optimal Trimming of High-Throughput mRNA Sequence Data." *Frontiers in Genetics* 5 (January): 13.
Mahgoub, Mohamed, Jacob Paiano, Melania Bruno, Wei Wu, Sarath Pathuri, Xing Zhang, Sherry Ralls, Xiaodong Cheng, André Nussenzweig, and Todd S. Macfarlan. 2020. "Dual Histone Methyl Reader ZCWPW1 Facilitates Repair of Meiotic Double Strand Breaks in Male Mice." *eLife* 9 (April). https://doi.org/10.7554/eLife.53360.
Marchler-Bauer, A. 2004. "CDD: A Conserved Domain Database for Protein Classification." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki069.
Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, et al. 2015. "CDD: NCBI's Conserved Domain Database." *Nucleic Acids Research* 43 (Database issue): D222–26.
Muñoz-Fuentes, Violeta, Anna Di Rienzo, and Carles Vilà. 2011. "Prdm9, a Major Determinant of Meiotic Recombination Hotspots, Is Not Functional in Dogs and Their Wild Relatives, Wolves and Coyotes." *PloS One* 6 (11): e25498.
Myers, Simon, Rory Bowden, Afidalina Tumian, Ronald E. Bontrop, Colin Freeman, Tammie S. MacFie, Gil McVean, and Peter Donnelly. 2010. "Drive against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination." *Science* 327 (5967): 876–79.
Narasimhan, Manjulaa, Caitlin Payne, Stephanie Caldas, John R. Beard, and Caitlin E. Kennedy. 2016. "Ageing and Healthy Sexuality among Women Living with HIV." *Reproductive Health Matters* 24 (48): 43–51.
O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44 (D1): D733–45.
Oliver, Peter L., Leo Goodstadt, Joshua J. Bayes, Zoë Birtle, Kevin C. Roach, Nitin Phadnis, Scott A. Beatson, Gerton Lunter, Harmit S. Malik, and Chris P. Ponting. 2009. "Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa." *PLoS Genetics* 5 (12): e1000753.
Pagel, Mark. 1994. "Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255 (1342): 37–45.
Pagel, Pagel, and Meade. 2006. "Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo." *The American Naturalist*. https://doi.org/10.2307/3844739.
Parvanov, Emil D., Petko M. Petkov, and Kenneth Paigen. 2010. "Prdm9 Controls Activation of Mammalian Recombination Hotspots." *Science* 327 (5967): 835.
Ponting, Chris P. 2011. "What Are the Genomic Drivers of the Rapid Evolution of PRDM9?" *Trends in Genetics: TIG* 27 (5): 165–71.
Powers, Natalie R., Emil D. Parvanov, Christopher L. Baker, Michael Walker, Petko M. Petkov, and Kenneth Paigen. 2016. "The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo." *PLoS Genetics* 12 (6): e1006146.
Qin, Su, and Jinrong Min. 2014. "Structure and Function of the Nucleosome-Binding PWWP Domain." *Trends in Biochemical Sciences* 39 (11): 536–47.
Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.
Schield, Drew R., Giulia I. M. Pasquesi, Blair W. Perry, Richard H. Adams, Zachary L. Nikolakis, Aundrea K. Westfall, Richard W. Orton, Jesse M. Meik, Stephen P. Mackessy, and Todd A. Castoe. 2020. "Snake Recombination Landscapes Are Concentrated in Functional Regions

despite PRDM9." *Molecular Biology and Evolution* 37 (5): 1272–94.
Schmieder, Robert, and Robert Edwards. 2011. "Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets." *PloS One* 6 (3): e17288.
Sievers, Fabian, and Desmond G. Higgins. 2014. "Clustal Omega." *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi0313s48.
Singhal, Sonal, Ellen M. Leffler, Keerthi Sannareddy, Isaac Turner, Oliver Venn, Daniel M. Hooper, Alva I. Strand, et al. 2015. "Stable Recombination Hotspots in Birds." *Science* 350 (6263): 928–32.
Smith, Stacey D., Matthew W. Pennell, Casey W. Dunn, and Scott V. Edwards. 2020. "Phylogenetics Is the New Genetics (for Most of Biodiversity)." *Trends in Ecology & Evolution*. https://doi.org/10.1016/j.tree.2020.01.005.
Spence, Jeffrey P., and Yun S. Song. 2019. "Inference and Analysis of Population-Specific Fine-Scale Recombination Maps across 26 Diverse Human Populations." *Science Advances* 5 (10): eaaw9206.
Thibault-Sennett, Sarah, Qi Yu, Fatima Smagulova, Jeff Cloutier, Kevin Brick, R. Daniel Camerini-Otero, and Galina V. Petukhova. 2018. "Interrogating the Functions of PRDM9 Domains in Meiosis." *Genetics* 209 (2): 475–87.
Tian, Hui, Timothy Billings, and Petko M. Petkov. 2018. "CXXC1 Is Not Essential for Normal DNA Double-Strand Break Formation and Meiotic Recombination in Mouse." *PLOS Genetics*. https://doi.org/10.1371/journal.pgen.1007657.
Úbeda, F., and J. F. Wilkins. 2011. "The Red Queen Theory of Recombination Hotspots." *Journal of Evolutionary Biology*. https://doi.org/10.1111/j.1420-9101.2010.02187.x.
Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2018. "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics." *Molecular Biology and Evolution* 35 (3): 543–48.
Wells, Daniel, Emmanuelle Bitoun, Daniela Moralli, Gang Zhang, Anjali Hinch, Julia Jankowska, Peter Donnelly, Catherine Green, and Simon R. Myers. 2020. "ZCWPW1 Is Recruited to Recombination Hotspots by PRDM9, and Is Essential for Meiotic Double Strand Break Repair." *eLife* 9 (August). https://doi.org/10.7554/eLife.53392.
Wu, Hong, Nikolas Mathioudakis, Boubou Diagouraga, Aiping Dong, Ludmila Dombrovski, Frédéric Baudat, Stephen Cusack, Bernard de Massy, and Jan Kadlec. 2013. "Molecular Basis for the Regulation of the H3K4 Methyltransferase Activity of PRDM9." *Cell Reports* 5 (1): 13–20.
Yang, Fang, Sigrid Eckardt, N. Adrian Leu, K. John McLaughlin, and Peijing Jeremy Wang. 2008. "Mouse TEX15 Is Essential for DNA Double-Strand Break Repair and Chromosomal Synapsis during Male Meiosis." *The Journal of Cell Biology* 180 (4): 673–79.
Yang, Ziheng. 1997. "PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/13.5.555.
Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91.

