For consideration as an *Article (Discoveries)*

Title: Snake recombination landscapes are concentrated in functional regions despite PRDM9

Drew R. Schield[1], Giulia I. M. Pasquesi[1], Blair W. Perry[1], Richard H. Adams[1], Zachary L. Nikolakis[1], Aundrea K. Westfall[1], Richard W. Orton[1], Jesse M. Meik[2], Stephen P. Mackessy[3], and Todd A. Castoe[1, §]

Affiliations:

1. Department of Biology, University of Texas at Arlington, Arlington, TX, USA

2. Department of Biological Sciences, Tarleton State University, Stephenville, TX, 76402 USA

3. School of Biological Sciences, University of Northern Colorado, Greeley, CO, USA

[§]To whom correspondence should be addressed: Todd A. Castoe, Department of Biology, University of Texas at Arlington, Arlington, TX 76010 USA. *Email:* todd.castoe@uta.edu *phone:* 817-272-9084 *fax:* 817-272-9615

Running title: Evidence for PRDM9 activity in snake meiotic recombination

Keywords: *Crotalus*, Linkage disequilibrium, microchromosomes, population genomics, PRDM9

Abstract

Meiotic recombination in vertebrates is concentrated in hotspots throughout the genome. The location and stability of hotspots have been linked to the presence or absence of PRDM9, leading to two primary models for hotspot evolution derived from mammals and birds. Species with PRDM9-directed recombination have rapid turnover of hotspots concentrated in intergenic regions (i.e., mammals), while hotspots in species lacking PRDM9 are concentrated in functional regions and have greater stability over time (i.e., birds). Snakes possess PRDM9, yet virtually nothing is known about snake recombination. Here we examine the recombination landscape and test hypotheses about the roles of PRDM9 in rattlesnakes. We find substantial variation in recombination rate within and among snake chromosomes, and positive correlations between recombination rate and gene density, GC content, and genetic diversity. Like mammals, snakes appear to have a functional and active PRDM9, but rather than being directed away from genes, snake hotspots are concentrated in promoters and functional regions – a pattern previously associated only with species that lack a functional PRDM9. Snakes therefore provide a unique example of recombination landscapes in which PRDM9 is functional, yet recombination hotspots are associated with functional genic regions – a combination of features that defy existing paradigms for recombination landscapes in vertebrates. Our findings also provide evidence that high recombination rates are a shared feature of vertebrate microchromosomes. Our results challenge previous assumptions about the adaptive role of PRDM9 and highlight the diversity of recombination landscape features among vertebrate lineages.

Introduction

Meiotic genetic recombination shapes the structure, composition, and variation of genomes (Eyre-Walker 1993; Posada et al. 2002; Kent et al. 2012). Accordingly, recombination also governs patterns of sequence evolution in populations, including the generation and maintenance of novel combinations of alleles (Nachman and Payseur 2012; Hunter 2015). Recombination erodes associations between physically linked loci through the decay of linkage disequilibrium, and is hypothesized to increase the efficacy of selection through Hill-Robertson effects (Hill and Robertson 1966) and to interact with hitchhiking and background selection to shape genetic diversity across the genome (Keinan and Reich 2010; Cutter and Payseur 2013; Haasl and Payseur 2016). Recombination can also act as a source of genetic variation by causing mutation (Lercher and Hurst 2002), which can drive regional increases in specific nucleotides or sequence motifs (Meunier and Duret 2004; McVean 2010; Pessia et al. 2012). Considering the broad relevance of recombination in shaping genomes of sexually reproducing organisms, mechanisms that

direct recombination and the resulting genome-wide recombination landscape are central for understanding genome structure, function, and evolution.

Across species with sexual reproduction, meiotic recombination is driven by the initiation of double-stranded breaks (DSBs) by SPO11 (Keeney 2007). The repair of these breaks results in crossover and non-crossover recombination events throughout the genome (Keeney 2007; Lam and Keeney 2014). DSBs initiated by SPO11 preferentially localize in regions of the genome that contain histone H3 lysine K4 trimethylation (H3K4me3) marks; these marks are typically associated with open chromatin, such as that found in open gene promoters (Borde et al. 2009). Further, high density of DSBs can lead to the formation of recombination hotspots, which are small genomic regions of disproportionately frequent recombination.

Although the presence of a universal DSB-generating mechanism and of recombination hotspots are common features of vertebrates studied to date, previous studies have demonstrated that the location of hotspots in the genome, and how rapidly they change over evolutionary time, varies in a bimodal fashion depending on whether a particular species has an active PRDM9 (Myers et al. 2005; Axelsson et al. 2012; Singhal et al. 2015; Baker et al. 2017; Kawakami et al. 2017; Schumer et al. 2018). Species with a partial complement of PRDM9 domains do not appear to have PRDM9-directed recombination (Baker et al. 2017), and the KRAB domain specifically is known to be required for proper function in mammals (Imai et al. 2017). Where present, and with the complete domain structure (e.g., in apes and mice), PRDM9 orchestrates the recombination landscape by the binding of its fast-evolving zinc-finger array to specific nucleotide motifs, which results in the alteration of H3K4me3 marks and in the diversion of recombination away from genes and functional regions (Myers et al. 2005; Berg et al. 2010; Brick et al. 2012; Lam and Keeney 2014). As a consequence of the rapid evolution of the PRDM9 binding site, species with PRDM9-directed recombination consistently exhibit rapid turnover of recombination hotspots, leading to major differences in recombination landscapes over short evolutionary timescales (Baudat et al. 2010; Myers et al. 2010). By contrast, vertebrates that either lack or have a nonfunctional PRDM9 (e.g., birds (Singhal et al. 2015; Baker et al. 2017) and canids (Axelsson et al. 2012)) possess recombination hotspots that are concentrated in promoter regions (Axelsson et al. 2012; Auton et al. 2013; Singhal et al. 2015; Kawakami et al. 2017), presumably due to the default targeting of SPO11 to open chromatin near gene promoters (Ramirez-Carrozzi et al. 2009; Thomson et al. 2010; Tock and Henderson 2018). This is further supported by studies of mice in which PRDM9 has been knocked out, shifting recombination hotspots to promoter regions (Brick et al. 2012). The genomic location of hotspots in species without PRDM9 also tends to be conserved over millions of years of evolution, leading to a

broad conservation of genomic recombination landscapes between populations and species (Singhal et al. 2015; Kawakami et al. 2017). Accordingly, the body of previous work on recombination landscapes in vertebrates has established a paradigmatic bifurcation of recombination control across vertebrates – one involving PRDM9 and one without PRDM9 – each of which leads to fundamentally divergent outcomes and evolutionary dynamics of genome-wide recombination. Evidence for the rapid evolution of PRDM9 through positive selection (Ponting 2011) suggests a potential adaptive role of PRDM9-directed recombination, possibly as a means of directing recombination away from functionally-important promoter regions and genes.

While recombination in mammals and birds has been well characterized, recombination landscapes have not been investigated in other divergent lineages of amniote vertebrates – for example, although they comprise more than 10,000 species (The Reptile Database; http://www.reptile-database.org), remarkably little is known about recombination landscapes in squamate reptiles (lizards and snakes)(Fujita et al. 2011; Baker et al. 2017). Squamates diverged from a common ancestor more than 200 million years ago and are important models for studying variation in recombination landscapes due to variation in presence or absence of a functional *PRDM9* gene among lineages (Baker et al. 2017). Snakes are particularly interesting in this regard because, like mammals, they appear to possess a functional PRDM9 (Baker et al. 2017), yet like birds possess both macro- and microchromosomes (Olmo 2005; Janes et al. 2010; Schield et al. 2019). Snakes also exhibit remarkable variation in genomic repeat content, even between closely related lineages, and appear to have reacquired GC isochores after a squamate ancestor with reduced isochore structure (Castoe et al. 2013; Pasquesi et al. 2018). Detailed studies of recombination landscapes in snakes have not been feasible to date due to the relatively poor quality of snake genome assemblies (Castoe et al. 2013; Vonk et al. 2013; Yin et al. 2016; Perry et al. 2018). However, the recent availability of a well-annotated chromosome-level genome assembly for the prairie rattlesnake (*Crotalus viridis*) (Schield et al. 2019) now provides an opportunity to study recombination rate variation in a squamate reptile.

In this study, we leverage whole genome resequencing of populations of two rattlesnake species, RNA-seq data from several snake lineages, and a chromosome-level rattlesnake genome assembly to characterize recombination landscapes in snakes. We use these datasets to address broad questions about the recombination landscape and the mechanisms underlying it within snakes, including: i) are snake recombination landscapes conserved over millions of years of evolution, as observed in birds, or do they instead resemble rapidly-shifting mammal landscapes; ii) what sequence features of snake genomes are associated with recombination hotspots; iii) is there evidence that PRDM9 is functional in snakes, iv)

does PRDM9 play a role in directing snake recombination, and v) do snake recombination landscapes adhere to one of the two well-characterized and divergent patterns of PRDM9-active or inactive vertebrate species?

Results

Recombination variation across the genome

We estimated linkage disequilibrium (LD)-based population scaled recombination rates across the genomes of two rattlesnake species, the prairie rattlesnake (*Crotalus viridis*) and the northern Pacific rattlesnake (*C. oreganus*), using whole genome resequencing data from natural populations mapped to the *C. viridis* reference genome (see Materials and Methods; Supplementary Fig. S1a; Supplementary Table S1). Our mapping and variant calling procedure yielded an average coverage of 36.7× per sample (range = 9 – 106.2×; Supplementary Table S1) and 14,102,890 SNPs in *C. viridis* and 18,233,083 SNPs in *C. oreganus*, which have roughly 1.22% sequence divergence. Using a Bayesian phylogenetic approach, we estimated the divergence time of the two species to be roughly 2.8 MYA (Supplementary Fig. S1b), and we estimated population-scaled genetic diversity (θ_W) to be 0.00057 in *C. viridis* and 0.00054 in *C. oreganus*. For simplicity, we refer to these species as 'CV' and 'CO', hereafter.

We observed broadly similar genome-wide patterns of recombination between the two rattlesnake species studied, both of which show substantial variation in recombination within and between chromosomes (Fig. 1). Variation in estimates of population scaled recombination rate, ρ/bp ($\rho = 4N_e r$, where r is the per generation recombination rate), spanned greater than eight orders of magnitude in both species (9.07×10^{-8} – 30.93 in CV, 3.86×10^{-7} – 41.95 in CO). Within macrochromosomes, we observed high recombination in telomeric regions (Fig. 1A). We characterized this variation further by calculating relationships between recombination rate and distance to chromosome end, and found significant negative correlations in both species (Supplementary Fig. S2; Pearson's correlation coefficients; r = -0.432, p-value = 2.4×10^{-12} in CV; r = -0.372, p-value = 2.76×10^{-9} in CO), consistent with a strong telomere effect on macrochromosomes. We also observed a high degree of within-chromosome variation in recombination rate on microchromosomes (Fig. 1a-c), similar to the pattern found in avian species (e.g., zebra finch (Backström et al. 2010)). For example, in CV the standard deviation for recombination rate in 1 Mb windows on Chromosome 1 was 0.0037, compared to 0.011 on Chromosome 9 (i.e., the largest microchromosome). Recombination rates in the two rattlesnake species were significantly greater on

microchromosomes relative to macrochromosomes (Welch's two-sample *t*-tests, *p*-values < 2.9×10^{-13}), and we found significant negative relationships between recombination rate and chromosome length (Pearson's correlation coefficients; $r = -0.681$, $p = 0.0063$ in CV; $r = -0.629$, $p = 0.0051$ in CO), indicating generally higher recombination rates on shorter chromosomes.

While population-scaled recombination rate was higher in CO (mean $\pm$ SD = 0.014 ± 0.016 ρ/bp) than CV (0.0063 ± 0.0082 ρ/bp), genome-wide patterns of recombination between the two species were highly correlated at broad scales (Fig. 1d). Our comparisons of mean population-scaled recombination rate estimates for 1 Mb and 100 kb windows between the two species yielded significant positive correlations at each of these resolutions (Pearson's correlation coefficients; 1 Mb $r = 0.819$, $p < 2.2 \times 10^{-16}$; 100 kb $r = 0.502$, $p < 2.2 \times 10^{-16}$), suggesting that the genomic landscape of recombination is conserved at broad scales between the two species (Fig. 1). Similar to patterns of rate variation across chromosomes and elevated recombination rates on smaller chromosomes, the broad-scale conservation of the recombination landscape between CV and CO is reminiscent of bird genomic landscapes.

Genomic correlates of recombination

To understand links between structural features of the genome, genetic diversity, and recombination, we compared broad scale patterns of recombination rate in 1 Mb genomic windows to measures of genomic nucleotide diversity (π), GC content, gene density, and repeat content (Fig. 1e-l, Table 1). We found strong positive relationships between recombination rate and π in both snake species (Fig. 1e-f), which correspond to significant pairwise correlations (Spearman's rank order correlation coefficients; CV $r = 0.586$, $p < 2.2 \times 10^{-16}$, CO $r = 0.546$, $p < 2.2 \times 10^{-16}$). These relationships were also significant after performing partial correlations to control for GC content, gene density, and repeat content ($r = 0.593$, $p = 1.87 \times 10^{-114}$ in CV; $r = 0.534$, $p = 4.95 \times 10^{-89}$). Partial correlations between nucleotide diversity and recombination were highest among all comparisons for both species (see below).

Among various features of genome structure, recombination was most positively correlated with GC content and gene density, and to a lesser degree with repeat content (Fig. 1g-j, Table 1); recombination rate and GC content were significantly correlated in both pairwise (CV $r = 0.293$, $p < 2.2 \times 10^{-16}$; CO $r = 0.261$, $p < 2.2 \times 10^{-16}$) and partial correlations controlling for gene density and repeat content (CV $r = 0.127$, $p = 1.03 \times 10^{-5}$; CO $r = 0.124$, $p = 1.72 \times 10^{-5}$). Recombination rate and gene density were also significantly correlated in pairwise tests (Fig. 1i-j, Table 1; CV $r = 0.289$, $p < 2.2 \times 10^{-16}$; CO $r = 0.238$, $p < 2.2 \times 10^{-16}$), and in partial correlation analyses (CV $r = 0.154$, $p = 8.46 \times 10^{-8}$; CO $r = 0.12$, $p = 3.95 \times$

10^{-5}). Finally, recombination rate and repeat content were also significantly correlated in pairwise tests (CV $r = 0.06$, $p = 0.047$; CO $r = 0.089$, $p = 0.002$) and partial correlation analyses (CV $r = 0.1$, $p = 5.33 \times 10^{-4}$; CO $r = 0.12$, $p = 2.39 \times 10^{-5}$). As a complement to 1 Mb-resolution analyses, we performed correlation analyses at 100 kb resolution (Table 1). These results were qualitatively similar to 1 Mb resolution, with significantly positive pairwise and partial correlations between recombination rate and nucleotide diversity, GC content, gene density, and repeat content. Collectively, positive associations between recombination rate, gene density, and nucleotide diversity are consistent with the expectations of genomic polymorphism being shaped by linked selection (Burri et al. 2015). The association between recombination rate and repetitive element content has also been observed in flycatchers (Kawakami et al. 2017), potentially due to the shared ability of the recombination machinery and repeat elements to preferentially access open chromatin regions.

