PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2

Maria Izabel A. Cavassim, Zachary Baker, Carla Hoge, Mikkel H. Schierup, Molly Schumer, and Molly Przeworski

Edited by Scott Edwards, Harvard University, Cambridge, MA; received August 4, 2021; accepted January 16, 2022

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 446 vertebrate species, inferring at least 13 independent losses. We then use the interdigitation of PRDM9 orthologs across vertebrates to test whether it coevolved with any of 241 candidate genes coexpressed with PRDM9 in mice or associated with recombination phenotypes in mammals. Accounting for the phylogenetic relationship among a subsample of 189 species, we find two genes whose presence and absence is unexpectedly coincident with that of PRDM9: ZCWPW1, which was recently shown to facilitate double strand break repair, and its paralog ZCWPW2, as well as, more tentatively, TEX15 and FBXO47. ZCWPW2 is expected to be recruited to sites of PRDM9 binding; its tight coevolution with PRDM9 across vertebrates suggests that it is a key interactor within mammals and beyond, with a role either in recruiting the recombination machinery or in double strand break repair.

Significance

We take a phylogenetic approach to search for molecular partners of PRDM9, a key meiotic recombination gene, by leveraging the fact that the complete PRDM9 gene has been lost at least 13 times independently in vertebrates. We identify two genes, ZCWPW1 and its paralog ZCWPW2, whose presence or absence across vertebrates is coupled to that of PRDM9. ZCWPW1 was recently shown to be recruited to sites of PRDM9 binding and to aid in the repair of double strand breaks. ZCWPW2 is likely recruited to sites of PRDM9 binding as well; its tight coevolution with PRDM9 across vertebrates suggests that it too plays an important role in mammals and beyond, either in double strand break repair or repair.
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 (11, 22). Therefore, the KRAB domain at least appears to be necessary for PRDM9 to direct recombination, likely by mediating interactions with other proteins (22, 23).

Conversely, the presence of a complete PRDM9 with a rapidly evolving ZF outside of mammals (11) suggests that PRDM9 also directs recombination to the genome in these species, as has been reported for rattlesnakes (24). 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 (25–27). 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 (27, 28). 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 (29, 30).

Although the role of PRDM9 in DSB repair is still poorly understood, recent papers report that it is facilitated by ZCWPW1, which binds H3K4me3 and H3K36me3 (25–27) and is expressed alongside PRDM9 in single-cell data from mouse testes (31). One line of evidence that led to the discovery of ZCWPW1 is its phylogenetic distribution: although it too has been lost numerous times in vertebrates, it is found in seven clades that carry an intact PRDM9 (26, 27). Thus, while the distribution of the two genes across species is not perfectly concordant, it is strongly suggestive of coevolution.

The example of ZCWPW1 highlights the potential power of coevolutionary tests to identify additional molecular partners of PRDM9. While not all partners of PRDM9 will coevolve with it, genes whose losses are coupled to those of PRDM9 are strong candidates for molecular interactors. We therefore applied this approach systematically; we considered a set of 241 candidate genes that are known to be involved in recombination in model organisms (32), associated with recombination phenotypes in a human genome-wide association study (33), or coexpressed with PRDM9 in single-cell data from mouse testes (31) and tested for their cooccurrence with PRDM9 across 189 vertebrate species. After verifying our initial gene status calls in whole-genome data and, for a subset of species, with RNA-seq data, we identified the paralog of ZCWPW1, ZCWPW2, as coevolving with PRDM9 and found more tentative evidence for two additional genes, TEX15 and FBXO47.

Results

A Revised Phylogeny of PRDM9. We previously reported that the complete PRDM9 gene, including the 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; Ref. 11). Here, we leverage the independent losses of PRDM9 in order to identify genes that are coevolving 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 (11). To this end, we created a curated data-set of 747 vertebrate PRDM9 sequences by analyzing publicly available protein sequences from RefSeq (34), whole-genome sequences, and RNA-seq data from testes samples, as well as four RNA-seq datasets from testes samples that we generated (see Methods; SI Appendix, Fig. S1 and Tables S1–S3). For this analysis, we defined PRDM9 orthologs as complete if they contain both the 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 nor the ZF array because its repetitive structure makes it difficult to sequence and assemble reliably.

Across 446 species, we identified 221 species with at least one complete PRDM9 ortholog and 225 species without a complete PRDM9 ortholog (Fig. 1; SI Appendix, 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 (11), including in the Tasmanian devil (Sarcophilus harrisii), the Atlantic cod (Gadus morhua), and the Atlantic herring (Clupea harengus), as well as in a handful of eutherian mammals that we had previously only investigated using RefSeq (SI Appendix, Table S4). We also found a complete PRDM9 ortholog in caecilians and in two species of frogs, suggesting that the previously reported loss of PRDM9 in amphibians (11) reflects at least one loss in salamanders and more than one independent loss in frogs. We note, finally, that by the approach taken here, the PRDM9 ortholog from the Australian ghostshark (Callorhinchus miliaris) is considered to be complete (uncontrast to in Baker et al., Ref. 11, where we also noted (on the SSXRD domain; SI Appendix, Table S4). Confirming our earlier finding (11, 15–17), there is a near-perfect correspondence between species carrying putatively complete PRDM9 orthologs (i.e., with KRAB domains) and those for which we identify rapidly evolving PRDM9 ZF arrays (Methods; SI Appendix, Tables S1 and S4).