Supplementary tables

Supplementary Table 1. Genes identified in this study. For *PRDM9* and each candidate gene that was initially identified as significantly coevolving with *PRDM9* (*ZCWPW1*, *ZCWPW2*, *TEX15*, *ANKRD31* and *MEI1*), we provide a table detailing, for each ortholog that we identified, which species it is from, how we identified it, its inferred domain architecture, its amino acid sequence, as well as various details about these domains (including their coordinates, e-values, and sequences). For *PRDM9* orthologs, we additionally report which amino acid residues align to three catalytic tyrosine residues in the human SET domain.

Supplementary Table 2. Description of genes found from whole genome sequences. For *PRDM9* and each gene initially identified as significantly coevolving with *PRDM9* (*ZCWPW1*, *ZCWPW2*, *TEX15*, *ANKRD31* and *MEI1*), we provide a table detailing where and how each ortholog identified in our analysis of whole genome assemblies was obtained.

Supplementary Table 3. Description of species for which a *de novo* assembly of testis transcriptomes was generated in order to verify the structure and expression of PRDM9 and four significant genes (*ZCWPW1*, *ZCWPW2*, *TEX15* and *ANKRD31*). To this end, we used publicly available RNA-seq data (downloaded from NCBI) (NCBI Resource Coordinators 2018) and in a subset of cases indicated with a star, generated our own data which are available from the NCBI sequence read archive (Bioproject PRJNA605699, SRA accessions: SRR11050679-SRR11050687, see Methods for further details).

Supplementary Table 4. The distribution of *PRDM9* orthologs across 379 vertebrate species. For each species, we describe how many *PRDM9* orthologs we identified of each unique domain architecture, the domain architecture of the most complete *PRDM9* ortholog from that species, and whether any *PRDM9* ortholog from that species with the most complete domain architecture has conserved three catalytic tyrosine residues in the SET domain. We additionally include columns comparing these results to those previously described in Baker et al. 2017, noting instances where we have revised our calls of domain architecture.

Supplementary Table 5. Description of candidate genes used in the phylogenetic tests. The 241 genes selected for the tests were based on three different sources: (1) the top 1 percent of genes most highly expressed in the same component as PRDM9 in mouse testis single cell analyses (Jung et al. 2019), (2) genes associated with variation in recombination phenotypes in humans (Halldorsson et al. 2019), and (3) genes known to have a role in mammalian meiotic recombination from functional studies (as summarized in the review by (Baudat, Imai, and de Massy 2013)). The genomic coordinates (column named ‘Start position’) of each gene were based on the GRCh38/hg38 human reference. The function of each candidate gene is described based on the definition from its source.

Supplementary Table 6. Presence and absence matrix computed for all candidate genes used in phylogenetic tests for coevolution with *PRDM9* (139 genes). We defined a gene as complete

("1") when it contained all the domains observed in four representative vertebrate species with a complete *PRDM9* sequence, and incomplete ("0") if the gene was not detected in the Refseq database or if it did not include all the domains shared across four species (see Methods for details).

Supplementary Table 7. Phylogenetic tests and p-values. P-values were computed by evaluating the patterns of presence or absence of *PRDM9* across 189 vertebrates against the patterns of presence or absence of candidate genes. Two models were tested using BayestraitsV3 (Barker and Pagel 2005): a null model in which *PRDM9* and a given candidate gene evolve independently of one another along the phylogeny versus an alternative model in which the gain ("1") and loss ("0") of the candidate gene is dependent on the status of *PRDM9* and vice versa. See Pagel, 1994 and the BayesTraitsV3.0.2 manual for further discussion of these models and rates description.

Supplementary Table 8. The distribution of PRDM9 and four genes initially found to be significantly coevolving with *PRDM9* (*ANKRD31*, *TEX15*, *ZCWPW1* and *ZCWPW2*) across 189 vertebrate species**.** *MEI1* is excluded here because it was found to be present in all species. (**A**) Curated calls for the presence or absence of complete genes based on searches of RefSeq, whole genome assemblies, and RNA-seq data (see **Tables S1-S3**). We additionally include the most complete domain architecture of orthologs from each species for each gene. (**B**) Summary of losses inferred for *PRDM9*, *ZCWPW1*, *ZCWPW2*, *TEX15* and *ANKRD31* among the 189 vertebrate species used in our co-evolutionary test.

Supplementary Table 9. Results of phylogenetic tests when considering the SSXRD domain in *PRDM9* classification. P-values were computed by evaluating the patterns of presence or absence of *PRDM9* across 189 vertebrates against the patterns of presence or absence of candidate genes. Two models were tested using BayestraitsV3 (Barker and Pagel 2005): a null model in which *PRDM9* and a given candidate gene evolve independently of one another along the phylogeny versus an alternative model in which the gain ("1") and loss ("0") of a gene is dependent on the status of *PRDM9* and vice versa. See Pagel, 1994 and the BayesTraitsV3.0.2 manual for further discussion of these models and rates description.

Supplementary Table 10. Tests for differences in the rates of amino acid evolution in three significant genes (*ZCWPW1*, *ZCWPW2*, *TEX15*) between representative species with and without a PRDM9 ortholog. To determine whether species lacking a *PRDM9* ortholog showed evidence for relaxed selection pressures in co-evolving genes, we estimated ω(dN/dS) using the Branch model within PAML (Yang 1997) under two models: a null model assuming the same ω across all branches of the phylogeny, and an alternative model in which there are two ω values allowed: one ω value in species lacking a functional PRDM9 and a second ω for the rest of the branches. The clades evaluated in each test are specified. The species used in the alignment for each test are also shown. The log likelihoods for each model, ω estimates and p-values are also provided. See Methods for details.