Recombination variation on the Z chromosome

There are broad predictions about how recombination manifests on sex chromosomes compared to the rest of the genome. First, sex-linked regions are expected to exhibit reduced recombination relative to autosomes in species with heteromorphic sex chromosomes (Barton and Charlesworth 1998; Bergero and Charlesworth 2009). Rattlesnakes have female heterogamety (ZW) with highly heteromorphic Z and W chromosomes (Baker et al. 1972; Matsubara et al. 2006), thus we predict reduced recombination within Z- and W-linked regions. Second, sex chromosomes include a region where recombination is unsuppressed (i.e., the pseudoautosomal region; PAR), in which recombination rates are expected to resemble those from autosomes (Bergero and Charlesworth 2009). These regions have been previously identified on the rattlesnake Z chromosome based on features of genome structure and comparative read mapping analyses from female and male individuals (Schield et al. 2019), but recombination has not yet been examined for snake sex chromosomes. Therefore, we addressed the above predictions using our recombination rate estimates across the previously identified Z chromosome regions.

Consistent with the expectation of reduced recombination in sex-linked regions, we observed significantly lower recombination rates on the Z chromosome compared to autosomes in both species (Fig. 2a-b; Welch's two-sample *t*-tests, *p*-values $< 2.2 \times 10^{-16}$). Further, consistent with the expectation of unsuppressed recombination between Z and W chromosomes within the PAR, recombination rates were significantly greater in the PAR than Z-linked regions of the Z chromosome (*t*-tests, CV *p*-value $= 7.7 \times 10^{-12}$, CO *p*-value $= 2.73 \times 10^{-9}$). PAR recombination rates were also significantly higher than autosomes overall (*t*-tests, CV *p*-value $= 2.39 \times 10^{-8}$, CO *p*-value $= 2.65 \times 10^{-6}$) and terminal regions of autosomes,

specifically (*t*-tests, CV *p*-value = 0.019, CO *p*-value = 0.027). Recombination rate estimates across the Z chromosome therefore broadly corroborate the previous identification of Z-linked and PAR regions in the prairie rattlesnake. We also observed low recombination rates in both species across the region immediately adjacent to the PAR (Fig. 2b). This region was previously identified as a recently recombination-suppressed evolutionary stratum between rattlesnake Z and W chromosomes ('Recent Stratum'; (Schield et al. 2019)) based on elevated female π (but not male π) and intermediate female coverage within the region. It was also hypothesized that there is ongoing degeneration of the gametologous region of the W chromosome. Accordingly, we find roughly equivalent recombination rates in this region compared to the remaining Z-linked region ('Older Strata'; *t*-tests, CV *p*-value = 0.051, CO *p*-value = 0.546), consistent with complete Z-W recombination suppression in the most recently established evolutionary stratum in this lineage.

Recombination hotspots evolve rapidly and are concentrated in functional genomic regions

Because recombination hotspots appear to be a common feature of vertebrate genomes regardless of PRDM9 activity, we explored fine-scale recombination rates to identify evidence for hotspots. We defined hotspots in snake genomes as small genomic regions with much higher (e.g., greater than 100-fold) recombination rate relative to regions immediately up- and downstream (Fig. 3, Supplementary Fig. S3), and identified 9,247 such regions in CV and 10,656 in CO (Fig. 3a-b, d). Only 2,517 hotspots were shared between the two species (Fig. 3c-d), corresponding to 27.2% and 23.6% of identified hotspots in CV and CO, respectively. This low frequency of shared hotspots suggests rapid turnover of hotspots between these two closely-related snake species, similar to patterns observed in PRDM9-active mammalian species (e.g., Ptak et al. 2005; Stevison et al. 2015), and in contrast to the conservation and evolutionary 'stability' observed in PRDM9-lacking bird species (Burri et al. 2015; Singhal et al. 2015; Kawakami et al. 2017).

Hotspots in both species were distributed throughout the genome with greatest density on autosomes (Supplementary Figs. S3, S4). Because power to identify hotspots has been shown to depend on background levels of recombination (Myers et al. 2005; Axelsson et al. 2012; Auton et al. 2013; Singhal et al. 2015), we expect that low hotspot density on the Z chromosome is likely due to low power within this large region of low background recombination rate. Simulations to determine our power to identify hotspots at various background recombination rates further suggest low power to identify Z-linked hotspots, as they consistently demonstrated low power in regions of low recombination (Supplementary

Fig. S5), such as the Z chromosome. Our results therefore likely provide a conservative estimate of the distribution of hotspots on the Z. Given evidence for a broad relationship between GC content and recombination across the genome (Fig. 1, Supplementary Fig. S6), we tested if rattlesnake hotspots also had higher GC content than the genomic background. We calculated GC content among all CV and CO hotspots and compared these to GC content measured in 2 kb windows across the prairie rattlesnake genome and found significantly higher GC content in hotspots of both species (Supplementary Fig. S6a; Welch's two-sample *t*-tests, *p*-values $< 2.2 \times 10^{-16}$). Fine-scale patterns of higher recombination in CpG islands (CGIs) were also consistent with a broad genomic relationship between recombination rate and CGI density in 1 Mb sliding windows (Supplementary Fig. S6b-d; Spearman's rank order correlation coefficients, CV $r = 0.314$, $p < 2.2 \times 10^{-16}$, CO $r = 0.257$, $p < 2.2 \times 10^{-16}$).

A key distinction between PRDM9-active and PRDM9-inactive recombination mechanisms in vertebrates studied to date is whether promoters and functional regions are hotspots for recombination, as is the case in PRDM9-inactive species. To examine the relationship between functional elements and fine-scale recombination rates, we examined ρ/bp in intervals of increasing distance from functional elements, including annotated promoters and CGIs, in the rattlesnake reference genome. We observed decreases in relative recombination rate with increasing distance from these features (Fig. 4a-d), with significant negative correlations between relative recombination rate and log-scaled distance from promoters (Spearman's rank order correlation coefficients; CV $r = -0.549$, $p < 2.2 \times 10^{-16}$; CO $r = -0.359$, $p < 2.2 \times 10^{-16}$) and CGIs (CV $r = -0.613$, $p < 2.2 \times 10^{-16}$; CO $r = -0.55$, $p < 2.2 \times 10^{-16}$). Further, 20.3% of CV hotspots and 19.6% of CO hotspots overlapped with promoter regions, and 42.3% of CV hotspots and 35.8% of CO hotspots were within CGIs. The proportion of shared hotspots in CGIs was even greater (48.7% of shared hotspots overlapped with CGIs), and CGIs were significantly enriched for shared hotspots (Fisher's exact tests of species-specific versus shared hotspots; CV $p = 4.02 \times 10^{-5}$, CO $p = 1.7 \times 10^{-15}$). This enrichment of shared hotspots may suggest that CGIs harbor a disproportionate number of hotspots that are stable over evolutionary time. Genes overlapping shared hotspots between CV and CO did not show evidence of functional enrichment after false discovery rate (FDR) correction (Supplementary Table S2).

Hotspot density in nine non-overlapping genomic features also provides consistent evidence for increased recombination near genes; hotspot density was highest in CGIs, promoters, and first exons (Fig. 4e; Supplementary Fig. 7a), and overall recombination rates were greatest in these regions (Fig. 4f; Supplementary Fig. 7b). Promoters with CGIs also had higher hotspot density than promoters lacking CGIs in both species, though separate comparisons of recombination rate estimates near promoters with

and without CGIs suggest that there is not a strong additive effect of CGIs on recombination rate with respect to promoters (Supplementary Fig. S8). As a comparison to candidate hotspots, we also calculated densities of coldspots with GC content matched to recombination hotspots from both species in genomic features. Distributions were distinct, with highest coldspot densities occurring in intergenic regions, transposable elements (TEs), and introns. However, the difference in density between functional regions and other features was not as pronounced as in hotspots, indicating a greater relative abundance of hotspots in CGIs and promoters (Fig. 4e, Supplementary Figs. S7, S9). Finally, we examined the distribution of recombination in upstream and downstream 500 kb regions of all genes and found evidence that recombination increases in genic regions (Fig. 4g; Supplementary Fig. 7c). Collectively, these findings illustrate that snake recombination hotspots are concentrated in functional regions associated with genes and open chromatin – a pattern otherwise typically associated with species that lack a functional PRDM9 (Axelsson et al. 2012; Singhal et al. 2015; Baker et al. 2017).

To further understand the composition of recombination hotspots, we identified DNA sequence motifs enriched in CV and CO hotspots compared to GC-matched coldspots. Enriched motifs varied in length and composition within and between species, and motifs in CV and CO hotspots were largely distinct, consistent with the large proportion of species-specific hotspots inferred (Supplementary Fig. S10). Searches of enriched motif sequences against the JASPAR binding motif database (Fornes et al. 2019) showed similarity of CV and CO hotspots to zinc-finger (ZF) binding proteins, including multiple transcription factors (Supplementary Table S3). While these searches identified ZF motifs in hotspots, none of these motifs were associated with known PRDM9 sequences, specifically. This result is not overly informative, however, because rapid evolution of binding motifs is a prominent feature of PRDM9 in most PRDM9-active lineages (e.g., mammals). Sequence motifs enriched in hotspots also included the binding motif associated with the insulator protein CTCFL (also known as BORIS; Hore et al. 2008). Below we investigate the recombination landscape in the context of both CTCFL and PRDM9 to determine the degree to which these proteins and their binding sites are associated with meiotic recombination in snakes.

Evidence for a role of CTCFL in snake meiotic recombination

CTCFL is a germline-expressed insulator protein and close paralog of CTCF, both of which share a highly conserved DNA binding motif (Sleutels et al. 2012). This group of proteins has been shown to play a role in maintaining genome integrity during double stranded breaks (Hilmi et al. 2017; Lang et al. 2017), and their binding sites are associated with regions of elevated recombination in humans (Kong et

al. 2014). We were therefore motivated to investigate evidence of associations between CTCF/CTCFL and recombination hotspots. We analyzed gene expression across germline and somatic tissues in three snake species and two mammals and found consistent evidence that CTCFL is expressed at comparatively high levels in germline tissues of snakes (Fig. 5a), consistent with CTCFL playing a role in meiotic recombination in this lineage.

Given evidence of germline expression and of hotspot enrichment for DNA motifs with similarity to that of CTCFL, we searched for matches to the JASPAR database consensus 14-mer CTCFL binding motif in hotspots and matched coldspots in both rattlesnake species. We found a large proportion of hotspots that contained the CTCFL motif (42.3% in CV and 40% in CO), corresponding to significant enrichment of CTCFL binding sites in species-specific and shared hotspots (Fig. 5b; Fisher's exact tests; CV-specific $p = 5.63 \times 10^{-82}$; CO-specific $p = 1.86 \times 10^{-25}$; shared $p = 1.41 \times 10^{-79}$). We also examined the prevalence of the consensus 19-mer CTCF binding site in recombination hotspots, and found evidence of enrichment in CV-specific and shared hotspots (Fisher's exact tests; p-values = 2.43×10^{-23} and 3.32×10^{-35}, respectively), but not CO-specific hotspots, despite roughly 28% of CO-specific hotspots containing predicted CTCF binding sites (Fisher's exact test; $p = 0.967$). We then predicted CTCFL binding motifs throughout the *C. viridis* genome to test if there was a relationship between recombination rate and proximity to predicted CTCFL binding sites. This analysis revealed a strong increase in recombination rate with proximity to predicted CTCFL binding sites (Fig. 5c-d; Spearman's rank-order correlation coefficients; CV $r = -0.43$, $p < 2.2 \times 10^{-16}$; CO $r = -0.38$, $p < 2.2 \times 10^{-16}$), similar to observed correlations for CGIs and promoters. To discern if this pattern is driven solely by an autocorrelation with increased recombination near functional regions, we examined pairwise relationships between distance from CGIs and promoters and CTCFL binding sites. These comparisons found weak positive associations unlikely to be fully explained by the relationship between recombination and functional regions (Spearman's rank-order correlation coefficients; CV CTCFL versus promoters $r = 0.148$, $p = 2.5 \times 10^{-6}$; CO CTCFL versus promoters $r = 0.139$, $p = 8.9 \times 10^{-6}$; CV CTCFL versus CGIs $r = 0.138$, $p = 1.05 \times 10^{-5}$; CO CTCFL versus CGIs $r = 0.037$, $p = 0.24$), arguing for a legitimate relationship between recombination and CTCFL.

Evidence for PRDM9-directed recombination

Multiple snake species have been shown to possess a putatively functional *PRDM9* ortholog that contains rapidly evolving DNA-binding ZF domains (Baker et al. 2017), yet no detailed tests of the relationship between PRDM9 and snake recombination landscapes have been conducted. Because the region of the *C.*

viridis genome assembly encoding the ZF exon of PRDM9 contained a gap, we augmented the genome assembly with additional data from *C. viridis* (10x Genomics linked-read data for genome subassembly and PacBio long read data; Supplementary Fig. S11; Supplementary Data Files S1-S3). This provided an inference of the PRDM9 coding sequence that included the full set of functional domains shown to be required for PRDM9 activity in recombination (Baker et al. 2017; Imai et al. 2017), and a partial tandem array of three ZFs. We confirmed orthology of the candidate rattlesnake PRDM9 to that of other vertebrates using a BLASTp search (Altschul et al. 1990), and aligned our sequence to PRDM9 sequences from other snakes: the five-pace viper (*Dienagkistrodon acutus)* and Burmese python (*Python bivittatus*); none of these snake PRDM9 sequences contained premature stop codons that would indicate loss of function (Supplementary Fig. S12). These orthologous comparisons, along with our own data for *C. viridis* (Supplementary Fig. S11), suggest that our *C. viridis* PRDM9 sequence is incomplete at the 3' end and lacks one or more ZFs that we were not able to resolve due to the lack of genome completeness and ambiguities in resolving this region. The strongest evidence for this is that our best *C. viridis* genome sequence ends in an open reading frame without a terminal stop codon (Supplementary Fig. S12).

Considering evidence for a putatively functional PRDM9 gene in multiple snake genomes, we next tested several major predictions of PRDM9 activity in meiotic recombination. First, if PRDM9 plays a role in directing meiotic recombination in snakes, it would have to be expressed in germline tissue, as observed in other vertebrates (Baker et al. 2017). Consistent with this expectation, we find evidence that PRDM9 is expressed at high levels in snake germline tissues, especially in testes (Fig. 6a). These findings support the view that not only is PRDM9 present in snake genomes, but that its expression pattern in snake tissues resembles patterns observed in PRDM9-active species, such as humans (Fig. 6a; Hayashi et al. 2005; Oliver et al. 2009).