Based on the phylogenetic relationships among species given by the TimeTree tool (http://timetree.org; Ref. 35), we inferred 23 putative complete or partial losses of PRDM9 across the 446 vertebrates considered (Fig. 1; SI Appendix, Table S4). These putative losses include six previously reported ones (11, 13–15), each observed in two or more closely related species: in percomorph fish, cypriniformes fish, characiformes and siluriformes fish, osteoglossomorpha fish, birds and crocodiles, and canids. In turn, the putative losses of PRDM9 in polypteriformes fish, salamanders, and three clades of frog species (Xenopus, Dicroglossidae, and Bufonidae) were each supported by the absence of PRDM9 in the genomes of two or more closely related species. We were further able to verify the absence of PRDM9 in two Xenopus frogs and in two salamanders using RNA-seq data from testes: despite sufficient power to detect a set of six highly conserved meiotic genes in each species, we did not detect the expression of any complete PRDM9 orthologs (SI Appendix, Table S3).

We also failed to find PRDM9 in ReSeq or the whole-genome sequence of the green anole (Anolis carolinensis). We verified this absence of PRDM9 by collecting RNA-seq data from testes in the green anole as well as in the fence lizard (Sceloporus undulatus), for which neither a ReSeq nor a genome sequence were available at the time, and did not detect PRDM9 expression in either species (SI Appendix, Figs. S2 and S3 and Table S3). Given the presence of a complete PRDM9 in bearded dragons (Pogona vicina), it appears that this loss of PRDM9 occurred in a lineage basal to the common ancestor of green anoles and fence lizards, over 99 Mya but less than 157 Mya (SI Appendix, Fig. S4).

The remaining 11 putative absences of PRDM9 are observed in single species; we were able to verify the call using testis RNA-seq data for the platypus (Ornithorhynchus anatinus), but not for the remaining 10 species, so their PRDM9 status remains uncertain. Thus, in total, we identified at least 13 independent PRDM9 losses in vertebrates, and possibly as many as 23 (Fig. 1; SI Appendix, Table S4). The 13 losses occur in five clades of ray-finned fish (percomorphs, cypriniformes, characiformes and siluriformes, osteoglossomorphs, and polypteriformes), in four clades...
of amphibians (Xenopus, Dicroglossidae, and Bufonidae frogs and salamanders), in two clades of reptiles/birds (birds and crocodiles and the clade of lizards comprised of anoles and fence lizards), and in two clades of mammals (platypus and canids). At least 12 of the 13 losses are relatively old (SI Appendix, Fig. S5); the most recent case manifest in these data is either the one that happened in the branch leading to platypus (sometime in the last 46 Mya) or the one in canids, which could be as recent as 14.2 Mya (SI Appendix, Fig. S5).

Identifying Genes Coevolving With PRDM9. We selected 193 candidate genes based on their coexpression with PRDM9 in single-cell RNA-seq data from mouse testes (specifically, in component 5; Methods; Ref. 31) (SI Appendix, Fig. S6 A and B). To this set, we added any gene associated with variation in recombination phenotypes in humans (33) as well as genes known to have a role in mammalian meiotic recombination from functional studies (summarized in Ref. 32). Together, these three sources provided a total of 241 genes to evaluate for possible coevolution with PRDM9 (SI Appendix, Table S5 and Fig. 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 driven by variation in genome quality (Methods; SI Appendix, Fig. S7). The phylogeny includes representative species for 11 of the 13 inferred PRDM9 losses (Methods). Species of
**Bufo**nidae frogs and salamanders were not included because of the absence of available gene annotations; moreover, because of the lack of gene annotations for frog species with PRDM9, within these 189 species, the losses in *Xenopus* and *Dicroglossidae* frogs cannot be distinguished from a single event. 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 (Methods). Many of the 241 genes are present in every sampled vertebrate and hence provide no information in our coevolutionary 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; the matrix of 189 × 139 gene status calls is presented in SI Appendix, Table S6.

We tested for the coevolution 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 model in which the state transition rates of the two genes are dependent on one another, using the maximum likelihood approach within Bayeastraisis3 (36, 37) (SI Appendix, Table S7 and Fig. S8). By this approach, we identified nine significant hits at the 5% level (uncorrected for multiple tests); in order of increasing *P* values, ZCWPW1, MEI1, ZCWPW2, TEX15, FBX047, ANKR3D1, NEKB1L1, SYCE1, and FMR1NB. We focused on the top five, for which the false discovery rate (FDR) value is below 50% (Table 1; Fig. 2A).

We sought to verify the phylogenetic distribution of the top five genes by developing curated datasets of high-confidence orthologs, as we had for PRDM9 (Methods; Fig. 3, SI Appendix, Tables S1 and S8 and Figs. S9 and S10). In doing so, we were able to identify MEI1 orthologs from the whole-genome assemblies in every species missing MEI1 in our initial dataset, resulting in the presence of MEI1 in every species considered (SI Appendix, Table S1); thus, it appears that its inferred coevolution with PRDM9 based on RefSeq calls is artificial (Methods). Renumering the phylogenetic test on the curated ortholog sets for the remaining four genes, TEX15 is no longer significant at the 5% level (P = 0.086), possibly because the curation uncovered an intact TEX15 ortholog in anoles. ZCWPW1 and ZCWPW2 are still highly significant; for FBXO47, the curation did not reveal any discrepancy with the initial calls, so the *P* value remains the same.

