DEVELOPING APPROACHES THAT LEVERAGE NEXT GENERATION SEQUENCING TO STUDY BROAD BIOLOGICAL PROBLEMS

by

NICOLE HALES

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

Dec 2019

Acknowledgements

I am grateful to the numerous people who contributed to my success during my dissertation. First, I'd like to thank the current and previous lab mates and friends of the Castoe Lab – Nicole Proctor, Jacobo-Reyes Valasco, Daren Card, Audra Aundrew, Giulia Pasquesi, Aundrea Westfall, Ricky Orton, Zach Nikolakis, and especially Drew Schield, who became the brother I never wanted. I consider myself lucky to have worked and befriended such an amazing group of people, and I couldn't have done this without them. Thank you to all the friends and colleagues (former and current) in the UTA Biology Department – especially, Shannon Beston, Danielle Rivera, Kathleen Currie, Matt Fujita, TJ Firneno, Jose Maldonado, Contessa Ricci, Misha Kazi, Kim Bowles, Melissa Muenzler, Jill Castoe, Linda Taylor, Ashley Priest, Mallory Roelke, Corey Roelke, Paul Chippendale, Esther Betran, Walter Schargel, Nick Pollock, John Weidanz, Kelly Sheff and Zibiao Guo. Thank you to my committee members, Drs. Jeff Demuth, Matt Walsh, Mark Pellegrino and Sen Xu for your support, knowledge and ideas for improving my research. A special thanks to my parents and grandparents for buying me groceries and always making sure I had a working vehicle. I also want to thank the ones who provided me with the emotional support required to stay in college for almost 10 years -- Angela Lemond, Mila Hunt, Amanda Linfield, Ola Moussa (and the Moussa family) and Princess Ananti. Lastly, I'd like to thank my advisor, Todd Castoe – thank you for your mentorship, friendship and shared love of candy for the last 6 years.

Nov. 8, 2019

Dedication

I dedicate this dissertation to the most amazing woman in the world, my mother, Rebecca Lynn Baebel.

Abstract

DEVELOPING NEW GENOMIC APPROACHES THAT LEVERAGE NEXT GENERATION SEQUENCING TO STUDY BROAD BIOLOGICAL PROBLEMS

Nicole Hales, PhD

The University of Texas at Arlington, 2019

Supervising Professor: Todd Castoe

The advent of Next Generation Sequencing has been an extremely powerful tool in transforming the way we answer biological questions today. With the price per base pair continuously decreasing, and throughput, sequencing speed and quality of sequence reads increasing, NGS has allowed scientists to develop novel biological applications that have led to significant findings. During my dissertation, I leveraged several non-model organisms across various projects to develop new approaches of NGS to study a broad range of biological questions, including (1) understanding the genetic processes underlying transgenerational plasticity in *Daphnia*, (2) using Hi-C sequencing to study vertebrate genome structure and how chromatin and transcription factors interact to regulate snake venom, and (3) resurrecting archived, very low-quality samples to understand patterns of transmission in parasites. Collectively, this worked serves as a demonstration of how NGS can be utilized across multiple systems to answer broad biological questions.

Table of Contents

Chapter 1

Introduction

The advent of Next Generation Sequencing has been an extremely powerful tool in transforming the way we answer biological questions today. With the price per base pair continuously decreasing, and throughput, sequencing speed and quality of sequence reads increasing, NGS has allowed scientists to develop novel biological applications that have led to really significant findings. The applications for NGS technologies are hugely broad and with the development of more user friendly computational tools, interpreting NGS data has become more and more tractable for researchers that are not necessarily specialized in genomics – and as a result has become increasingly available in diverse fields such as developmental biology, ecology, epidemiology, clinical studies etc. During my dissertation, I leveraged several non-model organisms across various projects to develop new approaches of NGS to study a broad range of biological questions, including (1) understanding the genetic processes underlying transgenerational plasticity in *Daphnia*, (2) using Hi-C sequencing to study vertebrate genome structure and how chromatin and transcription factors interact to regulate snake venom, and (3) resurrecting archived, very low-quality samples to understand patterns of transmission in parasites.

To fully understand the challenges in quantifying changes within the nucleus, it's important to appreciate that functional genomics goes far beyond a linear DNA sequence. For instance, during normal growth phase of a cell, spatial organization of a genome within a nucleus is vital in normal functioning. DNA and proteins organize together to form different conformations and structures that then ultimately drive the expression and regulation of genes. There are multiple regulatory elements involved in gene expression that need to be available at precisely the right time in order for expression to occur, and these factors can be located kilobases to megabases away from each other and thus require DNA looping in order for these elements to become physically close enough to interact. It is only when all these factors come together that genes can be expressed. Understanding genomic data at multiple scales is paramount in truly understanding how organisms respond in nature.

Chapter 2

Contrasting gene expression programs correspond with predator-induced phenotypic plasticity within and across-generations in *Daphnia*.

Nicole R. Hales (NRH)[1], Drew R. Schield (DRS)[1], Audra L. Andrew (ALA)[1], Daren C. Card (DCC)[1], Matthew R. Walsh (MRW)[1] and Todd A. Castoe (TAC)[1]

[1]Department of Biology, 501 S. Nedderman Dr., University of Texas at Arlington, Arlington, TX 76010 USA

Abstract

Research has shown that a change in environmental conditions can alter the expression of traits during development (i.e., 'within-generation phenotypic plasticity') as well as induce heritable phenotypic responses that persist for multiple generations (i.e., 'transgenerational plasticity'). It has long been assumed that shifts in gene expression are tightly linked to observed trait responses at the phenotypic level. Yet, the manner in which organisms couple within- and trans-generational plasticity at the molecular level is unclear. Here we tested the influence of fish predator chemical cues on patterns of gene expression within- and across-generations using a clone of *Daphnia ambigua* that is known to exhibit strong transgenerational plasticity but weak within-generation plasticity. *Daphnia* were reared in the presence of predator cues in generation 1 and shifts in gene expression were tracked across two additional asexual experimental generations that lacked exposure to predator cues. Initial exposure to predator cues in generation 1 was linked to ~50 responsive genes but such shifts were 3-4x larger in later generations. Differentially expressed genes included those involved in reproduction, exoskeleton structure, and digestion; major shifts in expression of genes encoding ribosomal proteins were also identified. Furthermore, shifts within the first generation and transgenerational shifts in gene expression were largely distinct in terms of the genes that were differentially expressed. Such results argue that the gene expression programs involved in within- versus trans-generation plasticity are fundamentally different. Our study provides new key insights into the plasticity of gene expression and how it relates to phenotypic plasticity in nature.

Introduction

It is now clear that organisms can respond to environmental signals by altering the expression of traits during development, as well as modifying traits across multiple generations (Bonduriansky, et al. 2012; Fox and Mousseau 1998; Jablonka and Raz 2009; Uller 2008). This 'within-generation phenotypic plasticity' and 'transgenerational plasticity' (TGP or across-generational plasticity) has been documented in a diverse array of taxa, including plants, bryozoans, rotifers, beetles, and birds (Charmantier, et al. 2008; Fox and Mousseau 1998; Galloway and Etterson 2007; Gilbert 2004; Marshall 2008) in response to numerous environmental stimuli (e.g., temperatures; Salinas and Munch 2012, food shortages; Bashey 2006, canopy shading; Galloway and Etterson 2007). Research has shown that both within-generation plasticity and TGP are often adaptive (Agrawal, et al. 1999; Bashey 2006; Galloway 2005; Dyer, et al. 2010; Galloway and Etterson 2007; Salinas and Munch 2012; Walsh, et al. 2015) and can evolve in response to divergent ecological conditions (Walsh, et al. 2016). The long-standing assumption is that underlying shifts in gene expression manifest as patterns of plasticity at the phenotypic level. Examples of environmentally induced changes in gene expression within- and across-generations are rapidly accumulating (Aubin-Horth and Renn 2009; Carone, et al. 2010; Herman, et al. 2014; Herman and Sultan 2011; Jablonka, et al. 1995; Miyakawa et al. 2010; Molinier, et al. 2006; Simons 2011; Tollrian and Leese 2010). Yet, the manner in which organisms couple within-generation plasticity in conjunction with TGP in response to a change in environmental conditions, especially at the molecular level, is unclear.

Recently developed theory predicts that divergent ecological conditions will select for divergent patterns of within- versus trans-generational plasticity (Kuijper and Hoyle 2015; Leimar and McNamara 2015; Uller, et al. 2015a). Here, variation in a key environmental selective pressure (i.e., temporal variation in environmental stability) is predicted to select for increased within-generation plasticity *or* increased transgenerational plasticity, but not both. Empirical research has indeed shown that organisms harbor extensive variation in the direction and magnitude of within- and across-generation plasticity (Donohue and Schmitt 1998; Schmitt, et al. 1992; Walsh, et al. 2015) and that environmentally induced within- and trans-generational responses can have synergistic (Galloway 2009; Lin and Galloway 2010; Sultan, et al. 2009) and antagonistic (Walsh, et al. 2015) effects on the traits of organisms. It follows logically that contrasting gene expression programs may be linked to divergent patterns of within- versus

trans-generational plasticity. Natural selection may alter the path from gene expression to phenotypic plasticity because selection for within- or across-generation plasticity acts either upon the same genes, but drives changes in the expression levels or in the direction of expression. Conversely, selection for within- versus trans-generational plasticity may act upon different genes or on different numbers of genes. These two hypotheses lie at opposite ends of a continuous spectrum and are therefore not necessarily mutually exclusive.

Studies of clonal and eco-responsive *Daphnia sp.* offer an opportunity to examine the manner in which natural selection modulates connections between gene expression and phenotype within and across generations. *Daphnia* are an ubiquitous feature of freshwater environments (Carpenter, et al. 1992), and they possess characteristics that make them ideal for experimental studies, including easy culturing, short generation times, parthenogenic reproduction, and many readily quantifiable traits (Miner, et al. 2012; Stollewerk 2010). *Daphnia* species are well-known to respond to changes in the environment by altering the expression of a multitude of traits (Riessen 1999; Stibor 1992). For example, exposure to predator chemical cues elicits dramatic shifts in morphology, behavior, and life history traits (Tollrian and Harvell 1999). Research has used these known patterns of plasticity to begin to consider the underlying molecular mechanisms for predator-induced plasticity (Rozenberg, et al. 2015; Schwarzenberger, et al. 2009). For example, Schwarzenberger et al. (2009) evaluated patterns of plasticity in gene expression of several candidate genes in *D. magna* that were exposed to chemical cues produced by fish and invertebrate predators. This approach revealed strong upregulation of cyclophylin, involved in protein folding in the presence of fish predator cues while exposure to invertebrate predator cues was associated with a downregulation of cyclophylin. Given that *Daphnia* differ in their life history responses to fish versus invertebrate predator cues (Riessen 1999; Stibor 1992), these contrasting gene expression responses could indicate that cyclophylin is linked to the expression of life history traits (Tollrian and Harvell 1999) (Rozenberg, et al. 2015; Schwarzenberger, et al. 2009). Such results provide a clear connection between phenotypic plasticity and gene expression. Still, the connections between within-generation responses and transgenerational plasticity at the molecular level remain largely unexplored.

Our previous work quantified patterns of within and transgenerational plasticity at the phenotypic level in multiple clones of *Daphnia ambigua* from lakes in Connecticut, USA. We found that *Daphnia*

respond to initial exposure to predator cues by shifting life history to mature slower and produce less embryos compared to the transgenerational change (Walsh, et al. 2015). We classify these life history responses that occur during development as 'within-generation plasticity'. We also found that *Daphnia* exposed to predator cues programmed future generations for faster development. Such transgenerational responses were apparent two generations following cue removal (Walsh, et al. 2015). That is, life history differences between parental *Daphnia* that were and were not exposed to predator cues were still observed in the grand-offspring. These patterns of transgenerational life history plasticity are correlated with shifts in methylation (Schield, et al. 2016). More importantly, phenotypic experiments have revealed extensive variation in the direction and magnitude of phenotypic responses to predator cues within and across generations. Such variation in these two forms of plasticity provides the raw material to test for variation in gene expression programs within and across generations. While our previous study (Walsh et al. 2015) measured life history traits, the current study complements previous work through the addition of gene expression analyses, thus providing new insight into the gene regulatory basis of these responses to environmental cues.

Here, we tested the influence of predator chemical cues on the patterns of gene expression within and across generations in a single clone of *Daphnia ambigua.* We reared *Daphnia* in the presence and absence of fish chemical cues in first-generation individuals and tracked shifts in gene expression across two additional asexual generations. Importantly, the clone of *Daphnia* used in these experiments responds to predator cues by strongly programming future generations for rapid development (i.e., strong transgenerational plasticity) but exhibits weak within-generation plasticity (Walsh, et al. 2015). These phenotypic data stem from our previous work (Walsh, et al. 2015) and thus set the foundation for comparisons with gene expression-based responses over multiple generations. Given these known divergent phenotypic responses to predator cues within and across generations for this clone, we predicted that the number genes that are differentially expressed across generations would exceed those that are differentially expressed within the first generation. Comparisons between patterns of predator-induced within- and trans-generational plasticity in gene expression responses will allow us to determine if *Daphnia* couple within- and trans-generational plasticity by altering the expression of the same sets of genes, or if these two forms of plasticity correspond with expression of distinct sets of genes.

Materials and Methods

Empirical experimental design

We used a single clone of *Daphnia ambigua* from Dodge Pond in Connecticut, USA (Post, et al. 2008). In June 2013, we isolated ephippia from a sediment sample that was originally collected via an Ekman grab in 2009. Upon hatching, cultures from this clone were maintained in 250-mL glass jars for several months prior to the start of the experiments. During this time, *Daphnia* cultures were maintained at moderate densities (<60 adults/L) and provided with fresh media and algae weekly. It is important to note that this clone of *Daphnia* reproduces asexually under benign conditions and reverts to sexual reproduction when stressed. However, all *Daphnia* were propagated asexually in the experiments described below. *Daphnia* rearing experiments, and all molecular laboratory experiments, were conducted at the same times and in parallel where possible to minimize experimental variation.

Our experimental approach consisted of rearing the focal clone of *Daphnia* in a common garden setting for two generations (Fig. 1), followed by three generations of experimental manipulation (Fig. 2). To initiate the multiple generations of common garden rearing, we isolated 30 adults from existing stock cultures and placed each adult in separate 90-mL containers containing COMBO media (Kilham, et al. 1998) and specified, non-limiting quantities of green algae (*Scendesmus obliquus;* concentration 0.8 mg x C L^{-1} x day^{-1}). For each isolated adult, a single neonate was immediately pulled from the first asexual clutch and these neonates were moved to new 90-mL containers containing the same media and algae; these individuals represent the first common garden generation. All individuals were transferred to fresh media and algae every day and were reared at 21°C and a 14 hour:10 hour light:dark schedule. To initiate the second common garden generation, we collected neonates from the second clutch of each replicate jar and these offspring were again transferred to fresh media and algae daily (see diagramed design in Fig. 1).

Our experiment began with third-generation lab born individuals. On day 1 of the experiment, we collected all neonates that were born over the previous 12 hours from each of the parental jars. This yielded ~180 newly-born *Daphnia* from the third clutch or later of the second generation lab reared parents and all neonates were placed into 250-mL jars containing COMBO media at a density of 40-48 *Daphnia*/L, or 18 jars with 10 *Daphnia* per jar (Kilham, et al. 1998). Each jar was randomly allocated to one of two treatments: (1) predator exposure during the first generation followed by two generations in the absence of predator cues (i.e., generation 1 = P, generation 2 = PN, generation 3 = PNN; Fig. 2), or (2) three generations in the absence of predator cues (i.e., generation 1 = N, generation 2 = NN, generation 3 = NNN). All *Daphnia* were transferred to fresh media, algae, and kairomones (see below details of kairomone collection) daily. We monitored jars daily for maturation (i.e., release of first clutch into the brood chamber) and for the production of new clutches. Based upon previous work (Walsh and Post 2012, 2011), we estimated that 10 days were needed for *Daphnia* to release their second clutch. We thus initiated the second experimental generation after 10 days of exposure to predator cues. This experimental generation was again initiated by collecting newborn individuals under the same conditions described above. The third experimental generation was collected and reared in this same manner. After collecting neonates to initiate the second and third experimental generations, all adults were flash-frozen (in liquid nitrogen) for the subsequent RNAseq analyses. *Daphnia* from the third experimental generation were frozen following 10 days of common garden rearing.

Kairomone collection

COMBO medium conditioned by the presence of planktivorous fish was collected daily from a tank containing 2 redbreast sunfish (*Lepomis auritus;* ~3-cm in total length) in 130-L of water. Each day, media containing fish chemical cues was filtered using membrane filters (47mm diameter, 0.45μm mesh) and added at a concentration of 0.0025 fish/L to the predator treatments. Injured *Daphnia* emit chemical cues that contribute to the magnitude of phenotypic response to predation (Laforsch, et al. 2006). We thus added filtered, macerated *Daphnia* (100 *Daphnia*/L) every day to the appropriate predator treatments to ensure that our predator treatments contained both fish kairomones and *Daphnia* alarm cues.

RNA isolation, library preparation, and sequencing

We extracted RNA using the Zymo Research Duet Kit from snap-frozen samples. Each generation included three biological replicates, with each replicate comprised of 30 clonal *Daphnia* individuals (Fig. 1 & 2). Appropriate amounts of RNA were not available from single individuals. We pulled 30 individuals per replicate for the purposes of library construction and sequencing, since all individuals have identical genetic backgrounds. A similar pooling approach has also been used in other studies of *Daphnia* differential gene expression (e.g., Roy Chowdhury, et al. 2015; Soetaert, et al. 2007). Isolated RNA was quantified using a Qubit fluorometer (Invitrogen) and mRNAseq libraries were constructed using Illumina TruSeq library kits. A total of 3 µg of total RNA from each replicate (representing a pool of 30 individuals) was used for RNAseq library preparation. Each of these samples were uniquely indexed and all 18 individual libraries were multiplexed into a single, pooled library and sequenced on a single Illumina MiSeq run using 150bp paired-end sequencing reads.

Assessing differential gene expression

Raw Illumina RNAseq reads were demultiplexed by index and quality trimmed using Trimmomatic v. 0.36 (Bolger, et al. 2014) with default settings. We used the BWA MEM algorithm v. 0.7.13 (Li and Durbin 2009), with default settings, to map quality trimmed reads to the complete annotated transcript set of *Daphnia pulex* (Colbourne, et al. 2011) obtained from Ensembl. On average, about 70% of trimmed reads mapped to the reference genome transcript set. Raw gene expression counts were estimated by counting the number of reads that mapped uniquely to a particular annotated transcript using SAMtools v. 1.3.1 (Li, et al. 2009). Raw expression counts were then normalized using the TMM normalization method in edgeR (Oshlack 2010; Robinson, et al. 2010), and all subsequent gene expression analyses used these normalized data. Using these normalized data, we identified genes that were significantly differentially expressed between treatments by conducting pairwise tests between replicated time point samples using an exact test of the binomial distribution estimated in edgeR (Robinson, et al. 2010), integrating both common and tagwise dispersion. To control for any responses that may be attributed to the experimental design, we only considered expression differences between experimental and control treatments within each generation (i.e., P vs. N, PN vs. NN, and PNN vs. NNN). All genes with evidence

of differential expression at an FDR value ≤ 0.05 were considered significantly differentially expressed between treatments. Significantly differentially expressed genes were visualized across all samples as heat maps that were generated in R (R Development-Core Team 2008) with genes clustered by expression pattern similarity using the R-package vegan (Dixon 2003); gene expression pattern clustering was calculated using average linkage hierarchical clustering based on Bray Curtis dissimilarity matrix (Fig. 3B-C, Fig. 6, and Fig. 7). We also used principle component analysis (PCA; using core functions in R) to identify the degree to which patterns of RNAseq variation could differentiate between generations and individuals by comparing the same normalized gene expression data for all samples (using a singular value decomposition of expression matrix). Significantly differentially expressed genes that overlap between generations were visualized in a Venn diagram (Fig. 3D). In order to test if the overlap of gene sets from different generations was more than expected by chance, we conducted a hypergeometric test using the stats package in R.

Analyses of trends in expression shifts and biological interpretations

Significantly differential genes were annotated using Blast2GO v3.3.4 (Conesa and Götz 2008; Conesa, et al. 2005; Götz, et al. 2011) and Ensembl BioMart (Kinsella, et al. 2011). From the Blast2GO annotation outputs, we grouped genes that were functionally similar and associated with traits including digestive function, reproductive function, epigenetic modifications, and proteolysis. Sequence IDs were then converted to DAPPUDRAFT IDs using the *Daphnia pulex* gene annotation list from Ensembl; these IDs were then used to assign Gene Ontology (GO) term identifiers. We performed GO enrichment analyses (Mi, et al. 2016) to determine if significantly differentially expressed gene sets were enriched for particular functional categories of genes (Ashburner, et al. 2000). Because our annotations were based on genes orthologous to *D. pulex*, we minimized bias in the GO enrichment analysis by including a background of only the genes we observed as expressed in any of our *D. ambigua* experiments. We considered GO term categories as significantly enriched if the ratio test resulted in a Bonferroni-corrected p-value ≤ 0.05. Enriched GO terms were summarized by removing redundancies using REViGO (Supek 2011) with allowed similarity of terms set to 0.1.

Results

Gene Expression Analyses

An average of 285,576 reads were mapped for each replicate. The numbers of raw reads obtained per library together with read mapping statistics, are provided in Table 1. Initial exposure to predator cues was associated with 48 significantly differentially expressed genes between experimental and control treatments in generation 1 (P vs. N; FDR ≤ 0.05). Following predator cue removal, we observed 223 differentially expressed genes in generation 2 (PN vs. NN), and 170 differentially expressed genes in generation 3 (PNN vs. NNN; Fig. 3A). Sets of responsive genes in each generation were mostly distinct; the data for generation 1 shared only five responsive genes that were differentially expressed with the patterns observed in generation 2, and zero genes with generation 3 (Fig. 3D). In contrast, of 223 responsive genes in generation 2, 121 (54%) were also differentially expressed in generation 3 (PNN vs. NNN; Fig 3D). Hypergeometric tests on the overlap revealed that the overlap between generations 1 and 2, as well as the overlap between generations 2 and 3 were significant (p values of 1.77×10^{-236} and 2.06×10^{-05} respectively). While the results of the hypergeometic test indicated that proportions of overlapping genes were greater than expected at random, it is notable that the vast majority of differentially expressed genes were distinct, considering the hypothesis that they may indeed be entirely the same set.

We conducted a principle components analysis (PCA) to further explore patterns of gene expression within and across generations (Fig. 4). The first principal component (PC1) explained 90.7% of the variance and clearly separated the control and experimental treatments in generation 2 and 3 (Fig. 3). PC1 therefore accounts for transgenerational shifts in gene expression related to predator-cue exposure. PC2 explained an order of magnitude less variation (9.3%) and primarily separated generation 2 (P & N) from generations 2 (PN & NN) and 3 (PNN & NNN); these results suggest some of the shifts in gene expression between generation 1 versus generations 2 and 3 were similar in both the control and experimental treatments (Fig. 4). However, it is important to note that none of these differences in gene expression across generations within control samples (e.g., N, NN, NNN) were statistically significant based on pairwise analyses of gene expression, and our analyses of gene expression in experimental samples take into account any such shift through comparisons with these negative controls.

Gene expression patterns tended to be consistent across all three biological replicates per generation, with the exception of a single replicate sample from the third experimental generation (PNN, replicate 1; Fig. 3C). Gene expression patterns in this particular sample were more similar to those in the negative control (Fig. 3C). Our PCA further confirmed this sample as having a unique replicate-specific transgenerational response compared to the other two PNN treatment replicates, as this sample clustered with control samples (with NN and NNN replicates).

Gene Function

To dissect the biological relevance of transgenerational shifts in gene expression, we grouped responsive sets of genes into functional categories to identify how gene expression shifts might be related to transgenerational phenotypic shifts, and how within- versus trans-generational transcriptional responses differ. Comparisons of Gene Ontology (GO) terms for differentially expressed gene sets per generation highlight the uniqueness of within-generation responsive gene functions (P vs. N response), and the broad similarities of functional categories of responsive genes in generations 2 and 3 (PN vs. NN, and PNN vs. NNN; Fig. 5). Within-generation responsive genes were associated with few enriched GO terms, all of which were related to lipid transport and lipid transporter activity (Fig. 5A-B); none of these functional classes were shared with across-generation responsive genes.

In contrast to limited responses in the first generation, responses in generations 2 and 3 show major shifts in the functional categories of genes responsive to predator cues, including the up-regulation of genes involved in cellular amide metabolism, translation, ribosome structure, ribosome biogenesis, biosynthesis, and cellular metabolism (Figs. 5A-B). GO terms enriched in transgenerational responsive gene sets were highly similar, and shared many biological process and molecular function terms. While broadly overlapping, generations 2 and 3 differed in the greater response of genes related to cellular amide metabolism in generation 2, and the greater relative abundance of responsive genes related to translation and ribosomes in generation 3 (Fig. 5). The only non-overlap identified between generations 2 and 3 was in cellular component GO terms, indicating that cytosolic ribosomes are enriched in generation 2, while ribosomal subunits are enriched in generation 3 (Fig. 5A – Cellular Component panel); this difference, however, may have been driven by related GO terms being derived from many of the same up-regulated genes. For these seemingly different categories, original GO enrichment terms were

identical but with differing p-values (Tables 2-3), and the difference in results is due largely to the differential summarization of terms by REViGO, which employed permissive thresholds for similarity (c=0.10) for visualization purposes (Tables 4-5). Thus, functional classes of responsive genes in generations 2 and 3 are, in fact, highly overlapping.