Supplementary Table 11. Testing the direction of dependency between *PRDM9* and *TEX15*, *ZCWPW1* and *ZCWPW2*. Here, we asked whether we could reject a model of independent state transitions of *PRDM9* and a given candidate gene (e.g. *TEX15*) in favor of a model in which state transitions of the candidate gene depend on those of *PRDM9* (model X). Next, we asked whether we could reject the null model in favor of a model in which the state transitions of *PRDM9* depend on those of the candidate gene (model Y). For comparison, we also provide results for the test shown in the main text, in which the alternative considered is that state transitions of *PRDM9* depend on those of the candidate gene and vice versa (also shown in Table 1). See Pagel, 1994 and the BayesTraitsV3.0.2 Manual for further discussion of these models and rates description.

Supplementary figures

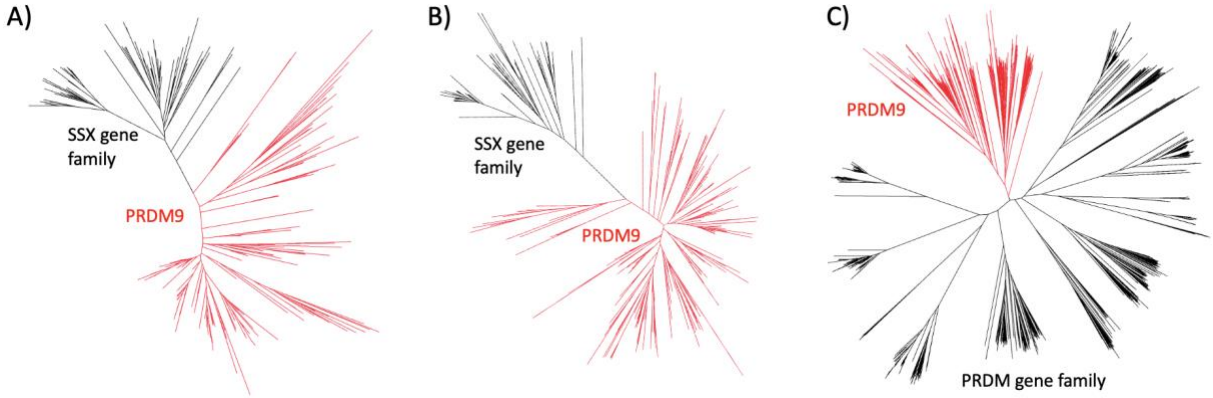

Figure S1. Guide trees created from our initial blastp search results for PRDM9 orthologs for **(A)** KRAB domains, **(B)** SSXRD domains and **(C)** SET domains. Genes were removed if they clustered with SSX genes in trees **(A)** or **(B)**, or if they clustered with PRDM gene family genes other than *PRDM9* or *PRDM7* in the tree **(C)**. Genes clustering with *PRDM9* and retained for subsequent analysis are shown in red.

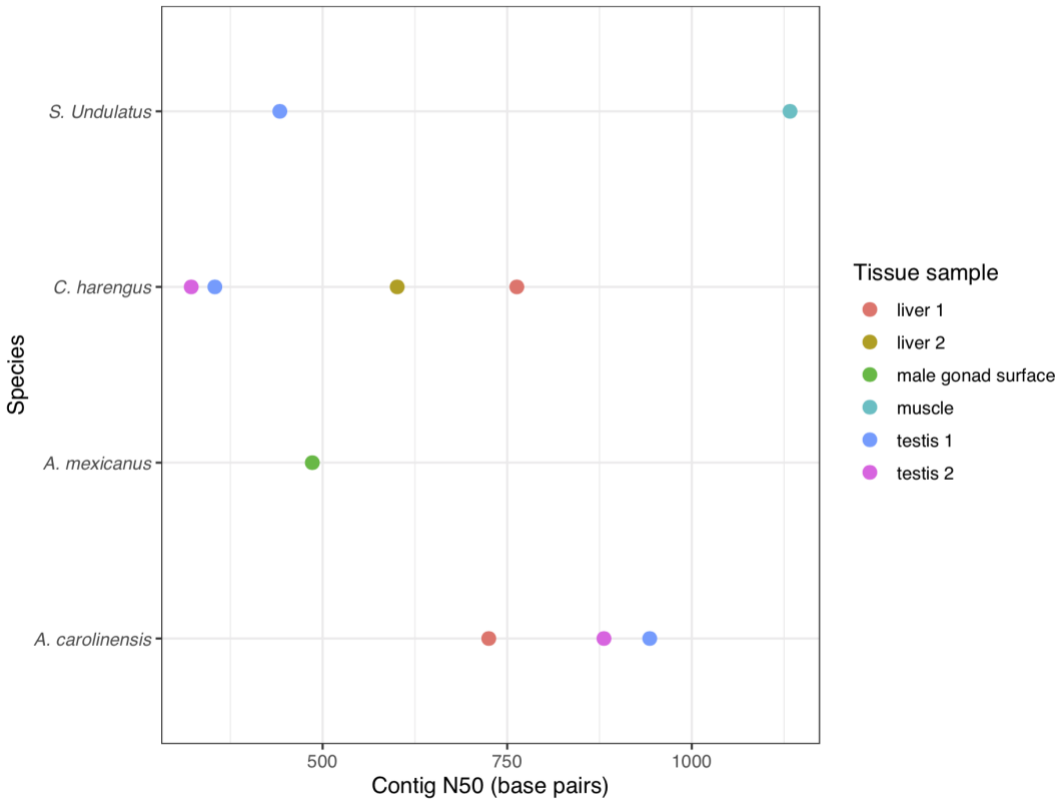

Figure S2. Contig N50 in base pairs as a statistic describing the quality of *de novo* transcriptome assemblies. Colors represent the different tissues used in the two lizard species (*S. undulatus* and *A. carolinensis*) and two fish species (*C. harengus* and *A. mexicanus*).