Our approach therefore uncovered two genes with clear-cut evidence of coevolution with PRDM9, the paralogs ZCWPW1 and ZCWPW2, and more tentative support for two others, TEX15 and FBXO47. ZCWPW1, ZCWPW2, and TEX15 were among our initial list of 241 candidate genes because they are coexpressed with PRDM9 in single-cell testis data from mouse (Fig. 2B; SI Appendix, Fig. S6) (31). FBXO47 was not included by that criterion but because missense mutations in the gene are associated with variation in the total genetic map length in humans in both sexes (33). In mice, the expression of FBXO47 is testis-specific (38), and the gene is expressed in the component in which PRDM9 had the highest loading (albeit with a smaller loading; Ref. 31; Fig. 2B; see also Ref. 39).

Like PRDM9, ZCWPW1, ZCWPW2, FBXO47, and TEX15 are inferred to have been present in the common ancestor of vertebrates. In the following sections, we describe the distribution of each of the four genes across the phylogeny of 189 species and the patterns that give rise to the evidence of statistical association with PRDM9—in particular, the correspondence between their distributions and that of 11 well-supported losses of PRDM9, as well as of nine species for which the status of PRDM9 is uncertain. Moreover, in lineages in which ZCWPW1, ZCWPW2, FBXO47, or TEX15 are present despite the absence of PRDM9, we evaluate if the genes appear to be under relaxed selective constraint, by testing whether *ω* = dn/ds is higher in lineages without a complete PRDM9 than in those that still carry PRDM9 (where dn is the rate of nonsynonymous substitutions and ds the rate of synonymous substitutions; see Methods).

**ZCWPW1 and PRDM9 Coevolution.** Our finding that ZCWPW1 is coevolving with PRDM9 (*P* = 0.0019 in the curated set; Table 1) is in line with previous reports of an association in vertebrates between the presence and absence of ZCWPW1 and PRDM9 orthologs (26, 27). Here, we found an even tighter coupling of PRDM9 and ZCWPW1 than previously documented. Specifically, we inferred 12 losses of ZCWPW1 among 189 species used in our phylogenetic test, distributed across 17 species that lack ZCWPW1 entirely and two species carrying partial ZCWPW1 genes (with the PWWP domain but not the zf-CW domain; SI Appendix, Table S8). Because of its known relationship with PRDM9, ZCWPW1 can be viewed as a positive control for our approach.

Seven of the ZCWPW1 losses occur among the 11 well-supported losses of PRDM9: in cypriniformes fish, percormorph fish (*Euacanthomorphaceae*), siluriformes fish, polypteriformes fish, osteoglossomorpha fish, birds, and *Dicroglossidae* frogs. An additional ZCWPW1 loss occurred in the denticle herring (*Denticips clupeoides*), a species for which the status of PRDM9 is uncertain. The remaining four losses of ZCWPW1 seem to break the pattern in that they occur in lineages containing a complete PRDM9 gene. However, three are observed only in a single species and may be spurious. Therefore, 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 (*Psodonaja textilis*) (SI Appendix, Table S8).

Inversely, ZCWPW1 has been retained in several lineages in which the absence of PRDM9 is well supported: in two

---

**Table 1. Results of phylogenetic tests**

| Gene      | Start position | Gene source | LogLik H0 | LogLik Ha | *P* value | FDR     | *P* value for improved status calls |
|-----------|----------------|-------------|-----------|-----------|-----------|---------|-----------------------------------|
| ZCWPW1    | chr7:100400826 | 1           | −65.941   | −53.35647 | 4.651e−05 | 0.0064  | 1.948e−03                        |
| MEI1      | chr22:41699503 | 3,1         | −54.200   | −44.779   | 8.442e−04 | 0.0586  | NA                                |
| ZCWPW2    | chr3:28348721  | 1           | −67.711   | −60.146   | 4.437e−03 | 0.2055  | 5.171e−06                        |
| TEX15     | chr8:30831544  | 1           | −138.430  | −131.764  | 9.760e−03 | 0.3391  | 8.682e−02                        |
| FBXO47    | chr17:38936278 | 2           | −53.678   | −47.559   | 1.566e−02 | 0.4354  | 1.566e−02                        |

We focused on the five genes that had an FDR ≤ 50%, 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 coexpressed with PRDM9 in single-cell mouse testes data (31), or 2) variants assigned to the gene are associated with variation in recombination phenotypes in humans (33), or 3) the gene was previously known to have a role in mammalian meiotic recombination from functional studies (32) (see Methods). Start positions are based on the human reference genome GRCh38/hg38. NA, not applicable.
PRDM9 is also absent from the closest evolutionary relative in this order of loss, ZCWPW2, both cypriniformes fish and siluriformes fish. Also suggestive of PRDM9 being lost at least twice within frogs, possibly coincident with passing cypriniformes fish and siluriformes fish, as well as the electric eel (Electrophorus electricus) which occurs in an additional species of frog with PRDM9 (Ranitomeya imitator). We were able to successfully identify a complete ZCWPW2 ortholog in this species, suggesting that ZCWPW2 has indeed been lost at least twice within frogs, possibly coincident with PRDM9 in each case. ZCWPW2 is also absent in a clad 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. Also suggestive of this order of loss, ZCWPW2 is absent in osteoglossomorpha fish (the Asian arowana, Scleropages formosus); in this case, the gene is also absent from the closest evolutionary relative in the tree, the elephantfish (Paramormyrus kingsleyae), which carries PRDM9.

