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,[1,2] Zachary L. Nikolakis,[1] Aundrea K. Westfall,[1] Richard W. Orton,[1] Jesse M. Meik,[3] Stephen P. Mackessy,[4] and Todd A. Castoe*,[1]

[1]Department of Biology, University of Texas at Arlington, Arlington, TX

[2]Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL

[3]Department of Biological Sciences, Tarleton State University, Stephenville, TX

[4]School of Biological Sciences, University of Northern Coloradooonl053 Tm1.6(l)-.8()Tj3Tm1.oe0.4(o)-.7*F0001eoM9E-mail:-255.6M

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 (ZF) 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 2015). 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

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 ($\rho = 4 \ _e$, where 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 region0.50mseiatdgme59u456.1259 .(104]TJ/F8 1 Tfa

correlation coefficients; $= -0.432$,

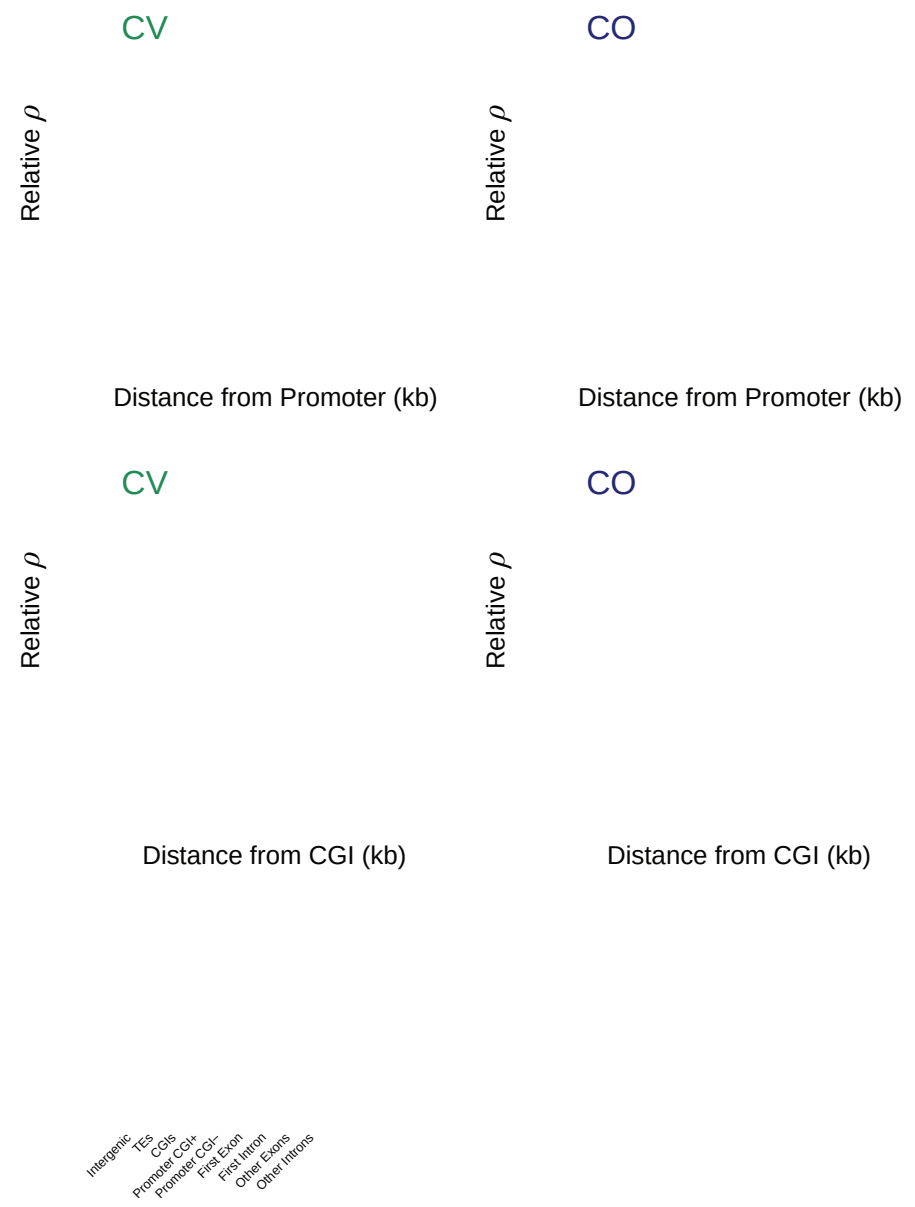
reduced recombination relative to autosomes in species with heteromorphic sex chromosomes (Barton and Charlesworth 1998; Bergero and Charlesworth 2009). Rattlesnakes have female heterogamety (ZW) with highly heteromorphic Z and W chromosomes (Baker et al. 1972; Matsubara et al. 2006), thus, we predict 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 with autosomes in both species (fig. 2 and ; Welch's two-sample -tests, values $<2.2 \times 10^{-16}$). Further, consistent with the expectation of unsuppressed recombination betweeun4msr7t8(n)3.8(e0 -

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,



CO did not show evidence of functional enrichment after false discovery rate correction (supplementary table S2, Supplementary Material online).