To complement our GO analyses and further dissect the links between gene expression and phenotypes associated with transgenerational plasticity, we broke down sets of responsive genes into functional categories linked to key phenotypic or molecular aspects of plasticity, including: epigenetic modification, reproduction, exoskeleton structure, digestion, and ribosomal protein synthesis (Figs. 6-7). Among genes relevant to epigenetic modifications, we identified a single responsive gene encoding a histone deacetylase (HDAC; a transcriptional silencer; Braunstein 1993) that was variably expressed across treatments, and only significantly differentially expressed between experimental and control treatments in generation 3 (PNN vs. NNN; Fig. 6); we provide plausible explanations for this observation in the discussion.

Genes encoding peptidases and other digestive enzymes exhibited a split pattern, with some genes in this class being responsive upon initial cue exposure (P vs. N), and others showing transgenerational responses. Genes that were significantly responsive exclusively between P and N were peptidases with serine-type endopeptidase activity (i.e., chymotrypsin and trypsin) or metalloendopeptidase activity (i.e., zinc metalloase; Fig. 5 and Fig. 8). Conversely, genes exclusively responsive in generations two and three (PN vs. NN and PNN vs. NNN) were carboxypeptidase D, peptidases functioning in cysteine-type endopeptidase activity (i.e., cathepsin and caspase) and digestive enzymes functioning in hydrolase and celluase activity (i.e., Cel7A fusion and lysosomal alpha-glucosidase-like, respectively; Fig. 6 and Fig. 8).

Up-regulation of genes involved in reproductive function (i.e., vitellogenins; VTG) was primarily associated with the within-generation response. Of the 48 responsive genes in generation 1 (Fig. 3B), five encoded proteins associated with VTG were associated with within-generational responses, (including VTG-like isoforms X2, VTG2, and vitelline membrane outer layer 1 homolog), and only a single gene annotated as ‘VTG-partial’ was responsive in the third generation treatment. GO enrichment analysis showed overrepresented genes involved in biological process terms for lipid transport and molecular

function for lipid transporter activity (Fig. 5), which is consistent with our finding of up-regulation of VTGs in generation 1. Genes encoding exoskeletal proteins were more responsive across-generations. Multiple exoskeleton-associated genes, including genes involved in the structural constituent of the cuticle, were exclusively differentially expressed between experimental and control treatments in generations 2 and 3 (PN vs. NN and PNN vs. NNN), while only a single gene (peritrophic matrix) involved in chitin binding was significantly responsive in the first generation (P vs. N; Fig. 6 and 8).

Among all responsive gene sets, the most pronounced example of a transcriptional program of functionally related genes exclusively linked to TGP was that of ribosomal protein-encoding genes; these genes were up-regulated in response to predator cues in generations 2 and 3 (Fig. 7). Of the 223 significantly differentially expressed genes in generation 2 (PN vs. NN; Fig. 3C), 52 (23%) were annotated as ribosomal protein components. Similarly, 40 (23%) out of 170 genes in the third generation were also annotated as ribosomal protein components (Fig. 7).

Discussion

Our results provide compelling evidence that within- versus across-generation responses may be driven by distinct gene expression programs, indicating these programs are likely regulated independently. Interest in the evolutionary drivers of transgenerational plasticity (TGP) has developed slowly and largely in parallel of the study of within-generation plasticity. Initially, theory predicted that similar ecological conditions favor the evolution of plastic responses that occur within- and across-generations (Day and Bonduriansky 2011; Ezard, et al. 2014; Fischer, et al. 2011; Hoyle and Ezard 2012; Jablonka, et al. 1992, 1989; Kuijper, et al. 2014; Levins 1968; Shea, et al. 2011). It is also hypothesized that varying environmental conditions that are consistent between parent and offspring generations are expected to favor simultaneous increases in phenotypic plasticity within and across generations (Ezard, et al. 2014; Hoyle and Ezard 2012), but this idea has recently been challenged. It is now becoming clear that organisms exhibit strong patterns of within-generation plasticity or across-generation plasticity but not both (Donohue and Schmitt 1998; Schmitt, et al. 1992; Walsh, et al. 2015). Additionally, new theory has identified ecological conditions that may independently select for within- versus trans-generational plasticity (Kuijper and Hoyle 2015; Leimar and McNamara 2015; Uller, et al. 2015b). These frameworks predict that high temporal variability selects for the evolution of within-generation plasticity, while low temporal variability (or high temporal stability) and slow rate of environmental change favors enhanced TGP. The decoupling of within- and across-generation responses, in turn, predicts that divergent ecological conditions favor divergent patterns of plasticity, and even divergent molecular mechanisms underlying plasticity.

Based upon recent theory and empirical work (Walsh, et al. 2015, 2016) illustrating a decoupling and even antagonism of within- versus across-generational phenotypic responses, it follows logically that within and across-generation phenotypic responses involve divergent programs of gene expression and even fundamentally different sets of responsive genes with distinct functions. Additionally, transcriptional responses (e.g., the number of responsive genes) are expected to be generally proportional to phenotypic responses within- versus across-generations. For example, lineages that exhibit strong patterns of TGP are expected to show enhanced transcriptional responses across generations. To test these predictions, we examined transcriptional responses to predator cues using a clone of *Daphnia* that is known to exhibit

strong TGP (Walsh, et al. 2015, 2016). We expected to observe more extensive shifts in gene expression across generations (versus within), especially for genes involved in phenotypically responsive life history traits (i.e., programming offspring for faster rates of development and production of larger clutch sizes; Walsh, et al. 2015).

Results of our gene expression analyses indicate that highly distinct gene expression programs may underlie within- versus across-generation responses, and that the magnitude of transcriptional responses appears to be linked to the magnitude of phenotypic responses. In this particular *Daphnia* lineage known to exhibit strong TGP responses (Walsh, et al. 2015), we found small within-generation transcriptional response upon initial exposure to predator cues (P vs. N) followed by a pronounced transgenerational transcriptional response across subsequent generations (i.e., PN vs. NN and PNN vs. NNN). Many sets of responsive genes were also linked to known phenotypic responses that have been shown to coincide with exposure to predator cues, including developmental rates, reproductive rates, and shifts in growth (Riessen 1999; Stibor 1992; Walsh, et al. 2015; Fig. 6). An 'informational' perspective might explain why organisms exhibiting strong TGP maintain adaptive responses to predator cues, even when the predator risk has ceased (Dall, et al. 2005). Assuming the evolution of mechanisms that allow different sources of information to be weighted differently, selection should favor some sources of information more than others (Dall, et al. 2005). The overall pattern suggests that this *Daphnia* clone appears to respond more to the environment experienced by their mother/grandmother rather than their own environment, which is logical if direct predation risk is likely to be experienced infrequently (Dall, et al. 2005).

Finally, multiple aspects of our results indicate a pattern of 'decay' in transgenerational programming in later generations (Fig. 3C and Fig. 4). That is, the number of differentially expressed genes decreased between generations 2 and 3. Such a trend is consistent with the decay in inherited epigenetic programming of subsequent generations, supporting the view that TGP is driven by epigenetic mechanisms (Kuijper and Hoyle 2015; Leimar and McNamara 2015; Uller, et al. 2015a).

Within-generation patterns of gene expression

Our data indicate that predator cues lead to consistent within-generation up-regulation of genes related to digestive function, including genes encoding the enzymes trypsin and chymotrypsin (serine-type endopeptidases) as well as genes associated with serine-type endopeptidase activity, metaloendopeptidase activity, and threonine-type endopeptidase activity. In addition to genes associated with peptidase activity, within-generation responses were also observed for genes that represent major precursor proteins involved in the production of egg yolk and embryo development (VTGs). Trypsin and chymotrypsin are known to represent major digestive proteases in the gut of *D. magna* (von Elert, et al. 2004). *D. pulex* have been shown to respond to metabolic shifts due to colder temperatures by down-regulating trypsins, chymotrypsins, and carboxypeptidases, and up-regulating VTG (Schwerin, et al. 2009). We observed up-regulation of these enzymes, which may correspond with the need to accommodate increased feeding rates to achieve increases rates of growth and development and larger reproductive investment in response to exposure to predator chemical cues (Riessen 1999; Stibor 1992).

Trans-generational patterns of gene expression

Distinct changes in gene expression persisted for two generations following predator cue removal. Transgenerational responses included 223 significantly differentially expressed genes in the second generation and 170 in the third generation (Fig. 3A and 3C), with an overlap of 121 responsive genes between these generations. These transgenerational responsive genes outnumbered those that were differentially expressed in generation 1 (i.e., within generation responses) by 2 to 4-fold. Hypergeometric tests on these areas of overlap (Fig. 3D) were performed to determine if the overlap in gene sets was more than expected by chance. In both cases, the overlap between generations 1 and 2 and the overlap between generations 2 and 3 were significant with p-values < 0.05. Despite this overlap, the degree to which within-and trans-generational responses were largely distinct, in terms of the genes that were differentially expressed, is notable. These contrasting gene expression responses within- and across-generations are consistent with new theory regarding the decoupling of these two forms of plasticity (Kuijper and Hoyle 2015; Leimar and McNamara 2015; Uller, et al. 2015b). Our previous phenotypic work showed that parents respond to initial exposure to predator cues by programming

offspring for earlier maturation and the production of larger clutch sizes (Walsh, et al. 2015). Though, within-generation responses focused on up-regulating a small set of genes related to reproductive efforts (Fig. 5B & Fig. 6), across-generation responses included many genes linked to components of the exoskeleton, ribosomal proteins, carboxypeptidase D and other peptidases functioning in cysteine-type endopeptidase activity, hydrolase activity and cellulose activity (Fig. 6). Chitin metabolism has been extensively studied in insects and in order for development to occur, cuticles forming the exoskeleton need to be continuously replaced during ecdysis. The ability for an arthropod to undergo morphogenesis is completely dependent on the constant destruction and reconstruction of chitin-containing structures (Merzendorfer 2003). Therefore, increasing the transcription of proteins involved in the cuticle in *Daphnia* is also likely indicative of more frequent molting.

Perhaps the most remarkable transcriptional evidence for a TGP-specific gene expression program is the observed up-regulation of 62 responsive genes encoding ribosomal proteins associated with 60S and 40S ribosomal subunits (Fig. 7). Despite a sensible explanation for this observation as being linked to an increase in translation, previous studies have shown increased transcription of ribosomal proteins without increased production of ribosomes (Sun, et al. 2015; Wang, et al. 2013). Proteomic data gathered on *D. magna* in response to predator cues show similar, but less extreme responses in ribosomal protein up-regulation (Otte, et al. 2015). Furthermore, it is known that ribosomal proteins have functions outside of ribosome assembly and translation in response to stress (i.e., oncoprotein suppression, immune signaling, and development; Zhou, et al. 2015). Although the functional significance of up-regulation of the ribosomal protein-coding genes observed in the current study unclear, it is notable that this class of responsive genes were tightly linked to TGP, and a greater understanding of this response may provide unique insight into TGP response programs.

Stability of transmission and epigenetic decay

Our understanding of the mechanistic basis of plasticity, especially TGP, has been historically limited. A major difficulty is that several non-exclusive mechanisms may underlie patterns of TGP (e.g., maternal effects, histone modification, RNA interference, DNA methylation; Bossdorf, et al. 2008; Jaenisch and Bird 2003; Vandegehuchte and Janssen 2011). Environmentally induced epigenetic shifts in DNA methylation can influence gene expression patterns (Kalisz and Purugganan 2004; Turck and

Coupland 2014) including TGP in gene expression (Boyko, et al. 2010; Carone, et al. 2010; Kooke, et al. 2015), and variation in patterns of DNA methylation among natural populations has been correlated with shifts in trait values and trait plasticity (Herrera and Bazaga 2010; Herrera, et al. 2012; Kooke, et al. 2015; Zhang, et al. 2013). Evolutionary theory connects environmental variation with the expression of TGP by predicting that evolutionary divergence in TGP may be linked to differences in the patterns and duration of environmentally-induced epigenetic effects (i.e., differences in rate of 'epigenetic resetting', (Kuijper and Hoyle 2015; Leimar and McNamara 2015; Uller, et al. 2015a). Additionally, our previous study of genome-wide methylation using the same clone of *D. ambigua* used in the present study found evidence for significant transgenerational shifts in genomic methylation patterns (Schield, et al. 2016). Collectively, available evidence supports DNA methylation as an important mechanism underlying both the transmission and evolution of TGP.

Motivated by existing links between epigenomic modification and TGP, we searched for evidence of predator cue responsive genes related to epigenetic modification in generations 2 and 3. We found distinguishable differences in histone deacetylace (HDAC) mRNA expression levels across treatments; HDAC is expressed consistently higher in the 'predator removal' treatments compared to controls (Fig. 6) and expression of HDAC differ significantly in generation 3 (PNN vs. NNN). These transcriptional silencers (Braunstein 1993) are involved in the epigenetic modifications of histones required to condense chromatin. We were somewhat surprised by this result, as we did not expect this gene to be responsive only in later generations (i.e., generation 3) because we have previously shown major epigenomic modifications resulting in shifts in genomic cytosine methylation patterns in generations 1-2 upon predator cue exposure (Schield, et al. 2016). It is also notable that we observed no significant changes in gene expression for DNA methyltransferases across generations and treatments, despite evidence for shifts in methylation in response to predator cue exposure (Schield, et al. 2016). We believe the most likely explanation for these findings is that HDACs and DNA methyltransferases don't necessarily require shifts in transcription to undergo epigenetic shifts (Law and Jacobsen 2010; Vandegehuchte, et al. 2010). An alternative explanation for the lack of transcriptional responses in genes encoding for epigenetic modifiers is that these marks occur early in development (i.e. *de novo* methylation), even as early as during embryonic development (Harris, et al. 2012; Robichaud, et al. 2012), and because our data was collected

from adult individuals, we may not have captured substantial latent signal of transcriptional regulation of these genes.

Because our transcriptome sampling design included pooling 30 individuals per replicate, it remains an open question how much variation in gene expression exists among individuals within a treatment. Our pooled sampling design should tend to average variation across individuals within a replicate, providing an underestimate of among-individual variation. Within our pooled sampling design, gene expression was highly replicable across treatments with the exception of one replicate in generation 3 (Gen: PNN, Rep: 1; Fig 3C); this replicate more closely resembles the expression patterns of the no-predator treatments (control group; Fig. 4). A plausible explanation for this replicate appearing more like the control group is that the stability of the transfer of non-genetic inheritance is variable (and/or unstable) in the generations following cue removal. In other words, failure of the mechanisms promoting a transgenerational transfer of information (i.e., epigenetic decay) could explain this discrepancy in third-generation responses among replicates, and also broadly explain shifts in TGP responses in later generations. However, more extensive tests need to be performed to confirm this possibility.

Conclusion

We examined the influence of predator cues on patterns of gene expression in a clone of *Daphnia ambigua*. Our results revealed divergent within- and trans-generational patterns of gene expression (Figs. 3-7), as shifts in gene expression in response to predator exposure were largely non-overlapping within and across-generations (Fig 3D). These contrasting gene expression programs are correlated with previously measured differences in patterns of phenotypic plasticity within- versus across-generations in this clone (Walsh, et al. 2015). These complementary data collectively indicate that the molecular mechanisms that underlie within- versus trans-generation plasticity are fundamentally distinct. Our results foreshadow that distinct molecular pathways determine the evolution of phenotypic plasticity within and across generations. A key next step is to determine how natural selection operates on the gene expression programs for within- and trans-generational plasticity in natural systems. Specifically, this poses the intriguing questions of what tradeoffs there might be in lineages with different phenotypic responses, and if these differences involve expression of fundamentally different sets of genes, or if

phenotypic differences instead stem from modulation based on which generation experiences up-regulation of particular gene expression programs.

Acknowledgments

Funding for this project was provided by the Research Enhancement Program and faculty startup funds from the University of Texas at Arlington to MRW and TAC.

Figures

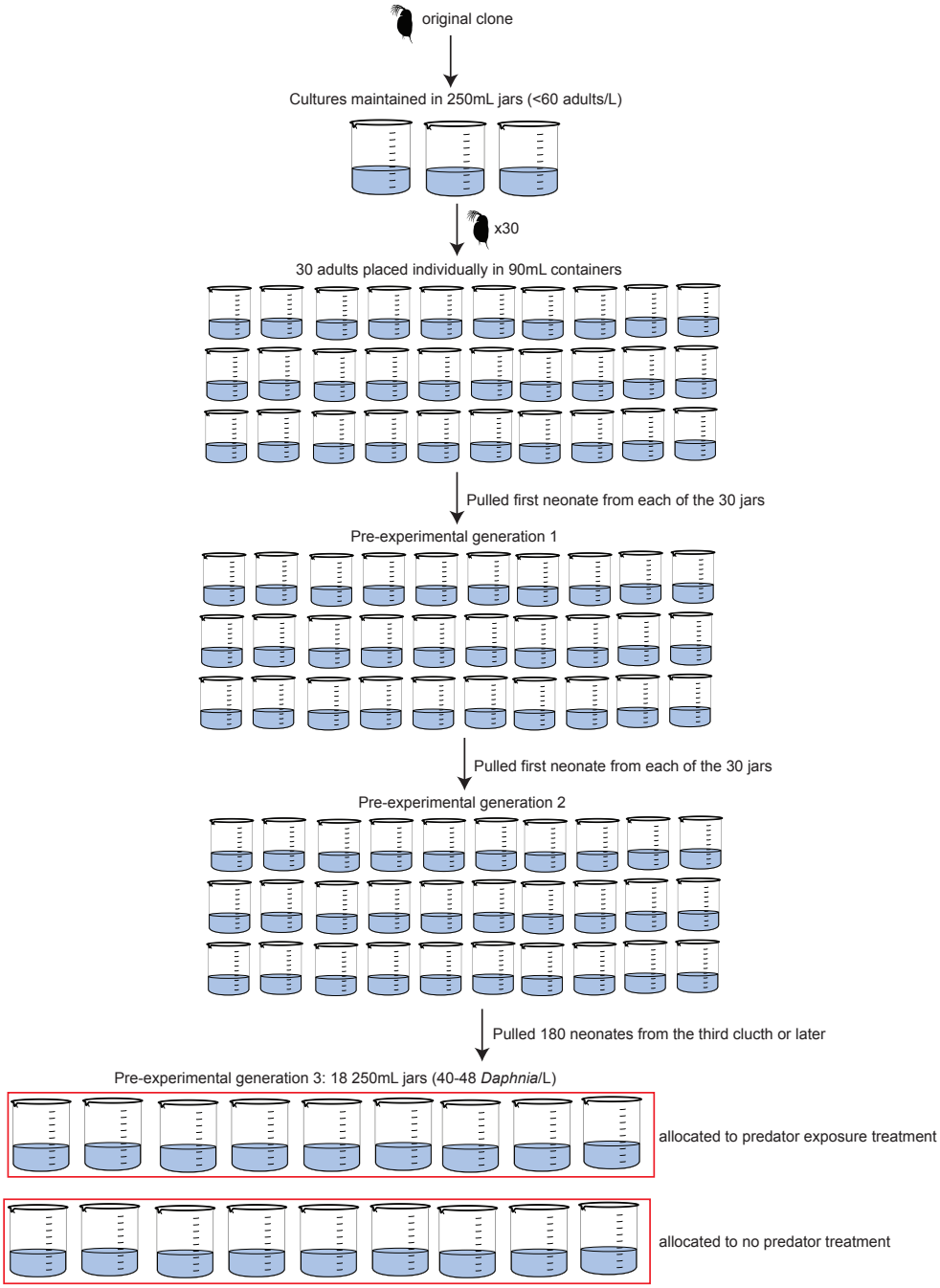

Figure 1. Our pre-experimental approach. A focal clone of Daphnia was reaered in a common garden setting for two generations followed by experimental treatments (described in Figure 1 in the text). The original clone was hatched in the lab and cultures from this clone were maintained in 250mL glass jars and maintained at densities of <60 adults/L. 30 adults were extracted from the existing stock cultures and individually placed into 30 90mL containers. One neotnate was immediately pulled from the first clutch produced by each adult and these were isolated into 90mL containers. The second pre-experimental generation was initiated by collecting neonates from the second cluth of each replicate jar. ~180 newborne Daphnia (from third clutch or later) were collected from the second gneration lab reared parents and placed into 18 250mL jars (densities maintained at 40-48 Daphnia/L). Each of the 18 jars were then randomly allocated to one of two treatments described in the text.

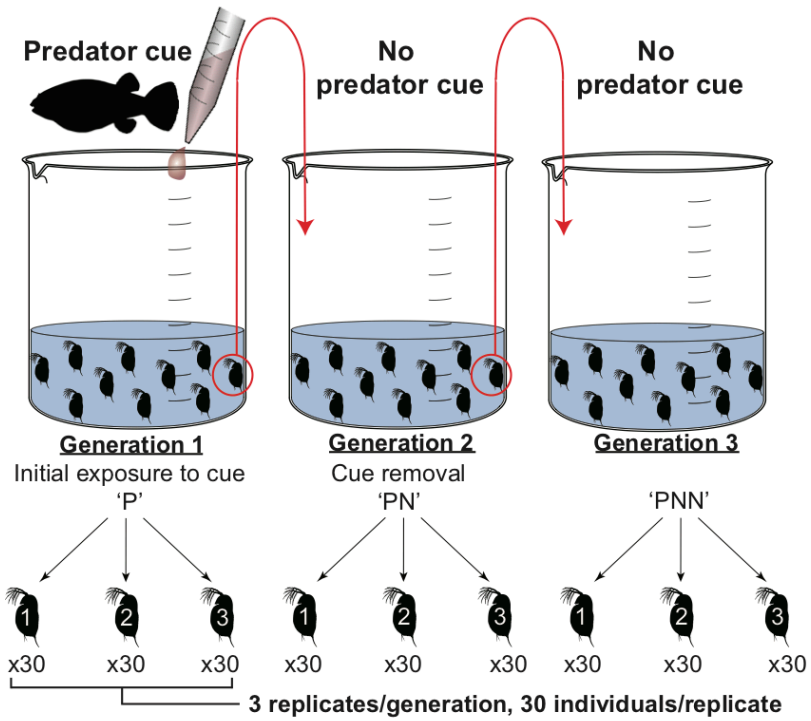

Figure 2. Experimental Design. A clonal generation representing the third common-garden generation of D. ambigua, was exposed to predator cues (Generation 1; 'P' treatment). A single neonate from the second clutch was transferred into a new jar (Generation 2). A neonate was then collected from Generation 2 at 10 days after initiation and placed in a new jar (Generation 3). Generations 2 and 3 were not exposed to any additional predator cues ('N' treatments), so any differentially expressed genes in these generations are a product of trans-generational plasticity stemming from the initial predator cues in Generation 1. RNAseq libraries were prepared from three replicates per generation, with 30 individual Daphnia composing each replicate. In addition to the predator cue removal experiment shown above (P, PN, PNN generations, respectively), a second control experiment was conducted in an identical manner only differing in the absence of any predator exposure (N, NN, NNN generations, respectively).

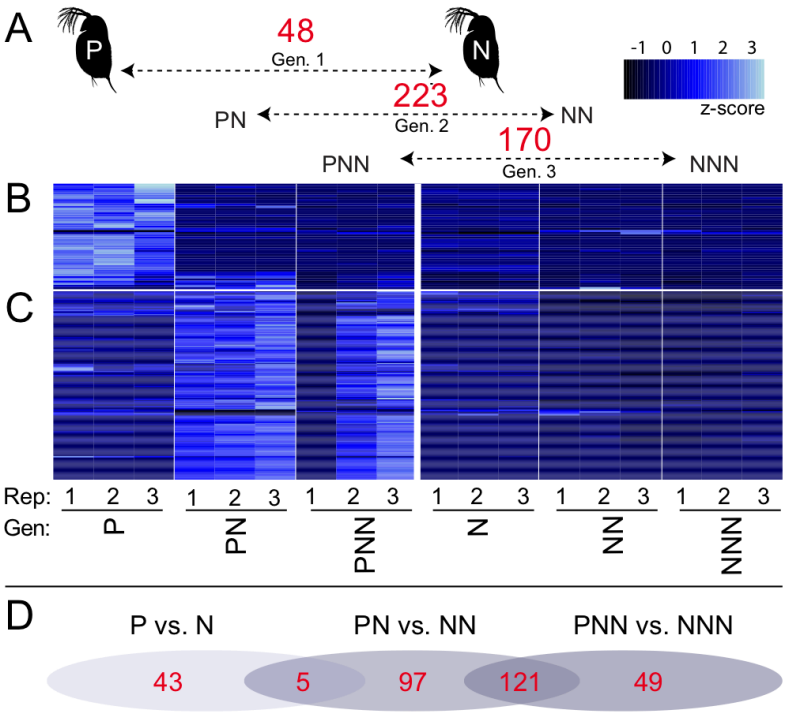

Figure 3. Analysis of gene expression (RNAseq) changes in response to predator cues. (A) Numbers of significantly differentially expressed genes between the experimental and control replicates in each generation. (B) Patterns of gene expression for 48 genes differentially expressed between first generation treatments (P vs. N). (C) Patterns of expression for 223 genes that differ significantly in expression between second-generation treatments (PN vs. NN). For B-C, lighter blue colors indicate high level of expression, while darker blue indicate low expression (this gradient is based on normalized count values); gene expression profiles are clustered by similarity. (D) Significantly differentially expressed genes that overlap between generations. There was no gene overlap between generation 1 (P vs. N) and generation 3 (PNN vs. NNN).