Among the nine species for which the status of PRDM9 is uncertain, ZCWPW2 is absent in the denticle herring (D. clupeoides). 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, the only species with ZCWPW2 but that clearly lack PRDM9 are canids and the platypus, the two lineages that experienced the most recent losses of PRDM9 (SI Appendix, Table S8).

This observation suggests that the retention of ZCWPW2 in the two mammalian lineages that lack PRDM9 could simply be a lag. Consistent with this hypothesis, there is statistical support for a relaxation of constraint on ZCWPW2 in lineages that lack PRDM9 (P = 0.0003; SI Appendix, Table S9) and in canid lineages, ZCWPW2 is no longer under any discernible selective constraint (testing against a model of ω = 1, P = 0.307; SI Appendix, Table S9).

Like its paralog, ZCWPW2 contains zf-CW and PWWP domains, predicted to bind H3K4me3 and H3K36me3, respectively (SI Appendix; Table 1). These domains are highly conserved, especially at residues for a lag. Consistent with this hypothesis, there is statistical support for a relaxation of constraint on ZCWPW2 in lineages that lack PRDM9 (P = 0.0003; SI Appendix, Table S9) and in canid lineages, ZCWPW2 is no longer under any discernible selective constraint (testing against a model of ω = 1, P = 0.307; SI Appendix, Table S9).

Six of the ZCWPW2 losses occur among the 11 well-supported losses of an intact PRDM9: in percomorph fish, polypteriformes fish, Xenopus frogs, Dicroglossidae frogs, birds, and the green anole. In order to distinguish whether the absence of ZCWPW2 in Xenopus and Dicroglossidae frogs reflects a single loss or multiple events, we investigated the status of ZCWPW2 in an additional species of frog with PRDM9 (Ranitomeya imitator). We were able to successfully identify a complete ZCWPW2 ortholog in this species, suggesting that ZCWPW2 has indeed been lost at least twice within frogs, possibly coincident with PRDM9 in each case. ZCWPW2 is also absent in a clad 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. Also suggestive of this order of loss, ZCWPW2 is absent in osteoglossomorpha fish (the Asian arowana, Scleropages formosus); in this case, the gene is also absent from the closest evolutionary relative in the tree, the elephantfish (Paramormyrus kingsleyae), which carries PRDM9.
Fig. 3. A summary of the phylogenetic distribution of PRDM9 and the four candidate genes across 189 species. Ortholog calls for candidate genes were based on a search of gene models within whole-genome sequences (Methods), and the phylogenetic test for coevolution with PRDM9 was rerun on these updated calls; updated $P$ values for the phylogenetic test are shown in Table 1. Solid white and black rectangles indicate whether PRDM9 is present or absent, respectively, and gray rectangles lineages for which the status of PRDM9 is uncertain (Methods). For candidate genes, white rectangles are instances in which the gene is present and complete and solid black rectangles indicate when loss of candidate gene is coincident with that of PRDM9. Solid blue rectangles point to instances in which loss of candidate gene occurred prior to that of PRDM9 and green rectangles when it occurred subsequent to that of PRDM9. Red rectangles denote cases in which loss of the candidate gene occurred in lineages with a complete PRDM9. The full phylogenetic distribution of PRDM9 and candidate genes is in SI Appendix, Table S9.

Discussion

Summary. By extending the reconstruction of PRDM9 to 446 vertebrate species, we identified 13 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 11 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 two genes that are clearly coevolving in their presence and absence with PRDM9, ZCWPW1 and its paralog ZCWPW2, and two for which the evidence is weaker, FBXO47 and, most tentatively, TEX15.

Dual Roles of PRDM9 Across Vertebrates. 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 ZF (11). Consistent with our prediction, there is tentative evidence for the influence of PRDM9 binding on recombination in rattle snakes (24). That a gene with a known role in recombination, ZCWPW1, coevolves with PRDM9 across vertebrates lends further support to this hypothesis.

The precise nature of the molecular interactions between PRDM9 and ZCWPW1 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 (27, 40) and exhibit defects in their ability to repair DSBs (25–27). In turn, the genomic locations of DSBs are not altered in mice and older female mice are sterile (27, 40) and exhibit defects in their ability to repair DSBs (25–27). In turn, the genomic locations of DSBs are not altered in mice and older female mice are sterile (27, 40) and exhibit defects in their ability to repair DSBs (25–27). In turn, the genomic locations of DSBs are not altered in mice and older female mice are sterile (27, 40) and exhibit defects in their ability to repair DSBs (25–27).
We note that although ZCWPW1 and ZCWPW2 are absent in the majority of lineages without complete PRDM9 genes, they are sometimes retained after the loss of PRDM9. Given the strong statistical evidence for coevolution, this retention could reflect pleiotropic constraints, if the genes play additional roles beyond their interaction with PRDM9, or simply that the genes have not yet been lost despite no longer being under constraint. In that regard, finding ZCWPW1 present and under selective constraint in siluriformes fish, *Xenopus* frogs, the green anole, and canids suggests that ZCWPW1 is not always dispensable in the absence of a complete PRDM9 gene. For instance, in a subset of species, ZCWPW1 might retain a role in DSB processing or repair regardless of whether DSBs are initiated at PRDM9-bound sites.

In turn, the only species for which ZCWPW2 is retained in the clear absence of a complete PRDM9 are platypus and canids. These two cases are the most recent losses of PRDM9 in our analysis, and there is a relaxation of constraint on ZCWPW2 in both lineages, with ZCWPW2 appearing to evolve neutrally in canids. These findings suggest that ZCWPW2 is dispensable in the absence of a complete PRDM9 ortholog.