We further evaluated evidence for a functional role of PRDM9 in meiotic recombination in snakes by testing for a relationship between predicted PRDM9 nucleotide binding sites and recombination. Our searches for putative PRDM9-binding motifs, however, were limited by having a partial array of three tandem ZFs for *C. viridis* (see Materials and Methods). First, we predicted the DNA-binding motifs of the first three tandem ZFs in PRDM9 from *C. viridis* and the two distantly-related snake species for which we also confirmed germline *PRDM9* expression (Fig. 6a-b), and the partial PRDM9 binding motif of *Boa constrictor* as an additional comparison. We used these species-specific predictions to identify candidate binding sites of these orthologous proteins in the *C. viridis* genome. We also compared distance from binding sites with recombination rate in CV and CO, annotated the density of predicted binding sites across the genome, and examined the proximity of species-specific binding sites to hotspots. We found

that relative recombination rate increased with proximity to predicted *C. viridis* PRDM9 binding sites (Fig. 6c-d), with a significant negative association between recombination rate and distance. This relationship was stronger in CV than CO (Spearman's rank order correlation coefficients, CV $r = -0.57$, $p < 2.2 \times 10^{-16}$, CO $r = -0.46$, $p < 2.2 \times 10^{-16}$).

Lineages with active PRDM9 are characterized by rapid turnover of recombination hotspots due to rapid evolution in PRDM9 binding sites across populations and species (Ségurel et al. 2011). We tested this prediction of PRDM9-activity by evaluating how well the divergent PRDM9 binding motifs from different snake species (*Python* and *Deinagkistrodon*; Fig. 6b) predicted recombination landscapes in rattlesnakes. The predicted PRDM9 binding sites from other snake species were poorer predictors of fine-scale recombination rate than those predicted from the *C. viridis* partial motif (Fig. 6e-h). Distance from PRDM9 binding site predictions from *Deinagkistrodon*, which diverged from *Crotalus* ~ 33 MYA, was positively correlated with recombination rate in the two species (CV $r = 0.148$, $p = 2.3 \times 10^{-6}$, CO $r = 0.142$, $p = 6.76 \times 10^{-6}$). Recombination rate also increased slightly with increasing distance from *Python* PRDM9 binding sites (~ 91 million years divergent; CV $r = 0.066$, $p = 0.04$, CO $r = 0.022$, $p = 0.48$), though binding sites matching the partial *Python* PRDM9 motif were comparatively rare, and this relationship was non-significant in the case of CO. *Boa* PRDM9 binding sites were even more rare (507 total), and did not exhibit any association with recombination in either rattlesnake species (Supplementary Fig. S13).

The rapid evolutionary turnover of hotspots expected in PRDM9-active species also predicts that hotspots should be enriched for PRDM9 binding sites in a species-specific fashion (i.e., hotspots of a particular species should be most highly enriched for the PRDM9 binding motif from that same species). Our comparisons with rattlesnake genome data support these patterns, demonstrating that CV-specific hotspots were significantly enriched for the *C. viridis* PRDM9 binding motif (Fisher's exact test; $p = 4.68 \times 10^{-17}$), while CO-specific hotspots were not (Fisher's exact test; $p = 1.0$).

Together, our recombination analyses indicate that recombination rate increases with proximity to promoters, CGIs, and *C. viridis* PRDM9 binding sites (Fig. 4, Fig. 6c-d). These results suggest that, if PRDM9 is involved in directing recombination, it does not appear to strongly direct recombination away from genes. Consistent with this prediction, we observed high densities of *C. viridis* PRDM9 binding sites in CGIs, promoters with CGIs, and exons (Fig. 6i). Predicted binding sites derived from PRDM9 orthologs from other snake species were predicted to occur rarely and relatively evenly across genomic features in the CV genome (Fig. 6j-k). To test if PRDM9 may target recombination towards functional

gene regions, we examined the relationship between hotspot density and PRDM9 binding site density across the nine non-overlapping genomic features used to annotate the density of hotstpost above. We found positive correlations between rattlesnake PRDM9 binding sites and hotspot density across features in CV and CO (Supplementary Fig. S14; Pearson's correlation coefficients, CV $r = 0.76$, $p = 0.017$, CO $r = 0.75$, $p = 0.019$). In an attempt to disentangle the effects of functional regions and the presence of PRDM9 binding sites, we tested if PRDM9 binding sites in intergenic regions were associated with increased recombination. These analyses did recover significant relationships, although weaker than those observed when we considered all PRDM9 binding sites (Supplementary Fig. S15; Spearman's rank-order correlation coefficients; CV $r = -0.143$, $p = 6.3 \times 10^{-6}$; CO $r = -0.157$, $p = 6.0 \times 10^{-7}$). Further supporting a role of PRDM9, we observed higher recombination rates in CGIs and promoters that contained predicted PRDM9 binding sites than those that did not (Supplementary Fig. S16; Wilcoxon-Mann-Whitney U tests; CGI p-values $< 2.2 \times 10^{-16}$; promoter p-values $< 1.4 \times 10^{-15}$).

Discussion

Snakes provide a unique example of a vertebrate recombination landscape with a strong concentration of recombination hotspots in functional regions, despite the presence of an apparently functional PRDM9. Using a complement of whole genome resequencing, gene expression, and comparative genomics analyses, we investigated recombination rate variation across snake genomes and found consistent evidence of increased recombination near gene regions as well as PRDM9 activity. This pattern of apparent PRDM9 activity, genomic hotspot location, and hotspot evolution deviates from the canonical patterns of recombination landscapes defined largely by birds and mammals. Our results also provide among the first estimates of recombination rates in squamate reptiles, highlighting substantial within- and between-chromosome variation in recombination rates. These inferences of high recombination rates and high rate variation in snake microchromosomes, similar to that observed in birds, suggest that these characteristics may be inherent general features of microchromosomes. Collectively, our findings illustrate the value of using non-traditional model systems capable of offering both novel and confirmatory perspectives on the diversity of mechanisms and modalities underlying fundamental processes, such as meiotic recombination.

PRDM9 function in snakes

Baker et al. (2017) found evidence to support a very deep (i.e., ancestral) origin and broad conservation of PRDM9 within vertebrates, but also repeated losses of the *PRDM9* gene and PRDM9 function. Among

squamates, it is known that the *PRDM9* gene has been lost at least once in lizards (i.e., *Anolis carolinensis*; Singhal et al. 2015; Baker et al. 2017), and there have been wholesale losses in other major reptilian lineages (e.g., crocodilians and birds). Previous support for PRDM9 function in snakes was based primarily on the presence of a *PRDM9* ortholog that contained a full complement of KRAB, SSXRD, SET, and C2H2 Zinc Finger (ZF) domains in the genomes of multiple divergent snake species, and on the rapid evolution of the PRDM9 ZF domain responsible for DNA-binding (Baker et al. 2017) – all hallmarks of PRDM9 activity (Myers et al. 2010; Axelsson et al. 2012; Schwartz et al. 2014; Stevison et al. 2015). However, links between the presence of PRDM9 and recombination in snakes have not been demonstrated due to a lack of high-quality genomic resources for snakes and population genomic data needed for estimation of recombination maps.

Here, we present new data consistent with PRDM9 function in snakes, including that the *PRDM9* gene is identifiable in multiple snake genomes (e.g., *Python*, *Deinagkistrodon*, and *Crotalus*), contains no missense mutations or premature stop codons in any available snake genome (Baker et al. 2017; Supplementary Fig. S12), and is expressed in the germline of each species for which RNAseq data were available (Fig. 6a). Because snakes included in this study represent multiple diverse lineages that span more than 90 MY of divergence (Zheng and Wiens 2016), PRDM9 function appears to be a conserved trait in snakes. Additionally, at a fine scale, evidence that a majority of identified recombination hotspots are species-specific when we compare CV and CO (which diverged ~3 MYA; Fig. 3d; Supplementary Fig. S1), is consistent with relatively rapid hotspot turnover due to PRDM9 activity. We found additional support for the activity of PRDM9 in snakes based on the relationships between recombination hotspots, predicted PRDM9 binding sites, and relatively high species-specificity of these relationships. Specifically, we identified strong correlations between species-specific PRDM9 DNA binding motif sequences and hotspots, as well as between binding motif sequences and fine-scale recombination rates (Fig. 5c-d). Consistent with the rapid evolution of PRDM9 (and its DNA binding motif) driving rapid turnover of hotspots, we found that only species-specific PRDM9 DNA binding motif sequences were good predictors of recombination rates in rattlesnakes, whereas binding motifs from other snake species were poor predictors of recombination (Fig. 6g-h). The *C. viridis* PRDM9 binding site was also only enriched in CV-specific hotspots, and not in CO-specific hotspots, further suggesting that rapid turnover of recombination hotspots between closely related rattlesnake species is related to the action of PRDM9. Importantly, our inferences of PRDM9 binding are limited by our recovery of a partial ZF binding motif for *C. viridis*, and further work is needed to fully investigate the binding of remaining ZFs in the PRDM9 array. Nonetheless, these combined lines of evidence together suggest that PRDM9 functions in snakes,

as it does in mammals and other vertebrates, as a mechanism for directing the genomic location of recombination hotspots.

Snake recombination occurs in functional regions despite PRDM9

How and where recombination hotspots arise in vertebrates are often explained by two divergent models of meiotic recombination that differ in whether PRDM9 is active or not. In species with a functional PRDM9, recombination hotspots form through double stranded breaks that occur due to PRDM9 binding to specific genomic motifs, leading to the recruitment of the recombination machinery (Keeney 2007; Borde et al. 2009; Lam and Keeney 2014). Hallmark features of PRDM9-active systems are that recombination is directed away from gene regions and that recombination hotspots experience rapid evolutionary turnover (Baudat et al. 2010; Myers et al. 2010; Baudat et al. 2013; Stevison et al. 2015). Alternatively, species lacking PRDM9 have hotspots localized to promoters, CGIs, and other H3K4me3-rich regions (Axelsson et al. 2012; Auton et al. 2013; Singhal et al. 2015; Baker et al. 2017; Kawakami et al. 2017; Schumer et al. 2018). Recombination hotspots in these species are also stable even over millions of years of evolution. This latter model of hotspot evolution has been observed repeatedly in anciently diverged lineages that have lost *PRDM9* or lost PRDM9 function (Baker et al. 2017), and has come to be considered the PRDM9-less 'default model'. Studies of canids, which lack PRDM9 function, and PRDM9 knock out mice lend further support to this hypothesis, as their hotspots localize to gene promoters, rather than intergenic regions (Brick et al. 2012; Auton et al. 2013).

We find that snakes represent a unique case in which PRDM9 function appears to enhance the default targeting of meiotic recombination to promoters and other functional elements (Figs. 4-5), rather than recruiting the recombination machinery to genomic regions away from genes. Our inferred recombination maps for snakes provide consistent evidence that snake hotspots are localized to promoters, CGIs, and other functional regions (e.g., first exons), with recombination rates increasing with proximity to genes. Combined with evidence of PRDM9 activity and positive correlations between hotspot and PRDM9 binding site densities in snakes, these findings suggest that PRDM9 does not direct recombination away from genes, and instead may facilitate or reinforce the targeting of recombination to promoters and functional elements (i.e., typically the default for PRDM9-inactive systems; Figs. 4,6, Supplementary Fig. S7). PRDM9 reinforcement of this default pattern is supported by our finding of higher recombination rates in promoters and CGIs that contain PRDM9 predicted binding sites (Supplementary Fig. S16).

Our conclusion that PRDM9 activity in snakes appears to enhance the presumed default recombination mechanism, by targeting promoters and functional elements, expands the known repertoire of PRDM9 function and raises the question of how widespread the targeting of functional regions by PRDM9 may be in other vertebrates. Evidence that PRDM9 is under positive selection in mammals (Oliver et al. 2009; Ponting 2011) has led to speculation for an adaptive role of PRDM9 in deflecting recombination hotspots away from function genomic regions (Brick et al. 2012). Our findings that PRDM9 in snakes may direct recombination toward functional regions challenges this hypothesis, and together with evidence for multiple losses of PRDM9 in vertebrates (Baker et al. 2017), raises further questions about the evolutionary significance of directing recombination away from genic regions. In an analogous departure from simple bifurcating models of recombination landscapes, a recent study in stickleback fish demonstrated rapid turnover in recombination hotspots, despite evidence for weak PRDM9 activity, and weak associations between PRDM9 binding and hotspot densities (Shanfelter et al. 2019). Our findings, along with those from sticklebacks, illustrate the potential for studies of diverse vertebrate systems to reveal novel deviations from the canonical models of vertebrate meiotic recombination.

A potential role for CTCFL in snake recombination

The CTCF gene family, including CTCF and CTCFL, plays central roles in directing chromatin loops that modulate the associations between genes and their regulators (Phillips and Corces 2009; Merkenschlager and Odom 2013; Ong and Corces 2014). They are also well known as ‘insulator proteins’, due to their role in precisely directing (or preventing) interactions between enhancers and promoters, and have been shown to be important in mammalian imprinting (Bell and Felsenfeld 2000; Rao et al. 2014). Recently CTCF has been shown to perform a secondary function of maintaining genomic stability in double stranded break regions in mammals (Lang et al. 2017; Hwang et al. 2019). While less well studied, CTCFL has been shown in mammals to be expressed more highly in germline tissues, and to bind to the recognition sites of CTCF (Loukinov et al. 2002), although it may bind differentially depending on nucleosome composition (Sleutels et al. 2012). The emerging hypothesis from mammalian studies is that CTCFL may play an important role in germline development (Sleutels et al. 2012), possibly due to its function in stabilizing double strand breaks during meiotic recombination.

We inferred that a large proportion of rattlesnake recombination hotspots contained CTCF and CTCFL binding sites, and further observed high relative expression of CTCFL in the snake germline (Fig. 5). Our findings that snake hotspots are enriched for CTCF family binding sites parallels previous findings that mammalian recombination hotspots exhibit an abundance of these binding sites (Wu et al. 2012; Kaiser

and Semple 2018), and provide additional evidence for a broad role of CTCF proteins in meiotic recombination in vertebrates. Specifically, considering the high frequency of double strand breaks associated with hotspots, the enrichment of CTCF/CTCFL binding sites in these regions argues further for the functional role of this group of proteins in genome stability in the context of meiotic recombination. These results also represent the first evidence for the importance of these proteins in recombination outside of mammals, suggesting that CTCFL may play similar functional roles across diverse vertebrate lineages. Altemose et al. (2017) also showed that PRDM9 can positively regulate the expression of CTCFL in the human germline, suggesting an intriguing secondary function of PRDM9 in general meiotic function, and a link between CTCFL activity and PRDM9 activity.