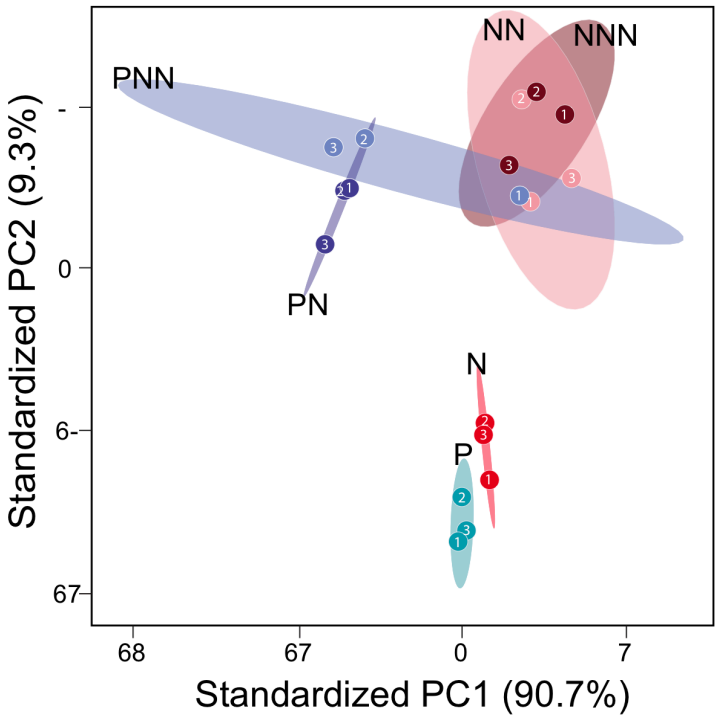

Figure 4. PCA analysis of gene expression profiles for 218 genes that differ in expression between second-generation treatments (PN vs. NN). Normal data ellipses were drawn for each group using 0.98 as the size of the ellipse in Normal probability. Red-shaded ellipses represent control groups (N, NN, and NNN) while blue-shaded ellipses represent the experimental group (P, PN, and PNN). Numbers (1, 2, and 3) within groups represent the replicates associated with Figure 2.

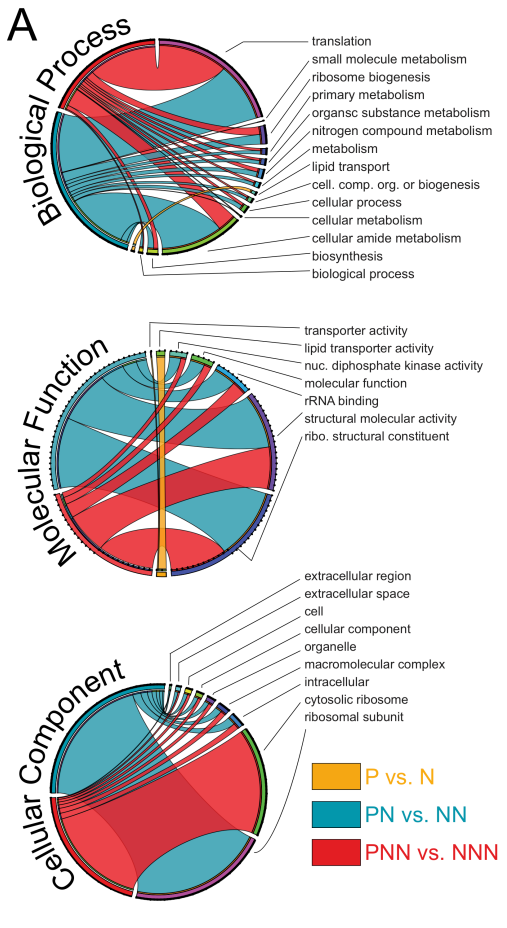

Figure 5. Summarized enrichment of GO biological terms. (A) CIRCOS plots of significant GO terms for biological processes, molecular functions, and cellular components (top to bottom, respectively). Yellow, blue, and red ribbons represent enrichment of significantly differentially expressed genes between generations (P vs. N, PN vs. NN, and PNN vs. NNN respectively). Colored peripheral rings represent GO terms; the larger the ribbon connecting to a peripheral ring (or GO term), the more genes described by that GO term. GO terms for Cellular component did not have any enriches terms for the first generations (P vs. N; yellow ribbon). (B) GO terms for biological process represented in a tree map provided by ReViGO. Large white letters denote the representative GO term for all the more descriptive terms (smaller, gray text) within each color block.

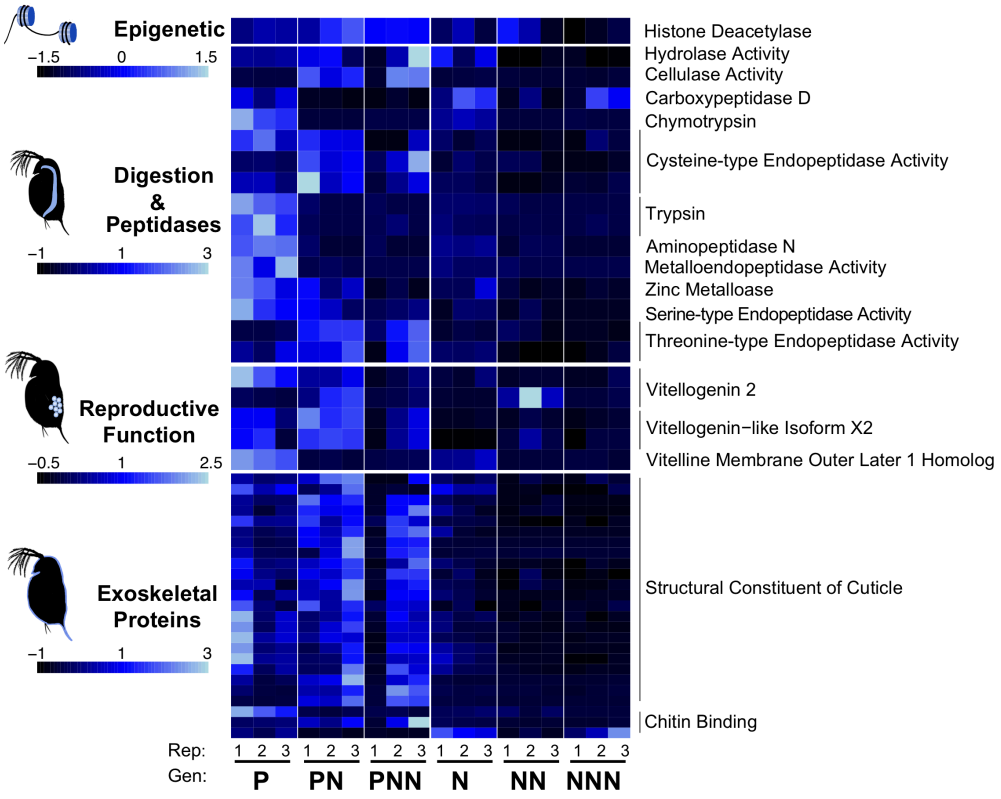

Figure 6. Patterns of responsive genes involved in digestion, proteolysis, reproductive, and exoskeletal function. Genes shown in heat map were found to be significantly differentially expressed across at least one pairwise time point comparison (FDR ≤ 0.05). The respective z-score gradient keys are provided for each heat map.

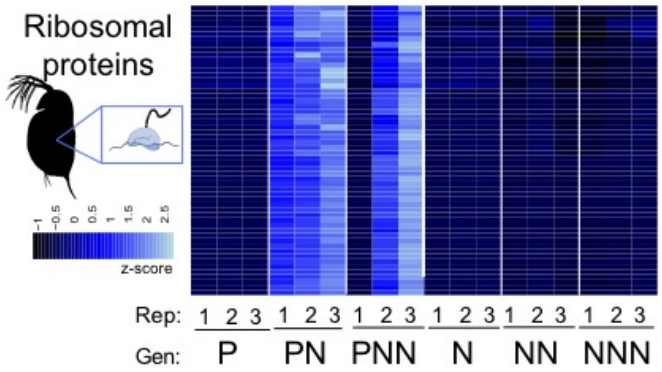

Figure 7. Heat map showing patterns of gene expression in ribosomal-related proteins. Genes shown were significantly differentially expressed across at least one pairwise time point comparison (FDR ≤ 0.05).

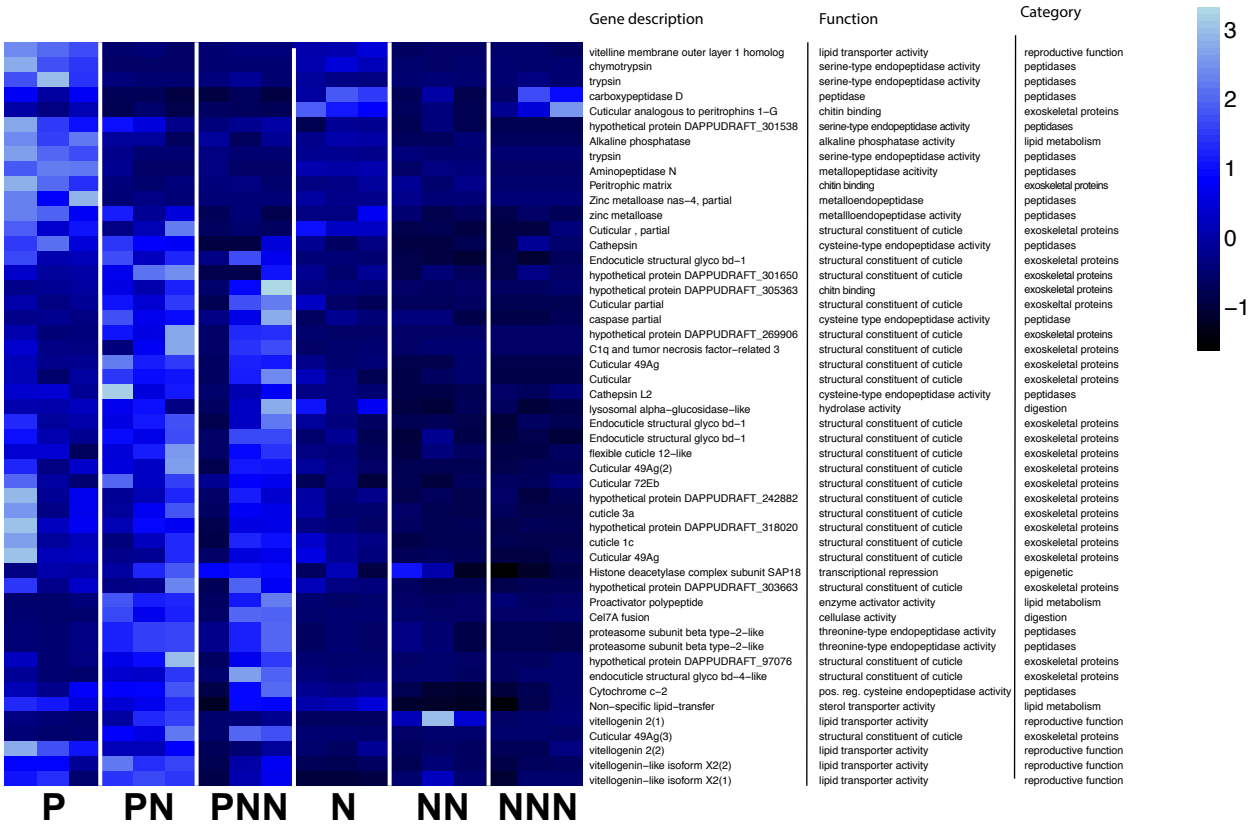

Figure 8. Heat map showing patterns of gene expression involved in digestion, proteolysis, reproductive function, lipid metabolism, and exoskeletal function. This is the same data shown in Figure 4 (aside from two 'lipid metabolism' genes), but here we clustered genes based on expression similarity across generations. A z-score gradient key is provided to the right.

Table 1. The total number of RNA-seq mapped reads per replicate. Each row, or replicate, is comprised of 30 clonal daphnia individuals. (14,018,539 total reads). The IDs given per individual are separated by a forward slash; the first ID is what we referred to these individuals as in the manuscript, while the second ID is representative of the uploaded data.

ID	Raw Reads	After Quality Trim	# Mapped	# Genes
N/1NP-1	302,789	221,953	161,867	15,894
N/1NP-2	326,199	245,356	180,841	16,174
N/1NP-3	297,645	225,298	178,936	16,117
NN/2NP-1	810,299	588,799	398,848	15,991
NN/2NP-2	834,144	666,698	475,725	16,149
NN/2NP-3	860,965	565,587	383,316	17,589
NNN/3NP-1	588,879	413,960	231,991	11,209
NNN/3NP-2	647,654	472,709	333,538	12,454
NNN/3NP-3	627,174	486,513	48,531	12,190
P/1X-1	689,184	514,884	412,073	19,207
P/1X-2	463,577	349,070	291,704	17,544
P/1X-3	510,465	384,320	306,439	18,246
PN/2X-1	781,222	567,821	371,617	13,774
PN/2X-2	880,648	658,245	468,548	14,629
PN/2X-3	789,832	564,721	401,972	13,137
PNN/3X-1	650,780	437,967	293,008	14,344
PNN/3X-2	608,689	484,733	336,212	12,948
PNN/3X-3	458,688	298,814	220,469	9,876
Average:	618,268.5	452,636	285576.1667	14859.55556

Table 2: PANTHER Overrepresentation Test of significantly differentially expressed genes between generation 2 (PN vs. NN) using the GO Ontology database released on 2016-07-29. All Daphnia pulex genes present in the database were used as a reference list. A Bonferroni correction was used.

GO cellular component complete	Daphnia pulex - REFLIST (30578)	upload_1 (210)	upload_1 (expected)	upload_1 (over/under)	upload_1 (fold Enrichment)	upload_1 (P-value)
cytosolic large ribosomal subunit (GO:0022625)	53	27	0.36	+	74.18	7.50E-39
cytosolic small ribosomal subunit (GO:0022627)	38	19	0.26	+	72.8	1.05E-26
cytosolic ribosome (GO:0022626)	94	46	0.65	+	71.26	4.38E-67
large ribosomal subunit (GO:0015934)	74	28	0.51	+	55.1	8.63E-37
ribosomal subunit (GO:0044391)	141	48	0.97	+	49.57	1.10E-62
cytosolic part (GO:0044445)	148	48	1.02	+	47.22	1.08E-61
small ribosomal subunit (GO:0015935)	66	20	0.45	+	44.12	6.60E-24
ribosome (GO:0005840)	174	51	1.19	+	42.68	1.39E-63
cytosol (GO:0005829)	363	50	2.49	+	20.06	2.72E-46
ribonucleoprotein complex (GO:1990904)	445	52	3.06	+	17.02	9.80E-45
intracellular ribonucleoprotein complex (GO:0030529)	445	52	3.06	+	17.02	9.80E-45
intracellular non-membrane-bounded organelle (GO:0043232)	1035	54	7.11	+	7.6	5.61E-29
non-membrane-bounded organelle (GO:0043228)	1036	54	7.11	+	7.59	5.88E-29
extracellular space (GO:0005615)	377	17	2.59	+	6.57	7.24E-07
cytoplasmic part (GO:0044444)	1681	71	11.54	+	6.15	9.03E-34
extracellular region part (GO:0044421)	415	17	2.85	+	5.96	2.95E-06
macromolecular complex (GO:0032991)	1844	65	12.66	+	5.13	5.35E-26
intracellular organelle part (GO:0044446)	1983	65	13.62	+	4.77	3.02E-24
organelle part (GO:0044422)	1987	65	13.65	+	4.76	3.38E-24
cytoplasm (GO:0005737)	2470	78	16.96	+	4.6	3.85E-29
extracellular region (GO:0005576)	635	20	4.36	+	4.59	1.03E-05
intracellular organelle (GO:0043229)	3619	77	24.85	+	3.1	6.54E-18
organelle (GO:0043226)	3645	77	25.03	+	3.08	1.00E-17
intracellular part (GO:0044424)	4508	84	30.96	+	2.71	2.20E-16
intracellular (GO:0005622)	4836	90	33.21	+	2.71	5.61E-18
cell part (GO:0044464)	5526	91	37.95	+	2.4	1.08E-14
cell (GO:0005623)	5564	91	38.21	+	2.38	1.68E-14
cellular_component (GO:0005575)	9210	125	63.25	+	1.98	5.35E-16
Unclassified (UNCLASSIFIED)	21368	85	146.75	-	0.58	0.00E+00

Table 3: PANTHER Overrepresentation Test of significantly differentially expressed genes between generation 3 (PNN vs. NNN) using the GO Ontology database released on 2016-07-29. All Daphnia pulex genes present in the database were used as a reference list. A Bonferroni correction was used.

GO cellular component complete	Daphnia pulex - REFLIST (30578)	upload_1 (162)	upload_1 (expected)	upload_1 (over/under)	upload_1 (fold Enrichment)	upload_1 (P-value)
cytosolic small ribosomal subunit (GO:0022627)	38	17	0.2	+	84.44	6.57E-25
cytosolic ribosome (GO:0022626)	94	38	0.5	+	76.3	1.66E-56
cytosolic large ribosomal subunit (GO:0022625)	53	21	0.28	+	74.79	4.70E-30
large ribosomal subunit (GO:0015934)	74	22	0.39	+	56.12	7.37E-29
ribosomal subunit (GO:0044391)	141	40	0.75	+	53.55	1.42E-53
small ribosomal subunit (GO:0015935)	66	18	0.35	+	51.48	1.20E-22
cytosolic part (GO:0044445)	148	38	0.78	+	48.46	4.14E-49
ribosome (GO:0005840)	174	41	0.92	+	44.48	9.54E-52
cytosol (GO:0005829)	363	39	1.92	+	20.28	4.30E-36
ribonucleoprotein complex (GO:1990904)	445	42	2.36	+	17.81	7.40E-37
intracellular ribonucleoprotein complex (GO:0030529)	445	42	2.36	+	17.81	7.40E-37
intracellular non-membrane-bounded organelle (GO:0043232)	1035	46	5.48	+	8.39	1.37E-26
non-membrane-bounded organelle (GO:0043228)	1036	46	5.49	+	8.38	1.42E-26
extracellular space (GO:0005615)	377	12	2	+	6.01	4.53E-04
cytoplasmic part (GO:0044444)	1681	50	8.91	+	5.61	1.91E-21
extracellular region part (GO:0044421)	415	12	2.2	+	5.46	1.21E-03
macromolecular complex (GO:0032991)	1844	50	9.77	+	5.12	1.05E-19
intracellular organelle part (GO:0044446)	1983	52	10.51	+	4.95	5.25E-20
organelle part (GO:0044422)	1987	52	10.53	+	4.94	5.74E-20
cytoplasm (GO:0005737)	2470	56	13.09	+	4.28	6.9319
extracellular region (GO:0005576)	635	13	3.36	+	3.86	1.76E-02
intracellular organelle (GO:0043229)	3619	60	19.17	+	3.13	5.77E-14
organelle (GO:0043226)	3645	60	19.31	+	3.11	8.05E-14
intracellular (GO:0005622)	4836	68	25.62	+	2.65	8.90E-13
intracellular part (GO:0044424)	4508	63	23.88	+	2.64	2.36E-11
cell (GO:0005623)	5564	70	29.48	+	2.37	7.77E-11
cell part (GO:0044464)	5526	69	29.28	+	2.36	1.90E-10
cellular_component (GO:0005575)	9210	96	48.79	+	1.97	7.21E-12
Unclassified (UNCLASSIFIED)	21368	66	113.21	-	0.58	0.00E+00

Table 4: Summarized GO term categories for cellular processes underlying signidicantly differentially expressed genes between generation 2 (PN vs. NN). Exported REViGO data used for the specific purpose of contruction TreeMap and Circos visualizations. This is not be used as a general list of non-redundant GO categories as it is extremely permissive (c = 0.1).

term_ID	description	frequencyIn Db	log10pvalue	uniqueness	dispensability	representative
GO:0005575	cellular_component	100.00%	-15.2716	1	0	cellular_component
GO:0005576	extracellular region	4.57%	-4.9872	0.928	0	extracellular region
GO:0005615	extracellular space	0.25%	-6.1403	0.88	0	extracellular space
GO:0044421	extracellular region part	2.87%	-5.5302	0.882	0.617	extracellular space
GO:0005623	cell	64.13%	-13.7747	0.973	0	cell
GO:0032991	macromolecular complex	14.46%	-25.2716	0.935	0	macromolecular complex
GO:0043226	organelle	16.72%	-17	0.937	0	organelle
GO:0044391	ribosomal subunit	1.36%	-61.9586	0.444	0	ribosomal subunit
GO:0015934	large ribosomal subunit	0.64%	-36.064	0.437	0.576	ribosomal subunit
GO:0005829	cytosol	0.81%	-45.5654	0.67	0.378	ribosomal subunit
GO:0022626	cytosolic ribosome	0.03%	-66.3585	0.526	0.691	ribosomal subunit
GO:0022625	cytosolic large ribosomal subunit	0.01%	-38.1249	0.504	0.637	ribosomal subunit
GO:0044445	cytosolic part	0.39%	-60.9666	0.608	0.35	ribosomal subunit
GO:0044444	cytoplasmic part	13.61%	-33.0443	0.716	0.189	ribosomal subunit
GO:0043228	non-membrane-bounded organelle	8.44%	-28.2306	0.597	0.44	ribosomal subunit
GO:0005737	cytoplasm	38.16%	-28.4145	0.701	0.316	ribosomal subunit
GO:0044424	intracellular part	43.66%	-15.6576	0.716	0.465	ribosomal subunit
GO:0043229	intracellular organelle	15.79%	-17.1844	0.483	0.69	ribosomal subunit
GO:0044422	organelle part	6.52%	-23.4711	0.606	0.573	ribosomal subunit
GO:0030529	ribonucleoprotein complex	6.09%	-44.0088	0.632	0.466	ribosomal subunit
GO:0005622	intracellular	46.14%	-17.251	0.841	0.063	intracellular
GO:0044464	cell part	64.13%	-13.9666	0.84	0.262	intracellular

Table 5: Summarized GO term categories for cellular processes underlying signidicantly differentially expressed genes between generation 3 (PNN vs. NNN). Exported REViGO data used for the specific purpose of contruction TreeMap and Circos visualizations. This is not be used as a general list of non-redundant GO categories as it is extremely permissive (c = 0.1).

term_ID	description	frequencyIn Db	log10pvalue	uniqueness	dispensability	representative
GO:0005575	cellular_component	100.00%	-11.1421	1	0	cellular_component
GO:0005576	extracellular region	4.57%	-1.7545	0.928	0	extracellular region
GO:0005615	extracellular space	0.25%	-3.3439	0.88	0	extracellular space
GO:0044421	extracellular region part	2.87%	-2.9172	0.882	0.617	extracellular space
GO:0005623	cell	64.13%	-10.1096	0.973	0	cell
GO:0022626	cytosolic ribosome	0.03%	-55.7799	0.526	0	cytosolic ribosome
GO:0015934	large ribosomal subunit	0.64%	-28.1325	0.437	0.576	cytosolic ribosome
GO:0005829	cytosol	0.81%	-35.3665	0.67	0.263	cytosolic ribosome
GO:0044391	ribosomal subunit	1.36%	-52.8477	0.444	0.427	cytosolic ribosome
GO:0022625	cytosolic large ribosomal subunit	0.01%	-29.3279	0.504	0.622	cytosolic ribosome
GO:0044445	cytosolic part	0.39%	-48.383	0.608	0.691	cytosolic ribosome
GO:0044444	cytoplasmic part	13.61%	-20.719	0.716	0.116	cytosolic ribosome
GO:0043228	non-membrane-bounded organelle	8.44%	-25.8477	0.597	0.44	cytosolic ribosome
GO:0005737	cytoplasm	38.16%	-18.1593	0.701	0.316	cytosolic ribosome
GO:0044424	intracellular part	43.66%	-10.6271	0.716	0.465	cytosolic ribosome
GO:0043229	intracellular organelle	15.79%	-13.2388	0.483	0.69	cytosolic ribosome
GO:0044422	organelle part	6.52%	-19.2411	0.606	0.573	cytosolic ribosome
GO:0030529	ribonucleoprotein complex	6.09%	-36.1308	0.632	0.466	cytosolic ribosome
GO:0032991	macromolecular complex	14.46%	-18.9788	0.935	0	macromolecular complex
GO:0043226	organelle	16.72%	-13.0942	0.937	0	organelle
GO:0005622	intracellular	46.14%	-12.0506	0.841	0.035	intracellular
GO:0044464	cell part	64.13%	-9.7212	0.84	0.262	intracellular

References

Agrawal AA, Laforsch C, Tollrian R (1999) Transgenerational induction of defenses in plants and animals. *Nature*, **401**, 60-63.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25-29.

Aubin-Horth N, Renn SPC (2009) Genomic reaction norms: using integrative biology to understand molecular mechanisms of phenotypic plasticity. *Molecular Ecology*, **18**, 3763-3780.

Bashey F (2006) Cross-generational environmental effects and the evolution of offspring size in the Trinidadian guppy *Poecilia reticulata*. *Evolution*, **60**, 348-361.

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

Bonduriansky R, Crean AJ, Day T (2012) The implications of nongenetic inheritance for evolution changing environments. *Evololutionary Applications*, **5**, 192-201.