The molecular function of ZCWPW2 is to our knowledge 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 (25–27) 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,

**Fig. 4.** Domain architecture 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 (61), and the star indicates the third Trp residue that is thought to stabilize the fold by hydrophobic interactions (61). The PWWP domain (yellow) is expected to bind to histone H3K36me3 through a hydrophobic cavity composed of three aromatic residues (purple) (62). 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. A similar plot for the conservation of residues in ZCWPW1 was previously reported (figure 1B in Ref. 27).

**Interpreting the Coevolution of ZCWPW1 and ZCWPW2 With PRDM9.**

If a gene interacts with PRDM9 by reading its histone modifications, as is the case for ZCWPW2 (25–27) and likely ZCWPW2 (Fig. 4), and has no other roles, we would expect that gene to be dispensable in species that no longer have an active PRDM9 SET domain. Previous papers reported that ZCWPW2 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 (26, 27). 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 (Fig. 3).

While this pattern suggests a dependence of ZCWPW1 and ZCWPW2 on the intact catalytic activity of PRDM9, the interpretation is complicated by the fact that all species with substitutions at the tyrosine residues in all PRDM9 copies are also carrying only partial PRDM9 orthologs lacking KRAB and SSXRD domains, and nearly all species with conserved tyrosine residues also carry a complete copy of PRDM9. In that regard, the few exceptions are informative: Among species with confident PRDM9 calls, the platypus and siluriformes fish carry PRDM9 orthologs putatively missing the KRAB domain but with intact tyrosine residues. ZCWPW2 is absent from all three siluriformes fish species analyzed here, while ZCWPW1 is absent from one. Moreover, our results suggest a reduction of selective constraint on both ZCWPW1 and ZCWPW2 in platypus. 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.
the link between histone modifications (specifically, H3K4me3) and the recruitment of Spool1 is made by Spool1 (42), in mammals, the ortholog of Spool1, CXXC1, is not essential for meiosis (43), 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. If ZCWPW2 does help recruit the recombination machinery, then losses of ZCWPW2 could drive changes in recombination strategy across vertebrates. Indeed, we previously hypothesized that changes in recombination strategy from PRDM9-directed recombination to recombination occurring preferentially around promoter-like features arise by PRDM9 loss of function (11). If the key molecular interactors of PRDM9 have no pleiotropic rules, as may be the case for ZCWPW2 and in some lineages at least ZCWPW1, then it is also possible that their loss could also result in a switch in the way recombination is directed to the genome.

**Possible Links With TEX15 and FBXO47.** The evidence for the coevolution of TEX15 and FBXO47 with PRDM9 is much weaker, and neither appears dispensable in the absence of a complete DSB ortholog. If these two genes are indeed coevolving with PRDM9, the relationship is likely to be indirect. As a possible example, recent work implicates TEX15 as an effector of piRNA-mediated transposable element (TE) methylation and silencing (44, 45). Mule mouse knockouts of Tex15 exhibit a meiotic arrest phenotype associated with the failure to repair DSBs and to undergo chromosomal synopsis (41), as well as the transcriptional activation of TEs (44, 45), a phenotype similar to those observed in mouse knockouts of other piRNA-pathway genes, such as Miwi or Dnmt3 (46). In Dnmt3 knockout mice, it has further been shown that TEs accumulate both H3K4me3 marks and Spool1-dependent DSBS, suggesting that the methylation of TEs serves not only to silence them but may also result in preventing their use as sites of recombination (46). Thus, TEX15 could conceivably play an indirect role in preventing the binding of PRDM9 to TEs.

In turn, FBXO47 is a member of the F-box protein family, which act as recognition subunits of Skp1-Cullin1-F-box protein E3 ubiquitin ligase complexes (38, 47). FBXO47 has recently been implicated as a key regulator of the telomere shuffling complex during meiosis. In mice and in mice it is necessary for telomere nuclear envelope attachment and subsequent events, including DSB repair (38). If this role of FBXO47 contributes to the formation of a chromatin environment that aids in the repair of PRDM9-dependent DSBS, or possibly in the recruitment of ZCWPW1, it might lead to increased conservation of FBXO47 in the presence of PRDM9.

**Which Loss Came First?** While PRDM9 has two distinct roles—in specifying the location of DSBS and in facilitating their repair—the 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 could be asymmetric. For instance, if we ignore possible pleiotropic roles of ZCWPW1 and assume it is necessary for the repair of PRDM9 DSBS (but not DSB localization), we might predict that ZCWPW1 is likely to be lost after PRDM9 (as appears to have been the case in Tachysurus fulvidraco; SI Appendix, Table S88), as otherwise DSBS would go unrepairered. Whereas if ZCWPW2 is involved in DSB localization but not repair, it could be lost either before PRDM9 (as was seemingly the case in two lineages of ray-finned fish; SI Appendix, Table S88) or potentially after.

More generally, 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 ZCWPW1, ZCWPW2, FBXO47, and TEX15 on PRDM9 as well as vice versa (in all tests, maximum $P < 0.07$, testing the null model of no dependence against either dependence as an alternative model; SI Appendix, Table S11). 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.