Recombination variation within and among chromosomes

Microchromosomes are a widespread feature of vertebrate genomes. All major amniote lineages, except mammals and crocodilians, possess microchromosomes (O'Connor et al. 2018), as do most fish and non-anuran amphibians (Voss et al. 2011; Braasch et al. 2016). A recurring theme among these species is that recombination rates on microchromosomes tend to be higher than rates on macrochromosomes (Backström et al. 2010; Roesti et al. 2013; Burri et al. 2015; Singhal et al. 2015). In the zebra finch, for example, microchromosomes exhibit consistently elevated recombination rates reminiscent of telomeric regions of macrochromosomes (Backström et al. 2010). To date, recombination rate estimates from microchromosome-possessing species have been almost entirely based on analyses of bird genomes (e.g., Backström et al. 2010; Singhal et al. 2015; Kawakami et al. 2017) – our results therefore corroborate that high recombination rates may be a consistent feature of vertebrate microchromosomes in general. Snake microchromosomes also have greater GC richness, density of CGIs, gene content, and genetic diversity than macrochromosomes (Schield et al. 2019), features which are broadly reminiscent of avian microchromosome structure and genetic diversity (McQueen et al. 1996; Smith et al. 2000; Hillier et al. 2004; Backström et al. 2010; Warren et al. 2010). The finding that features of microchromosome composition and recombination show similar patterns in snakes and birds suggests the existence of common links between microchromosome structure, function, and evolution between these divergent groups that would be interesting to evaluate across other vertebrate lineages.

Why recombination rates are elevated on microchromosomes is not entirely clear. Previous explanations have implicated mechanistic and structural factors, such as an obligate chiasma per meiosis regardless of chromosome length (Jones and Franklin 2006), the unique compositional properties of microchromosomes described above, and the observation that microchromosomes are rich in open

chromatin that may favor recombination (McQueen et al. 1996; McQueen et al. 1998). Our findings demonstrate that mean recombination rate in snakes is negatively correlated with chromosome length, and positively correlated with gene density (Fig. 1, Table 1) – broadly supporting multiple hypotheses for mechanisms that may explain elevated recombination on microchromosomes, as well as the potential synergistic effects of an obligate chiasma per meiosis and open chromatin density. Together, evidence for the distinct structural and evolutionary properties of microchromosomes and their prevalence across vertebrates pose broad questions about the evolutionary significance of microchromosomes and the genes that they contain, which should be far less constrained by genetic linkage than genes on macrochromosomes.

Previous studies have illustrated associations between recombination rates, targets of selection (i.e., genes), and polymorphism across the genomes of diverse taxa, supporting the hypothesis that natural selection and recombination shape the distribution of genetic diversity across the genome, and ultimately shape processes of speciation and adaptation (Ellegren et al. 2012; McGaugh et al. 2012; Burri et al. 2015; Wang et al. 2016). Our findings support this hypothesis in snakes, as we found significant genome-wide associations between recombination rates, gene density, and polymorphism (Fig. 1), suggesting that genetic diversity in snakes is also shaped by linked selection. Linked selection has also been shown to generate genomic islands of differentiation between lineages and species (Nadeau et al. 2012; Burri et al. 2015; Martin et al. 2019). Further investigation into the roles of recombination and natural selection in driving genomic differentiation between snake lineages will be required to test this hypothesis, but would be valuable as a comparison to other vertebrate systems to evaluate the degree to which linked selection drives speciation.

Within chromosomes, our results demonstrate that recombination in snakes is highly heterogeneous and is concentrated toward the ends of chromosomes (Fig. 1, Supplementary Fig. S2). This pattern is similar to patterns observed in birds and mammals (Jensen-Seaman et al. 2004), with more pronounced telomere effects in birds (Backström et al. 2010). In humans, this pattern has been attributed to the greater relative rate of male recombination concentrated in telomeres (Broman et al. 1998), and it is suggested that differences in the magnitude of the telomere effect between mammal species may be driven by differences in male meiosis (Jensen-Seaman et al. 2004). While our sex-averaged estimates of recombination rate in snakes preclude us from drawing conclusions about male-biased recombination, the similarity of telomere-effect patterns between divergent vertebrate species suggests that the concentration of recombination in telomeres may be driven by mechanisms that are conserved across amniote vertebrates.

Conclusion

We examined snake recombination landscapes for the first time using population genomic data from two rattlesnake species. We found evidence that snakes have recombination hotspots, and these hotspots evolve rapidly among lineages, consistent with observed patterns of PRDM9 activity in directing meiotic recombination. Snake recombination hotspots and higher recombination in general are focused in promoters and genes, which is otherwise associated with vertebrates that lack an active or functional PRDM9. Multiple lines of evidence suggest that snakes represent an outlier among vertebrates because, unlike vertebrates studied to date, they appear to have a functional PRDM9 that may reinforce, rather than counteract the targeting of recombination to functional regions. We also find evidence that CTCFL binding sites are enriched in snake recombination hotspots, providing further evidence for a potential role of CTCFL in maintaining genome stability in regions of frequent recombination. Finally, broad patterns across the genome illustrate the potential for variation in recombination rate to shape the distribution of genetic diversity, and suggest that elevated recombination rates on microchromosomes may be a common feature across vertebrate lineages. Our conclusions that snake recombination mechanisms break with existing paradigms for vertebrates highlight the value of investigating diverse vertebrate lineages to understand variation in recombination mechanisms, and how this variation may shape vertebrate genome evolution.

Materials and Methods

Reference genome and annotation

Previously, we assembled and annotated a reference genome for the prairie rattlesnake (*Crotalus viridis*; CroVir3.0, (Schield et al. 2019)). The assembly includes scaffolds corresponding to the 18 chromosomes in the rattlesnake karyotype (2n = 36), and 17,352 annotated protein-coding genes. We also previously annotated repeat elements throughout the genome using RepeatMasker (Smit et al. 2015). We used this genome assembly as the reference for read mapping, and used gene and repeat annotations from the previous study for downstream analyses. The reference genome is available through NCBI BioProject accession PRJNA413201.

We identified CpG islands (CGIs) in the prairie rattlesnake genome we used the EMBOSS v.6.6.0 ‘cpgplot’ function (Larsen et al. 1992). We specified a search window of 500 bp and a minimum CGI

length of 250 bp. CGIs were only called if the observed versus expected ratio of CpG content exceeded 0.6 and the proportion of GC bases within the window was greater than 0.5. Overlapping predicted CGIs were then collapsed into single contiguous CGI regions, which resulted in 43,538 CGIs in the rattlesnake genome. We measured CGI density throughout the genome using a Python script 'window_quantify_CGIs.py' (https://github.com/drewschield/recombination/), which we ran on the GFF file generated by cpgplot. Specifically, we measured the density of CGIs as the proportion of bases within a window annotated as CGIs. We defined promoters as the 2 kb region upstream of transcription start sites (TSSs), based on the prairie rattlesnake genome annotation described above. To annotate the recombination landscape, we divided the genome into intervals of nine non-overlapping feature categories: intergenic regions, repeat elements, CGIs, promoters with CGIs, promoters without CGIs, first exons, first introns, other exons, and other introns. For intergenic regions, we extracted all regions that were not already annotated as genes, promoters, CGIs, or repeat elements using the bedtools 'complement' program (Quinlan and Hall 2010). We identified CGIs not associated with genes or promoters in the same way, and used bedtools 'intersect' to identify promoters containing CGIs. We used a custom Python script 'parse_first_exon_intron.py' (https://github.com/drewschield/recombination/) to parse first exons and introns from the genome GFF annotation file, then used bedtools 'complement' using the output from our script and the GFF to obtain remaining exons and introns from all genes.

Whole genome resequencing, mapping, and variant calling

We sampled populations of the prairie rattlesnake (*Crotalus viridis*; n = 21) and the northern Pacific rattlesnake (*C. oreganus*; n = 17) for whole genome resequencing. For simplicity, we refer to these taxa as CV and CO throughout this study. We also sampled an individual western diamondback rattlesnake (*C. atrox*), red diamondback rattlesnake (*C. ruber*), and Mojave rattlesnake (*C. scutulatus*; Supplementary Table S1) as outgroup taxa for phylogenetic and ancestral allele inferences. DNA was extracted from blood and liver tissue that was either snap frozen or preserved in DNA lysis buffer. We extracted DNA using a standard phenol-chloroform-isoamyl alcohol extraction and precipitation. Genomic sequencing libraries were generated from purified DNA using KAPA HyperPlus and Illumina Nextera DNA Flex kits, multiplexed together, and sequenced on multiple Illumina NovaSeq 6000 S4 lanes using 150 bp paired-end reads. Raw sequencing reads were adapter trimmed and processed using the Illumina BaseSpace distributions of FastQC v1.0.0 and Fastq Toolkit v2.2.0 (BaseSpace Labs). Bases on 5' and 3' ends of reads with quality scores less than 20 were trimmed, and any reads with a final length less than 36 bp, or with an average quality score below 30 were removed. We aligned trimmed and filtered reads to the prairie rattlesnake reference genome using the BWA v0.7.1 'mem' algorithm (Li and Durbin 2009)

with default settings, and marked duplicate mappings for downstream removal. These mappings yielded an average coverage of 36.7× per sample (range = 9 – 106.2×; Supplementary Table S1).

We called indel and single nucleotide polymorphism (SNP) variants from individual mappings using the Illumina BaseSpace distribution of the Dragen Germline pipeline v3.0 variant caller (Edico Genomics), which generated a genomic variant call file (gVCF) for each individual. We then used the GATK v3.8.1 'genotypeGVCFs' module (McKenna et al. 2010) to call population variants within CV and CO separately. We performed several post-processing steps using BCFtools (Li et al. 2009) and VCFtools (Danecek et al. 2011) to remove potential artifacts from low quality variant calls. Specifically, we filtered indels from downstream analysis and retained only biallelic SNPs with a genotype quality score greater than 10, filtered all SNPs within 5 bp of an indel, and avoided variant calls in highly repetitive regions by filtering all SNPs that overlapped with repeat elements annotated in Schield et al. (2019). We also removed all variant sites on the Z chromosome (other than the PAR) that had erroneous heterozygous calls in known females after previous filtering steps, as these are the likely product of variation between Z and W gametologs. After variant calling and filtering, we retained 14,102,890 SNPs in CV and 13,471,063 SNPs in CO for downstream analyses of divergence times, linkage disequilibrium (LD)-based population-scaled recombination rates, polymorphism, and allele frequency distributions. We used the number of filtered SNPs to estimate within-lineage diversity (θ_W), according to Watterson (1975), based on a total sequence length of 1.34 Gbp, with 2n = 36 chromosomes.

Divergence time estimation

To obtain an estimate of the divergence time between CV and CO, we performed divergence date estimation using the Bayesian framework implemented in SNAPP (Bryant et al. 2012). To obtain an input SNP alignment, we sampled at random two individuals from our sampling of CV and CO, and also included data from *C. atrox* and *C. ruber* (see above). We then called variants using the same procedures and settings as above for variant calling in CV and CO, but to make analyses computationally tractable we also thinned the SNP matrix so that SNPs were at least 100 kb apart, and filtered to remove sites where data was present in fewer than four individuals after previous filtering steps. This procedure resulted in 15,712 SNPs used in phylogenetic analyses. To generate a SNAPP input XML file for divergence dating, we followed the protocol of (Stange et al. 2018), using their provided Ruby script ('snapp_prep.rb'). In SNAPP, we constrained two nodes in the species phylogeny using priors from Reyes-Velasco et al. (Reyes-Velasco et al. 2013): i) the ancestral node of *Crotalus atrox* and *C. ruber* (3.2 MYA offset, SD 1), and ii) the root ancestor for the group (6.1 MYA offset, SD 1). We ran two parallel

analyses, and ran each MCMC chain for 1 x 10^7 generations, sampling every 1×10^4 generations. We combined the posterior samples of both into a single posterior distribution after removing the first 25% of iterations from each chain as burn-in, then generated a maximum clade credibility consensus tree using TreeAnnotator (Bouckaert et al. 2014).

Haplotype phasing

Because singleton variants may make haplotype phasing difficult and lead to spurious downstream results, we first identified and filtered out singleton variants from CV and CO using BCFtools (Li et al. 2009), leaving 9,751,928 SNPs in CV and 6,373,627 SNPs in CO. We phased variants for each chromosome independently using SHAPEIT v2.904 (Delaneau et al. 2013), after identifying phase-informative reads (PIRs) with the 'extractPIRs' extension distributed alongside the software package. PIRs are sequencing reads that span at least two heterozygous SNPs, which extractPIRs identified from our mapping files. We first separated VCF files by chromosome and extracted PIRs, specifying a mapping quality score > 20, then ran SHAPEIT using the parameters –states 1000 –burn 200 –prune 210 –main 2000, and assessed switch error rate using the VCFtools 'diff-switch-error' function (Danecek et al. 2011), in order to determine if the SHAPEIT MCMC had converged between individual runs. Using the parameter settings above, our runs resulted in low mean switch error (1.2% in CV and 2.2% in CO) between independent runs of the phasing algorithm.

Estimation of linkage disequilibrium-based recombination rate

We estimated the recombination maps for CV and CO from phased haplotypes using the linkage disequilibrium-based approach implemented in LDhelmet (Chan et al. 2012). In addition to haplotypes, LDhelmet takes an estimate of population-scaled genetic diversity (i.e., Watterson's θ), and prior estimates of ancestral allele states and a 4×4 mutation transition matrix as input. Here, we used an average of the Watterson's θ estimates for CV and CO described above (0.005), and generated ancestral allele and mutation matrix priors for each chromosome using the Perl script described in Shanfelter et al. (2019), using *C. atrox* and *C. scutulatus* as outgroup taxa for ancestral allele inference with the topology (*C. atrox*, (*C. scutulatus*, (*C. viridis*, *C. oreganus*))) (Supplementary Table S1). For each SNP in CV and CO, we assigned an ancestral prior if the SNP was present in both *C. atrox* and *C. scutulatus* and if both outgroup species were homozygous for the same allele. Following Singhal et al. (2015) and Shanfelter et al. (2019), we assigned the inferred ancestral state a prior probability of 0.91, and each other allele state a

uniform prior of 0.03. Where we could not determine the ancestral allele, the prior probability per allele was set equal to its frequency across the entire chromosome.

We generated input SNP sequence and position files for LDhelmet using a custom Python script, modified from Shanfelter et al. (2019), and generated full sequence fasta inputs using the vcflib component 'vcf2fasta' (https://github.com/vcflib/vcflib), which were formatted using an additional Python script ('change_fasta_header.py'). We generated haplotype configuration files for each chromosome using the 'find_confs' module, setting the window size to 50 SNPs. We then used the 'table_gen' module to produce a likelihood lookup table per chromosome, with the grid of population-scaled recombination rate per base pair (ρ/bp) values recommended in the LDhelmet manual (-r 0.0 0.1 10.0 1.0 100.0; (Chan et al. 2012)). We generated padé coefficient files to be used in the reverse-jump MCMC procedure using the module 'pade'. For table_gen and pade steps, we specified the population-scaled diversity parameter –t = 0.005. We then estimated recombination rates in CV and CO using the rjmcmc module, setting a window size of 50 SNPs, a burnin of 100,000 generations, and 1,000,000 sampled generations per run. We performed rjmcmc analyses under two block penalties, 10 and 100 – lower block penalties are shown to more reliably capture fine-scale recombination variation, and larger block penalties are useful for characterizing the broad genomic landscape of recombination (Singhal et al. 2015). Finally, we converted the output of the rjmcmc module using the 'post_to_text' module, and used a custom Python script to calculate ρ/bp in 10 kb, 100 kb, and 1 Mb sliding windows as the mean value of ρ for all sampled positions in per window. For these steps, we masked assembled centromere regions identified in Schield et al. (2019), as these exhibited spurious recombination rates in preliminary runs, likely due to local over- and under-assembly. We then combined windowed results from each chromosome per species using custom scripts. Bash and Python scripts used for LDhelmet analysis and for processing MCMC results are available at https://github.com/drewschield/recombination.