Bossdorf O, Richards CL, Pigliucci M (2008) Epigenetics for ecologists. *Ecology Letters*, **11**, 106-115.

Boyko A, Belvins T, Yao Y, Golubov A, Bilichak A, *et al.* (2010) Transgenerational adaptation of *Arabidopsis* to stress requires DNA methylation and the function of Dicer-like proteins. *PLoS One*, **5**, e9514..

Braunstein M, Rose, AB, Holmes, SG, Allis CD, Broach JR (1993) Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes and Development*, **7**, 592-604.

Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, Bock C, Li C, Gu H, Zamore PD, *et al.* (2010) Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell*, **143**, 1084-1096.

Carpenter SR, Fisher SG, Grimm NB, Kitchell JF (1992) Global change and freshwater ecosystems. *Annual Review of Ecology and Systematics*, **23**, 119-139.

Charmantier A, McCleery RH, Cole LR, Perrins C, Kruuk LEB, Sheldon BC (2008) Adaptive phenotypic plasticity in response to climate change in a wild bird population. *Science*, **320**, 800-803.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, *et al*. (2011) The ecoresponsive genome of *Daphnia pulex*. *Science*, **331**, 555-561.

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, **2008**, 1-13.

Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674-3676.

Day T, Bonduriansky R (2011) A unified approach to the evolutionary consequences of genetic and nongenetic inheritance. *American Naturalist*, **178**, E18-E36.

Dixon P (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, **14**, 927-930.

Donohue K, Schmitt J (1998) *Maternal Environmental Effects in Plants: Adaptive Plasticity?* Oxford University Press. Oxford, UK.

Dyer AR, Brown CS, Espeland EK, McKay JK, Meimberg H, Rice KJ (2010) SYNTHESIS: The role of adaptive trans-generational palsticity in biological invasions of plants. *Evolutionary Applications*, **3**, 179-192.

von Elert E., Agrawal MK, C. G, Jaensch H, Bauer U, Zitt A (2004) Protease activity in gut of *Daphnia magna*: evidence for trypsin and chymotrypsin enzymes. *Comparative Biochemistry and Physiology*, **137**, 287-296.

Ezard THG, Prizak R, Hoyle RB (2014) The fitness costs of adaptation via phenotypic plasticity and maternal effects. *Functional Ecology*, **28**, 693-701.

Fischer B, Taborsky B, Kokko H (2011) How to balance the offspring quality–quantity tradeoff when environmental cues are unreliable. *Oikos*, **120**, 258-270.

Fox CW, Mousseau TA (1998) *Maternal Effects as Adaptations for Transgenerational Phenotypic Plasticity in Insects*. In: Maternal Effects as Adaptations (Mousseau TA and Fox CW eds.) pp. 159-177, Oxford University Press. Oxford, UK.

Galloway LF (2005) Maternal effects provide phenotypic adaptation to local environmental conditions. *New Phytologist*, **166**, 93-100.

Galloway LF (2009) Plasticity to canopy shade in a monocarpic herb: within-and between-generation effects. *New Phytologist*, **182**, 1003-1012.

Galloway LF, Etterson JR (2007) Transgenerational plasticity Is adaptive in the wild. *Science*, **318**, 1134-1136.

Schröder T, Gilbert JJ (2004) Transgenerational plasticity for sexual reproduction and diapause in the life cycle of monogonont rotifers: intraclonal, intraspecific and interspecific variation in the response to crowding. *Functional Ecology*, **18**, 458-466.

Götz S, Arnold R, Sebastián-León P, Martín-Rodríguez S, Tischler P, Jehl MA, Dopazo J, Rattei T, Conesa A (2011) B2G-FAR, a species centered GO annotation repository. *Bioinformatics*, **27**, 919-924.

Herman JJ, Spencer HG, Donohue K, Sultan SE (2014) How stable 'should' epigenetic modifications be? Insights from adaptive plasticity and bet hedging. *Evolution*, **68**, 632-643.

Herman JJ, Sultan SE (2011) Adaptive transgenerational plasticity in plants: case studies, mechanisms, and implications for natural populations. *Frontiers in Plant Science*, **6**, 1-10.

Herrera CM, Bazaga P (2010) Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet *Viola cazorlensis*. *New Phytologist*, **187**, 867-876.

Herrera CM, Pozo MI, Bazaga P (2012) Jack of all nectars, master of most: DNA methylation and the epigenetic basis of niche width in a flower-living yeast. *Molecular Ecology*, **21**, 2602-2616.

Hoyle RB, Ezard THG (2012) The benefits of maternal effects in novel and in stable environments. *Journal of the Royal Society Interface*, **9**, 2403-2413.

Jablonka E, Lachmann M, Lamb MJ (1992) Evidence, mechanisms and models for the inheritance of acquired characters. *Journal of Theoretical Biology*, **158**, 245-268.

Jablonka E, Lachmann M, Lamb MJ (1989) The inheritance of acquired epigenetic variations. *Journal of Theoretical Biology*, **139**, 69-83.

Jablonka E, Oborny B, Molnar I, Kisdi E, Hofbauer J, Czaran T (1995) The adaptive advantage of phenotypic memory in changing environments. *Philosophical Transactions of the Royal Society B*, **350**, 133-141.

Jablonka E, Raz G (2009) Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Quarterly Review of Biology*, **84**, 131-176.

Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, **33**, 245-254.

Kalisz S, Purugganan MD (2004) Epialleles via DNA methylation: consequences for plant evolution. *Trends in Ecology and Evolution,* **19**, 309-314.

Kilham SS, Kreeger DA, Lynn SG, Goulden CE, Herrera L (1998) COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia*, **377**, 147-159.

Kinsella R, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Flicek P (2011) Ensembl BioMarts: a hub for data retrieval across the taxonomic space. *Database*, **2011**.

Kooke R, Johannes F, Wardenaar R, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJB (2015) Epigenetic basis for morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. *Plant Cell*, **27**, 337-348.

Kuijper B, Hoyle RB (2015)a. When to rely on maternal effects and when on phenotypic plasticity? *Evolution*, **69**, 950-968.

Kuijper B, Hoyle RB (2015)b. When to rely on maternal effects and when on phenotypic plasticity? *Evolution*, **69**, 950-968.

Kuijper B, Johnstone RA, Townley S (2014) The evolution of multivariate maternal effects. *PLoS Computational Biology*, **10**, e1003550.

Laforsch C, Beccara L, Tollrian R (2006) Inducible defenses: The relevance of chemical alarm cues in *Daphnia*. *Limnology and Oceanography,* **51**, 1466-1472.

Leimar O, McNamara JM (2015) The evolution of transgenerational integration of information in heterogeneous environments. *American Naturalist* **185**, E55-E69.

Levins R (1968) *Evolution in Changing Environments*. Princeton University Press. Princeton, New Jersey.

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 (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* **25**, 2078-2079.

Lin SM, Galloway LF (2010) Environmental context determines within-and potential between-generation consequences of herbivory. *Oecologia*, **163**, 911-920.

Marshall DJ (2008) Transgenerational plasticity in the sea: context-dependent maternal effects across the life history. *Ecology,* **89**, 418-427.

Merzendorfer H, Zimoch, L (2003) Chitin metabolism in insects: structure, function and regulation of chitin synthases and chitinases. *The Journal of Experimental Biology*, **206**, 4393-4412.

Miner BE, De Meester L, Pfrender ME, Lampert W, Hairston NG (2012) Linking genes to communities and ecosystems: *Daphnia* as an ecogenomic model. *Proceedings of the Royal Society B-Biological Sciences*, **279**, 1873-1882.

Miyakawa H, Imai M, Sugimoto N, Ishikawa Y, Ishikawa A, Ishigaki H, Okada Y, Miyazaki S, Koshikawa S, Cornette R, Miura T (2010) Gene up-regulation in response to predator kairomones in the water flea, *Daphnia pulex*. *BMC Developmental Biology*, **10**.

Molinier J, Ries G, Zipfel C, Hohn B (2006) Transgeneration memory of stress in plants. *Nature* **442,** 1046-1049.

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.

Otte KA, Schrank I, Frohlich T, Arnold GJ, Laforsch C (2015) Interclonal proteomic responses to predator exposure in *Daphnia magna* may depend on predator composition of habitats. *Molecular Ecology*, **24**, 3901-3917.

Post DM, Palkovacs EP, Schielke EG, Dodson SI (2008) Intraspecific variation in a predator affects community structure and cascading trophic interactions. *Ecology,* **89,** 2019-2032.

Riessen HP (1999) Predator-induced life history shifts in *Daphnia:* a synthesis of studies using meta-analysis. *Canadian Journal of Fisheries and Acquatic Sciences*, **56,** 2487-2494.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* **26,** 139-140.

Rozenberg A, Parida M, Leese F, Weiss LC, Tollrian R, Manak JR (2015) Transcriptional profiling of predator-induced phenotypic plasticity in *Daphnia pulex*. *Frontiers in Zoology,* **12**.

Salinas S, Munch SB (2012) Thermal legacies: transgenerational effects of temperature on growth in a vertebrate. *Ecology Letters*, **15,** 159-163.

Schield DR, Walsh MR, Card DC, Andrew AL, Adams RH, Castoe TA (2016) EpiRADseq: scalable analysis of genomewide patterns of methylation using next-generation sequencing. *Methods in Ecology and Evolution*, **7**, 60-69.

Schmitt J, Niles J, Wulff RD (1992) Norms of reaction of seed traits to maternal environments in *Plantago lanceolata*. *American Naturalist*, **139**, 451-466.

Schwarzenberger A, Courts C, von Elert E (2009) Target gene approaches: gene expression in *Daphnia magna* exposed to predator-borne kairomones or to microcystin-producing and microcystin-free *Microcysts aeruginosa*. *BioMed Central Genomics*, **10**, 527-541.

Schwerin S, Zeis B, Lamkemeyer T, Paul RJ, Koch M, Madlung J, Fladerer C, Pirow R (2009) Acclimatory responses of the *Daphnia pulex* proteome to environmental changes. II. Chronic exposure to different temperatures (10 and 20°C) mainly affects protein metabolism. BMC Physiology **9**.

Shea N, Pen I, Uller T (2011) Three epigenetic information channels and their different roles in evolution. *Journal of Evolutionary Biology,* **24**, 1178-1187.

Simons AM (2011) Adaptive transgenerational plasticity in plants: case studies, mechanisms, and plications for natural populations. *Frontiers in Plant Science*, **2**.

Stibor H (1992) Predator induced life-history shifts in freshwater cladoceran. *Oecologia,* **92,** 162-165.

Stollewerk A (2010) The water flea *Daphnia*- a 'new' model system for ecology and evolution? *Journal of Biology*, **9**.

Sultan SE, Barton K, Wilczek AM (2009) Contrasting patterns of transgenerational plasticity in ecologically distinct congeners. *Ecology,* **90**, 1831-1839.

Sun J, Li C, Wang S (2015) The Up-regulation of ribosomal proteins further regulates protein expression profile in female *Schistosoma japonicum* after pairing. *PLOS*, **10**, e0129626.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS*, **6**, e21800.

Tollrian R, Harvell DC (1999) *The Ecology and Evolution of Inducible Defenses*. Princeton University Press. Princeton, New Jersey.

Tollrian R, Leese F (2010) Ecological genomics: steps towards unraveling the genetic basis of inducible defenses in *Daphnia*. BMC Biology, **5**.

Turck F, Coupland G (2014) Natural variation in epigenetic gene regulation and its effects on plant developmental traits. *Evolution*, **68**,620-631.

Uller T (2008) Developmental plasticity and the evolution of parental effects. *Trends Ecolology and Evolution*, **23**, 432-438.

Uller T, English S, Pen I (2015)a. The evolution of transgenerational integration of information in heterogeneous environments. *Evolutionary Applications*, **185**, E55-E69.

Uller T, English S, Pen I (2015)b. When is incomplete epigenetic resetting in germ cells favoured by natural selection? *Proceedings of the Royal Society of London B*. **282**, 20150682.

Vandegehuchte MB, Janssen CR (2011) Epigenetics and its implications for ecotoxicology. *Ecotoxicology*, **20**, 607-624.

Walsh MR, Castoe TA, Holmes J, Packer M, Biles K, Walsh MJ, Munch SB, Post DM (2016) Local adaptation in transgenerational responses to predators. *Proceedings of the Royal Society of London B.* **283**, 20152271.

Walsh MR, Cooley F, Biles K, Munch SB (2015) Predator-induced phenotypic plasticity within- and across-generations: a challenge for theory? *Proceedings of the Royal Society B*, **282**, 20142205.

Walsh MR, Post DM (2012) The impact of intraspecific variation in a fish predator on the evolution of phenotypic plasticity and investment in sex in *Daphnia ambigua*. *Journal of Evolutionary Biology*, **25**, 80-89.

Walsh MR, Post DM (2011) Interpopulation variation in a fish predator drives evolutionary divergence in prey in lakes. Proceedings of the Royal Society B-Biological Sciences 278:2628-2637.

Wang J, Lan P, Gao H, Zheng L, Li W, Schmidt W (2013) Expression changes of ribosomal proteins in phosphate- and iron-deficient *Arabidopsis* roots predict stress-specific alterations in ribosome composition. *BMC Genomics*, **14**.

Wickstrom M, Larsson R, Nygren P, Gullbo J (2010) Aminopeptidase N (CD13) as a target for cancer chemotherapy. *Cancer Science*, **102**, 501-508.

Zhang Y, Fischer M, Colot V, Bossdorf O (2013) Epigentic variation creates potential for evolution of plant phenotypic plasticity. *New Phytologist*, **197**, 314-322.

Zhou X, Liao W-J, Liao J-M, Liao P, Lu H (2015) Ribosomal proteins: functions beyond the ribosome. *Journal of Molecular Cell Biology* **7**, 92-104.

Chapter 3

A chromosome-level prairie rattlesnake genome provides new insight into reptile genome biology and gene regulation in the venom gland.

Drew R. Schield[1], Daren C. Card[1], Nicole R. Hales[1], Blair W. Perry[1], Giulia M. Pasquesi[1], Heath Blackmon[2], Richard H. Adams[1], Andrew B. Corbin[1], Balan Ramesh[1], Jeffrey P. Demuth[1], Marc Tollis[3], Jesse M. Meik[4], Stephen P. Mackessy[5], and Todd A. Castoe[1, §]

[1]Department of Biology & Amphibian and Reptile Diversity Research Center, University of Texas at Arlington, Arlington, TX, USA

[2]Department of Biology, Texas A&M University, College Station, TX, USA

[3]School of Life Sciences, Arizona State University, Tempe, AZ, USA

[c]Department of Biological Sciences, Tarleton State University, 1333 W. Washington Street, Stephenville, TX, 76402 USA

[5]School of Biological Sciences, University of Northern Colorado, Greeley, CO, USA

Abstract

Here we present the genome sequence of the prairie rattlesnake (*Crotalus viridis viridis*), the first chromosome-level vertebrate genome generated using only next-generation sequencing, and use this genome to study key biological features of reptiles, and venomous snakes specifically. We identify the full length rattlesnake Z chromosome, including the recombining pseudoautosomal region, and demonstrate remarkable similarities in Z chromosome evolution and structure between snake and avian species. We also find evidence for incomplete dosage compensation, and identify multiple mechanisms that appear to contribute to the incomplete dosage on the snake Z chromosome. This genome also provides some of the first clear insight into the origins, structure, and function of reptile microchromosomes, which we find have markedly different structure and function compared to macrochromosomes. Rattlesnake microchromosomes harbor elevated gene density, show substantial variation in regional GC content, and (based on analysis of the 3D chromatin structure) appear to interact with other chromosomes at a much higher frequency than do macrochromosomes. This genome assembly also allowed, for the first time, identification of the chromosomal locations of all rattlesnake venom gene families. We find that microchromsomes are particularly enriched for venom genes, which we show have evolved through multiple tandem duplication events of multiple gene families. By overlaying 3D chromatin structure information and gene expression data we identify specific transcription factors that direct expression on venom genes, and demonstrate how chromatin structure guides precise expression of multiple venom gene families. Together, analyses of the prairie rattlesnake genome reveal multiple key features of reptile genome biology, and provide insight into the origins, structure, and regulation of a complex and dynamic phenotype - snake venom.

Introduction

Snakes have become important model systems for understanding the evolution of specialized and extreme phenotypes, such as limblessness (Cohn and Tickle 1999), metabolic adaptation (Castoe et al. 2008), and a spectrum of adaptations linked to prey acquisition. Indeed, snakes possess several of the

most striking examples of adaptations for feeding known in vertebrates, including a highly kinetic skull allowing snakes to feed on large prey (Gans 1961), extreme physiological and metabolic fluctuations in response to feeding (Secor and Diamond 1998), and the evolution of toxic venoms for prey capture together with highly-specialized structures for storing and delivering venom (i.e., venom glands and fangs). Snakes and other squamate reptiles have also played increasingly prominent roles in studies of genomic repeat element evolution (Castoe et al. 2011; Castoe et al. 2013; Pasquesi et al. In Review), GC isochore structure (Fujita et al. 2011; Castoe et al. 2013), and the evolution of sex chromosomes (Matsubara et al. 2006; Vicoso et al. 2013). In particular, snakes are a valuable system for understanding the evolutionary trajectories of sex chromosome evolution because snakes have evolved both ZW and XY sex chromosomes independently several times (Gamble et al. 2017), and snake species exhibit a spectrum of sex chromosome differentiation (Matsubara et al. 2006), ranging from karyologically indistinguishable homomorphic (e.g., in boas), to highly heteromorphic differentiated sex chromosomes in vipers.

Limited genomic resources and fragmented genome assemblies have been a barrier to fully leveraging snakes as model systems for studying the genomic basis of extreme adaptations and the evolution genome structure (Bradnam et al. 2013; Castoe et al. 2013; Vonk et al. 2013; Yin et al. 2016). To address this, we constructed a high-quality genome of the prairie rattlesnake (*Crotalus viridis viridis*) using a combination of high-throughput sequencing and Hi-C scaffolding (Lieberman-Aiden et al. 2009). The prairie rattlesnake is a pitviper native to North America, which possesses potent and complex venom. This species, and viperid snakes in general, have highly-differentiated sex chromosomes and remarkable genome-wide variation in transposable element abundance and diversity. Here we use the rattlesnake genome, which is the first chromosome-level genome assembly for a reptile, to address multiple hypotheses regarding snake and vertebrate genome evolution, and to provide new insight into the regulatory mechanisms underlying venom production in the rattlesnake venom gland.

Methods

Prairie rattlesnake Genome Sequencing and Assembly

A male prairie rattlesnake (Crotalus viridis viridis) collected from a wild population in Colorado was used to generate the genome sequence. This specimen was collected and humanely euthanized according to University of Northern Colorado Institutional Animal Care and Use Committee protocols 0901C-SM-MLChick-12 and 1302D-SM-S-16. Colorado Parks and Wildlife scientific collecting license 12HP974 issued to S.P. Mackessy authorized collection of the animal. Genomic DNA was extracted using a standard Phenol-Chloroform-Isoamyl alcohol extraction from liver tissue that was snap frozen in liquid nitrogen. Multiple short-read sequencing libraries were prepared and sequenced on various platforms, including 50bp single-end and 150bp paired-end reads on an Illumina GAII, 100bp paired-end reads on an Illumina HiSeq, and 300bp paired-end reads on an Illumina MiSeq. Long insert libraries were also constructed by and sequenced on the PacBio platform. Finally, we constructed two sets of mate-pair libraries using an Illumina Nextera Mate Pair kit, with insert sizes of 3-5Kb and 6-8Kb, respectively. These were sequenced on two Illumina HiSeq lanes with 150bp paired-end sequencing reads. Short and long read data were used to assemble the previous genome assembly version CroVir2.0 (NCBI accession SAMN07738522). Details of these sequencing libraries are in Table 1. Prior to assembly, reads were adapter trimmed using BBmap (Bushnell 2014) and we quality trimmed all reads using Trimmomatic v0.32 (Bolger et al. 2014). We used Meraculous (Chapman et al. 2011) and all short-read Illumina data to generate a contig assembly of the prairie rattlesnake. We then performed a series of scaffolding and gap-filling steps. First, we used L_RNA_scaffolder (Xue et al. 2013) to scaffold contigs using the complete transcriptome assembly (see below), SSPACE Standard (Boetzer et al. 2010) to scaffold contigs using mate-pair reads, and SSPACE Longread to scaffold using long PacBio reads. We then used GapFiller (Nadalin et al. 2012) to extend contigs and fill gaps using all short-read data cross five iterations. We merged the scaffolded assembly with a contig assembly generated using the de novo assembly tool in CLC Genomics Workbench (Qiagen Bioinformatics, Redwood City, CA, USA).

We improved the CroVir2.0 assembly using the Dovetail Genomics HiRise assembly method, leveraging both Chicago and Hi-C sequencing. Chicago assembly requires large amounts of high molecular weight

DNA from a very fresh tissue sample. We thus extracted high molecular weight genomic DNA from a liver of a closely related male to the CroVir2.0 animal (i.e., from the same den site). This animal was collected and humanely euthanized according to the Colorado Parks and Wildlife collecting license and UNC IACUC protocols detailed above. Hi-C sequencing data were derived from the venom gland of the same animal (see details below on venom gland Hi-C and RNAseq experimental design). Dovetail Genomics HiRise assembly resulted in a highly contiguous genome assembly (CroVir3.0) with a physical coverage of greater than 1,000x. We estimated the size of the genome using k-mer frequency distributions (17, 19, and 21mers) quantified using Jellyfish (Marçais and Kingsford 2011).

We generated transcriptomic libraries from RNA sequenced from 16 different tissues: two venom gland tissues; 1 day and 3 days post-venom extraction (see Hi-C and RNA sequencing of Venom Gland section below), one from pancreas, and one from tongue were taken from the Hi-C sequenced genome animal. Additional samples from other individuals included a third venom gland sample from which venom had not been extracted (‘unextracted venom gland’), three liver, three kidney, two pancreas, and one each of skin, lung, testis, accessory venom gland, shaker muscle, brain, stomach, ovaries, rictal gland, spleen, and blood tissues. Total RNA was extracted using Trizol, and we prepared RNAseq libraries using an NEB RNAseq kit for each tissue, which were uniquely indexed and run on multiple HiSeq 2500 lanes using 100bp paired-end reads (Table 6). We used Trinity v. 20140717 (Grabherr et al. 2011) with default settings and the ‘--trimmomatic' setting to assemble transcriptome reads from all tissues. The resulting assembly contained 801,342 transcripts comprising 677,921 Trinity-annotated genes, with an average length of 559 bp and an N50 length of 718 bp.

Repeat Element Analysis

Annotation of repeat elements was performed using homology-based and *de novo* prediction approaches. Homology-based methods of transposable element identification like *RepeatMasker* cannot recognize elements that are not in a reference database, and have low power to identify fragments of repeat elements belonging to even moderately diverged repeat families (Platt et al. 2016). Since the current release of the Tetrapoda RepBase library (Bao et al. 2015) (v.20.11, August 2015) is unsuitable

for detailed repeat element analyses of most squamate reptile genomes, we performed *de novo* identification of repeat elements on 6 snake genomes (*Crotalus viridis*, *Crotalus mitchellii*, *Thamnophis sirtalis*, *Boa constrictor*, *Deinagkistrodon acutus*, and *Pantherophis guttatus*) in RepeatModeler v.1.0.9 (Smit and Hubley 2015) using default parameters. Consensus repeat sequences from multiple species were combined into a large joint snake repeat library that also includes previously identified elements from an additional 12 snake species (Castoe et al. 2013). All genomes were annotated with the same library with the exception of the green anole lizard, for which we used a lizard specific library that includes *de novo* repeat identification for *Pogona vitticeps*, *Ophisaurus gracilis*, and *Gekko japonicus*. To verify that only repeat elements were included in the custom reference library, all sequences were used as input in a BLASTx search against the SwissProt database (UniProt 2017), and those clearly annotated as protein domains were removed. Finally, redundancy and possible chimeric artifacts were removed through clustering methods in CD-HIT (Li and Godzik 2006) using a threshold of 0.85.

Homology-based repeat element annotation was performed in RepeatMasker v.4.0.6 (Smit et al. 2015) using a PCR-validated BovB/CR1 LINE retrotransposon consensus library (Castoe et al. 2013), the Tetrapoda RepBase library, and our custom library as references. Output files were post-processed using a modified implementation of the ProcessRepeat script (RepeatMasker package).