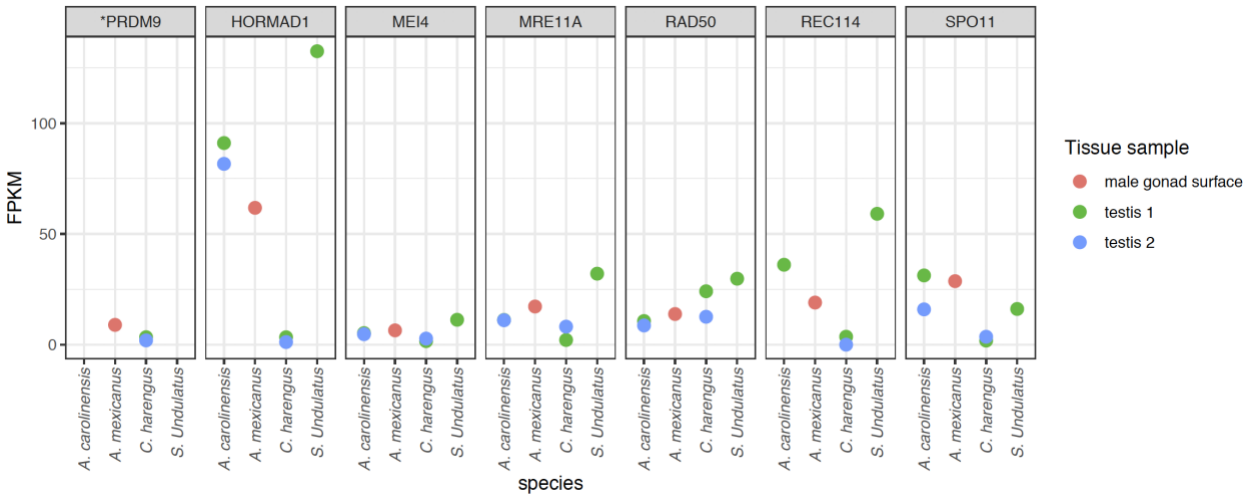

Figure S3. Expression levels of six core meiosis-related genes (Baudat, Imai, and de Massy 2013) across species and tissues. The y-axis corresponds to fragments per kilobase of transcript per million mapped reads (FPKM). Despite evidence for expression of the other six core meiotic genes, PRDM9 expression is not detected in *S. undulatus* and *A. carolinensis*.

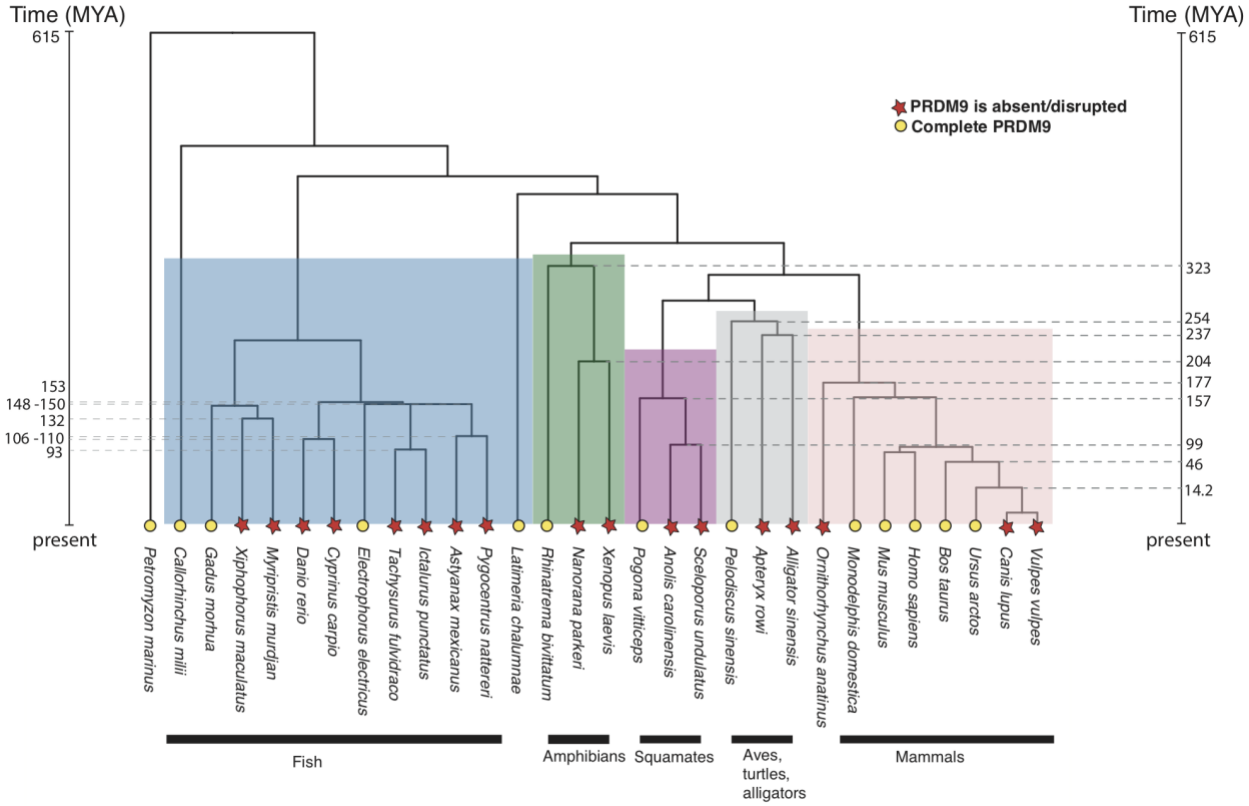

Figure S4. Phylogenetic distribution of *PRDM9* orthologs in vertebrates, using the phylogenetic tree and divergence dates obtained from Timetree (Kumar et al. 2017). A complete *PRDM9* was found in species marked with yellow circles. Species marked with a red star are ones for which we were unable to identify PRDM9 expression in testis samples. The highlighted dates indicate the inferred timing of the multiple *PRDM9* losses. Dashed lines indicate the earliest and latest possible dates for each loss of *PRDM9*. The minimum date reflects the time to the most recent common ancestor amongst species without PRDM9, whereas the 'maximum date' is the time to the first common ancestor between species without *PRDM9* and the most closely related species with *PRDM9*. The most recent loss of *PRDM9* occurred either in the branch leading to canids, between 14.2 and 46 million years ago (Mya), or potentially the branch leading to platypus.