Recombination variation and relationships with other genomic features

To characterize within-chromosome variation in recombination rate, we identified candidate telomere regions based on Schield et al. (Schield et al. 2019), and using sliding window measures of GC and repeat content. Here, we calculated GC content in 1 Mb genomic windows of the *C. viridis* reference genome as the proportion of G+C bases over total non-ambiguous bases in each window, and we measured repeat density in 1 Mb sliding windows as the total number of repeat element bases annotated by RepeatMasker divided by the total window length. Because patterns on microchromosomes were less clear, we limited our qualitative comparison of recombination rates in centromere and telomere regions to

macrochromosomes. Following Bäckstrom et al. (2010), we compared recombination rate and distance to chromosome end by sampling the 15 Mb end regions of each macrochromosome, and we tested the relationship between variables using Pearson's correlation coefficients in R (R Core Team 2017).

We then compared LD-based recombination rate to genetic diversity and other features of the rattlesnake genome using pairwise and partial Spearman's rank order correlation coefficient analyses between recombination rate estimates per species and nucleotide diversity (π), GC content, gene density, and repeat content. For these comparisons, GC content and repeat content were measured as described above, and gene density was calculated as the number of annotated genes in a given 1 Mb sliding window. We measured π using the VCFtools 'window-pi' function (Danecek et al. 2011). All pairwise correlation analyses were performed using the 'cor.test' function in R, and partial correlation analyses were done using the R package 'ppcor' (Kim 2015). In partial correlation analyses, we controlled for non-predictor variables by including them as potential confounding variables. For example, for partial correlations between recombination rate and π, we accounted for GC content, gene density, and repeat content. For partial correlations between recombination rate and GC content, gene density, and repeat content, we specified the two respective non-predictor variables as potentially confounding variables (Table 1). Finally, to explore relationships at multiple resolutions, we repeated all analyses using measurements in 100 kb windows. We further visualized the pairwise genomic relationships between variables using the 'smooth.spline' function in R, using the settings 'spar' = 0.6 and 'nknots' = 10.

We explored recombination variation across distinctive regions previously identified in Schield et al. (Schield et al. 2019) that may constitute evolutionary strata of the Z chromosome, including the pseudoautosomal region (PAR), the recent stratum, and older strata. For comparison to sliding window recombination rate estimates across the Z in CV and CO, we used measure of gene density and GC content detailed above, and also used measures of normalized measure of female and male π from the previous study. Here, normalized π was calculated by dividing the value of π in sliding windows by the autosomal median π value for each sex. These measures were used previously to characterize the recent stratum, which was inferred to have been recently recombination suppressed between the Z and W chromosomes, and which bears a unique pattern of high normalized female π and normalized male π that is roughly equivalent to the older strata. We used Welch's two-sample *t*-tests to compare distributions of recombination rate between the PAR and other Z-linked regions and autosomes.

Power to detect recombination hotspots

We performed a simulation study in order to determine our power to identify hotspots at various background recombination rates, hotspot 'heats' relative to flanking regions, and block penalty parameters. We used the coalescent simulator MaCS (Chen et al. 2009) to simulate 1 Mb sequences under different background recombination rates (ρ/bp = 0.00002, 0.0002, 0.002, 0.02, 0.2) for 38 haplotypes (i.e., 19 diploid individuals, the average number of individuals in our empirical datasets), setting population scaled mutation rate equal to –t = 0.005. We simulated two sets of 2.5 kb recombination hotspots with 5×, 10×, 20×, and 40× relative heat in each replicate, and performed at total of 50 replicates for each ρ/bp value. We converted the results of coalescent simulations to sequence alignments using Seq-Gen (Rambaut and Grass 1997) under the HKY substitution model, and estimated recombination rates per replicate using LDhelmet across multiple block penalties (5, 10, 20, and 50). Other LDhelmet analysis parameters matched those described above for our empirical data.

We then searched recombination maps for each parameter set in 2 kb windows, comparing recombination rate in each window to 40 kb up- and downstream regions in order to identify hotspots. Regions within the specified hotspot intervals set during coalescent simulations with greater than twice the background recombination rate were counted as true positives, and windows in flanking regions outside of specified hotspots with greater than ten times the background recombination rate were counted as false positives. We repeated these steps for each background recombination rate and block penalty parameter set in order to compare our power to detect true hotspots and false positive rates under different settings. These simulations demonstrate that we have higher power to detect hotspots at lower block penalties, and when background recombination rate is intermediate. For example, analyses using a block penalty of 50 consistently failed to identify true hotspots at high frequency, and we were unable to detect hotspots reliably when background recombination rate was very high or low (i.e., 0.02 or 0.00002). We note that, while power to detect hotspots was consistently highest under a block penalty of 5, these analyses also produced higher false positive rates. Analyses under a block penalty of 10, however, had lower overall power, but consistently lower false positive rates at different background recombination rates and hotspot heats (Supplementary Fig. S5). We therefore used a block penalty of 10 in our empirical identification of hotspots, to reduce the likelihood of inferring spurious hotspots. We also specified that putative hotspots must have at least ten-times the background recombination rate.

Recombination hotspot identification

Based on the results of our simulation study, and following the procedures of Singhal et al. (2015) and Kawakami et al. (2017), we used an operational definition based on the magnitude of relative population-scaled recombination rate (ρ) to identify candidate recombination hotspots in each species. Specifically, we defined potential hotspots as regions which had greater than ten-fold ρ/bp compared to 40 kb upstream and downstream regions. To search for potential hotspots, we calculated mean ρ/bp within 2 kb sliding windows, where we slid each window 1 kb per iteration. We calculated the ratio of relative 'heat' by dividing mean ρ/bp per window by the mean rate of the up- and downstream regions using a Python script 'identify_hotspots.py'. After identifying candidate hotspots, we filtered using a Python script 'filter_hotspots.py' to remove hotspots within 5 kb of another hotspot by assigning the window with the highest heat as the candidate hotspot for downstream analyses. We then characterized whether hotspots were specific to CV or CO, or if they were shared (i.e., < 5 kb apart), and examined mean relative recombination rate across all hotspots and their flanking regions in 1 kb windows using the deepTools2 'plotProfile' function (Ramírez et al. 2016).

For comparison to candidate hotspots, we identified a set of matched coldspots for CV and CO using several search criteria. First, we characterized all putative coldspots as genomic windows with a background ρ/bp between 0.001 and 1 and with heat between 0.9 and 1.1 using a Python script 'identify_coldspots.py'. Then, we matched hotspots to coldspots by identifying the candidate coldspot that was physically closest, but at least 25 kb away from the nearest hotspot, and that also had similar GC content (i.e., within 2%) to the nearest hotspot. We identified 9,253 and 10,662 matched coldspots in CV and CO, respectively, using these criteria. For further comparison of candidate hotspots and coldspots to the genomic background, we generated a random background sequence set by randomly sampling with replacement 2 kb genomic windows equal to the average number of hotspots identified in CV and CO, and repeated this procedure ten times. We refer to these sequences as the 'random background' set, which contains 99,520 random background sequences from the *C. viridis* genome. We then compared GC content between hotspots and the genomic background using Welch's two-sample *t*-tests, and measured Pearson's correlation coefficients between mean ρ/bp and CGI density in 1 Mb windows.

Fine-scale recombination and hotspot annotation

As detailed above, we calculated fine-scale recombination rate for CV and CO in 10 kb sliding windows. To study relative recombination rate as a function of distance from functional regions, we compared ρ in

each 10 kb window to the physical distance to the nearest promoter or CGI annotated in the rattlesnake genome. For each window, we calculated the distance to the nearest up or downstream feature using the bedtools 'closest' function (Quinlan and Hall 2010). We then used a Python script 'bin_rho_distances.py' to calculate the mean and standard deviation recombination rate in increasing distance intervals of 100 bp, for a maximum distance of 100 kb. We measured relative recombination rate in binned distance intervals by dividing the mean ρ/bp value per interval by the median ρ/bp among all intervals. We then used Spearman's rank order correlation coefficients in R to measure the relationships between distance from promoters and CGIs and recombination rates in CV and CO. To explore the potential interaction of CGIs in promoters, we also repeated our promoter analysis, but first separated promoters with and without overlapping CGIs, then calculated mean ρ/bp per distance interval from CGI promoters and non-CGI promoters.

We used the nine non-overlapping genomic features described in the 'Reference genome and annotation' section to investigate the density of recombination hotspots in each feature category. For each feature, we determined the overlap between hotspots and feature intervals in base pairs, then calculated hotspot density as the proportion of overlapping bases in total feature interval bases for hotspots in both species. We then repeated these analyses for matched coldspots from both species. To compare hotspot density to overall recombination rate across genomic features, we also measured the distribution of recombination rate in each feature category.

Given evidence for higher recombination in gene-associated regions and a general concentration of hotspots in promoters and CGIs, we also measured recombination rate in 500 kb up- and downstream regions of all genes. First, we calculated the distance of 10 kb windowed recombination rate estimates from the nearest gene using bedtools 'closest', retaining the up- or downstream orientation of the window relative to the orientation of the gene. We then calculated mean recombination rate in regions 500 kb up- and downstream of genes in 5 kb intervals using the script 'bin_rho_upstream_downstream.py' (Schield github). We tested for functional enrichment of genes that overlapped shared hotspots between CV and CO using WebGestalt (Liao et al. 2019), with all annotated genes with an assigned orthologous human ID as the background, and using default program parameters.

Identification of DNA motifs in recombination hotspots

We used components of the MEME suite (Bailey et al. 2009) to identify DNA sequence motifs enriched in recombination hotspots, using matched coldspots as control sequences. We first used MEME v5.1.0

(Bailey and Elkan 1994) to identify enriched motifs in hotspots using the 'zoops' option in the differential enrichment mode. This option ignores repeat motif occurrences within the same sequences, and therefore avoids reported on repetitive motifs. We performed analyses for all CV and CO hotspots versus coldspots, and for CV-specific and CO-specific hotspots versus coldspots, specifically. In each analysis, we set the number of motif bases between 6 and 30 in order to confine the motif search space. We compared enriched hotspots motifs from MEME searches to known motifs in the JASPAR database using Tomtom v5.1.0 (Gupta et al. 2007) to identify homology to known binding motifs. MEME runs indicated that CV and CO hotspots had motifs with similarity to CTCF/CTCFL binding motifs, so we further tested for enrichment of the *C. viridis* CTCF/CTCFL binding site (JASPAR motifs MA0139.1 and MA1102.1) in hotspots of each species using AME v5.1.0 (McLeay and Bailey 2010), again setting coldspots as control sequences, and tested for significant enrichment using Fisher's exact tests with a Bonferroni correction for multiple comparisons.

Snake *PRDM9* expression and identification of rattlesnake PRDM9 binding sites

Snakes possess a potentially functional *PRDM9* gene, with a complement of KRAB, SSXRD, SET, and zinc-finger (ZF) domains, and with fast-evolving ZFs (Baker et al. 2017). We identified a candidate rattlesnake *PRDM9* gene in the *Crotalus viridis* genome assembly detailed in the 'Reference genome and annotation' section using a BLASTp homology search with the candidate *PRDM9* from the Burmese Python (*Python bivittatus*) (Castoe et al. 2013). The prairie rattlesnake *PRDM9* gene model included KRAB, SSXRD, and SET domains, but lacked ZFs due to a gap in the assembly.

To construct a rattlesnake *PRDM9* gene model that includes ZFs, we first used Trinity (Grabherr et al. 2011) to generate a transcriptome from testis RNA-seq data. Total RNA was extracted from ~100 mg of snap-frozen testis tissue using Trizol reagent (Invitrogen). We then performed phase separation using BCP, followed by precipitation of RNA by isopropanol. We constructed an Illumina mRNAseq library using the Illumina TruSeq RNAseq kit, which included poly-A selection, RNA fragmentation, cDNA synthesis, and indexed Illumina adapter ligation. The constructed mRNAseq was then sequenced on an Illumina HiSeq4000 using 150 bp paired-end reads. We quality filtered and adapter-trimmed the raw RNA-seq data using Trimmomatic v0.36 (Bolger et al. 2014). We then performed *de novo* transcriptome assembly using forward and reverse paired reads in Trinity v2.2.0 (Grabherr et al. 2011) using program default settings. Following *de novo* transcriptome assembly, we translated all assembled transcripts and performed a second BLASTp homology search against the translated CDS database for the Burmese python and identified a single best-hit candidate PRDM9 protein sequence, which was also identical to

the annotated PRDM9 protein from the prairie rattlesnake genome assembly, except for the missing ZFs. We then queried the NCBI database with the candidate PRDM9 for *C. viridis* using an additional BLASTp search, which identified putative KRAB, SSXRD, SET, and C_2H_2 ZF domains (four total ZFs), and confirmed orthology to PRDM9 from other vertebrates.

Because *de novo* transcriptome assembly could produce potentially spurious results in the case of complex genes such as *PRDM9*, we separately used multiple approaches to assess the validity of the mRNAseq derived transcript, including Sanger sequencing, gap-filling approaches, and Illumina sequencing of amplicons. Approaches that provided evidence of additional sequence within the ZF-encoding exon of *C. viridis PRDM9* included the generation of two separate 10x Genomics linked-read sequencing libraries for a female *C. viridis* from the same population as the genome animal and PacBio long read data from Schield et al. (2019). Linked-read libraries were generated at the Texas A&M Genomics Core and sequenced on an Illumina NovaSeq 6000 using 150 bp paired-end reads. We assembled linked-read assemblies from these data using Supernova v.2.1.1 (Weisenfeld et al. 2017), setting the number of input reads to 560 million, as recommended by the program manual for the estimated *C. viridis* genome size. This produced two assemblies, *C. viridis* 10x assembly 'A' and 'B', with which we output scaffolds using 'pseudo-hap2' format.

We then aligned DNA and translated protein sequences from *Deinagkistrodon acutus* and the *C. viridis PRDM9* putative transcript to the *PRDM9* region of the *D. acutus* genome using the Exonerate models 'protein2genome' and 'coding2genome' (Slater and Birney 2005). This confirmed conservation of exon and intron boundaries and identified the extent of the incomplete portions of the *C. viridis* transcript. Using the same models, the *D. acutus* PRDM9 transcript was then aligned to the *C. viridis* genome region predicted to contain PRDM9 to identify exon and intron boundaries. This revealed the entire beginning portion of exon 1, from the start codon to the beginning of the *C. viridis* putative transcript. In addition, it placed exon 9, containing the zinc finger array, entirely within the gapped region of *C. viridis*. The alignment of *D. acutus* exons to each genome were used to anchor a sequence alignment between the *D. acutus* and *C. viridis* genome regions (Supplemental Fig. S11; Supplementary Data Files S1-S3), and intron alignments were refined using both Muscle (Edgar 2004) and manual edits. Additional genome sequence from *D. acutus* beyond the end of the protein-coding region was included to span the gapped portion of the *C. viridis* genome region, producing alignments for both sequences flanking the gap. While *D. acutus* exon 9 was contained entirely within the gap region, there was not a stop codon present in frame until the aligned region in the sequence flanking the gap, suggesting that the end portion of the *C. viridis* zinc finger array was present in the assembly.