Gene Annotation

We used MAKER v. 2.31.8 (Cantarel et al. 2008) to annotate protein-coding genes in an iterative fashion. Several sources of empirical evidence of protein-coding genes were used, including the full *de novo C. viridis* transcriptome assembly and protein datasets consisting of all annotated proteins from NCBI for *Anolis carolinensis* (Alfoldi et al. 2011), *Python molurus bivittatus* (Castoe et al. 2013), *Thamnophis sirtalis* (Perry et al. In Review), and *Ophiophagus hannah* (Vonk et al. 2013), and from GigaDB for *Deinagkistrodon acutus* (Yin et al. 2016). We also included 422 protein sequences for 24 known venom gene families that were used to infer *Python* venom gene homologs in a previous study (Reyes-Velasco et al. 2015). Prior to running MAKER, we used BUSCO v. 2.0.1 (Simão et al. 2015) and the full *C. viridis* genome assembly to iterative train AUGUSTUS v. 3.2.3 (Stanke and Morgenstern 2005) HMM models based on 3,950 tetrapod vertebrate benchmarking universal single-copy orthologs

(BUSCOs). We ran BUSCO in the 'genome' mode and specified the '--long' option to have BUSCO perform internal AUGUSTUS training. We ran MAKER with the 'est2genome=0' and 'protein2genome=0' options set to produce gene models using the AUGUSTUS gene predictions with hints supplied from the empirical transcript and protein sequence evidence. We provided the coordinates for all interspersed, complex repetitive elements for MAKER to perform hard masking before evidence mapping and prediction, and we set the 'model_org' option to 'simple' to have MAKER soft mask simple repetitive elements. We used default settings for all other options, except 'max_dna_len' (set to 300,000) and 'split_hit' (set to 20,000). We iterated this approach an additional time and we manually compared the MAKER gene models with the transcript and protein evidence. We found very little difference between the two gene annotations and based on a slightly better annotation edit distance (AED) distribution in the first round of MAKER, we used our initial round as the final gene annotation. The resulting annotation consisted of 17,486 genes and we ascribed gene IDs based on homology using reciprocal best-blast (with e-value thresholds of 1e-5) and stringent one-way blast (with an e-value threshold of 1e-8) searches against protein sequences from NCBI for *Anolis*, *Python*, and *Thamnophis*.

Hi-C and RNA Sequencing of the Venom Gland

We dissected the venom glands from the Hi-C *Crotalus viridis viridis* 1 day and 3 days after venom was initially extracted in order to track a time-series of venom production. A subsample of the 1-day venom gland was sent to Dovetail Genomics where DNA was extracted and replicate Hi-C sequencing libraries were prepared according to their protocol (see above). We also extracted total RNA from both 1-day and 3-day venom gland samples, as well as tongue and pancreas tissue from the Hi-C genome animal (see Sequencing and Assembly and Annotation sections above). mRNAseq libraries were generated and sequenced at Novogene on two separate lanes of the Illumina HiSeq 4000 platform using 150 bp paired-end reads (Table 6).

Chromosome Identification and Synteny Analyses

Genome assembly resulted in several large, highly-contiguous scaffolds with a relative size distribution consistent with the karyotype for *C. viridis* (Baker et al. 1972), representing nearly-complete

chromosome sequences. We determined the identity of chromosomes using a BLAST search of the chromosome-specific markers linked to snake chromosomes from (Matsubara et al. 2006), downloaded from NCBI (accessions SAMN00177542 and SAMN00152474). We kept the best alignment per cDNA marker as its genomic location in the Prairie Rattlesnake genome, except when a marker hit two high-similarity matches on different chromosomes. The vast majority of markers linked to a specific macrochromosome (i.e., chromosomes 1-7; Table 4) in *Elaphe quadrivirgata* mapped to a single genomic scaffold; only 6 of 104 markers did not map to the predicted chromosome from *E. quadrivirgata*. All snake microchromosome markers mapped to a single 139Mb scaffold, which was later broken into 10 microchromosome scaffolds (scaffold-mi1-10; see below).

We identified a single 114Mb scaffold corresponding to the Z chromosome, as 10 of 11 Z-linked markers mapped to this scaffold. To further vet this as the Z-linked region of the genome, we mapped reads from male and female *C. viridis* (Table 7) to the genome using BWA (Li and Durbin 2009) using program defaults. Male and female resequencing libraries were prepared using an Illumina Nextera prep kit and sequenced on an Illumina HiSeq 2500 using 250bp paired-end reads. Adapters were trimmed and low-quality reads were filtered using Trimmomatic (Bolger et al. 2014). After mapping, we filtered reads with low mapping scores and quantified per-base read depths using SAMtools (Li et al. 2009). We then binned read depths into 100Kb windows and normalized female and male windowed-coverage by calculating the log2(female/male) ratio. Here, the expectation is that a hemizygous locus will show roughly half the normalized coverage, which we observe for females over the majority of the Z chromosome scaffold length, and not elsewhere in the genome (Fig. 9). To demonstrate Z chromosome conservation among pit vipers and to further determine the identity of this scaffold, we mapped male and female pygmy rattlesnake (*Sistrurus catenatus*) reads from (Vicoso et al. 2013) to the genome using the same parameters detailed above (Fig. 9). *Anolis* chromosome 6 is homologous with snake sex chromsomes (Srikulnath et al. 2009), thus we aligned *Anolis* Chromosome 6 (Alfoldi et al. 2011) to the Prairie Rattlesnake genome using BLASTn. As expected, we found a large quantity of high-similarity hits to the rattlesnake Z chromosome scaffold, specifically, which were organized in a sequential manner across the entire Z scaffold (Figs. 1b, 2a). In Fig. 2a, ‘high-stringency’ hits refer to alignments with e-

values < $1e^{-250}$ and bit-scores > 200. The gray circles, which appear as a solid line due to their density, correspond with hits with an e-value < 0.00001.

We used multiple sources of information to identify the best candidate breakpoints between microchromosomes within the 139Mb fused microchromosome scaffold in the initial Hi-C assembly. Because reptile microchromosomes are highly syntenic (Alfoldi et al. 2011), we aligned the microchromosome scaffold to microchromosome scaffolds from chicken (Hillier et al. 2004) and *Anolis* using LASTZ (Harris 2007) to determine likely chromosomal breakpoints. To retain only highly similar alignments per comparison, we set the 'hspthresh' option equal to 10,000 (default is 3,000). We also set a step size equal to 20 to reduce computational time per comparison. This approach delineated candidate boundaries between rattlesnake microchromosomes based on clear breaks in cross-species synteny (Fig. 3d). We further validated candidate break points using genomic features that consistently vary at the ends of chromosomes. Here, we specifically evaluated if candidate breakpoints exhibited regional shifts in GC content and repeat content, similar to the ends of macrochromosomes (Fig. 1). For each candidate breakpoint, we then determined if there was a junction between two Chicago assembly scaffolds (i.e., two contiguous pieces of sequence that were not assembled using Hi-C) within the breakpoint region. Finally, if no annotated genes spanned this junction, we considered it biologically plausible. There were nine candidate breakpoints that met each of these criteria, equaling the number of boundaries expected given ten microchromosomes (Fig. 3d). Importantly, this approach assumes that the ten microchromosomes were assembled in a contiguous fashion per chromosome. Intrachromosomal chromatin contacts are far more frequent than contacts between chromosomes (Lieberman-Aiden et al. 2009). The ten candidate microchromosomes match this expectation, and show clear signal of consistent intrachromosomal contact frequencies across their entire length (the same as macrochromosomes; Fig. 14).

To explore broad-scale structural evolution across reptiles, we used the rattlesnake genome to perform in silico painting of the chicken (*Gallus gallus* version 5) and *Anolis carolinensis* (version 2) genomes. Briefly, we divided the rattlesnake genome into 2.02 million potential 100 bp markers. For each of these markers, we used BLAST to record the single best hit in the target genome requiring an alignment length of at least 50 bp. This resulted in 41,644 potential markers in *Gallus* and 103,801 potential markers in *Anolis*. We then processed markers on each chromosome by requiring at least five

consecutive markers supporting homology to the same rattlesnake chromosome. We consolidated each group of five consecutive potential markers as one confirmed marker. In *Gallus*, we rejected 12.4% of potential markers and identified 7,291 confirmed merged markers. In *Anolis*, we rejected 39.7% of potential markers and identified 12,511 confirmed merged markers.

This approach demonstrates considerable stability at the chromosomal level despite 158 million years of divergence between *Anolis* and *Crotalus* (Fig. 1b), and between squamates and birds, despite 280 million years of divergence between squamates and *Gallus*. This stability is evident not only in the macrochromosomes but also in the microchromosomes. In fact, 7 of 10 *Crotalus* microchromosomes had greater than 80% of confirmed markers associated with a single chromosome in the chicken genome (Fig. 1b, microchromosome inset). Comparisons among the three genomes suggest that the *Crotalus* genome has not experienced some of the fusions found in *Anolis*. Specifically, we infer that *Anolis* chromosome 3 is a fusion of *Crotalus* chromosome 4 and 5. Likewise, *Anolis* chromosome 4 is a fusion of *Crotalus* chromosome 6 and 7. Divergence time estimates discussed above and shown in Fig. 1b were taken from the median of estimates for divergence between *Crotalus* and *Gallus* and between *Crotalus* and *Anolis* from Timetree (www.timetree.org; (Kumar et al. 2017).

Genomic Patterns of GC Content

We quantified GC content in sliding windows of 100Kb and 1Mb across the genome using a custom Python script (https://github.com/drewschield/Comparative-Genomics-Tools/blob/master/slidingwindow_gc_content.py). GC content in 100Kb windows is presented in Fig. 1.

To determine if there is regional variation in nucleotide composition consistent with isochore structures across the rattlesnake genome, we quantified GC content and its variance within 5, 10, 20, 40, 80, 160, 240, and 320-kb windows. The variation (standard deviation) in GC content is expected to decrease by half as window size increases four-fold if the genome is homogeneous (i.e., lacks isochore structures; (Consortium 2001). By comparing the observed variances of GC content across spatial window scales to those from 11 other squamate genomes, including lizards (*Anolis* has been shown to lack isochore structure (Alfoldi et al. 2011), henophidian snakes, and colubroid snakes, we were able to determine the relative heterogeneity of nucleotide composition in the rattlesnake (Table 8). To reduce

potential biases from estimates from small scaffold sizes, we filtered to only retain scaffolds greater than the size of the window analyzed (e.g., only scaffolds longer than 10 Kb when looking at the standard deviation in GC content over 10 Kb windows) and for which more there was less than 20% of missing data for all analyzed genomes.

To explore trajectories of GC content evolution among squamates, we generated whole genome alignments for the species in Table 8 using Multi-Z (Blanchette et al. 2004), using program defaults. We then filtered the multi-species whole genome alignment to retain only blocks for which information for all 12 species was available, and concatenated blocks according to their organization in the *Anolis* lizard genome. We then calculated GC content within consecutive 50 Kb windows of this concatenated alignment using the 'slidingwindow_gc_content.py' script detailed above.

Hi-C analysis

Raw Illumina paired-end reads were processed using the Juicer pipeline (Durand et al. 2016) to produce Hi-C maps binned at multiple resolutions, as low as 5kb resolution, and for the annotation of contact domains. These data were aligned against the CroVir3.0 assembly. All contact matrices used for further analysis were KR-normalized in Juicer. We identified topologically-associated chromatin domains (TADs) using the Hi-C Explorer 'hicFindTADs' function (Ramírez et al. 2018), using default parameters and specifying a Bonferroni correction for multiple comparisons.

We compared intra and interchromosomal contact frequencies between the rattlesnake venom gland and various tissues from mammals. To do this we quantified the total intra- and interchromosomal contacts between chromosome positions from the rattlesnake and the following Hi-C datasets: human lymphoblastoma cells (Rao et al. 2014) and human retinal ephithelial cells, mouse kidney, and Rhesus monkey tissue (Darrow et al. 2016). To investigate patterns of intra- and interchromosome contact frequency, we normalized contact frequencies by chromosome length. In the case of the mouse, we removed the Y chromosome due to its small size and relative lack of interchromosomal contacts. We then performed linear regressions of chromosome length and normalized intra- and interchromosomal contact frequencies (i.e., contact frequency/chromosome length). In all cases we observed a positive relationship

between normalized intrachromosomal contacts and chromosome size and a negative relationship between normalized interchromosomal contacts and chromosome size (Fig. 3b).

Sex Chromosome Analysis

We identified the Prairie Rattlesnake Z chromosome using methods described in section 1.X above. We localized the candidate pseudoautosomal region (PAR) based on normalized female/male coverage (Fig. 2a; the PAR is the only consistent region of the Z with equal female and male coverage. We quantified gene content, GC content, and repeat content across the Z chromosome and PAR (Supplementary Figs. 10, 11, and 12), and tested for gene enrichment in the PAR using a Fisher's exact test, where we compared the number of genes within each region to the total length of the region.

To compare within individual nucleotide diversity across the genome between male and female *C. viridis*, we called variants (i.e., heterozygous sites) from the male and female reads used in coverage analysis detailed above. With the mappings from coverage analysis, we used SAMtools (Li et al. 2009) to compile all mappings into pileup format, from which we called variant sites using BCFtools. We filtered sites to retain only biallelic variants using VCFtools (Danecek et al. 2011) and calculated the proportion of heterozygous sites (i.e., within-individual nucleotide diversity) using a custom pipeline of scripts. First, calcHet (https://github.com/darencard/RADpipe) outputs details of heterozygous site and window_heterozygosity.py (https://github.com/drewschield/Comparative-Genomics-Tools/blob/master/window_heterozygosity.py) uses this output in conjunction with a windowed .bed file generated using BEDtools 'make_windows' tool to calculate the proportion of heterozygosity within a given window size.

Evolutionary patterns of the Z chromosome were also analyzed by examining transposable element age and composition along the whole chromosome, and across the three inferred evolutionary strata (see Main Text). Since the length of the PAR is significantly smaller than the combined length of Strata 1 and 2, to rule out potential biases due to unequal sample size we also independently analyzed fragments of the other strata with lengths equal to the PAR (total of 15 7.18 Mbp fragments). Each region was analyzed in RepeatMasker using a single reference library that included the squamate fraction of the RepBase Tetrapoda library, and the snake specific library clustered at a threshold of 0.75. The age

distribution of TE families was estimated by mean of the Kimura 2-parameter distance from the consensus sequence per element (CpG corrected) calculated from PostProcessed.align outputs (see section 1.X above). We then merged estimates of repeat content from each of these regions for comparison to the PAR region, specifically.

To quantify gene expression on the rattlesnake Z chromosome and across the genome, we prepared RNAseq libraries from liver and kidney tissue from two males and females and sequenced them on an Illumina HiSeq using 100bp paired-end reads (Table 6). Samples and libraries were prepared following the methods of (Andrew et al. 2017). After filtering and adapter trimming using Trimmomatic v. 0.32 (Bolger et al. 2014), we mapped RNAseq reads to the *C. viridis* genome using STAR v. 2.5.2b (Dobin et al. 2013) and counts were determined using featureCounts (Liao et al. 2013). We normalized read counts across tissues and samples using TMM normalization in edgeR (Robinson et al. 2010) to generate both counts per million (CPM) for use in pairwise comparisons between males and females, and fragments per kilobase million (FPKM) normalized counts for comparisons of chromosome-wide expression within samples. We tested for differential gene expression between males and females using pairwise exact tests in edgeR followed by independent hypothesis weighting (IHW) p-value correction (Ignatiadis et al. 2016) and quantified normalized gene expression across the Z chromosome in 100Kb windows, based on the location of each gene in the genome annotation. Per gene female-to-male ratios of normalized expression were generated by dividing the average female expression level by that of the male, only including genes with expression information in both the male and female (>1 avg. FPKM in each sex). Two-sided student's t-tests in R were used to compare of median female-to-male ratios between chromosomes and/or chromosomal regions (i.e. the PAR). To explore regional variation in dosage across the Z chromosome, we performed a sliding window analysis of the F/M log2 normalized expression ratio with a window size of 30 genes and a step size of 1 gene.

A possible mechanism for upregulation of certain Z-linked genes in females is regulation through estrogen response elements (EREs), which can enable binding of enhancers and promote transcription of genes over long distances (Lin et al. 2007). Rice et al. (2017) identified that the binding domain of *ESR1* is completely conserved among humans, chickens, and alligators, thus we used the *ESR1* binding motif of humans ('GGTCAnnnTGACC'; (Lin et al. 2007) and a regular expression motif finding script

(https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder) to predict *ESR1* binding motifs (ER motif) throughout the rattlesnake genome. Using BEDtools 'closest' function (Quinlan and Hall 2010), we calculated the distance from each gene to the nearest predicted ER motif. We considered a gene to be a candidate for ERE-based upregulation if it was within 100Kb of a predicted ERE. We calculated the number of genes with evidence of partial dosage (i.e., genes with a F/M expression ratio greater than the lower bound of the autosomal 95% quantile), and used a Fisher's Exact test to determine if 'dosed' genes were enriched for proximity to EREs, which was not significant.

Comparative Microchromosome Genomics

To understand evolutionary shifts in microchromosome composition among amniotes, we compared measures of gene density, GC content, and repeat content of macro- and microchromosomes between the rattlesnake, anole (Alfoldi et al. 2011), bearded dragon (Georges et al. 2015; Deakin et al. 2016), chicken (Hillier et al. 2004), and zebra finch (Warren et al. 2010) genomes. These species were chosen because their scaffolds are ordered into chromosomes and because their karyotypes contain microchromosomes. For each genome, we quantified the total number of genes per chromosome, total number of G+C bases, and total bases masked as repeats in RepeatMasker. We then normalized each measure by the total length of macrochromosome and microchromosome sequences in each genome, then calculated the ratio of microchromosome:macrochromosome proportions. We then used Fisher's Exact Tests determine if one chromosome set possessed a significantly greater proportion of each measure. We generated a phylogenetic tree (Fig. 7) for the five species based on divergence time estimates from TimeTree (Kumar et al. 2017), and plotted the ratio values calculated above onto the tree tips for between-species comparisons.

Venom Gene Annotation and Analysis

We took a multi-step approach toward identifying venom gene homologs in the rattlesnake genome. We first obtained representative gene sequences for 38 venom gene families from Genbank (Table 9), comprising known enzymatic and toxin components of snake venoms. We then searched our transcript set using the venom gene family query set using a tBLASTx search, defining a similarity cutoff

e-value of $1x10^{-5}$. For each candidate venom gene transcript identified in this way, we then performed a secondary tBLASTx search against the NCBI database to confirm its identity as a venom gene. In the case of several venom gene families, such as those known only from elapid snake venom, we did not find any candidate genes. Three venom gene families that are especially abundant, both in terms of presence in the venom proteome (Fig. 4a) and in copy number, in the venom of *C. viridis* are phospholipases A2 (PLA2s), snake venom metalloproteinases (SVMPs), and snake venom serine proteases (SVSPs). Rattlesnakes possess multiple members of each of these families (Mackessy 2008; Casewell et al. 2011; Dowell et al. 2016), and the steps taken above appeared to underestimate the total number of copies in the *C. viridis* genome. Therefore, for each of these families, we performed an empirical annotation using the FGENESH+ (Solovyev et al. 2006) protein similarity search. We first extracted the genomic region annotated for each of these families above plus and minus a 100 Kb flanking region. We used protein sequences from Uniprot (PLA2: APD70899.1; SVMP: Q90282.1; and SVSP: F8S114.1) to query the region and confirm the total number of copies per family. Each gene annotated in this way was again searched against NCBI to confirm its identity and manual searches of aligned protein sequences (see phylogenetic analyses below) further confirmed their homology to each respective venom gene family. Genomic locations and details of annotated venom genes in the rattlesnake genome are provided in Table 10.

We used LASTZ (Harris 2007) to align the genomic regions containing PLA2, SVMP, and SVSP genes to themselves. We used program defaults, with the exception of the 'hspthresh' command, which we set to 8,000. This was done to only return very high similarity matches between compared sequences. Here the expectation is that when alignments are plotted against one another, we will observe a diagonal line demonstrating perfect matches between each stretch of sequence and itself. In the case of segmental duplications, we also expect to see parallel and perpendicular (if in reverse orientation) segments adjacent to the diagonal 'self' axis. We plotted LASTZ results for each of the regions using the base plotting function in R (R Core Team 2017).

We then performed Bayesian phylogenetic analyses to further evaluate evidence of tandem duplication and monophyly among members of the PLA2, SVMP, and SVSP venom gene families. We generated protein alignments of venom genes with their closest homologs using MUSCLE (Edgar 2004)

with default parameters, with minor manual edits to the alignment to remove any poorly aligned regions. We analyzed the protein alignments using BEAST2 (Bouckaert et al. 2014), setting the site model to 'WAG' for each analysis. We ran each analysis for a minimum of 1 x 10^8 generations, and evaluated whether runs had reached stationarity using Tracer (Drummond and Rambaut 2007). After discarding the first 10% of samples as burnin, we generated consensus maximum clade credibility trees using TreeAnnotator (distributed with BEAST2).

Raw Illumina RNAseq reads (Table 6) were quality trimmed using Trimmomatic v. 0.36 (Bolger et al. 2014) with default settings. We used STAR (Dobin et al. 2013) to align reads to the genome. Raw expression counts were estimated by counting the number of reads that mapped uniquely to a particular annotated transcript using HTSeq-count (Anders et al. 2013). These raw counts were then normalized and filtered in edgeR using TMM normalization (Oshlack et al. 2010; Robinson et al. 2010), and all subsequent analyses were done using these normalized data. We used two-sided student's t-tests in R to compare gene expression between venom gland samples and body tissues to test for evidence of genes exhibiting a significantly upregulated signature of expression in the venom gland, specifically.

To identify candidate transcription factors regulating venom gene expression, we searched the genome annotation for all genes included on the UniProt (http://www.uniprot.org) reviewed human transcription factor database, by specifying species = 'Homo sapiens' and reviewed = 'yes' in the advanced search terms. Using this list, we parsed the rattlesnake genome for all matching gene IDs and compared their expression across rattlesnake tissues. We then identified likely candidate venom gland transcription factors, which showed a pattern of overall low body-wide expression and statistically significant evidence of higher expression in the venom gland, specifically. We found 13 candidates using this approach, including four members of the *CTF*/*NFI* family of RNA polymerase II core promoter-binding transcription factors (*NFIA,* two isoforms of *NFIB*, and *NFIX*). *NFI* binding sites have been identified upstream of venom genes in several venomous snake taxa, including viperids, elapids, and colubrids (e.g., crotamine/myotoxin in *Crotalus durissus* (Rádis-Baptista et al. 2003) and three finger toxins in *Naja sputatrix* (Lachumanan et al. 1998) and *Boiga dendrophila* (Pawlak and Kini 2008). *NFI* family members were also found to be expressed in the venom glands of several species in a previous study exploring

putative venom gland transcription factors (Hargreaves et al. 2014), but information about whether they showed venom gland-specific expression was not provided.

Because four transcription factors of the *NFI* family each showed evidence of venom gland-specificity, we tested the hypothesis that their binding motifs are also upstream of venom genes more than they are other genes. We obtained the TRANSFAC position weight matrix for each transcription factor from the CIS-BP database (Weirauch et al. 2014), scanned a 1 Kb region upstream of each gene in the snake venom PLA2, SVMP, and SVSP gene families for predicted transcription factor binding sites per upstream region using PoSSuM Search (Beckstette et al. 2006), setting a p-value cutoff of $1x10^{-6}$ for each search. We then performed the same analysis on 1 Kb regions of the closest related non-venom homologs per venom gene family, as well as the 1 Kb upstream regions of five independent random samples of 100 genes per sample. For each analyzed set of upstream regions, we performed a Fisher's Exact test of significant enrichment upstream of venom genes by comparing 1) the number of predicted binding motifs divided by the number of upstream regions, 2) the number of predicted binding motifs divided by the total combined length of upstream regions, and 3) the total length of predicted binding motifs divided by the total combined length of upstream regions.

Results and Discussion

A chromosome-level rattlesnake genome

We sequenced and assembled the genome of a male Prairie Rattlesnake (*Crotalus viridis viridis*) at 1,658-fold physical coverage using multiple high-throughput sequencing approaches combined with the Dovetail Genomics HiRise sequencing and assembly method (Tables 1 and 2), which combines long-range Chicago data (Rice et al. 2017) with 3D chromatin contact information from Hi-C. This approach resulted in the most contiguous reptile genome to date (*CroVir3.0*), with a scaffold N50 of 179.9 Mbp, represented by 3 scaffolds. We estimate the total genome size to be between 1.25 and 1.34 Gbp based on k-mer frequency distributions and the assembled genome size, respectively. The genome annotation contains 17,352 predicted protein-coding genes, and an annotated repeat element content of 39.49% (Table 3).

The rattlesnake is the first vertebrate genome to achieve chromosome-level assembly using only next-generation sequencing technology, and the first ever snake chromosome-level assembly, including all chromosomes in the rattlesnake karyotype (2n = 36). Chromosome identities of large scaffolds were further confirmed using chromosome-specific gene markers (Matsubara et al. 2006), which mapped uniquely to large scaffolds corresponding to macrochromosomes (Chromosomes 1 through 7; Table 4). Microchromosomes were originally over-assembled into a single large scaffold, which was manually split based on multiple lines of evidence. The corrected assembly resulted in microchromosome scaffolds with lengths matching the size predictions of the rattlesnake karyotype (Baker et al. 1972). Finally, we identified the rattlesnake Z chromosome using multiple lines of evidence, which we discuss further below. The chromosomal sequences include assembled telomeric and centromeric regions, with centromeres containing an abundant 164 bp monomer with 42% GC content (Fig. 6).