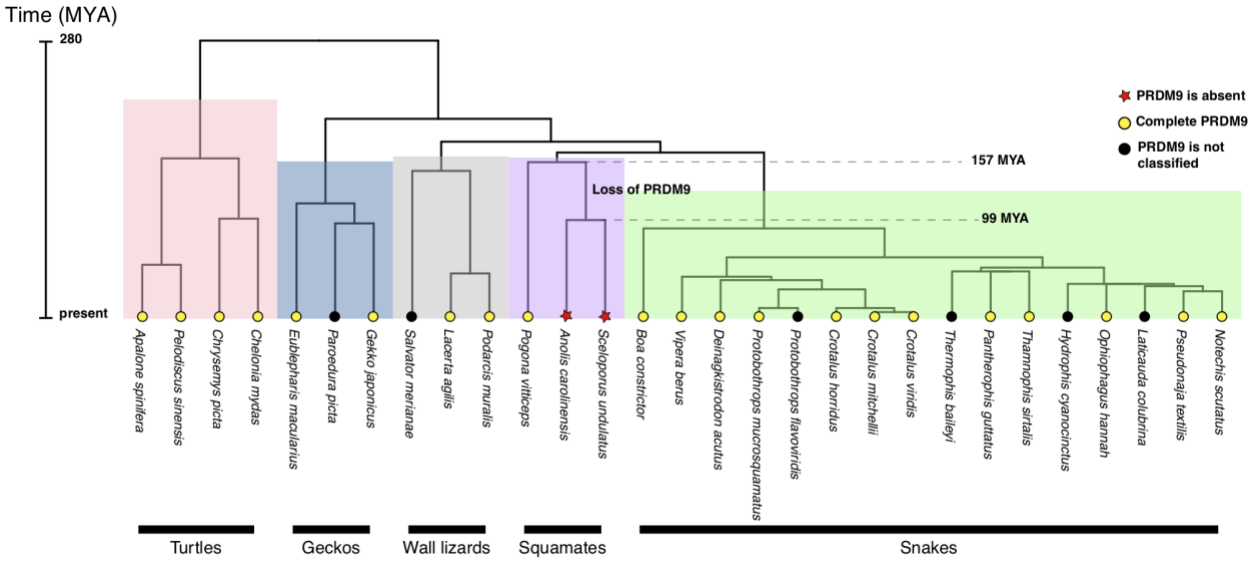

Figure S5. Phylogenetic distribution of *PRDM9* orthologs in reptiles, using the phylogenetic tree and divergence dates obtained from Timetree (Kumar et al. 2017). Species assigned with yellow circles were previously found to have a complete *PRDM9* (Baker et al. 2017). Species indicated with a red star are ones for which we were unable to identify PRDM9 expression in testis samples. Species indicated with a black circle are species for which Refseq is not available and *PRDM9* classification was therefore not conducted. Based on the phylogenetic relationship between *Anolis carolinensis* and *Sceloporus undulatus*, the *PRDM9* loss shared by these two species likely occurred between 99 and 157 million years ago (Mya).

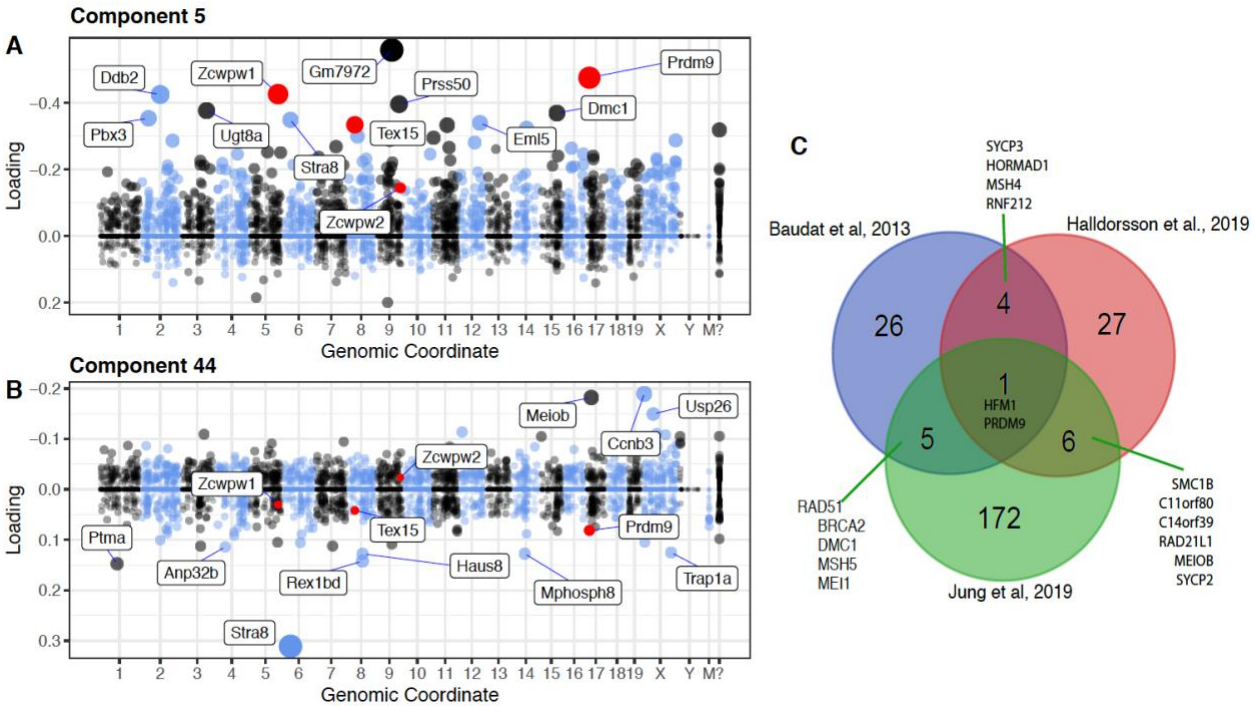

Figure S6. Meiosis-specific candidate genes. (Jung et al. 2019) inferred 46 principal components from single cell expression patterns during mouse spermatogenesis, which are thought to loosely correspond to regulatory programs. Shown in A-B are the two components in which PRDM9 is most highly expressed. The dot sizes are proportional to the square of the absolute value of the loading, so are indicative of higher expression within each component. PRDM9 and the five genes with $p<0.05$ in our phylogenetic analysis are shown in red. Mouse genomic coordinates are displayed. (**A**) Component 5 is the one in which PRDM9 has its highest loading; it is associated with double strand break formation and active during (pre)leptotene (Jung et al. 2019). (**B**) Component 44 is the component in which PRDM9 has its second highest loading; this component is active during zygotene (Jung et al., 2019). (**C**) Intersection of candidate genes from three sources: (i) the top 1 percent of genes most highly expressed in component 5 (ii) genes associated with variation in recombination phenotypes in humans (Halldorsson et al. 2019) and (iii) genes known to have a role in mammalian meiotic recombination from functional studies (as summarized in the review by (Baudat, Imai, and de Massy 2013)).