We then performed a BLASTn search (Altschul et al. 1990; Camacho et al. 2009) of the 1,587 bp region of *D. acutus* that spanned the *C. viridis* gap against 10x *C. viridis* assembly A, 10x *C. viridis* assembly B, and the un-aligned, error-corrected PacBio reads. The *D. acutus* and translated *de novo* transcriptome predicted *C. viridis* protein sequences of the exon contained in the gap were then aligned to each BLAST result using Exonerate 'protein2genome' to confirm its presence and verify homology. Each result was then aligned to the *C. viridis* genome sequence flanking the gap regions, to each other, and to the *D. acutus* region spanning the gap to confirm sufficient amount of overlap between sequences with high sequence identity. 'Scaffold 468599' from *C. viridis* 10x assembly A aligned and covered the two exons preceding the gap, and extended into the *C. viridis* gap region, with 3,887/5,491 bases covering the region flanking the gap. A second scaffold ('Scaffold 332766') from *C. viridis* 10x assembly B then produced BLAST results with substantial overlap (with 'Scaffold 468599' from the other assembly coverage over the new 10x scaffold for 1,023/1,090 bp, and extended further into the gap region. Finally, a single long PacBio read had high-similarity hits to both ends of the gap and the new region filled by the 10x scaffolds, but contained a large number of erroneous bases that precluded the prediction of the ZF array beyond the consensus sequence from the *C. viridis* genome, *C. viridis* 10x assemblies, and the PacBio sequence upstream of the erroneous region (Supplementary Datasets S1, S2). When annotated using FGENESH+ (Solovyev et al. 2006), this *C. viridis* consensus sequence included three tandem ZFs, the first two of which matched exactly those recovered using *de novo* transcriptome assembly (Supplementary Dataset S3). The third ZF differed between sequences produced by these approaches, which, given the overlapping evidence of the correct sequence from our genomic data, was likely a spurious result introduced when attempting to *de novo* assemble a transcript from this highly complex region. The sequence produced by partially gap-filling the *C. viridis PRDM9* gene region was then used for downstream gene expression and motif prediction. We provide the alignment of the sequences described above, the consensus *C. viridis* partially gap-filled *PRDM9* gene region sequence, and FGENESH+ PRDM9 annotation as Supplementary Datasets S1-S3.

We compared gene expression of *PRDM9* orthologs from the prairie rattlesnake to other vertebrate species, including two snake species, in order to examine if *PRDM9* expression is consistent with a functional role in directing recombination. We first compiled RNA-seq data from testis, ovary, brain, heart, kidney, liver, muscle, and small intestine from the zebrafish (*Danio rerio*; (Hu et al. 2015)), clawed frog (*Xenopus laevis*; Session et al. 2016), human (*Homo sapiens*; Fagerberg et al. 2014), mouse (*Mus musculus*; Mouse ENCODE consortium), Burmese python (*Python bivittatus*; Castoe et al. 2013), five-pace viper (*Deinagkistrodon acutus*; Yin et al. 2016) and the prairie rattlesnake (*Crotalus viridis*; Perry et al. 2019; Schield et al. 2019). For *D. acutus*, only testis, ovary, brain, and liver RNA-seq data were

available. All RNA-seq data were generated previously and were available through the NCBI short-read archive (SRA) or have been submitted to the NCBI SRA (Supplementary Table S4). We quality and adapter trimmed raw reads using Trimmomatic v0.36 (Bolger et al. 2014), and assessed processed read quality using FastQC (Andrews 2010; available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc). We then mapped processed reads using default parameters in STAR v2.7.0f (Dobin et al. 2013), using the latest version of the reference genome and gene annotation per species available on NCBI (Supplementary Table S4). For human and mouse, only primary genome assemblies and corresponding annotation files were used. We normalized raw read counts from STAR using DESeq2 v1.20 (Love et al. 2014). We then extracted normalized counts across tissues for the *PRDM9* ortholog of each species, and compared expression across tissues within each species using Log2 normalized counts in R.

We aligned *PRDM9* nucleotide sequences from *C. viridis*, *D. acutus*, and *P. bivittatus* using MUSCLE v3.8.31 (Edgar 2004), translated the resulting alignment to protein, and made minor manual edits to examine evidence of premature stop codons in the *PRDM9* ortholog of each species. We visualized the alignment using AliView v1.18.1 (Larsson 2014). The translated alignment contained no premature stop codons. We used a web-based DNA binding site predictor (http://zf.princeton.edu/logoMain.php) (Persikov and Singh 2013) for C_2H_2 ZF proteins to identify a binding motif for the four prairie rattlesnake PRDM9 ZFs using the expanded linear SVM prediction model. This analysis yielded a position weight matrix (PWM) for a putative 10-mer PRDM9 binding motif (Fig 6b). We used PoSSuM Search v2.0 (Beckstette et al. 2006) to computationally-predict binding sites in the prairie rattlesnake genome using this PWM, specifying a *p*-value threshold of 1×10^{-6}, which identified 52,522 putative binding sites across the genome. For downstream comparisons, we also identified PWMs and predicted binding sites for PRDM9 from *P. bivitattus* and *D. acutus* based on gene models from (Castoe et al. 2013; Yin et al. 2016), and using the settings detailed above for binding site prediction and computational identification of putative binding sites in the *C. viridis* genome; here we also included the *Boa constrictor* PRDM9 binding motif as an additional comparison (Supplemental Fig. S13). We tested for occurrence and enrichment of PRDM9 binding sites in species-specific recombination hotspots with the PWM described above using AME v5.1.0, using matched coldspots as background controls sequences for each species, and tested for significant enrichment using Fisher's exact tests.

Acknowledgments

We are grateful to the California Academy of Sciences and Jens Vindum for tissue loans. We thank Alice Shanfelter for helpful scripts and guidance on ancestral allele and mutation matrix inferences. This work was supported by the National Science Foundation (grant numbers DEB-1655571 to TAC, and DEB-1501886 to DRS and TAC); and the University of Northern Colorado Research Dissemination and Faculty Development grant to SPM. All procedures using animals or animal tissue were performed according to the University of Colorado Institutional Animal Care and Use Committee (IACUC) protocols 0901C-SM-MLChick-12 and 1302D-SM-S-16. The datasets supporting the conclusions of this article are available from the NCBI short-read archive under accessions PRJNA476794 and PRJNA593834). The genome assembly is available via NCBI GenBank under accession SAMN10416089. The assembly and annotation files are also available at https://figshare.com/projects/Prairie_rattlesnake_Crotalus_viridis_genome_assembly_and_annotation/66560. Inferred recombination maps are available at https://figshare.com/articles/Rattlesnake_Recombination_Maps/11283224. The repository with scripts used in analyses is available at https://github.com/drewschield/recombination.

References

Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, Aricescu AR, Myers SR. 2017. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *Elife* 6:e28383.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Greally JM, Wang J, et al. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet* 9:e1003984

Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K, Consortium L. 2012. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res* 22:51–63.

Backström N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, Webster MT, Öst T, Schneider M, Kempenaers B. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res* 20:485–495.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37:W202–W208.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.

Baker RJ, Bull JJ, Mengden GA. 1972. Karyotypic studies of 38 species of North-American snakes. *Copeia*: 257-265.

Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *Elife* 6:e24133.

Barton NH, Charlesworth B. 1998. Why sex and recombination? *Science* 281:1986-1990.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, De Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327:836-840.

Baudat F, Imai Y, De Massy B. 2013. Meiotic recombination in mammals: Localization and regulation. *Nat Rev Genet* 14:794-806.

Beckstette M, Homann R, Giegerich R, Kurtz S. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* 7:389.

Bell AC, Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* 405:482.

Berg IL, Neumann R, Lam K-WG, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* 42:859.

Bergero R, Charlesworth D. 2009. The evolution of restricted recombination in sex chromosomes. *Trends Ecol Evol* 24:94–102.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Borde V, Robine N, Lin W, Bonfils S, Geli V, Nicolas A. 2009. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J* 28:99–111.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10:e1003537.

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* 48:427.

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

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 63:861–869.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* 29:1917–1932.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Res* 25:1656–1665.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Castoe TA, de Koning APJ, Hall KT, Card DC, Schield DR, Fujita MK, Ruggiero RP, Degner JF, Daza JM, Gu WJ, et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc Natl Acad Sci USA* 110:20645–20650.

Chan AH, Jenkins PA, Song YS. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genet* 8:e1003090.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res* 19:136–142.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet* 14:262-274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. 2013. Haplotype estimation using sequencing reads. *Am J Hum Genet* 93:687–696.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. *Nature* 491:756.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc R Soc London B* 252:237–243.

Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K. 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 13:397–406.

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D. 2019. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. Soi: 10.1093/nar/gkz1001.

Fujita MK, Edwards S V, Ponting CP. 2011. The Anolis lizard genome: an amniote genome without isochores. *Genome Biol Evol* 3:974–984.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* 8:R24.

Haasl RJ, Payseur BA. 2016. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol Ecol* 25:5–23.

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

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* 8:269-294.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695.

Hilmi K, Jangal M, Marques M, Zhao T, Saad A, Zhang C, Luo VM, Syme A, Rejon C, Yu Z. 2017. CTCF facilitates DNA double-strand break repair by enhancing homologous recombination repair. *Sci Adv* 3:e1601898.

Hore TA, Deakin JE, Graves JAM. 2008. The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. *PLoS Genet* 4:e1000169.

Hu P, Liu M, Zhang D, Wang J, Niu H, Liu Y, Wu Z, Han B, Zhai W, Shen Y. 2015. Global identification of the genetic networks and cis-regulatory elements of the cold response in zebrafish. *Nucleic Acids Res* 43:9198–9213.

Hunter N. 2015. Meiotic recombination: the essence of heredity. *Cold Spring Harb Perspect Biol* 7:a016618.

Hwang SY, Kang MA, Baik CJ, Lee Y, Hang NT, Kim B-G, Han JS, Jeong J-H, Park D, Myung K. 2019. CTCF cooperates with CtIP to drive homologous recombination repair of double-strand breaks. *Nucleic Acids Res* 47:9160–9179.

Imai Y, Baudat F, Taillepierre M, Stanzione M, Toth A, de Massy B. 2017. The PRDM9 KRAB domain is required for meiosis and involved in protein interactions. *Chromosoma* 126:681–695.

Janes DE, Organ CL, Fujita MK, Shedlock AM, Edwards S V. 2010. Genome evolution in Reptilia, the sister group of mammals. *Annu Rev Genomics Hum Genet* 11:239–264.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14:528-538.

Jones GH, Franklin FCH. 2006. Meiotic crossing-over: obligation and interference. *Cell* 126:246–248.

Kaiser VB, Semple CA. 2018. Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline. *Genome Biol* 19:101.

Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, Ellegren H. 2017. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Mol Ecol* 26:4158-4172.

Keeney S. 2007. Spo11 and the formation of DNA double-strand breaks in meiosis. *In: Eleg R, Lankenau D-H, editors. Recombination and meiosis*. Berlin: Springer. p. 81–123.

Keinan A, Reich D. 2010. Human population differentiation is strongly correlated with local recombination rate. *PLoS Genet* 6:e1000886.

Kent CF, Minaei S, Harpur BA, Zayed A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc Natl Acad Sci* 109:18012–18017.

Kim S. 2015. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun Stat Appl Methods* 22:665.

Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, Villemoes R, Magnusdottir E, Olafsdottir SB, Thorsteinsdottir U, Stefansson K. 2014. Common and low-frequency variants associated with genome-wide recombination rate. *Nat Genet* 46:11.

Lam I, Keeney S. 2014. Mechanism and regulation of meiotic recombination initiation. *Cold Spring Harb Perspect Biol* 7:a016634.

Lang F, Li X, Zheng W, Li Z, Lu D, Chen G, Gong D, Yang L, Fu J, Shi P. 2017. CTCF prevents genomic instability by promoting homologous recombination-directed DNA double-strand break repair. *Proc Natl Acad Sci* 114:10912–10917.

Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CpG islands as gene markers in the human genome. *Genomics* 13:1095-1107.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30:3276–3278.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* 18:337–340.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

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

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* 47:W199-W205.

Loukinov DI, Pugacheva E, Vatolin S, Pack SD, Moon H, Chernukhin I, Mannan P, Larsson E, Kanduri C, Vostrov AA. 2002. BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proc Natl Acad Sci* 99:6806–6811.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol* 17:e2006288.

Matsubara K, Tarui H, Toriba M, Yamada K, Nishida-Umehara C, Agata K, Matsuda Y. 2006. Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *Proc Natl Acad Sci USA* 103:18190–18195.

McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MAF. 2012. Recombination modulates how selection affects linked sites in Drosophila. *PLoS Biol* 10:e1001422.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303.

McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 11:165.

McQueen HA, Fantes J, Cross SH, Clark VH, Archibald AL, Bird AP. 1996. CpG islands of chicken are concentrated on microchromosomes. *Nat Genet* 12:321.

McQueen HA, Siriaco G, Bird AP. 1998. Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. *Genome Res* 8:621-630.

McVean G. 2010. What drives recombination hotspots to repeat DNA in humans? *Philos Trans R Soc B* 365:1213–1218.

Merkenschlager M, Odom DT. 2013. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* 152:1285–1297.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* 21:984-990.

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

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327:876-879.

Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. *Philos Trans R Soc B* 367:409-421.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philos Trans R Soc B* 367:343–353.

O'Connor RE, Romanov MN, Kiazim LG, Barrett PM, Farré M, Damas J, Ferguson-Smith M, Valenzuela N, Larkin DM, Griffin DK. 2018. Reconstruction of the diapsid ancestral genome permits chromosome evolution tracing in avian and non-avian dinosaurs. *Nat Commun* 9:1883.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet* 5:e1000753.

Olmo E. 2005. Rate of chromosome changes and speciation in reptiles. *Genetica* 125:185–203.

Ong C-T, Corces VG. 2014. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 15:234–246.

Pasquesi GIM, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, Reyes-Velasco J, Ruggiero RP,

Vandewege MW, Shortt JA, et al. 2018. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat Commun* 9:2774.

Perry BW, Andrew AL, Kamal A, Card DC, Schield DR, Pasquesi GIM, Pellegrino MW, Mackessy SP, Chowdhury SM, Secor SM, et al. 2019. Multi-species comparisons of snakes identify coordinated signalling networks underlying post-feeding intestinal regeneration. *Proc R Soc B* 286:20190910.

Perry BW, Card DC, McGlothlin JW, Pasquesi GIM, Hales NR, Corbin AB, Adams RH, Schield DR, Fujita MK, Demuth JP, et al. 2018. Molecular adaptations for sensing and securing prey, and insight into amniote genome diversity, revealed by the garter snake genome. *Genome Biol Evol* 10:2110-2129.