This chromosome-level assembly provides new insight into the genome-wide distribution of key features and homology across amniote genomes. In the rattlesnake, we find the microchromosomes to contain the highest and most variable GC content. We also find rattlesnake microchromosomes have particularly high gene density (100 Kb windows; p-values < 0.00001) and reduced repeat element content compared to macrochromosomes (p < 0.00001; Fig. 1a), similar to patterns observed in the Chicken (Fig. 7). Rattlesnake chromosomes show high degrees of synteny with those from *Anolis*, except for an apparent fusion/separation of *Anolis* chromosome 3 into snake chromosomes 4 and 5 (Fig. 1b). Microchromosomes also appear largely homologous across squamates (although this is limited by resolution of microchromosome linkage groups in *Anolis*). We were also able to further validate previous inferences that *Anolis* chromosome 6 is homologous to the sex chromosomes of rattlesnakes. Despite conservation of squamate microchromosome homology, patterns of chicken-squamate homology suggest that there have been major shifts between macro- and microchromosome locations for large syntenic regions of the genome. The chicken has a large number of microchromosomes, and we find that about half of these are syntenic with squamate microchromosomes (Fig. 1b), but that other microchromosomes are syntenic with blocks of squamate macrochromosomes (i.e., squamate chromosome 2 is an amalgam of chicken microchromosomes and the Z chromosome). We also observe large regions of synteny between chicken and squamate macrochromosomes. For example, squamate chromosome 1 shares

large syntenic tracks with chicken chromosomes 3, 5, and 7 (Fig. 1b). Surprisingly, the largest chicken macrochromosomes (1 and 2) show synteny patterns that are scattered across multiple squamate macrochromosomes, including the rattlesnake Z chromosome, indicating multiple exchanges of genomic regions between macro- and microchromosomes early in amniote evolution.

Squamate reptiles have become particularly important for studying the evolution of genomic GC content and isochore structure, due to the loss of GC isochores in *Anolis* yet the apparent re-emurgence of isochore structure in snakes (Fujita et al. 2011; Castoe et al. 2013). To visualize genomic GC variation, we compared orthologous aligned genomic regions across 12 squamates, which demonstrates that there have been two major transitions in genomic GC content, including a reduction in GC content from lizards to snakes, and a secondary further reduction in GC content within the colubroid lineage of snakes that includes the rattlesnake and cobra (Fig. 1c). This suggests that higher genome-wide GC content was likely the ancestral squamate condition, and that snakes have evolved increased GC variation through an increase in genomic AT content (i.e., AT isochores), rather than a buildup of GC-rich islands, as was suggested by the finding of AT-biased substitutions from lizard to python and cobra genomes (Castoe et al. 2013). Interestingly, the negative relationship between genomic GC content (Fig. 1c) and GC isochore structure across squamate evolution (Fig. 1d; Table 8) further suggests that GC-biased gene conversion cannot explain GC variation in snakes – a finding that has broad ramifications for understanding the mechanisms underlying shifts in genomic nucleotide content and variation, and the spatial structure of this variation. In addition to genomic GC content variation, squamate reptiles are notable because they appear to have remarkably variable, active, and rapidly-evolving genomic repeat element content across lineages, which is surprising given the relatively small and conserved genome size of squamates (Castoe et al. 2011; Pasquesi et al. In Review), and our analyses here confirm these trends (Fig. 1e). We find the genomes of colubroid snakes are dominated largely by several DNA elements (e.g., hAT and Tc1) and by non-LTR retrotransposons, and CR1-L3 LINEs in particular. The rattlesnake genome, specifically, has the highest abundance of CR1-L3s among sampled colubroid genomes (Fig. 1e), and low divergence of the majority of rattlesnake CR1-L3s suggests that these elements are quite active in the genome (Fig. 8).

Sex chromosome evolution mechanisms of dosage compensation

The contiguity of the rattlesnake genome facilitates new perspectives into the structure of the pseudoautosomal region (PAR) and evolutionary strata of snake sex chromosomes, as well as new information on patterns of dosage compensation in snakes that provide new parallels across amniotes for understanding sex chromosome evolution. Recent studies have shown that snake sex chromosomes have evolved multiple times, apparently from different autosomal chromosomes (Gamble et al. 2017), and have suggested that colubroid Z/W chromosomes are homologous with *Anolis* chromosome 6 (Srikulnath et al. 2009; Vicoso et al. 2013). We identified a single 114 Mb scaffold as the rattlesnake Z chromosome, which was confirmed by its broad synteny with *Anolis* chromosome 6, the presence of multiple known Z-linked markers ((Matsubara et al. 2006); Table 4), and with coverage of mapped genomic reads that match expectations of hemizosity based on additional genomic data we collected from female individuals (Matsubara et al. 2006); Fig. 2a; Fig. 9). We further identified the Z/W recombining PAR as the distal 7.2 Mb region of the Z chromosome that shows equal male-female genomic read depth (Fig. 2a) – this rattlesnake PAR is GC-rich relative to the genomic background and the non-PAR Z-chromosome regions (42.9%; Fig. 10), similar to the pattern observed in the PAR of the Collard Flycatcher (Smeds et al. 2014). This suggests that common processes (e.g., GC-biased gene conversion) may drive increased GC content in the recombining regions of the independently evolved snake and avian sex chromosomes. The rattlesnake PAR also exhibits distinctive patterns of repeat element content compared to the Z, with lower levels of divergence among particular repeat elements in the PAR (e.g., CR1 and Bov-B LINEs), suggesting more recent element activity and insertion (Fig. 11). We also find higher gene density in the rattlesnake PAR than elsewhere in the Z chromosome (Fisher's Exact Test: $p = 4.46 \times 10^{-7}$; Fig. 12).

The existence of evolutionary strata on snake sex chromosomes has been suggested (Vicoso et al. 2013; Yin et al. 2016), but prior analyses have lacked the important context of a contiguous Z chromosome assembly. In addition to the PAR (i.e., Stratum 3), we identified a secondary evolutionary stratum situated between the PAR and the remaining Z chromosome. This region (Stratum 2) shows near-autosomal levels of female-male ratios of mapped genomic reads (Fig. 2a,b). We hypothesize based on its location between the PAR and the oldest recombination-suppressed region (Stratum 1), combined with observed intermediate female:male read depth, that Stratum 2 represents a recombination-suppressed region that has retained substantial homology between Z and W chromosomes (Fig. 2b).

Consistent with this hypothesis, a comparison of within-individual nucleotide diversity between females and males revealed elevated diversity in the female across Stratum 2, likely explained by the mapping of reads to divergent Z and W-linked paralogs in Stratum 2 in females (Fig. 2a,b). This suggests that a number of W-linked gene copies have been retained over the course of W chromosome degeneration and divergence from the Z chromosome, as has been hypothesized for birds (Bellott et al. 2017). The oldest evolutionary stratum (Stratum 1) is characterized by half female coverage relative to that of males, and roughly zero female nucleotide diversity, consistent with female hemizogosity across Stratum 1 (Fig. 2b). Based on our Hi-C data, we also find that the evolutionary strata on the Z chromosome broadly coincide with the boundaries of inferred topologically-associated domains (TADs; Fig. 2a), which provides the first precise demonstration that chromatin organization co-evolves with recombination suppression and sex chromosome differentiation.

Dosage compensation in organisms with differentiated sex chromosomes is of broad interest, especially due to the surprising diversity of mechanisms by which dosage is accomplished (Graves 2016). Colubroid snakes have been shown to exhibit partial dosage compensation (Vicoso et al. 2013; Yin et al. 2016), yet no mechanisms for compensation have been proposed. The absence of complete dosage compensation is also supported by our data, which demonstrate that the overall ratio of female:male gene expression is significantly lower on the Z-chromosome compared to that of autosomes ($p < 10^{-16}$; Fig. 2c; Fig. 13). Intriguingly, we identified patterns of partial or incomplete dosage compensation that varied widely across regions of the Z-chromosome, ranging from a total lack of compensation to equal expression in females and males (Fig. 2a and c), further raising the question of what mechanisms drive such variation.

To address mechanisms that might underlie partial compensation, we analyzed gene expression data from males and females for two different tissues (liver and kidney) in a stratum-specific fashion. In Stratum 3, we find that gene expression ratios between sexes largely match those on autosomes for both tissues (Fig. 2c; liver $p = 0.366$, kidney $p = 0.453$). This further confirms the identification of this region as the PAR, where compensation is achieved by the Z and W being homologous and effectively autosomal. In Stratum 2, in addition to intermediate female:male genomic read coverage, we find evidence for intermediate dosage (Fig. 2c), consistent with partial dosage compensation in females due to what we

hypothesize represents effective diploidy through retained W-linked Z-chromosome homologs. Indeed, 24.5% of genes with female:male expression ratios greater than the 5th quantile of autosomal female:male ratio are within this region; these genes combined with genes in the PAR constitute 46.9% of dosed genes on the Z. Therefore, our results suggest that a substantial proportion of 'dosage' is driven by effective diploidy in females for genes in Strata 2-3. Finally, Stratum 1 showed the most variation in dosage, ranging from nearly complete to absent. We tested for evidence of a female-biased transcriptional regulatory mechanism (estrogen response elements; EREs) that could explain regional or gene-specific compensation, and find that this mechanism may only account (at best) for a small number (8.5%) of dosed Stratum 1 genes, which we estimated were linked (i.e., within 100 Kb) to a predicted ERE (Fig. 2a). These findings suggest that additional unidentified dosage compensation mechanisms likely exist in snakes, which may include post-transcriptional mechanisms, as have been implicated in partial chicken dosage compensation (Uebbing et al. 2015).

Hi-C reveals unique microchromosome biology

Our analyses of the first available 3D chromatin contacts for a non-mammalian vertebrate (Fig. 3a) provide new perspectives on high-order genome organization and contact structure in reptiles and unique features of microchromosome biology. Patterns of intra- and interchromosomal chromatin contacts across rattlesnake macrochromosomes are broadly consistent with patterns observed in mammals, such that when interchromosomal contact frequencies are normalized by chromosome length, they show a consistent negative linear relationship across species (Fig. 3b). However, rattlesnake microchromosomes show a much steeper negative slope, deviating significantly from expectations based on macrochromosome contact frequencies. These data indicate an unexpected higher degree of contact between microchromosomes and other chromosomes (Fig. 3a), and a surprisngly high degree of interchromosomal contact among microchromosomes (Fig. 3c). In fact, the initial misassembly of microchromosomes into a single scaffold was likely driven by unexpected high frequencies of contact among microchromosomes, which significantly exceed assumptions of genome assembly based on mammalian macrochromosomes ($t = 13.38$, $p < 2.2 \times 10\text{-}16$, Fig. 3d) – this assembly error was later corrected using complementary information from chromatin contact frequencies, reptile

microchromosome synteny, and patterns of GC and repeat content. Importantly, this first demonstration of Hi-C assembly of microchromosomes indicates that similar steps may need to be taken in future Hi-C sequencing and assembly projects for organisms with microchromosomes, and highlights the uniqueness of microchromosome interactions within the nucleus of at least snakes, if not other amniotes.

Microchromosomes are present in most birds and reptiles, but tend to be poorly represented and characterized in existing assembled genomes. Further, much of what we understand about microchromosome biology comes from studies of birds, and limited comparisons with other species (e.g., *Anolis* (Alfoldi et al. 2011) and *Pogona* (Georges et al. 2015) lizards) suggest that genomic features of microchromosomes may differ among species, despite the existence of considerable reptile microchromosome synteny (Fig. 1b). A comparison of compositional features between micro- and macrochromosomes of other species suggests that the rattlesnake exhibits patterns remarkably similar to chicken and zebra finch (i.e., significantly higher GC and gene content and lower repeat content on microchromosomes than on macrochromosomes), with the exception of higher repeat content in zebra finch microchromosomes (Fig. 7). Lizards are more variable, with lower gene density in microchromosomes than rattlesnake and the birds (microchromosome gene density in *Pogona* is lower than in macrochromosomes), however, consistencies among species possessing microchromosomes suggest that ancestral amniote microchromosomes likely exhibited patterns similar to those observed in both the rattlesnake and chicken (Fig. 1a, Fig. 3a-c, Fig. 7).

Insight into the origins, evolution, and regulation of snake venom and its production

Snake venoms and venom systems are intriguing examples for studying the evolution of biological novelty and represent topics of intense study and medical relevance (Mackessy 2010; Arnold 2016). The rattlesnake genome provides the first clear insight into the genomic location, organization, and broader genomic context for snake venom gene family evolution (Fig 4a). Our localization of rattlesnake venom genes to chromosomes revealed that venom gene families are enriched for being located on microchromosomes ($p = 0.0017$). Moreover, microchromosome-linked families include three of the most abundant, well-characterized, and medically-relevant components of prairie rattlesnake venom (Fig. 4a; snake venom metalloproteinases, SVMPs; snake venom serine proteinases, SVSPs; and type IIA

phospholipases A2, PLA2s) – each of these families is located on a different microchromosome. The only remaining major component of prairie rattlesnake venom, myotoxin (crotamine), is located on Chromosome 1 (Fig. 4a). The intriguing location and abundance of venom genes on microchromosomes suggests intimate associations between microchromosome biology and venom evolution. To identify the origins and mechanisms underlying the evolution of these venom families we conducted phylogenetic estimates of each of the microchromosome-linked families listed above (including non-venom members) and inferred that each venom family represents a distinct set of tandemly-duplicated genes derived from a single duplication that gave rise to a monophyletic cluster of venom paralogs (Fig. 4b). While this mechanism has been proposed previously (Ikeda et al. 2010; Vonk et al. 2013), the contiguituy of our genome assembly provides the first definitive proof of this representing a repeated mechanism underlying the origin of snake venom gene clusters.

Using gene expression data from multiple venom gland samples and a diversity of other tissues, we find that genes in these venom clusters can be further readily demarcated by their distinctive venom gland-specific expression patterns (Fig. 5), which also highlights marked expression differences between the venom cluster versus flanking non-venom genes of PLA2, SVMP, and SVSP gene families. Such discrete expression patterns of adjacent venom and non-venom genes raises the intriguing question of how venom genes are uniquely regulated and targeted for expression in venom glands.

To understand mechanisms underlying venom-gland-targeted expression of venom genes we combined Hi-C, gene expression, and genome information, and took advantage of the fact that snake venom glands are paired (one on the left, one on the right side). First, to investigate the chromatin architecture of venom production, we extracted venom from one venom gland of the genome animal two days prior to the other gland, then dissected the venom glands one day after the second gland was extracted – this accomplished staggering the process of venom expression in these two glands, providing a 1 and 3 day post-extraction design. Hi-C sequencing of the 1-day post-extraction venom gland then enabled us to capture the chromatin contacts underlying venom production, which we further investigated by comparing gene expression between the two glands. Based on our Hi-C data, we find that the precise genomic regions containing venom clusters show a highly specific chromatin structuresituated within discrete high-frequency contact regions representing distinct topologically-associated domains (TADs;

(Dixon et al. 2016) of open chromatin (Fig. 5). Genes adjacent to, and outside of these venom-specific TADs exhibit significantly lower expression in the venom gland, indicating a remarkably strong insulating regulatory effect of TAD boundaries surrounding venom cluster regions (Fig. 5b), which also serves to block the spread of positive regulators to non-venom regions.

To identify transcription factors that may be responsible for directing venom-gland specific expression of venom genes, we compared gene expression levels of all annotated transcription factors in the rattlesnake genome between venom glands and other tissues. Here, we specifically tested for significant evidence that transcription factors exhibit an expression profile similar to the observed profiles for the venom gene clusters (Fig. 5a). This analysis identified a set of candidate transcription factors of interest that were significantly more highly expressed in the venom gland, including six with specific DNA binding function: *FOXC2, SREBF2*, and four members of the *CTF/NFI* family of DNA-binding transcription factors (*NFIA*, two isoforms of *NFIB,* and *NFIX*; Table 5). To narrow this candidate set of putative venom-driving transcrion factors, we tested for evidence that predicted binding site sequences for these transcription factors were over-represented specifically in venom genes. We find that the upstream regions of genes in each of the three main venom clusters are significantly enriched for predicted *NFI* transcription factor binding sites (p-values < 0.05), but not *FOXC2* or *SREBF2* (p-values > 0.05). The combination of venom gland expression specificity and binding site enrichment analyses thus imply a central role of the *NFI* family of transcription factors in regulating expression of snake venom. Additionally, *NFI* has low binding affinity for nucleosomal DNA (Chikhirzhina et al. 2008), and the inferred open chromatin state within venom clusters should further enable efficient binding of highly-expressed *NFI* isoforms to their predicted binding sites, indicating the important complementary roles of both regional open chromatin state and venom-gland-specific transcription factors in the regulation of snake venom production.

While not directly involved in venom gene regulation, we also find evidence that other transcription factors that exhibit significant upregulation in the venom gland play important roles in in venom production, such as those involved in the unfolded protein response of the endoplasmic reticulum (e.g., *ATF6* and *CREB3L2*) and in glandular epithelial development and maintenance (e.g., *ELF5* and *GRHL1*). While not immediately obvious, the increased activity of each of these categories of transcription

factor in the venom gland makes sense considering the distinct features and requirements of venom production. For example, the rapid and immediate production of venom proteins following the release of venom (Luna et al. 2009) is expected to place incredibly high demands (and stress) on the endoplasmic reticulum as proteins are packaged and secreted into the venom gland lumen, and increased expression of factors involved in protein-folding chaperone recruitment would be critical during punctuated bursts of venom production post envenomation, as the venom store in the gland is rapidly replenished. Similarly, transcription factors involved in epithelial development that show increased expression in the venom gland are undoutebly linked to demands to maintain the venom gland lumen during venom production. Collectively, our findings raise the possibility that a core set of venom gland-specific transcription factors function to co-regulate venom production in venom gene clusters of open chromatin, and illustrate that venom production may be made possible through increased activity of other transcription factors involved in cellular stress responses and development.

Conclusion

Our analysis of the prairie rattlesnake genome provides new, and in some cases, surprising insight into the structure and function of reptilian and snake genomes, and broadly argues for the importance of studying diverse vertebrate lineages to understand the scope of vertebrate genome structure and function. For example, it appears that snakes have re-evolved genomic isochore structure not through an accumulation of GC content as observed in mammals and birds, but rather through the accumulation of AT content, suggesting a distinct GC isochore generative mechanism in snake genomes. Evidence for distinct evolutionary strata and the pseudoautosomal region of a snake sex chromosome, which bear unique hallmarks of the evolutionary trajectory from an ancestral autosomal chromosome pair, provide key comparative evidence to explain mechanisms underlying at least a majority of the partial dosage compensation observed in snakes. As the first species with microchromosomes to be analyzed at the nuclear organizational level using Hi-C, we show the surprising degree to which rattlesnake microchromosomes physically contact and interact with other chromosomes in the nucleus, suggesting that microchromosomes may operate in a fundamentally different way than macrochromosomes. Finally, in addition to the medical importance of studying snake venom, snake venom systems represents an

intriguing model for understanding how evolution can direct the organization and regulation of a novel organ system – the venom gland – one of nature's most dynamic trophic adaptations. The excellent contiguity of our genome assembly enabled the definitive chromosomal localization of venom gene clusters, most of which are found on microchromosomes, and illustrates clearly the mechanistic process of tandem duplication that has given rise to venom gene diversity multiple times across venom gene families. Our results also demonstrate many new and exciting mechanisms that underlie the tight regulation of venom genes, and the coordinated roles of chromatin and specific transcription factors in this process, as well as the co-evolution of other cellular mechanisms required to meet the extreme demands of bursts of venom production. Despite the key perspectives that the rattlesnake genome provides, many open questions remain, such as the evolutionary mechanisms by which snakes have accumulated AT content, how venom genes have gained venom gland-specific transcription factor binding sites, and the degree to which chromatin state is modulated in other tissues to prevent toxic venom gene expression. Conclusions from this and other studies consistently point to the unique and extreme biology of snakes that also extends to the unique biology of their genomes, highlighting the value of snakes and other non-traditional models in delivering new and often surprising perspectives into vertebrate biology and evolution.

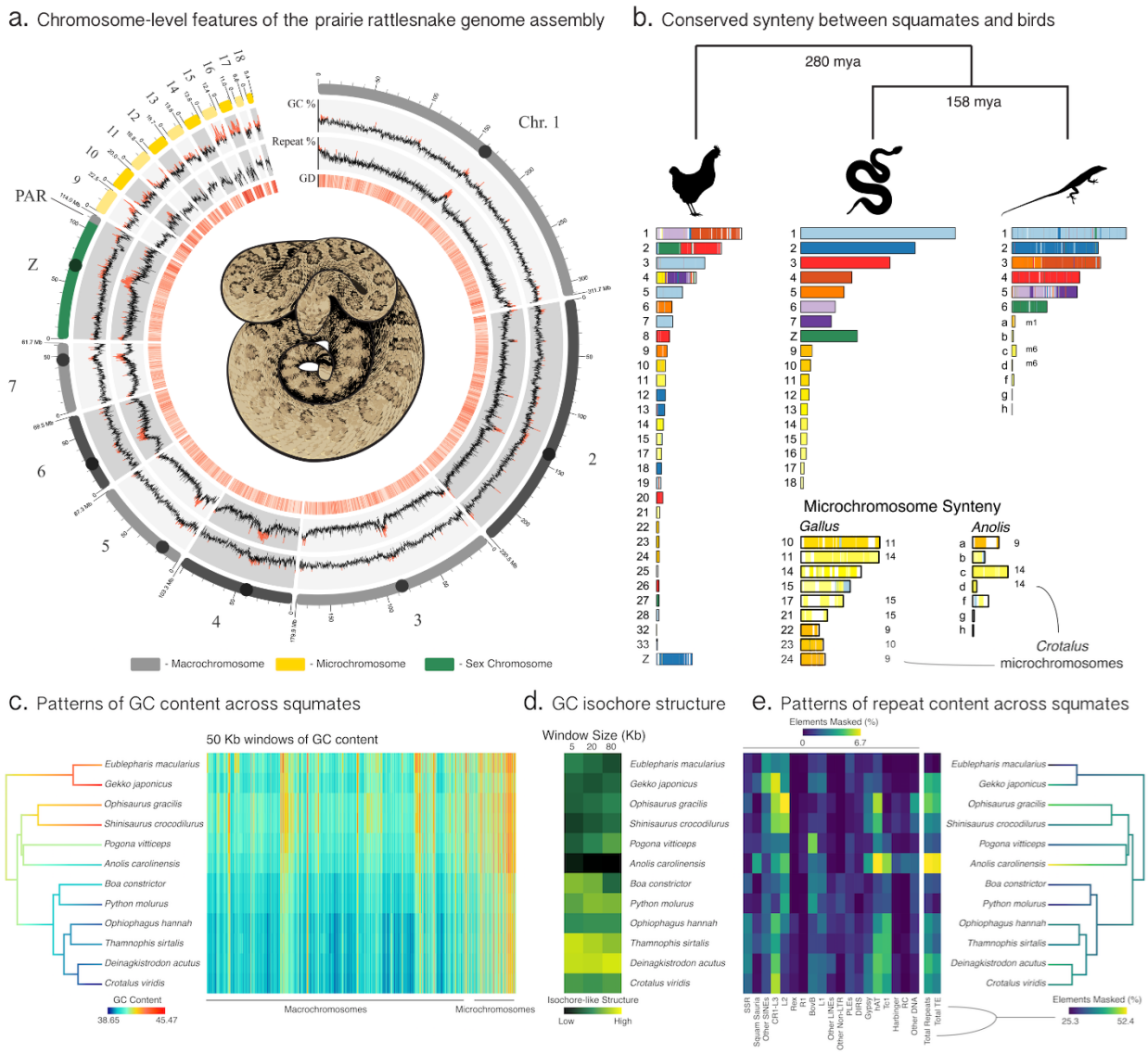

Figure 1. The chromosome-level prairie rattlesnake genome assembly. a, Diagram of genome-wide statistics. Chromosomes are show to scale and tick marks represent megabases of sequence. Chromosome ideogram is shown in the outer band, where circles represent inferred centromere locations. The grey segment of the Z chromosome represents the candidate pseudoautosomal region. GC% is the proportion of GC in 100Kb windows and Repeat % is the proportion of bases annotated as repeat content within each 100Kb window. Values above the genome-wide median are in red. GD is gene density, or the number of genes per 100Kb window; higher density is represented by darker red bands. **b**, Synteny between the rattlesnake and chicken and anole lizard genomes. Colors on chicken and anole chromosomes correspond with homologous sequence in rattlesnake. In the microchromosome inset, numbers to the right of chromosomes represent rattlesnake microchromosomes with which a given chicken or anole chromosome was syntenic for greater than 80% of its length. **c,** Tree branches are colored according to genomic GC content. The heatmap to the right depicts GC content in consecutive 50 Kb windows from a whole genome alignment, with macro- and microchromosome regions delineated. **d,** Genomic GC isochore structure measured by the standard deviation in GC content among 5, 20, and 80 Kb windows. **e,** Repeat content among 12 squamate species. Tree branches are colored by genomic repeat content.