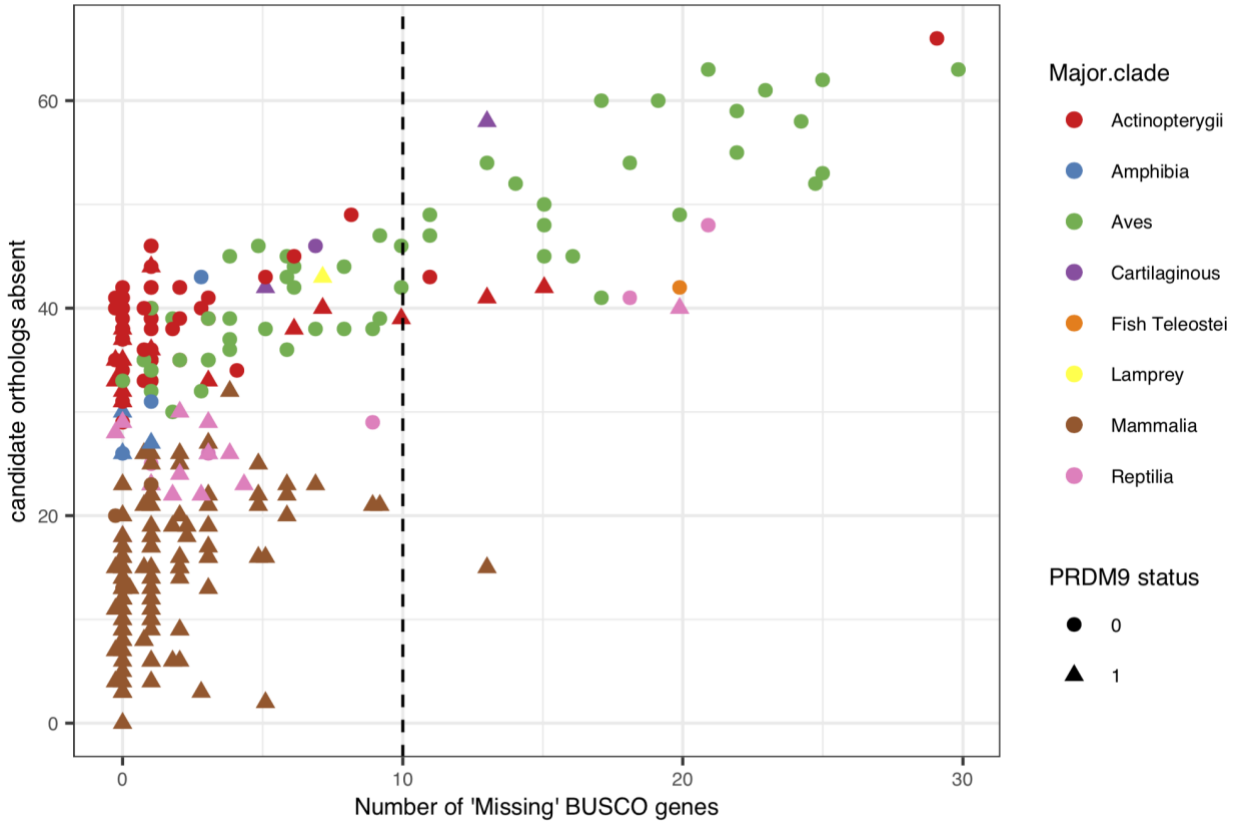

Figure S7. The relationship between the number of candidate genes that were absent in a genome assembly and the number of 'Missing' BUSCO genes (Waterhouse et al. 2018) for that assembly, across species. BUSCO statistics were computed for the genomes of 339 species. The relationship is significant (Spearman's rank correlation ρ = 0.5, p-value < 2.2e-16), suggesting that orthologs of candidate genes of interest might be missed in genomes with low BUSCO scores. In the phylogenetic tests, we therefore considered only species with 10 or fewer missing BUSCO genes (dashed line), leading 32 species to be excluded.