**Outlook.** Our phylogenetic analysis allowed us to identify 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, from 11 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, when possible, de novo assemblies from RNA-seq data, leading us to realize that in one case (MEI1), the apparent coevolution 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 (SI Appendix, Fig. S7) and downsamples our tree to include at most three species below every inferred PRDM9 loss. As genome qualities improve and as their assemblies become more uniform (e.g., Ref. 48), these issues should be alleviated. Moreover, as species are added to the phylogeny, additional losses will be identified; as one example, our identification of two species of frogs with a complete PRDM9 revealed that PRDM9 had not been lost once in the common ancestor, as had been inferred using fewer species by Baker et al. (11), but has instead been lost more than once within amphibians. This discovery also suggests that frogs may be an interesting clade within which to study the steps by which PRDM9 and its partners are lost.

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 (49). 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 online (e.g., Refs. 48–50), 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.

**Methods**

**Identification of PRDM9 Orthologs.** We characterized the distribution of PRDM9 in vertebrates following the same general approach as in our previous analysis (11, 12); a full description is provided in the SI Appendix. In brief, we first identified putative PRDM9 orthologs using a blastp search (30) against the RefSeq database and confirmed the orthology of each by visually inspecting where these genes clustered in neighbor-joining trees built with Clustal Omega (51) for identified KRAB, SSXRD, and SET domains. For each species for which we could not initially identify a PRDM9 ortholog with KRAB and SET domains from RefSeq, we sought to identify PRDM9 orthologs from the nonredundant protein sequence database, whole-genome sequences, or testis RNA-seq datasets when available. We additionally searched whole-genome sequences or testis RNA-seq datasets from select species not represented in RefSeq in order to better time putative orthologs 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, when possible, de novo assemblies from RNA-seq data, leading us to realize that in one case (MEI1), the apparent coevolution 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 (SI Appendix, Fig. S7) and downsamples our tree to include at most three species below every inferred PRDM9 loss. As genome qualities improve and as their assemblies become more uniform (e.g., Ref. 48), these issues should be alleviated. Moreover, as species are added to the phylogeny, additional losses will be identified; as one example, our identification of two species of frogs with a complete PRDM9 revealed that PRDM9 had not been lost once in the common ancestor, as had been inferred using fewer species by Baker et al. (11), but has instead been lost more than once within amphibians. This discovery also suggests that frogs may be an interesting clade within which to study the steps by which PRDM9 and its partners are lost.

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 (49). 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 online (e.g., Refs. 48–50), 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.

**Identification of PRDM9 Orthologs.** We characterized the distribution of PRDM9 in vertebrates following the same general approach as in our previous analysis (11, 12); a full description is provided in the SI Appendix. In brief, we first identified putative PRDM9 orthologs using a blastp search (30) against the RefSeq database and confirmed the orthology of each by visually inspecting where these genes clustered in neighbor-joining trees built with Clustal Omega (51) for identified KRAB, SSXRD, and SET domains. For each species for which we could not initially identify a PRDM9 ortholog with KRAB and SET domains from RefSeq, we sought to identify PRDM9 orthologs from the nonredundant protein sequence database, whole-genome sequences, or testis RNA-seq datasets when available. We additionally searched whole-genome sequences or testis RNA-seq datasets from select species not represented in RefSeq in order to better time putative orthologs 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, when possible, de novo assemblies from RNA-seq data, leading us to realize that in one case (MEI1), the apparent coevolution 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 (SI Appendix, Fig. S7) and downsamples our tree to include at most three species below every inferred PRDM9 loss. As genome qualities improve and as their assemblies become more uniform (e.g., Ref. 48), these issues should be alleviated. Moreover, as species are added to the phylogeny, additional losses will be identified; as one example, our identification of two species of frogs with a complete PRDM9 revealed that PRDM9 had not been lost once in the common ancestor, as had been inferred using fewer species by Baker et al. (11), but has instead been lost more than once within amphibians. This discovery also suggests that frogs may be an interesting clade within which to study the steps by which PRDM9 and its partners are lost.
domains; 19 species for which we identify PRDM9 orthologs containing KRAB and SET domains, but not SSXRD domains; 215 species for which we have evidence for the absence of a complete PRDM9 gene; and 10 species for which we were unable to make a confident determination (Fig. 1; SI Appendix, Tables S1–S4). 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 SET domain (10, 52) and examined the evidence for positive selection acting on the DNA-binding specificity of PRDM9 ZF arrays following our previous approach (11, 12); see SI Appendix for details.

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 A. 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, S. 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. Detailed information about these analyses can be found in SI Appendix.

Similarly, in two species in which we had originally identified only a partial ortholog of PRDM9 (A. mexicanus and C. harenegus), we wanted to verify the incomplete domain structure inferred from the genome sequence by conducting a de novo transcriptome assembly in C. harenegus, 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 A. mexicanus testis tissue and liver and testis from C. harenegus. See SI Appendix for more details.

Using the same approach to de novo assembly and gene detection, we also analyzed publicly available RNA-seq datasets from testis for 28 additional species (SI Appendix, Table S3), either to verify the absence of PRDM9 or of one of the candidate genes; see SI Appendix for details.

Choice of Candidate Genes and Orthology Assignments. To identify a set of genes that may coevolve with PRDM9, we relied on three publicly available datasets, namely 1) 39 genes associated with variation in recombination phenotypes in a genome-wide association study in humans (33). 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); 2) 215 species for which we coexpressed 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 has the highest loading (31); and 3) 36 genes known to have a role in mammalian meiotic recombination based on functional studies (32).