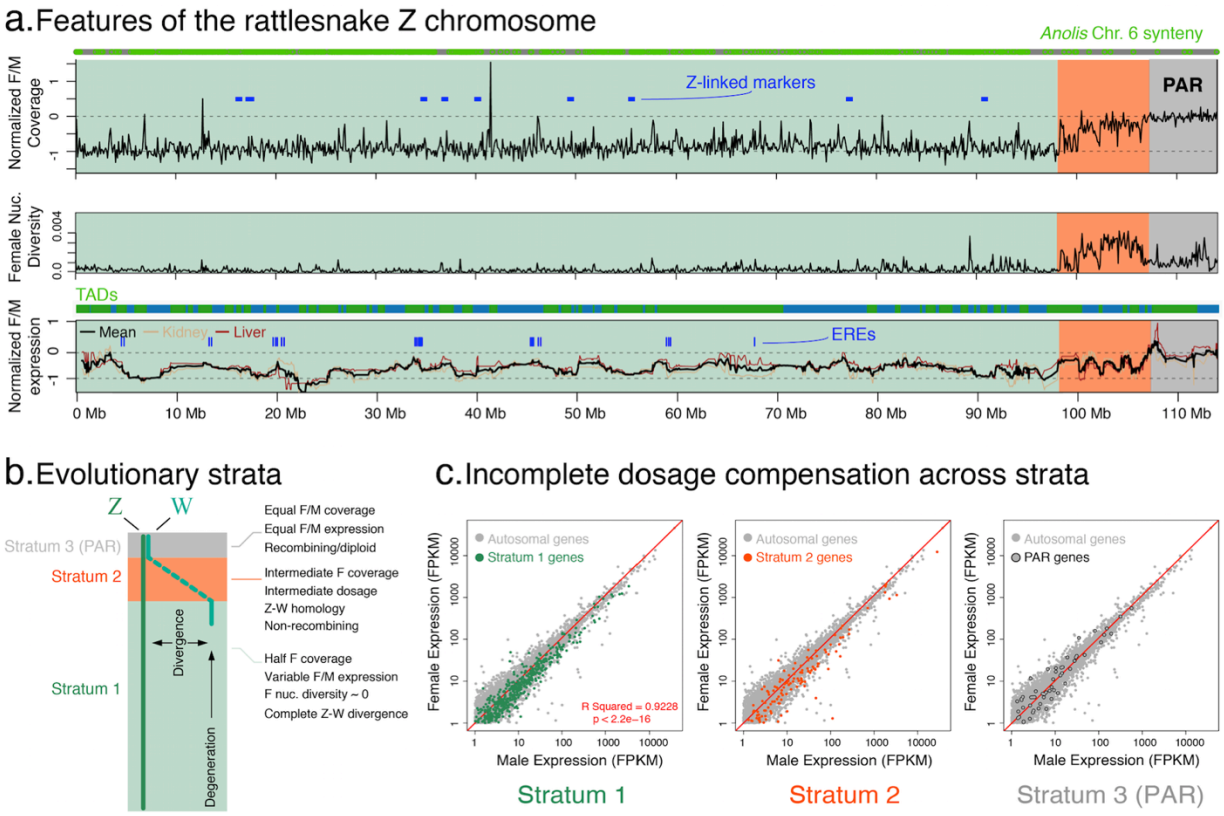

Figure 2. The Z chromosome of the prairie rattlesnake. a, Chromosomal landscape of log2 normalized female/male coverage, female nucleotide diversity, and log2 normalized female/male gene expression. High similarity BLAST hits to *Anolis* chromosome 6 are shown at the top as grey and green circles (high-stringency hits are in green; see Supplementary Methods). The positions of Z-linked cDNA markers from Matsubara et al. (2006) are shown as blue blocks, and the intervals of chromatin domains (TADs) are depicted as alternating green and blue blocks above the normalized expression plot. In the normalized expression plot, blue vertical lines represent the positions of predicted estrogen response elements (EREs) within 100 Kb of a dosed gene. On each plot, the pseudoautosomal region (PAR) and evolutionary Stratum 2 are highlighted in grey and orange, respectively, and Stratum 3 is highlighted in green. **b,** Schematic of the hypothesized evolutionary strata on the rattlesnake Z chromosome, with features that define them to the right. The dashed blue-green line representing the W chromosome depicts the inferred intermediate level of divergence between the Z and W chromosomes along Stratum 2. **c,** Patterns of relative female and male gene expression in each evolutionary stratum, plotted against the autosomal background (grey).

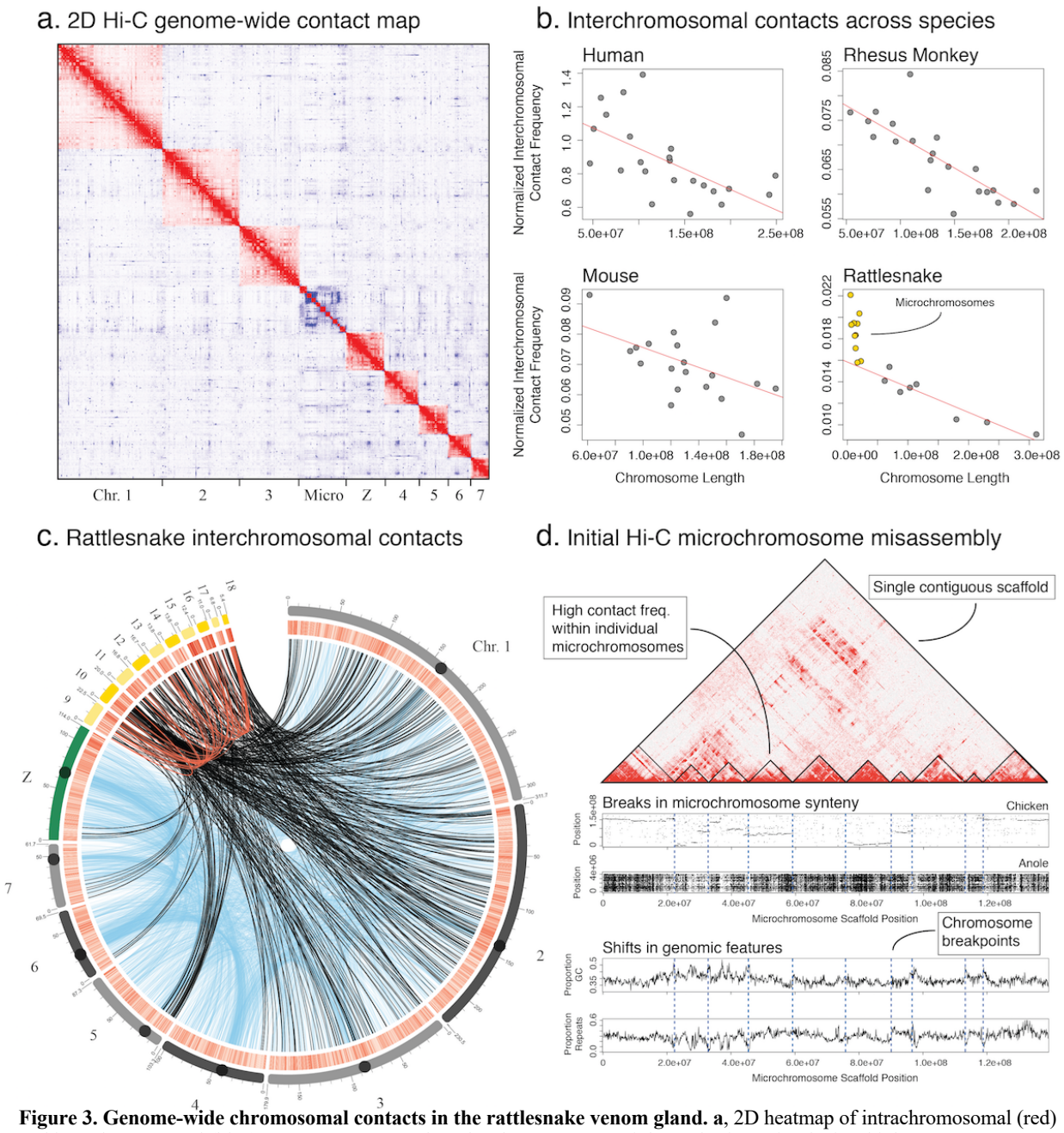

Figure 3. Genome-wide chromosomal contacts in the rattlesnake venom gland. a, 2D heatmap of intrachromosomal (red) and interchromosomal (blue) contacts among rattlesnake chromosomes. Higher color intensity depicts higher contact frequency. **b,** Comparison of interchromosomal contacts, normalized by chromosome length, and chromosome length between mammalian Hi-C datasets and the rattlesnake. Red lines depict the negative linear relationship between interchromosomal contacts and chromosome length for macrochromosomes. **c,** Locations of high-frequency interchromosomal contacts among rattlesnake chromosomes. Blue lines represent inter-macrochromosome contacts, black lines represent micro-to-macrochromosome contacts, and red lines represent inter-microchromosome contacts. **d,** Schematic of the initial misassembled microchromosome scaffold. The heatmap panel at the top depicts the high frequency inter- and intrachromosomal contacts among microchromosomes, and black triangles depict boundaries between microchromosomes. The middle two panels show synteny alignments between rattlesnake, chicken, and anole microchromosomes. The bottom two panels show windowed GC and repeat content across microchromosomes. Blue dashed lines in the lower panels show breakpoints between individual microchromosomes.

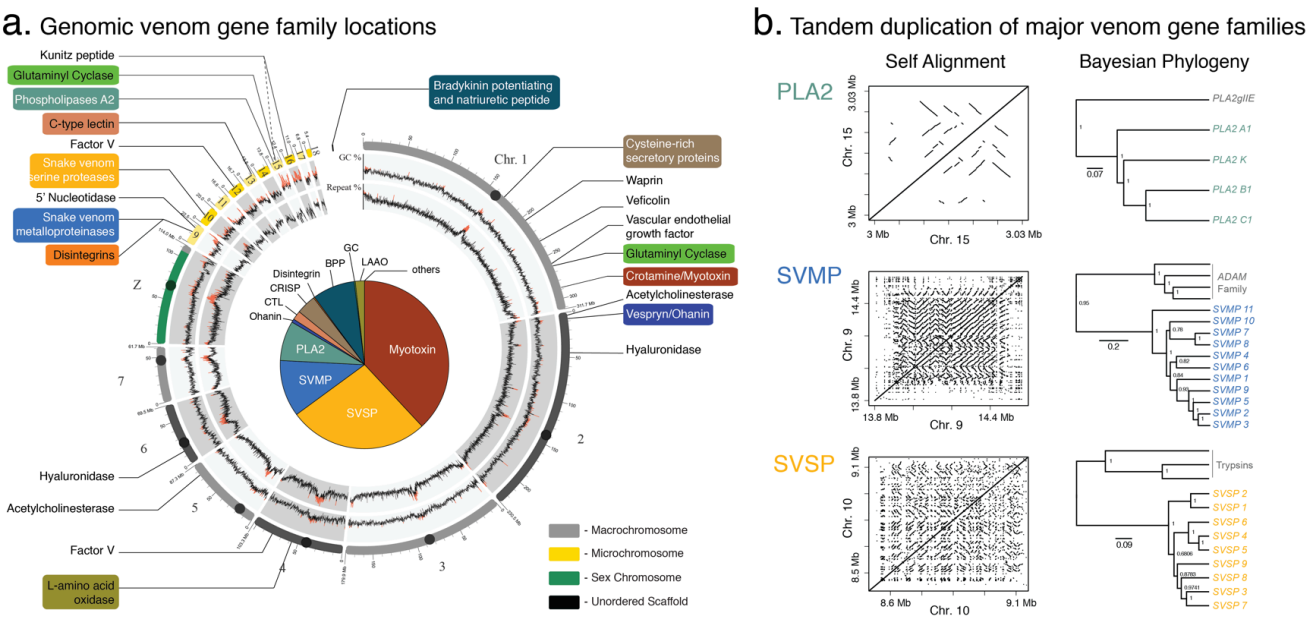

Figure 4. Genomic location of venom gene families and evidence for venom gene evolution through tandem duplication. **a,** The pie chart on the inside of the circularized genome ideogram represents the prairie rattlesnake proteome, redrawn from Saviola et al. 2015. The genome ideogram, GC content, repeat content, and legend at the bottom right follow the description in figure 1. Outside labels point to the genomic location of each venom gene family. **b,** Regional self alignment of phospholipase A2 (PLA2), snake venom metalloproteinase (SVMP), and serine proteinase (SVSP) venom gene clusters (left). Parallel and perpendicular lines off of the central diagonal line indicate segmental duplications. Bayesian phylogenetic tree estimates for each of the three gene families (right). Values at nodes represent posterior probabilities.

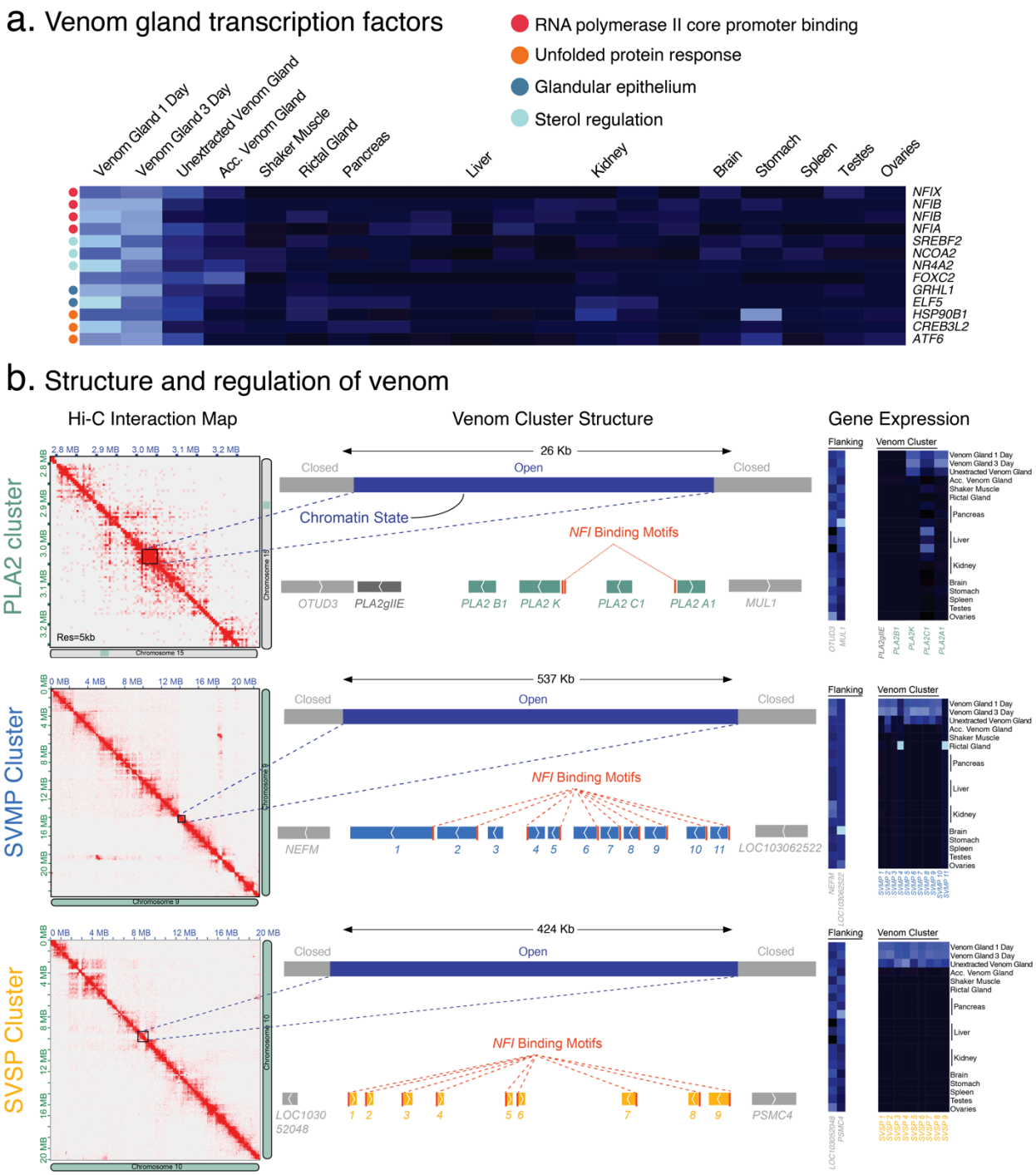

Figure 5. Mechanisms of venom gene regulation. a, Gene expression of transcription factors found to be significantly upregulated in venom glands, with expression values shown across tissues. Brighter colors show higher gene expression. The colored dots to the left of heatmap rows correspond to UniProt classifications of each transcription factor, which generally fell into the four categories in the legend in the top right. **b,** Genomic structure and regulation of the PLA2, SVMP, and SVSP venom gene families. 2D Hi-C contact maps are shown to the left, and boxes are used to show the bounds of each venom gene region. The schematics in the center depict the inferred chromatin state of each venom gene region (i.e., open chromatin) in the venom gland, the structure of each venom gene family and the non-venom genes flanking them. Predicted *NFI* transcription factor binding sites are shown as orange boxes upstream of genes. Gene expression profiles are shown to the right for each venom gene family and the flanking genes.

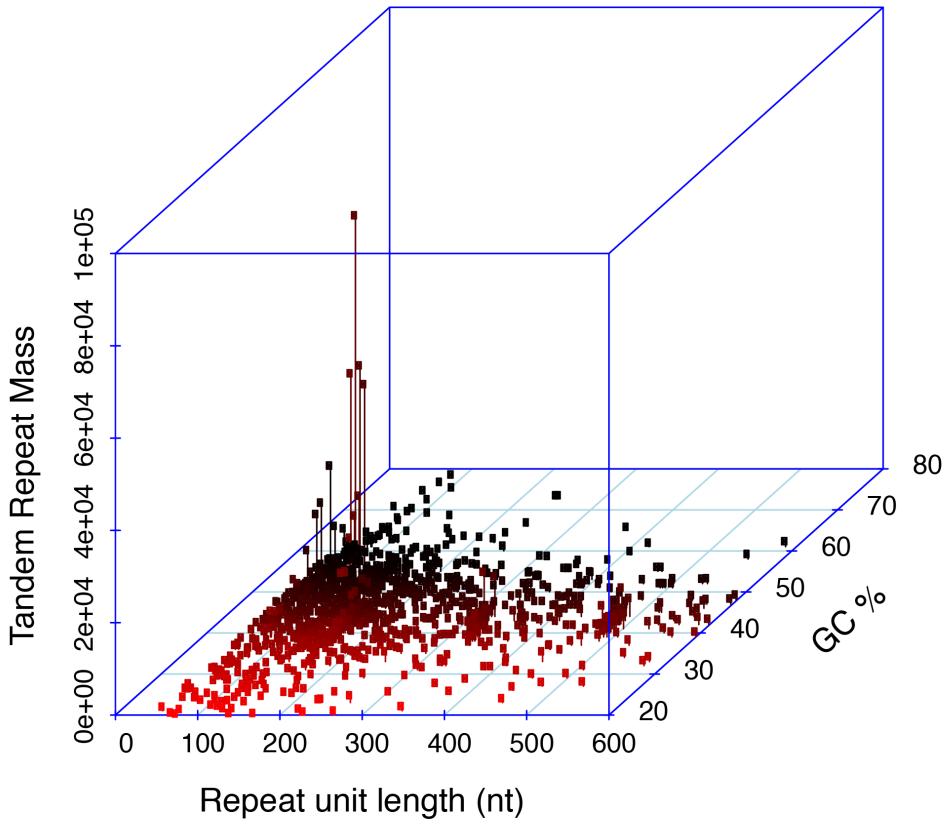

Figure 6. Centromeric tandem repeat motif characterized using tandem repeats finder. Analysis of high frequency tandem repeats identified a 164-mer with high relative GC to the genomic background. The y-axis, tandem repeat mass, represents the relative abundance of tandem repeats of a given unit length and GC content.

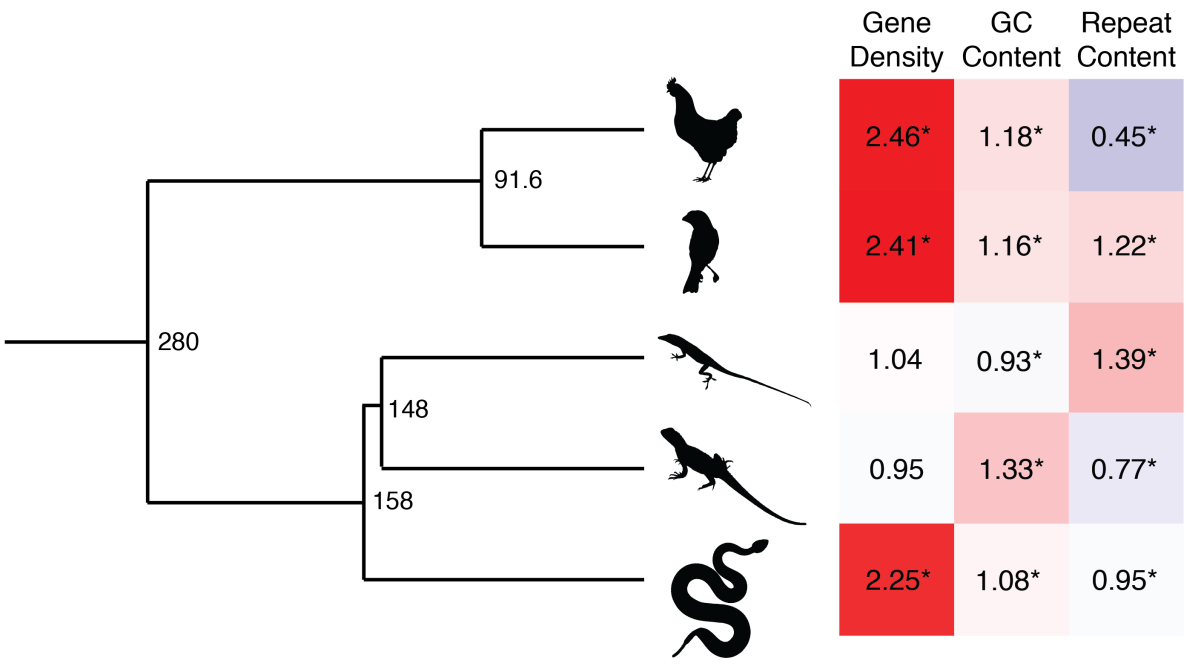

Figure 7. Evolutionary patterns of genomic features of microchromosomes among reptiles. Values at nodes on the phylogenetic tree represent the node age in millions of years, and were obtained using median estimates from TimeTree. The heatmap to the right represents the relative abundance of a given measure on microchromosomes versus macrochromosomes within each species (blue values represent greater abundance on macrochromosomes and red values represent greater abundance on microchromosomes). Values in each heatmap cell equal the ratio of each measure on microchromosomes:macrochromosomes, and values with asterisks represent significant differences between microchromosomes and macrochromosomes.

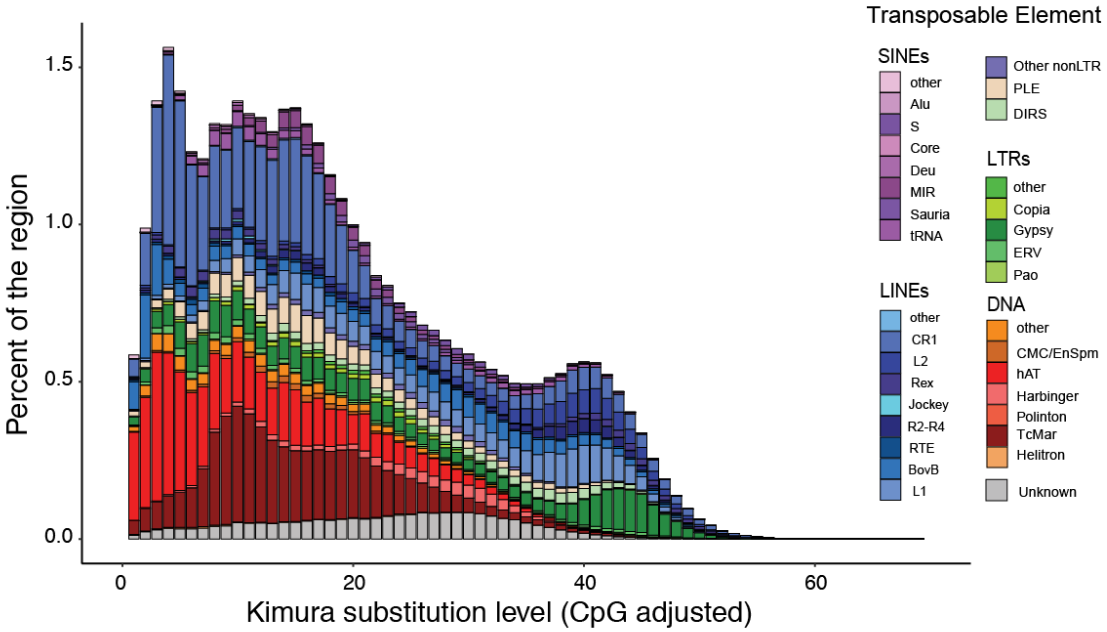

Figure 8. Genomic repeat element abundance at a range of relative age values. Age is measured using the Kimura substitution level of transposable elements when compared to a consensus sequence.