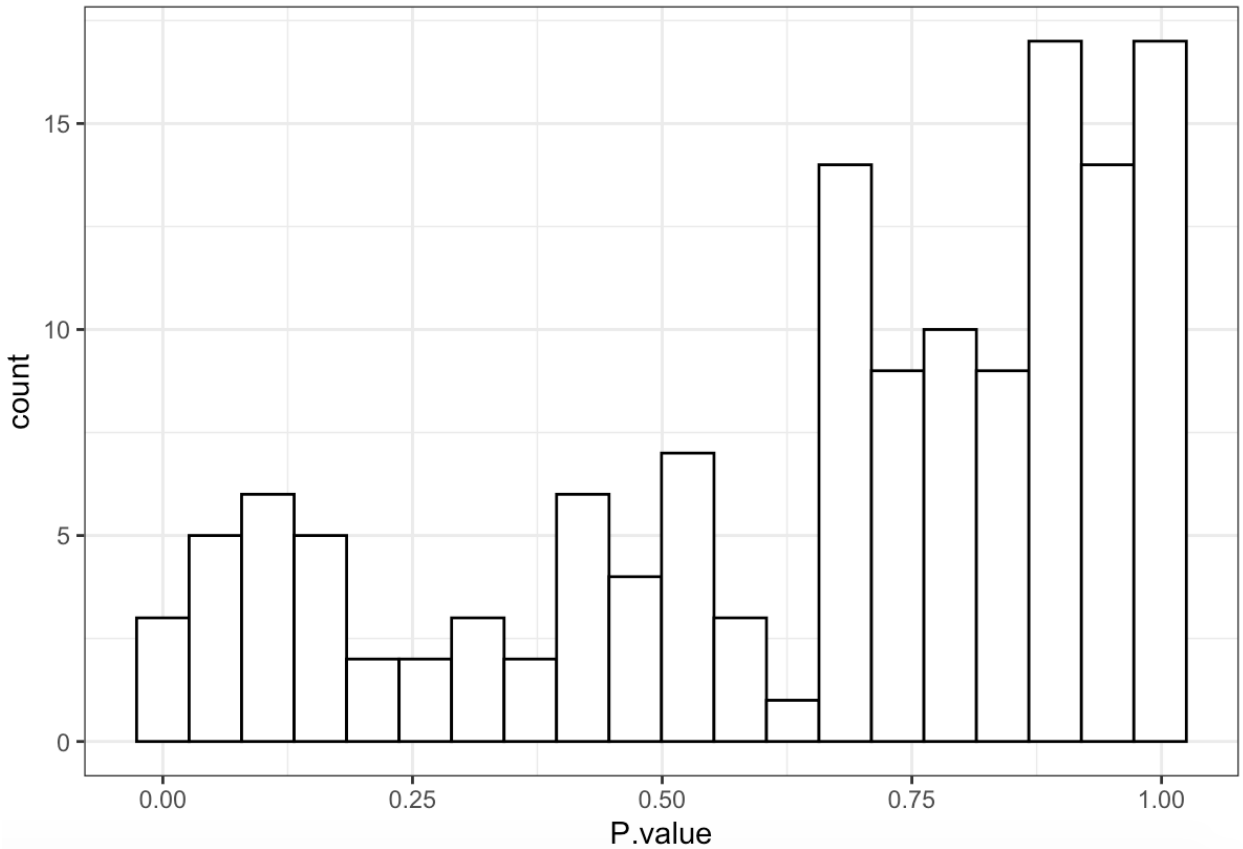

Figure S8. The distribution of p-values obtained across the 131 genes included in phylogenetic tests (individual p-values are available in **Table S7**).

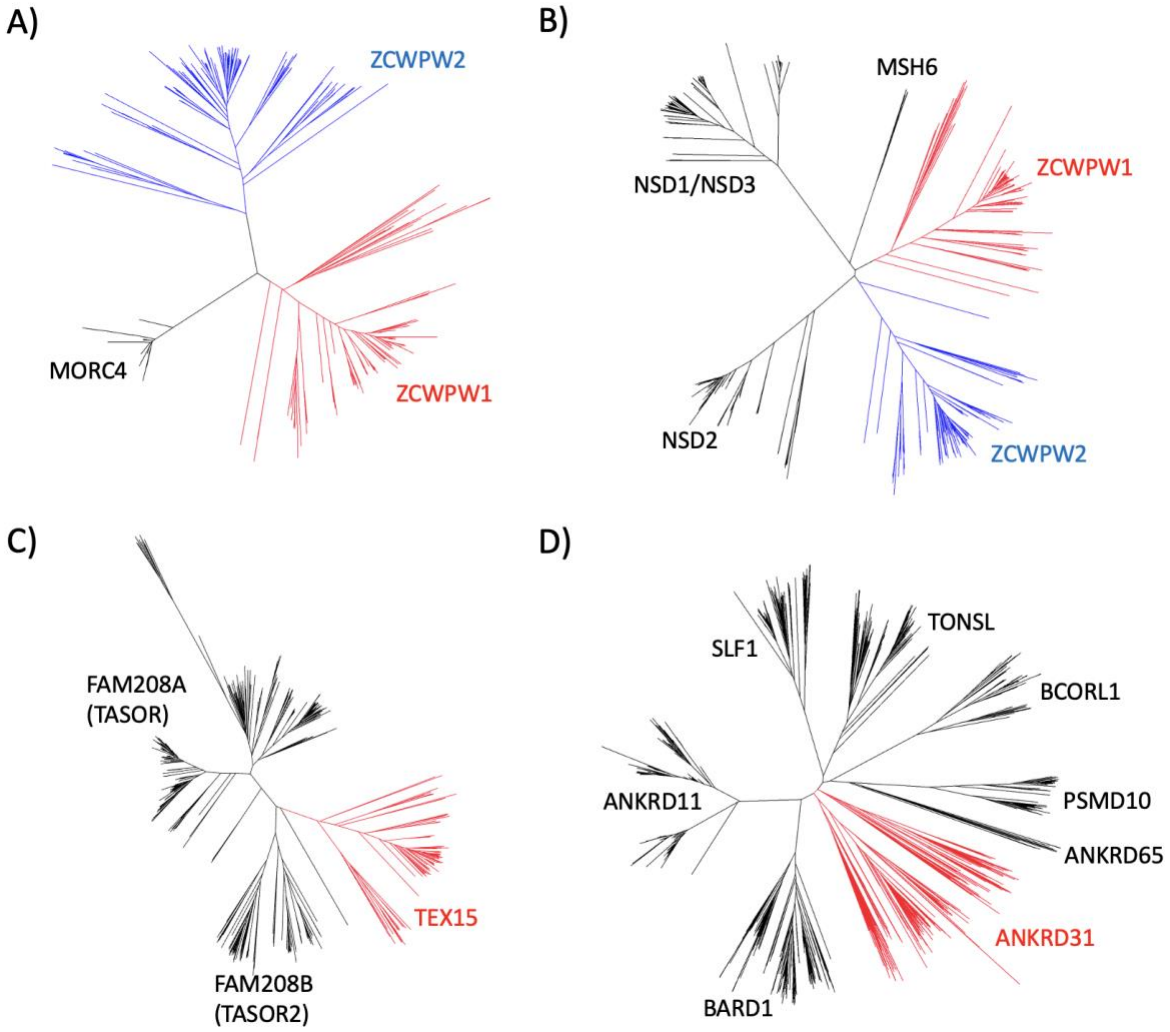

Figure S9. Guide trees created from our initial blastp search results for the zf-CW (**A**) and PWWP (**B**) domains of *ZCWPW1* and *ZCWPW2* orthologs, the DUF3715 domain of *TEX15* orthologs (**C**), and the ANKYR domain of *ANKRD31* orthologs (**D**). Genes were removed if they clustered with *MORC4* in tree **A**, *MSH6*, *NSD1*, *NSD2*, or *NSD3* in tree **B**, *FAM208A* or *FAM208B* in tree **C**, or *ANKRD11*, *ANKRD65*, *BARD1*, *BCORL1*, *PSMD10*, *SLF1*, or *TONSL* in tree **D**. Genes clustering with *ZCWPW1*, *ZCWPW2*, *TEX15* or *ANKRD31* and retained for subsequent analysis are shown in red or blue.