Genes coexpressed with PRDM9 in mouse spermatogenesis were converted to human gene symbols using the package biomart in R (53). Fifteen of these genes did not have an orthologous human gene symbol (Gm7972, H2-K1, Gm4653, Gm3534, Gm3684, Gm3685, Gm3686, Gm3687, Gm3688, Aev9, Gm4349, ZCWPW2, D1, Trap1a, Xr14a, 2310035C23Rik, and Tmem5) and eight other genes coexpressed with PRDM9 in single-cell data from mouse testes. Specifically, 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 coevolution with PRDM9 (SI Appendix, Fig. S6). A supplementary file describing each meiosis candidate gene is available in SI Appendix, Table S5.

For the 241 genes, we characterized whether the orthology 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 253 genes) (54), reasoning that their genomes were sufficiently incomplete to prevent pollyomavirus observation in the phylogeny: specifically, we removed the minimal number of species necessary to remove each pollyomavirus while preserving any transitions in the state of PRDM9. Moreover, to minimize possible phylegetic 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 that may be missing orthologs by chance (SI Appendix, Fig. S7). Of the remaining 307 species, we further excluded 29 species in order to remove polytomies observed in the phylogeny; for the remaining 278 species, 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 1 jawless fish. This phylogeny includes representative species for 11 of the 13 inferred PRDM9 losses; species of Bunoideae frogs and salamanders were not included because of the absence of available gene annotations; also because of the lack of gene annotations for frog species with PRDM9. Within these 189 species, the losses in Xenopus and Dicroglossidae frogs cannot be distinguished from a single event.

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 protein using the Batcat CD-Search (55) 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 vertebrate phylogeny carrying a complete PRDM9 (Homo sapiens (human), Exon luscius (fish), Geotrypetes saraphin (caecilian), and Pseudonaja textilis (snake)), at an e-value threshold of 1e−4. For the 15 genes (FANC8, FMNT1B, GPR137C, HAUS88, M1AP, MEI1, SPAT22, CLSNI, FBX047, HMCG2, HSZPB, IQCBI, LRRXC2, 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 (SI Appendix, Table S6).

Testing for the Coevolution of PRDM9 and Candidate Genes. For each gene identified by the above pipeline, we tested whether coevolution of PRDM9 and the candidate gene occurred. To this end, we used the package bayestraitsV3 (56). We did so by comparing the statistical support for the 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 (“+”) and loss (“−”) 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 four degrees of freedom and reported a P value corrected for multiple tests (SI Appendix, Table S4). For the 100 maximum likelihood tries computed, and the maximum likelihood value was retained. A quantile-quantile plot was drawn to access the distribution of P values, and the R package “Haplin” was used to compute point-wise confidence intervals. To control for the FDR, we computed q-values using the R package “qvalue” and set a 50% FDR threshold.

Given the phylogenetic distribution of PRDM9, it is likely that a PRDM9 ortholog was present in the common ancestor of vertebrates (11, 12). 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 nine species that lack PRDM9 but in which the loss was not supported by a closely related species (i.e., for which we considered the status uncertain).

For FBX047, TEK15, 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 depend 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 two degrees of freedom. For each gene and model, 100 maximum likelihood tries were computed, and the maximum likelihood value was retained.

We also explored whether redefining a complete PRDM9 ortholog as containing not only the KRAB and SET domain but also the SSXRD domain would change the statistical significance. By using the improved calls (see below), only ZCWPW2 remains significant (P = 0.004) and ZCWPW1 marginally so (P = 0.056) (SI Appendix, Table S10 B and C).

Cavassim et al. PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2
PNAS | 9 of 11 https://doi.org/10.1073/pnas.2114401119

Covered by copyright notice on 07/2022.
In addition, we considered the coevolution of PRDM9 with genes highly expressed in the component in which PRDM9 had its second highest loading in single-cell data from mouse testes (31). For this set, the P values were roughly uniform, as expected under the null model of no coevolution, and no gene stood out as a promising candidate (SI Appendix, Fig. S12).

Improving Gene Status Calls of Top Candidate Genes. For the five genes with an FDR ≤ 50% (Fig. 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. We provide an overview of the steps taken for each candidate gene in the SI Appendix. For each gene, we provide descriptions of identified orthologs and how they were identified in SI Appendix, Table S1; specific details about orthologs identified from whole-genome assemblies in SI Appendix, Table S2; our improved calls per species in SI Appendix, Table S8A; and a summary of loss events in SI Appendix, Table S8B.

Conservation of Residues in ZCWPW2. We carried out a residue conservation analysis using an approach proposed by Ref. 37, using code score conservation analysis (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) (35, 38). As recommended, the overall background amino acid distribution was drawn based on the BLOSUM62 amino acid substitution matrix provided by the software (57). 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.

Evidence for Relaxed Selective Constraint in the Absence of PRDM9. To test for relaxed constraint in species without a complete PRDM9, we used the program codeml within PAML (59, 60). Codeml uses protein coding sequences to estimate the ratio of nonsynonymous to synonymous substitution rates (ω = dN/dS). Values of ω significantly 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 coded 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; Refs. 35, 58) and inspected the codon-aware alignment visually to ensure that the same isoforms were used across species. For each multispecies alignment, we tried two approaches. 1) 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 as internal branches on which PRDM9 may have been lost). 2) 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 1), significance was assessed using a likelihood ratio test with one degree of freedom; for 2), by the number of degrees of freedom corresponded to the number of distinct ω values minus one. If ω values were found to be significantly higher in species without a complete PRDM9, we tested whether or not we could reject ω = 1 for these species. For two cases in which we could not obtain a multispecies alignment that included the whole coding sequence (ZCWPW1 in fish and TEXT15 in amphibians), we instead used the pairwise model (runmode: -2 with an FDR ≤ 50%). On alignment of a pair of species and tested whether we could reject ω = 1 for species lacking PRDM9 by comparing a model allowing ω to vary versus a null model fixing the ω value at 1, with one degree of freedom.