Persikov A V, Singh M. 2013. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res* 42:97–108.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* 4:675–682.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* 137:1194–1211.

Ponting CP. 2011. What are the genomic drivers of the rapid evolution of PRDM9? *Trends Genet* 27:165–171.

Posada D, Crandall KA, Holmes EC. 2002. Recombination in evolutionary genomics. *Annu Rev Genet* 36:75–97.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* 37:429.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

R Core Team. 2017. R: A language and environment for statistical computing. https://www.R-project.org.

Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13:235–238.

Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M, Smale ST. 2009. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* 138:114–128.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44:W160–W165.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159:1665–1680.

Reyes-Velasco J, Meik JM, Smith EN, Castoe TA. 2013. Phylogenetic relationships of the enigmatic longtailed rattlesnakes (*Crotalus ericsmithi*, *C. lannomi*, and *C. stejnegeri*). *Mol Phylogenet Evol* 69:524–534.

Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome—patterns and consequences. *Mol Ecol* 22:3014–3027.

Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GIM, Blackmon H, Adams RH, Corbin AB, Smith CF, Ramesh B, et al. 2019. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* 29:590-601.

Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* 360:656-660.

Schwartz JJ, Roach DJ, Thomas JH, Shendure J. 2014. Primate evolution of the recombination regulator PRDM9. *Nat Commun* 5:4370.

Ségurel L, Leffler EM, Przeworski M. 2011. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biol* 9:e1001211.

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538:336.

Shanfelter AF, Archambeault SL, White MA. 2019. Divergent fine-scale recombination landscapes between a freshwater and marine population of threespine stickleback fish. *Genome Biol Evol* 11:1573–1585.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN, et al. 2015. Stable recombination hotspots in birds. *Science* 350:928-932.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC *Bioinformatics* 6:31.

Sleutels F, Soochit W, Bartkuhn M, Heath H, Dienstbach S, Bergmaier P, Franke V, Rosa-Garrido M, van de Nobelen S, Caesar L. 2012. The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner. *Epigenetics Chromatin* 5:8.

Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. http//repeatmasker. org.

Smith J, Bruley CK, Paton IR, Dunn I, Jones CT, Windsor D, Morrice DR, Law AS, Masabanda J, Sazanov A, et al. 2000. Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim Genet* 31:96–103.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* 7:S10.

Stange M, Sánchez-Villagra MR, Salzburger W, Matschiner M. 2018. Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Syst Biol* 67:681–699.

Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Project GAG, Bustamante CD, Hammer MF, Wall JD. 2015. The time scale of recombination rate evolution in great apes. *Mol Biol Evol* 33:928–945.

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*

464:1082.

Tock AJ, Henderson IR. 2018. Hotspots for initiation of meiotic recombination. *Front Genet* 9:521.

Vonk FJ, Casewell NR, Henkel C V, Heimberg AM, Jansen HJ, McCleary RJR, Kerkkamp HME, Vos RA, Guerreiro I, Calvete JJ, et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc Natl Acad Sci USA* 110:20651–20656.

Voss SR, Kump DK, Putta S, Pauly N, Reynolds A, Henry RJ, Basa S, Walker JA, Smith JJ. 2011. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Res* 21:1306-1312.

Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related populus species. *Genetics* 202:1185-1200.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al. 2010. The genome of a songbird. *Nature* 464:757–762.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7:256–276.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* 27:757–767.

Wu M, Kwoh C-K, Przytycka TM, Li J, Zheng J. 2012. Epigenetic functions enriched in transcription factors binding to mouse recombination hotspots. In: *Proteome SCi* 10 (Supple 1): S11.

Yin W, Wang ZJ, Li QY, Lian JM, Zhou Y, Lu BZ, Jin LJ, Qiu PX, Zhang P, Zhu WB, et al. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nat Commun* 7:13107.

Zheng Y, Wiens JJ. 2016. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. *Mol Phylogenet Evol* 94:537–547.

Figure Captions

Figure 1. Genome-wide recombination landscapes in rattlesnakes. a Linkage disequilibrium-based estimates of population-scaled recombination rate (ρ/bp) for *Crotalus viridis* (CV) and *Crotalus oreganus* (CO) across chromosomes in 1 Mb windows. Identified centromere regions are shown as violet bars. **b** Box and whisker plots of ρ/bp variation within and between chromosomes in CV. **c** Box and whisker plots of ρ/bp variation within and between chromosomes in CO. **d** Scatterplot of CV and CO mean ρ/bp estimates in 1 Mb windows. The dashed line summarizes the correlation between estimates in each species. Relationships between genome-wide recombination rate and **e**, **f** nucleotide diversity (π), **g**, **h** GC content, **i**, **j** gene density, and **k**, **l** repeat content in CV (top panels) and CO (bottom panels). Grey circles are pairwise values measured in 1 Mb windows. Green and blue lines depict smoothed splines of relationships between CV and CO measures, respectively.

Figure 2. Genomic variation on the Z chromosome. a Schematic of Z chromosome regions identified in (Schield et al. 2019), with approximate boundaries in Mb. PAR stands for pseudoautosomal region. **b** Variation in statistics across the Z chromosome in 100 kb windows, including ρ/bp in CV and CO, gene density, GC content, and normalized female and male π. In recombination rate panels for CV and CO, mean ρ/bp estimates for each window are shown as dots, and the lines show a smoothened spline across windowed estimates. Black dashed lines represent mean autosomal macrochromosome recombination rates per species. Gene density and GC content were measured from the prairie rattlesnake reference genome. Normalized π in both sexes was calculated by dividing the value of π in each window by the autosomal median of each sex, respectively.

Figure 3. Relative recombination rate in recombination hotspots in rattlesnake genomes. Green lines represent estimates for CV and blue lines represent estimates for CO. Relative recombination rates in hotspots were calculated by dividing candidate hotspot ρ/bp by ρ/bp in the surrounding 80 kb region, and averaged across all identified hotspots in each species. Relative recombination rates are shown for hotspots found in CV (**a**), CO (**b**), and shared between both species (**c**). The numbers of identified species specific and shared hotspots are shown in (**d**).

Figure 4. Fine-scale recombination rate near functional regions and annotation of the genomic recombination landscape. a, **b** Relative recombination rate with increasing distance in kb from promoters in CV (green) and CO (blue). **c**, **d** Relative recombination rate with increasing distance from

CpG islands (CGIs) in CV and CO. Points depict mean population scaled recombination rate in intervals of 100 bp and bold lines are smoothed splines. The opacity and size of points reflects the number of measurements per distance interval, with darker and larger points representing higher numbers. Relative rates were calculated by dividing rate per interval by the median rate across all intervals. **e** Hotspot density in nine non-overlapping genomic features. For each feature, hotspot density was calculated by dividing the number of hotspot bases overlapping with each feature by the total length of each feature. **f** Recombination rate in each of the nine genome feature categories. The height of each bar is equal to the mean rate within each feature, and black lines show standard error. **g** Recombination rate in 500 kb up- and downstream regions of genes. Points depict mean recombination rates in 500 bp sliding windows. As in **a**-**d**, darker and larger points representing higher numbers of measurements within a distance interval. Results for CV are shown in **e**-**g**. Results for CO are provided in the Supplementary Material.

Figure 5. Germline expression of *CTCFL* and associations between CTCFL-binding and the recombination landscape. a Log_2 gene expression of *CTCFL* across tissues from amniote species, including mammals (*Homo sapiens* and *Mus musculus*) and snakes (*Python bivittatus*, *Deinagkistrodon acutus*, and *Crotalus viridis*). Grey boxes in the heatmap represent missing data from *Deinagkistrodon* (RNAseq was not available for these tissues). The phylogeny to the left shows relationships among the sampled taxa, and nodes are labeled with estimated divergence times in MYA. **b** Numbers and proportions of species-specific and shared hotspots with CTCFL binding sites. **c**, **d** Relative recombination rate with increasing distance in kb from predicted CTCFL binding sites in CV (**c**) and CO (**d**). Points depict mean population scaled recombination rate in intervals of 100 bp and bold lines are smoothed splines. Relative rates were calculated by dividing rate per interval by the median rate across all intervals. The size and opacity of points correspond to the number of measurements from the recombination map for a given distance interval - larger, darker points depict more measurements.

Figure 6. Association between PRDM9 and recombination in snakes. a Log_2 gene expression of *PRDM9* homologs across tissues from vertebrate species, including mammals (*Homo sapiens* and *Mus musculus*) and snakes (*Python bivittatus*, *Deinagkistrodon acutus*, and *Crotalus viridis*). The phylogeny to the left shows relationships among the sampled taxa, and nodes are labeled with estimated divergence times in MYA. **b** Computationally predicted PRDM9 binding motifs from the first three zinc-fingers in the DNA-binding arrays of *Python bivittatus*, *Deinagkistrodon acutus*, and *Crotalus viridis*. **c**, **d** Relative recombination rate with increasing distance in kb from predicted *C. viridis* PRDM9 binding sites in CV (**c**) and CO (**d**). Points depict mean population scaled recombination rate in intervals of 100 bp and bold

lines are smoothed splines. The size and opacity of points correspond to the number of measurements from the recombination map for a given distance interval – larger, darker points depict more measurements. Relative rates were calculated by dividing rate per interval by the median rate across all intervals. **e**, **f** Relative recombination rates in CV and CO in distance intervals from *Deinagkistrodon acutus* PRDM9 binding sites predicted in the *C. viridis* genome. **g**, **h** Relative recombination rates in CV and CO in distance intervals from *Python bivittatus* PRDM9 sites predicted in the *C. viridis* genome. **i** Annotation of *C. viridis* PRDM9 binding site density in nine non-overlapping genomic features of the *C. viridis* genome. Bar heights are equal to the total number of binding site bases in each category divided by the total length of features within the category. **j** Annotation of predicted *Deinagkistrodon* PRDM9 binding site density in *C. viridis* genomic features. **k** Annotation of predicted *Python* PRDM9 binding site density in *C. viridis* genomic features.

Table 1. Pairwise and partial Spearman's rank order correlation coefficients between population-scaled recombination rate (ρ) and genomic measures in CV and CO at 1 Mb and 100 kb windowed resolution.

Resolution	Species	ρ vs. π		ρ vs. GC Content		ρ vs. Gene Density		ρ vs. Repeat Content	
		Pairwise	Partial[a]	Pairwise	Partial[b]	Pairwise	Partial[c]	Pairwise	Partial[d]
1 Mb	CV	0.586^{***}	0.593^{***}	0.293^{***}	0.127^{**}	0.289^{***}	0.154^{**}	0.06^{*}	0.1^{**}
	CO	0.546^{***}	0.534^{***}	0.261^{***}	0.124^{**}	0.238^{***}	0.12^{**}	0.089^{*}	0.12^{**}
100 kb	CV	0.6^{***}	0.606^{***}	0.242^{***}	0.153^{***}	0.224^{***}	0.14^{***}	0.006	0.032^{**}
	CO	0.522^{***}	0.509^{***}	0.202^{***}	0.132^{***}	0.166^{***}	0.08^{***}	0.022^{*}	0.037^{**}

[a]Partial correlation controls for GC content, gene density, and repeat content. [b]Partial correlation controls for gene density and repeat content. [c]Partial correlation controls for GC content and repeat content. [d]Partial correlation controls for GC content and gene density.

$^{*}p < 0.05$.

$^{**}p < 0.0001$.

$^{***}p < 2.2 \times 10^{-16}$.

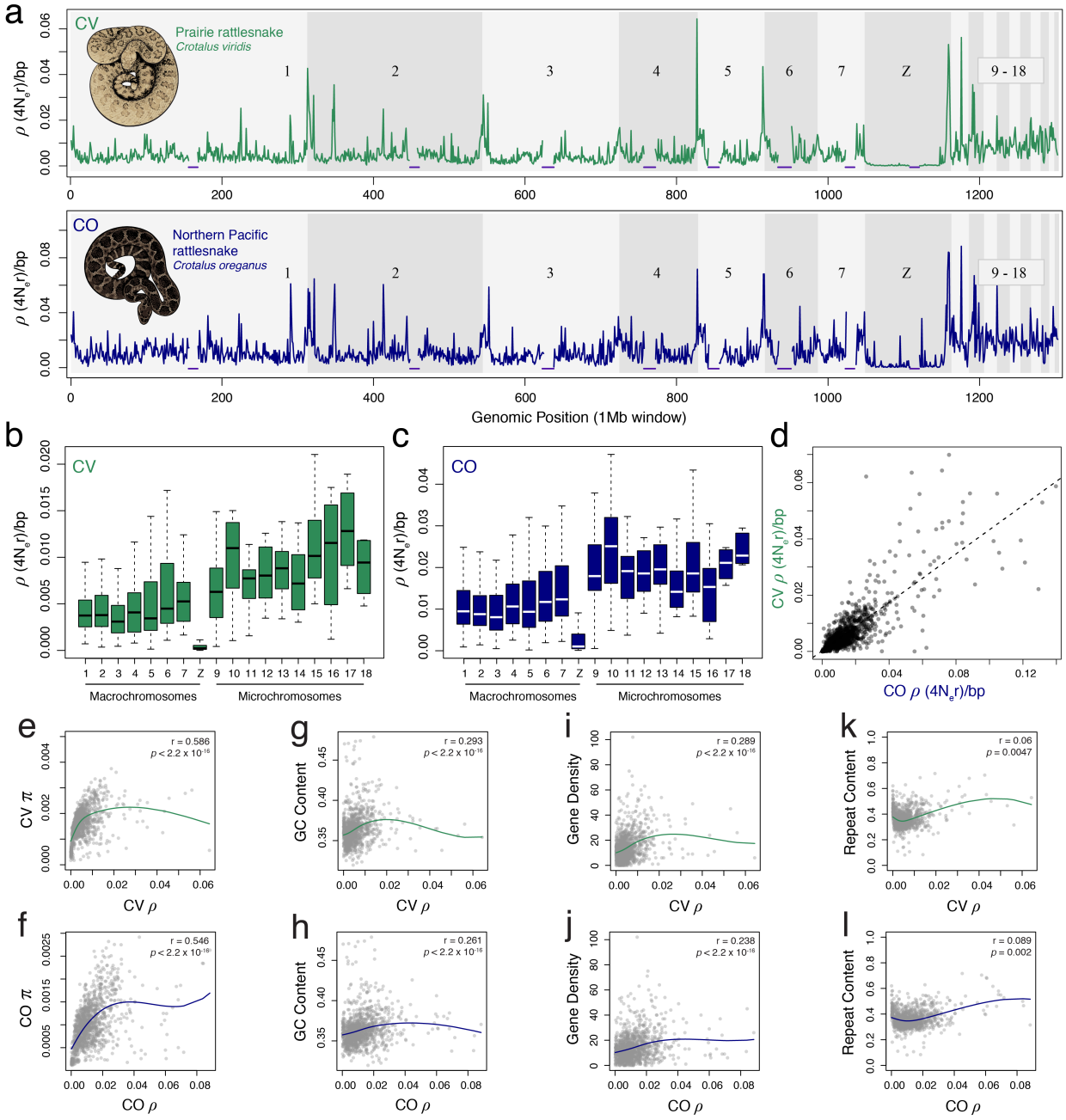

Figure 1.