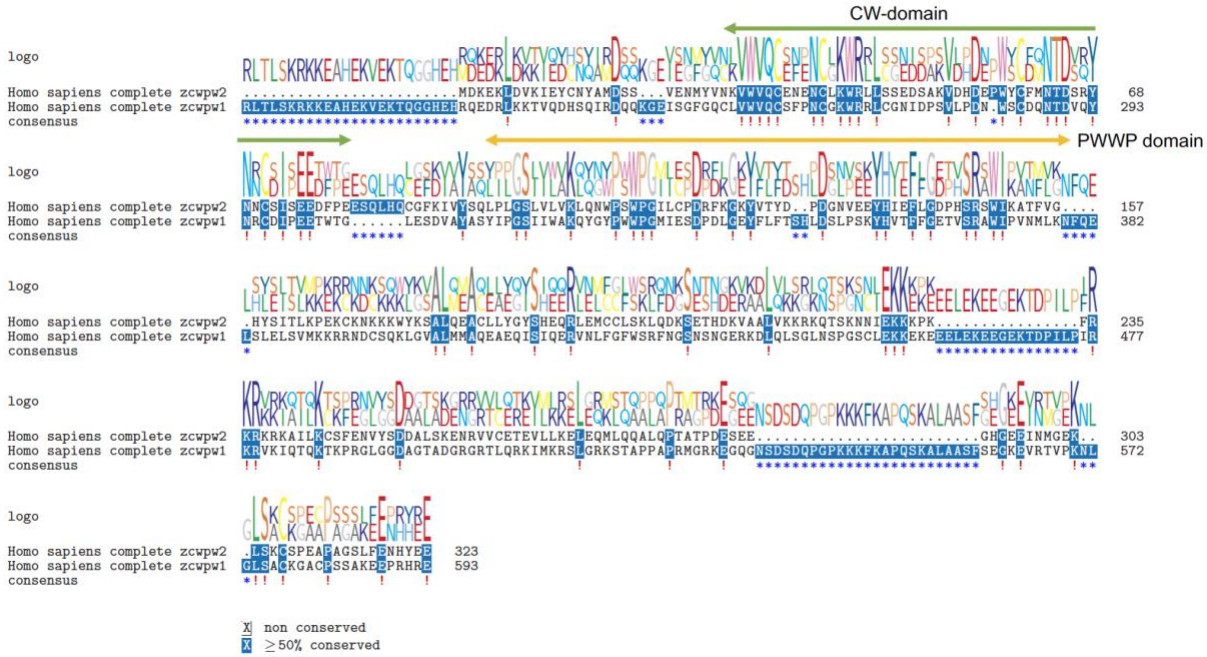

Figure S10: Amino acid sequence alignment between ZCWPW1 and ZCWPW2 proteins from humans. Superfamily domains are marked. In mice, the CW-domain (green arrow) recognizes different methylated states of lysine 4 on histone H3 (H3K4me) (He et al. 2010), while the PWWP domain (yellow arrow) recognizes methylated H3K36 histone tail (Qin and Min 2014). The SET domain of PRDM9 tri-methylates both histones H3K4 and H3K36 (Wu et al. 2013).

References Supplementary Figures

Baker, Zachary, Molly Schumer, Yuki Haba, Lisa Bashkirova, Chris Holland, Gil G. Rosenthal, and Molly Przeworski. 2017. “Repeated Losses of PRDM9-Directed Recombination despite the Conservation of PRDM9 across Vertebrates.” eLife 6 (June). https://doi.org/10.7554/eLife.24133.

Barker, Daniel, and Mark Pagel. 2005. “Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes.” PLoS Computational Biology 1 (1): e3.

Baudat, Frédéric, Yukiko Imai, and Bernard de Massy. 2013. “Meiotic Recombination in Mammals: Localization and Regulation.” Nature Reviews. Genetics 14 (11): 794–806.

Halldorsson, Bjarni V., Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, et al. 2019. “Characterizing Mutagenic Effects of Recombination through a Sequence-Level Genetic Map.” Science 363 (6425). https://doi.org/10.1126/science.aau1043.

He, F., Y. Muto, M. Inoue, T. Kigawa, M. Shirouzu, T. Terada, S. Yokoyama, and RIKEN Structural Genomics/Proteomics Initiative (RSGI). 2010. “Complex Structure of the Zf-CW Domain and the H3K4me3 Peptide.” https://doi.org/10.2210/pdb2rr4/pdb.

Jung, Min, Daniel Wells, Jannette Rusch, Suhaira Ahmad, Jonathan Marchini, Simon R. Myers, and Donald F. Conrad. 2019. “Unified Single-Cell Analysis of Testis Gene Regulation and Pathology in Five Mouse Strains.” eLife 8 (June). https://doi.org/10.7554/eLife.43966.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” Molecular Biology and Evolution 34 (7): 1812–19.

NCBI Resource Coordinators. 2018. “Database Resources of the National Center for Biotechnology Information.” Nucleic Acids Research 46 (D1): D8–13.

Pagel, Mark. 1994. “Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters.” Proceedings of the Royal Society of London. Series B: Biological Sciences 255 (1342): 37–45.

Qin, Su, and Jinrong Min. 2014. “Structure and Function of the Nucleosome-Binding PWWP Domain.” Trends in Biochemical Sciences 39 (11): 536–47.

Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2018. “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.” Molecular Biology and Evolution 35 (3): 543–48.

Wu, Hong, Nikolas Mathioudakis, Boubou Diagouraga, Aiping Dong, Ludmila Dombrovski, Frédéric Baudat, Stephen Cusack, Bernard de Massy, and Jan Kadlec. 2013. “Molecular Basis for the Regulation of the H3K4 Methyltransferase Activity of PRDM9.” Cell Reports 5 (1): 13–20.

Yang, Ziheng. 1997. “PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood.” Bioinformatics. https://doi.org/10.1093/bioinformatics/13.5.555.