Data Availability. RNA-seq data have been deposited in NCBI, ID: PRJNA605699. Code generated for this study can be found in GitHub at https://github.com/zabelcavassim/PRDM9_analyses. All other study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank Arild Folvkord and Leif Andersson (C. harenaurus), Clif Finn (A. mexicanus), Sonia Schwartz and Tracy Langkilde (S. undulatus), and Athanasia Tzika (A. carolinesis) for providing tests, 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 laboratory space and sequencing facilities and to Andrew Meade for helpful discussions about the software Bayestrains. In addition, we thank Todd Jackman and Daniel Wells for pointing out mistakes in the previous version of our manuscript, which are now corrected. This work was partly supported by R01 GM83098 to M.P. and NIH 1R35GM133774 to M.S. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.
31. M. Jung et al., Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. elife 8, e43966 (2019).
32. F. Baudat, Y. Imai, B. de Massy, Meiotic recombination in mammals: Localization and regulation. Nat. Rev. Genet. 14, 794–806 (2013).
33. B. V. Hallidorsson et al., Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363, eaau1043 (2019).
34. N. A. O’Leary et al., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44 (D1), D733–D745 (2016).
35. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34, 1812–1819 (2017).
36. D. Barker, M. Pagel, Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1, e3 (2005).
37. M. Pagel, Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B Biol. Sci. 255, 37–45 (1994).
38. R. Hua et al., FBXO47 regulates telomere-inner nuclear envelope integration by stabilizing TRF2 during meiosis. Nucleic Acids Res. 47, 11755–11770 (2019).
39. Y. Chen et al., Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. Cell Res. 28, 879–896 (2018).
40. M. Li et al., The histone modification reader ZCPW1 is required for meiosis prophase I in male but not in female mice. Sci. Adv. 5, eaax1101 (2019).
41. F. Yang, L. Acquaviva, J. Drogat, P.-M. Dehe, C. de La Roche Saint-André, V. Géli, Spp1 at the crossroads of H3K4me3 regulation and meiotic recombination. Epigenetics 8, 355–360 (2013).
42. H. Tian, T. Billings, P. M. Petkov, CXXC1 is not essential for normal DNA double-strand break repair and chromosomal synapsis during male meiosis. J. Cell Biol. 180, 673–679 (2008).
43. L. Acquaviva, J. Drogat, P.-M. Dehe, C. de La Roche Saint-André, V. Géli, Spp1 at the crossroads of H3K4me3 regulation and meiotic recombination. Epigenetics 8, 355–360 (2013).
44. H. Tian, T. Billings, P. M. Petkov, CXXC1 is not essential for normal DNA double-strand break formation and meiotic recombination in mouse. PLoS Genet. 14, e1007657 (2018).
45. F. Yang et al., TEX15 is an essential executor of MIWI2-directed transposon DNA methylation and silencing. Nat. Commun. 11, 3739 (2020).
46. N. Zamudio et al., DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. Genes Dev. 29, 1256–1270 (2015).
47. K. I. Nakayama, K. Nakayama, Ubiquitin ligases: Cell-cycle control and cancer. Nat. Rev. Cancer 6, 369–381 (2006).
48. A. Rhee et al., Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746 (2021).
49. S. D. Smith, M. W. Pennell, C. W. Dunn, S. V. Edwards, Phylogenetics is the new genetics (for most of biodiversity). Trends Ecol. Evol. 35, 415–425 (2020).
50. K.-P. Koepfl, B. Paten, S. J. O’Brien; Genome 10K Community of Scientists, The genome 10K project: A way forward. Annu. Rev. Anim. Biosci. 3, 57–111 (2015).
51. F. Sievers, D. G. Higgins, Clustal Omega. Curr. Protoc. Bioinform. 48, 3.13.1–3.13.16 (2014).
52. H. Wu et al., Molecular basis for the regulation of the H3K4 methyltransferase activity of PRDM9. Cell Rep. 5, 13–20 (2013).
53. S. Durins, J. Bullard, P. T. Spellman, S. Dudoit, GenomereGraphics: Integrated genomic data visualization with R. Bioinformatics 10, 2 (2009).
54. R. M. Waterhouse et al., BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35, 543–548 (2018).
55. A. Marchler-Bauer et al., CDD: NCBI’s conserved domain database. Nucleic Acids Res. 43, D222–D226 (2015).
56. M. Pagel, A. Meade, Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. 167, 808–825 (2006).
57. J. A. Capra, M. Singh, Predicting functionally important residues from sequence conservation. Bioinformatics 23, 1875–1882 (2007).
58. S. Kumar, G. Stecher, K. Tamura, MEGAF: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874 (2016).
59. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591 (2007).
60. Z. Yang, PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556 (1997).
61. F. He et al., Structural insight into the zinc finger CW domain as a histone modification reader. Structure 18, 1127–1139 (2010).
62. S. Qin, J. Min, Structure and function of the nucleosome-binding PWWP domain. Trends Biochem. Sci. 39, 536–547 (2014).