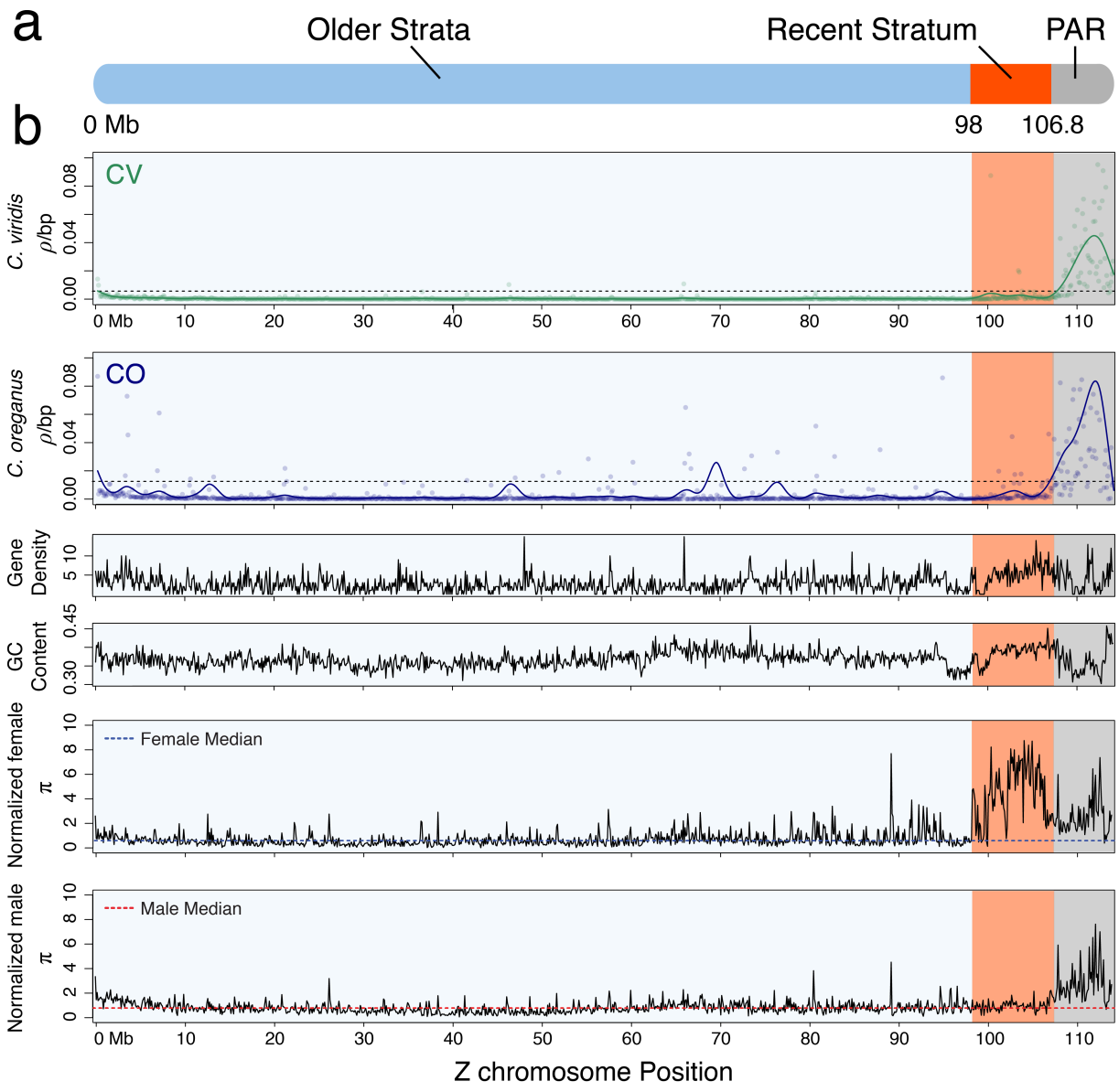

Figure 2.

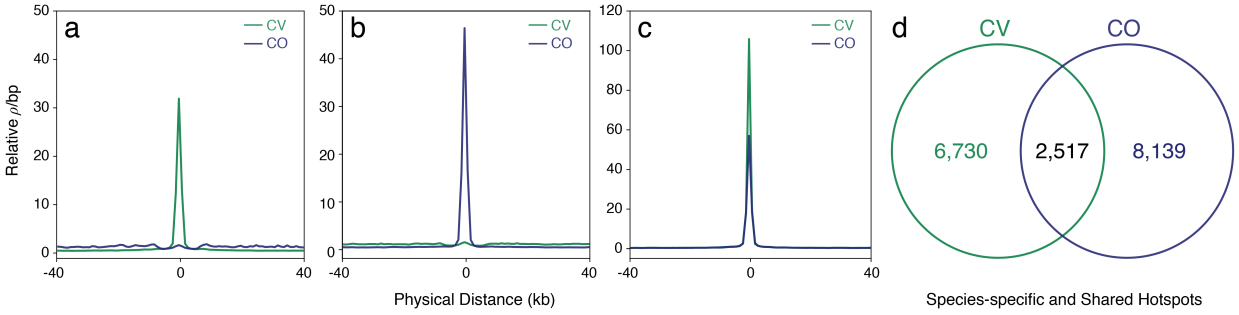

Figure 3.

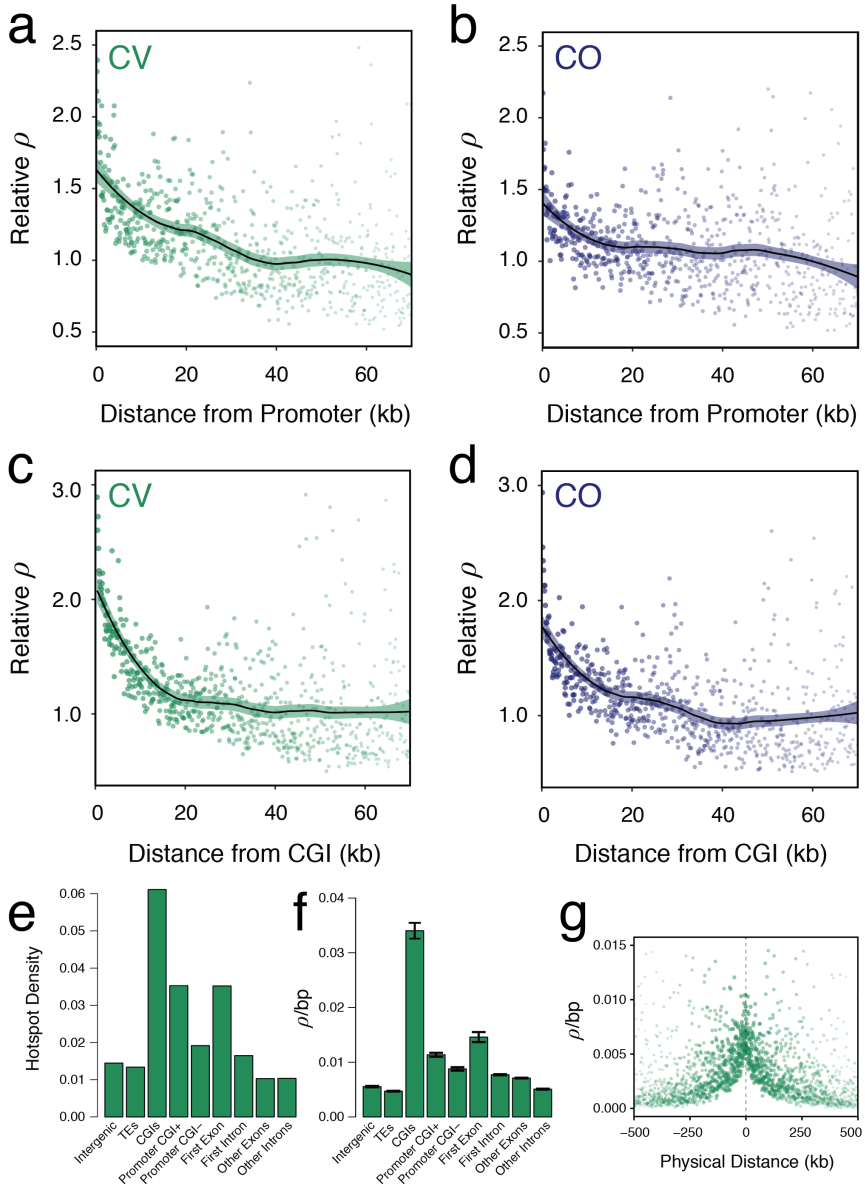

Figure 4.

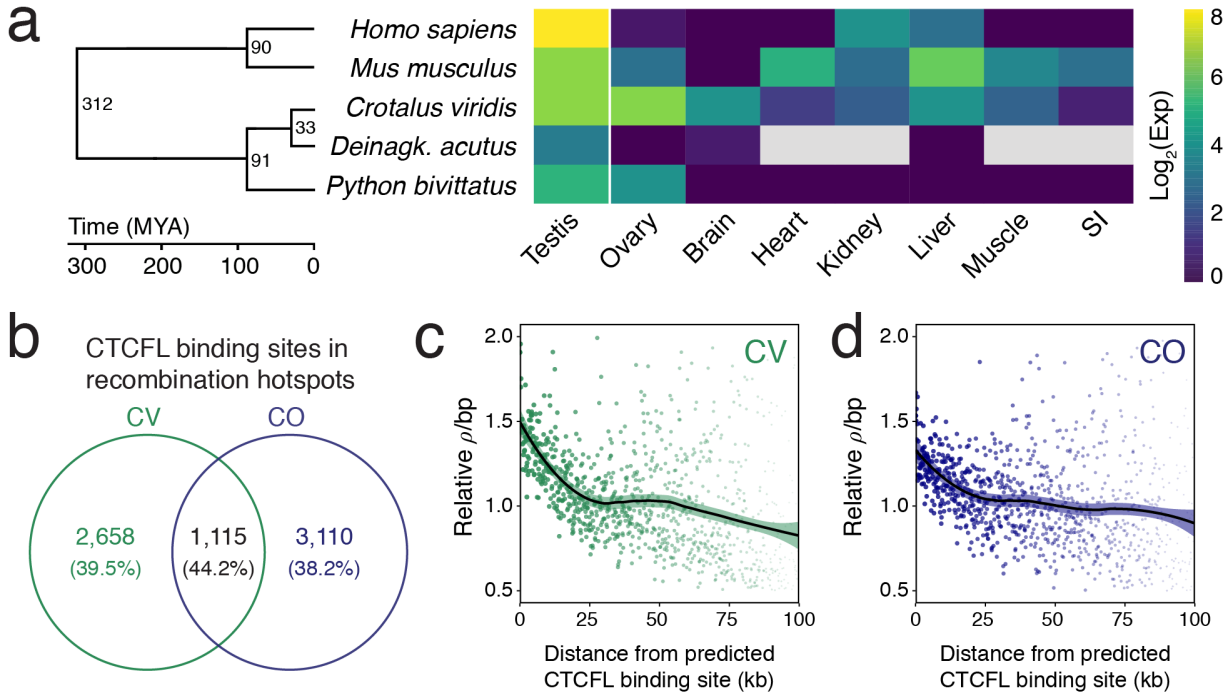

Figure 5.

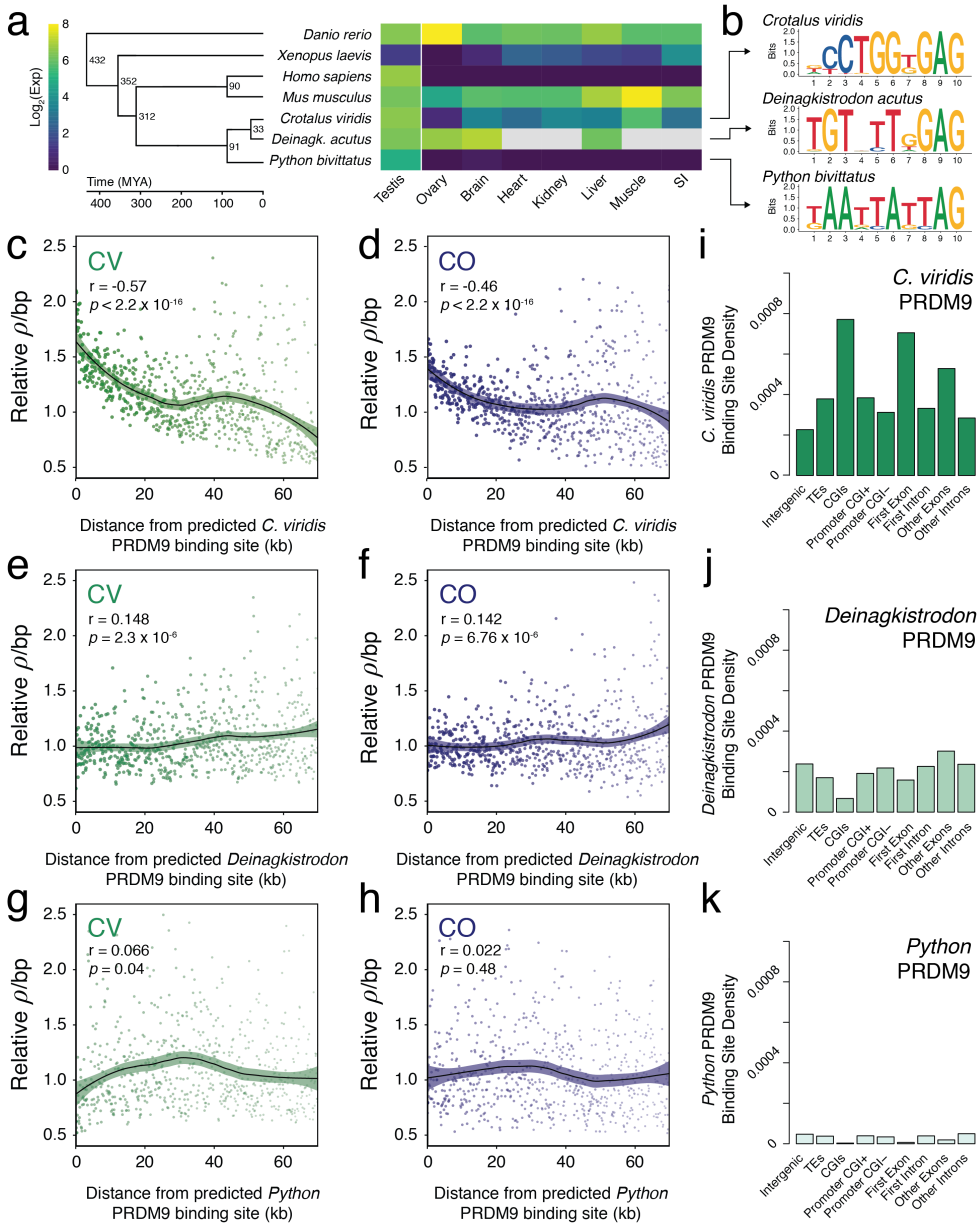

Figure 6.