Hotspot density in nine nonoverlapping genomic features also provides consistent evidence for increased recombination near genes; hotspot density was highest in CGIs, promoters, and first exons (fig. 4 and supplementary fig. 7 , Supplementary Material online), and overall recombination rates were greatest in these regions (fig. 4 and supplementary fig. 7 , Supplementary Material online). 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, Supplementary Material online). As a comparison to candidate hotspots, we also calculated densities of coldspots with GC content matched to recombination hotspots

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. 5 ; Fisher's exact tests; CV-specific $= 5.63 \times 10^{-82}$; CO-specific $= 1.86 \times 10^{-25}$; shared $= 1.41 \times 10^{-79}$). 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; values $= 2.43 \times 10^{-23}$ and 3.32×10^{-35}, respectively), but not CO-specific hotspots, d

t2.1(O)8 1 Tf1.636 0 TDD-.0261 s5001 2.(n)1.4(s Tfg(y)2.hs Tf274.3(bu)-1.5(ot2

predictions from

$= 0.04$, CO $= 0.022$, $= 0.48$), though binding sites matching the partial PRDM9 motif were comparatively rare, and this relationship was nonsignificant in the case of CO. PRDM9-binding sites were even more rare (507 total), and did not exhibit any association with recombination in either rat-

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. 5 and). 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. 6 and). The 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 , 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 vertebraa suggtrbrsnrwp-2.5(9.4(i)-2.3(e)-;4(b)2.8(r)-71-.2(16.2.3(e)72(a)2.49n.9(s)(n).8(r)-w(

divergent vertebrate species suggests that the concentration of recombination in telomeres may be driven by mechanisms that are conserved across amniote vertebrates.

C

We examined snake recombination landscapes for the first

Genomics), which generated a genomic variant call file

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 underassembly. 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 (last accessed January 22, 2020).

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. (

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, although 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, Supplementary Material online). 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.

Soti69(uri-206.2.6(io)1.3(50.452t).72ot2.6(ioia)4(l)-)-]TJ-.0221 T(Ho7(Wa)-3((r)-631(ec7.5(a)).72otla)4(lc)1.2(u-(o)l).72ot2.-4-.02otd(ti6e)]T2.6(i

(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

suggesting that the end portion of the ZF 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 that spanned the gap against 10× assembly A, 10× assembly B, and the unaligned, error-corrected PacBio reads. The and translated de novo transcriptome predicted 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 genome sequence flanking the gap regions, to each other, and to the region spanning the gap to confirm sufficient amount of overlap between sequences with high sequence identity. "Scaffold 468599" from 10× assembly A aligned and covered the two exons preceding the gap, and extended into the gap region, with 3,887/5,491 bases covering the region flanking the gap. A second scaffold ("Scaffold 332766") from 10× assembly B then produced BLAST results with substantial overlap (with "Scaffold 468599" from the other assembly) over the new 10× 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 10× scaffolds, but contained a large number of erroneous bases that precluded the prediction of the ZF array beyond the consensus sequence from the genome, 10× assemblies, and the PacBio sequence upstream of the erroneous region (supplementary data sets S1 and S2, Supplementary Material online). When annotated using FGENESH+ (Solovyev et al. 2006), this consensus sequence included three tandem ZFs, the first two of which matched exactly those recovered using de novo transcriptome assembly (supplementary data set S3, Supplementary Material online). 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 gene region was then used for downstream gene expression and motif prediction. We provide the alignment of the sequences described earlier, the consensus partially gap-filled gene region sequence, and FGENESH+ PRDM9 annotation as supplementary data sets S1–S3, Supplementary Material

January 22, 2020). Inferred recombination maps are available at https://figshare.com/articles/Rattlesnake_Recombination_Maps/11283224 (last accessed January 22, 2020). The repository with scripts used in analyses is available at https://github.com/drewschield/recombination (last accessed January 22, 2020).

Supplementary data are available at online.

A

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 Nos. DEB-1655571 to T.A.C. and DEB-1501886 to D.R.S. and T.A.C.) and the University of Northern Colorado Research Dissemination and Faculty Development grant to S.P.M. 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.

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. 6:e28383.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. . 215(3):403–410.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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. . 9(12):e1003984.

Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K, Consortium L, The LUPA Consortium. 2012. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. . 22(1):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 genome. . 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.

McVean G. 2010. What drives recombination hotspots to repeat DNA in humans? ^. . 365(1544):1213–1218.
Merkenschlager M, Odom DT. 2013. CTCF and cohesin: linking gene regulatory elements with their targets. 152(6):1285–1297.

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. 464(7291):1082–1086.

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

Vonk FJ, Casewell NR, Henkel CV, 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. 110(51):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. . 21(8):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 species. . 202(3):1185–1200.