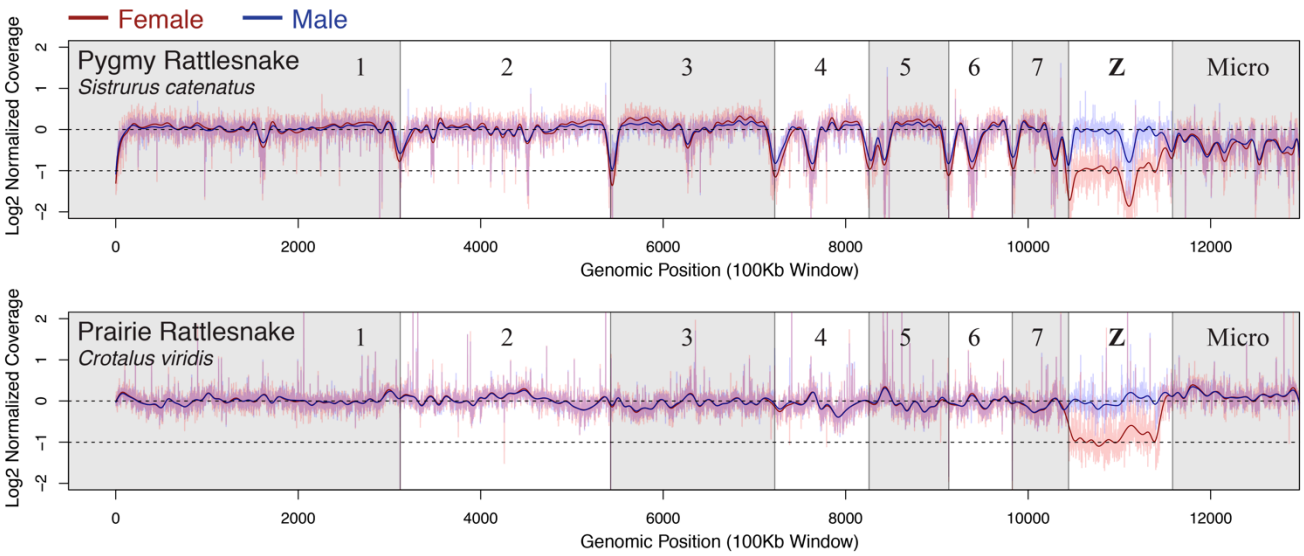

Figure 9. Log2 normalized female (red) and male (blue) coverage of two rattlesnake species (*Sistrurus catenatus* and *Crotalus viridis*), when mapped to the prairie rattlesnake reference genome. The dashed line at zero represents the normalized coverage expectation for diploid loci, and the dashed line at -1 represents the expectation of a hemizygous locus. The transparent lines show values for each 100 Kb window in a sliding window analysis of coverage, and bold lines show a smoothened spline of relative coverage across the genome.

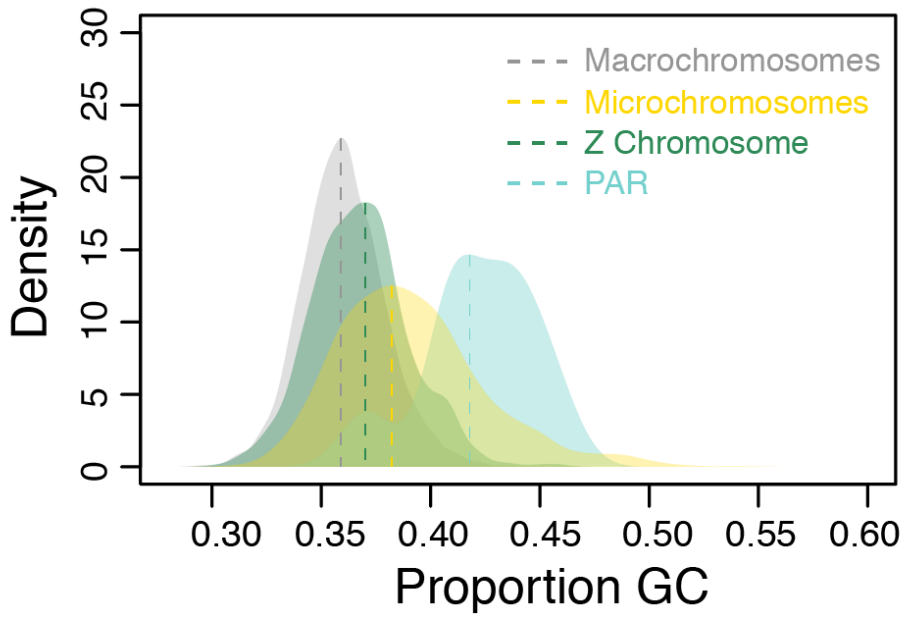

Figure 10. Density distributions of GC content across prairie rattlesnake chromosomes, showing specific distributions of macrochromosomes, microchromosomes, the Z chromosome, and the pseudoautosomal region (PAR) of the sex chromosomes, specifically.

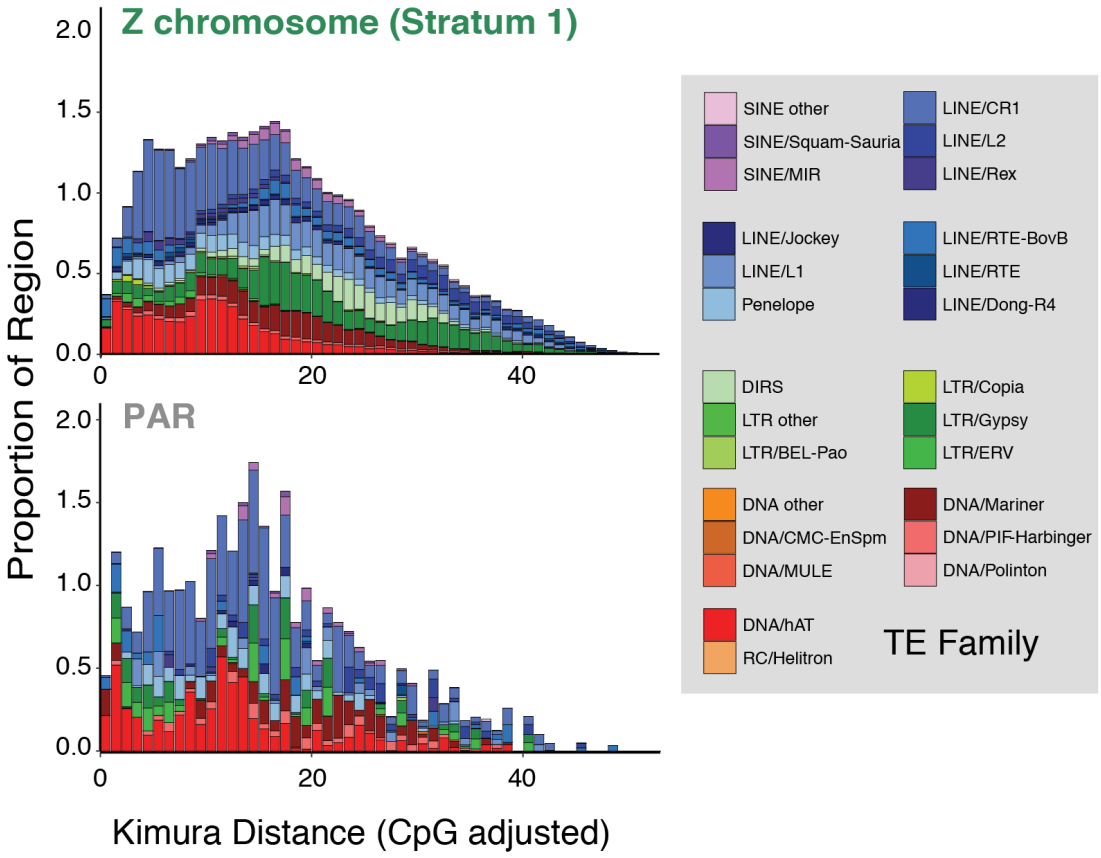

Figure 11. Comparative age distributions of proportions of transposable elements (TEs) across Stratum 1 (upper) and the pseudoautosomal region (PAR; lower) of the rattlesnake Z chromosome.

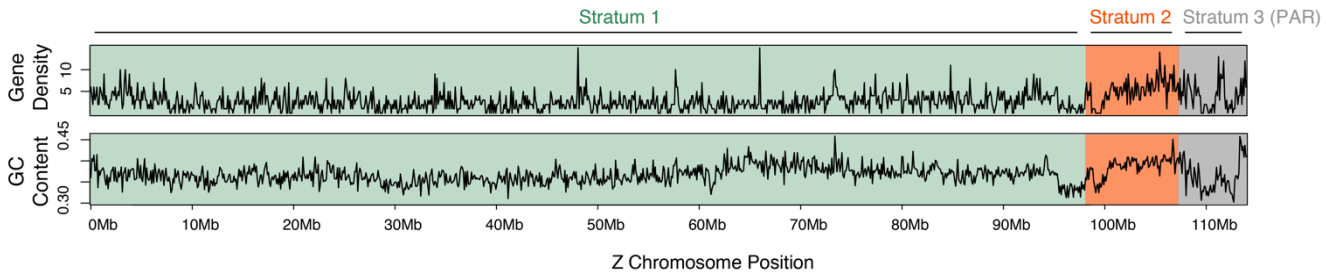

Figure 12. 100 Kb windowed scans of gene density (measured as number of genes per window) and GC content (i.e., proportion of GC bases within each window) across the Z chromosome of the prairie rattlesnake. The evolutionary strata are denoted by green (Stratum 1), orange (Stratum 2), and grey (Stratum 3; PAR) backgrounds.

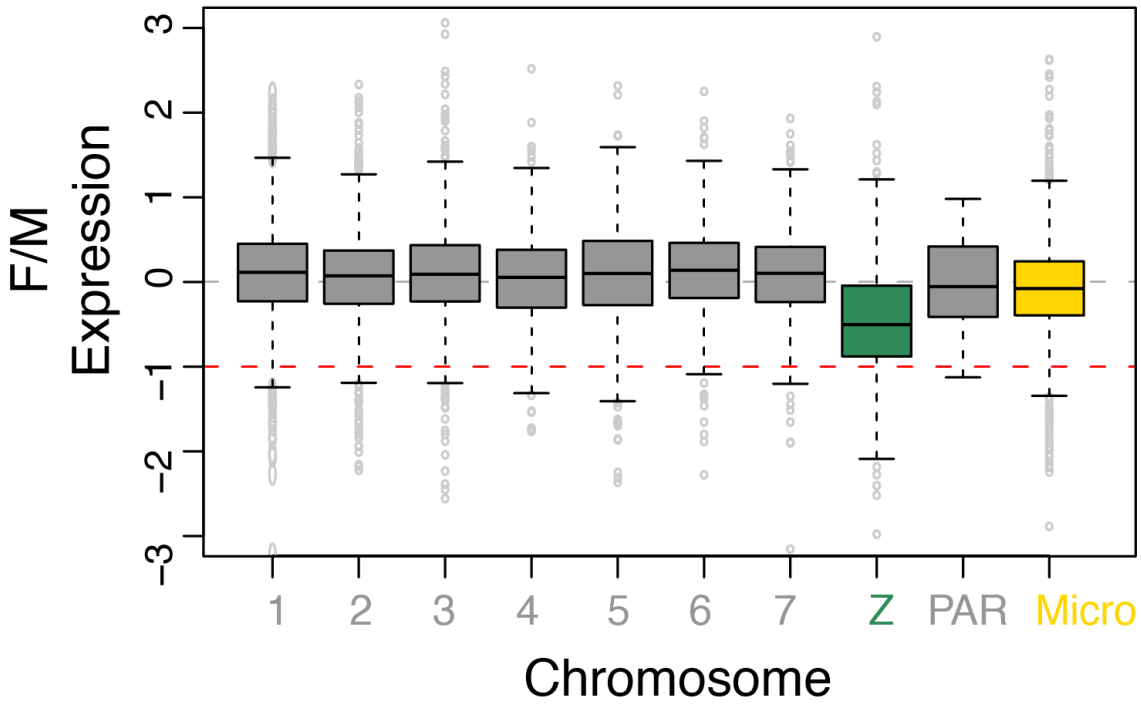

Figure 13. Boxplots of relative female:male gene expression on autosomal macrochromosomes (grey), the Z chromosome (green), the pseudoautosomal region (PAR; also grey), and microchromosomes. Outliers per chromosome are shown as small grey circles. The grey horizontal dashed line represents the expected value for autosomal loci, and the red dashed line represents gene expression for a hemizygous locus in the absence of any dosage compensation mechanisms.

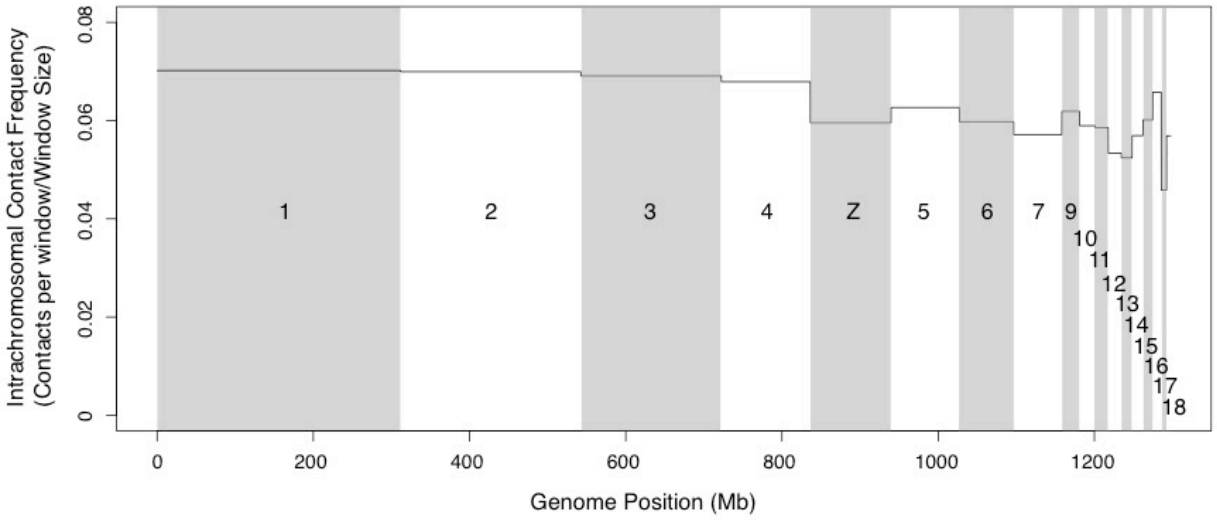

Figure 14. Intrachromosomal contact frequencies (i.e., the number of observed contacts between a chromosomal window and all other intrachromosomal windows divided by the size of the window) measured using Hi-C of the venom gland across rattlesnake chromosomes, demonstrating a constant level of intrachromosomal contact within each assembled chromosome. Individual chromosomes are labeled.

Table 1. Sequencing libraries used in the prairie rattlesnake genome assembly. Where noted, various libraries were used in the previous assembly (CroVir2.0), published in Pasquesi et al. (in review).

Library	Read Type	Number of Reads	Assembly Version
50bp short read	single end	9,536,384	CroVir2.0
100bp short read	paired end	449775645	CroVir2.0, CroVir3.0
150bp short read	paired end	41,211,014	CroVir2.0
150bp long insert mate pair (3-5Kb)	paired end	188,532,564	CroVir2.0
150bp long insert mate pair (6-8Kb)	paired end	189,928,342	CroVir2.0
PacBio long reads	-	1,027,365	CroVir2.0
Chicago long range proximity ligation library 1 (150bp)	paired end	251,689,106	CroVir3.0
Chicago long range proximity ligation library 2 (150bp)	paired end	206,176,028	CroVir3.0
Hi-C library 1 (150bp)	paired end	230,083,402	CroVir3.0
Hi-C library 2 (150bp)	paired end	160,673,944	CroVir3.0

Table 2. Basic information about assembly versions for the prairie rattlesnake genome.

	Input Assembly (CroVir2.0)	Chicago Assembly	HiRise (Chicago + Hi-C) Assembly
Longest Scaffold (bp)	1,184,546	11,576,738	311,712,589
Number of Scaffolds	47,782	8,183	7,034
Number of Scaffolds > 1Kb	47,658	8,059	6,910
Contig N50 (Kb)	15.81	14.91	14.96
Scaffold N50 (Kb)	139	2,472	179,898
Number of Gaps	112,369	158,269	159,024
Percent of Genome in Gaps	5.84%	6.15%	6.16%

Table 3. Genome-wide annotated repeat proportions identified using RepeatMasker.

	# elements	length masked (bp)	% of sequence	% element masked
Total masked	2966274	489373735	38.91	100.00
Total interspersed repeats	2348232	463237605	36.83	79.16
Retroelements	1139213	295244109	22.81	38.41
SINEs	173332	22894322	1.82	5.84
Squam1/Sauria	19230	3376458	0.27	0.65
Other SINEs	126898	15602678	1.24	4.28
LINEs	621859	170275973	13.54	20.96
CR1-Like	359387	91177000	7.25	12.12
CR1/L3	288888	74285822	5.91	9.74
L2	53219	12036490	0.96	1.79
Rex	19032	5339363	0.42	0.64
R1/LOA/Jockey	3272	854611	0.07	0.11
R2/R4/NeSL	35256	9045775	0.72	1.19
RTE/Bov-B	101958	32795496	2.61	3.44
L1/CIN4	78926	28358227	2.25	2.66
Other LINEs	154019	16472232	0.64	5.19
Other nonLTR	10119	1572442	0.13	0.34
DIRS	28657	13553057	1.08	0.97
PLEs	120162	19278497	1.53	4.05
LTR elements	156427	54116761	4.30	5.27
BEL/Pao	4007	1927682	0.15	0.14

Ty1/Copia	9160	3340874	0.27	0.31
Gypsy	77793	35080772	2.79	2.62
Retroviral	16727	5393228	0.43	0.56
Other LTR	48740	8374205	0.67	1.64
DNA transposons	850487	125287793	9.96	28.67
hobo-Activator	428247	60243144	4.79	14.44
Tc1-IS630-Pogo	283367	48888185	3.89	9.55
En-Spm	12485	1964905	0.16	0.42
MuDR-IS905	1300	383077	0.03	0.04
PiggyBac	131	22504	0.00	0.00
Tourist/Harbinger	80904	7193605	0.57	2.73
P elements	155	45074	0.00	0.01
Rolling-circles	3736	635885	0.05	0.13
SPIN	253	26640	0.00	0.01
Other DNA	39909	5884774	0.47	1.35
Unclassified	358532	48493199	3.86	12.09
Total interspersed repeats	2348232	463237605	36.83	79.16
Small RNA	2054	174940	0.01	0.07
Satellites	4952	1104344	0.09	0.17
Simple repeats	540288	28572170	2.27	18.21
Low complexity	70748	4755565	0.38	2.39

Table 4. Mapping of cDNA markers from Matsubara et al. 2006 to the prairie rattlesnake genome. Locations of best BLAST hits of each cDNA marker to the genome are reported. Where noted, cDNA markers mapped with exceptional similarity to multiple locations in the genome, or did not map to the chromosome as predicted by Matsubara et al. 2006. Markers for which there were two high-similarity hits on multiple chromosomes are denoted with italics.

Marker	Accession	Chromosome	Scaffold	e-value	bit-score	Start Position	End Position	Notes
OMG	BW999947	1p	scaffold-ma1	6.00E-115	398	309337082	309336564	
XAB1	AU312353	1p	scaffold-ma1	2.00E-46	122	297437298	297437486	
MGC15407	AU312344	1p	scaffold-ma1	2.00E-65	92.3	288097081	288097206	
XPO1	AU312325	1p	scaffold-ma1	2.00E-113	153	289547707	289547901	
DEGS	AU312341	1p	scaffold-ma1	5.00E-106	356	269312409	269311948	
KIAA0007	AU312332	1p	scaffold-ma1	5.00E-50	120	265943692	265943841	
EPRS	AU312324	1p	scaffold-ma1	2.00E-91	174	270708945	270709160	
ARID4B	AU312346	1p	scaffold-ma1	1.00E-129	333	252059286	252059699	
QKI	AU312356	1p	scaffold-ma1	5.00E-112	124	246094729	246094887	
MDN1	AU312339	1p	scaffold-ma1	7.00E-60	109	211517498	211517349	
AFTIPHILIN	AU312311	1p	scaffold-ma1	5.00E-75	112	170752748	170752888	
SF3B1	AU312337	1q	scaffold-ma1	7.00E-95	215	150078848	150078576	
CACNB4	BW999948	1q	scaffold-ma1	1.00E-47	102	127283965	127283819	
ZFHX1B	BW999949	1q	scaffold-ma1	6.00E-93	204	123301385	123301101	
UMPS	AU312331	1q	scaffold-ma1	8.00E-95	198	113761458	113761724	
TCIRG1	BW999950	1q	scaffold-ma1	2.00E-72	164	102088882	102089094	
TSG101	AU312316	1q	scaffold-ma1	4.00E-76	113	88358887	88359054	
M11S1	AU312350	1q	scaffold-ma1	4.00E-31	94.5	70777673	70777560	
GPHN	AU312327	1q	scaffold-ma1	5.00E-68	116	60249829	60249644	
DNCH1	AU312310	1q	scaffold-ma1	1.00E-71	145	25060055	25059885	
HSPCA	BW999951	1q	scaffold-ma1	2.00E-123	149	25029984	25030184	
ISYNA1	AU312338	1q	scaffold-ma1	2.00E-89	178	7770987	7771196	
TUBGCP2	AU312343	1q	scaffold-ma1	4.00E-74	136	9697568	9697377	
ZFR	AU312309	2p	scaffold-ma2	8.00E-110	208	222653709	222653461	
PHAX	AU312322	2p	scaffold-ma2	3.00E-99	224	189308026	189307715	

VPS13A	BW999952	2p	scaffold-ma2	9.00E-70	109	179725513	179725656	
UBQLN1	BW999953	2p	scaffold-ma2	2.00E-87	132	182156077	182156238	
C9orf72	AU312326	2p	scaffold-ma2	5.00E-91	203	164760033	164760347	
KIAA0368	BW999954	2p	scaffold-ma2	1.00E-56	116	161287251	161287397	
TOPORS	BW999955	2p	scaffold-ma2	8.00E-118	410	162258381	162257809	
FAM48A	BW999956	2cen	scaffold-ma2	1.00E-45	102	157286823	157286680	
UNQ501	AU312305	2cen	scaffold-ma2	6.00E-118	284	142895238	142895636	
DCTN2	AU312317	2q	scaffold-ma2	4.00E-80	122	122527271	122527110	
EXOC7	BW999957	2q	scaffold-ma2	3.00E-93	121	92952368	92952526	
DDX5	BW999958	2q	scaffold-ma2	7.00E-112	144	108253948	108253775	
CCNG1	AU312308	2q	scaffold-ma2	6.00E-70	173	80553964	80553731	
CPEB4	AU312333	2q	scaffold-ma2	3.00E-119	250	72297563	72297874	
FLJ22318	AU312329	2q	scaffold-ma2	2.00E-105	194	51908839	51908582	
DCTN4	AU312349	2q	scaffold-ma2	4.00E-50	99.6	58962806	58962928	
C5orf14	AU312304	2q	scaffold-ma2	4.00E-120	329	64853582	64853127	
NOSIP	AU312303	2q	scaffold-Z	1.00E-51	93.6	92988551	92988661	Did not map to predicted chromosome
RBM5	BW999960	2q	scaffold-mi8	6.00E-78	90.4	9620291	9620181	Mapped to multiple chromosomes with high similarity
RBM5	BW999960	2q	scaffold-ma2	7.00E-13	76.1	130725514	130725606	Mapped to multiple chromosomes with high similarity
ITPR1	BW999961	2q	scaffold-ma2	9.00E-53	135	23858424	23858585	
ENPP2	BW999962	3p	scaffold-ma3	6.00E-90	121	9756367	9756209	
YWHAZ	BW999963	3p	scaffold-ma3	2.00E-99	180	16759896	16760114	
LRRCC1	BW999964	3p	scaffold-ma3	4.00E-83	150	21993774	21993565	
LYPLA1	BW999965	3p	scaffold-ma3	3.00E-107	149	31673258	31673440	
SS18	AU312302	3p	scaffold-ma3	1.00E-83	126	36811554	36811724	
MBP	AU312318	3p	scaffold-ma3	7.00E-111	179	49049170	49049382	
EPB41L3	BW999966	3p	scaffold-ma3	3.00E-84	141	40222999	40222808	
TUBB2A	BW999967	3p	scaffold-ma3	8.00E-91	155	59187732	59187532	
LRRC16	BW999968	3p	scaffold-ma3	2.00E-100	144	51025171	51025350	
SERPINB6	BW999969	3p	scaffold-ma5	5.00E-99	130	36540937	36540755	Mapped to multiple chromosomes with high similarity
SERPINB6	BW999969	3p	scaffold-ma3	2.00E-76	113	60484038	60483865	Mapped to multiple chromosomes with high similarity

BPHL	BW999970	3p	scaffold-ma3	1.00E-87	118	59199779	59199621	
KIF13A	BW999971	3p	scaffold-ma3	3.00E-78	139	53681516	53681349	
TPR	BW999972	3q	scaffold-ma3	6.00E-83	122	93408800	93408636	
AKR1A1	BW999973	3q	scaffold-ma3	9.00E-75	153	133869419	133869619	
ZNF326	BW999974	3q	scaffold-ma2	2.00E-77	120	224940437	224940586	Did not map to predicted chromosome
YIPF1	BW999975	3q	scaffold-ma3	6.00E-52	112	127724189	127724353	
BCAS2	AU312354	3q	scaffold-ma3	3.00E-51	141	151621402	151621229	
KIAA1219	BW999976	3q	scaffold-ma3	4.00E-101	158	155122635	155122844	
STAU1	BW999977	3q	scaffold-ma3	2.00E-116	169	165663812	165663594	
RBM12	BW999978	3q	scaffold-ma3	2.00E-152	406	154706304	154705780	
TPT1	BW999979	4p	scaffold-ma4	2.00E-68	148	1006155	1006349	
EIF2S3	AU312306	4p	scaffold-ma4	1.00E-111	126	49115724	49115885	
SYAP1	AU312328	4p	scaffold-ma4	3.00E-96	121	46147275	46147135	
DSCR3	AU312319	4q	scaffold-ma4	1.00E-74	119	60873037	60872873	
DCAMKL1	BW999980	4q	scaffold-ma4	8.00E-49	110	86291138	86291302	
ELMOD1	BW999981	4q	scaffold-ma4	1.00E-56	147	93207704	93207522	
BCCIP	AU312307	5q	scaffold-ma5	1.00E-46	148	32597249	32597061	
SH3MD1	AU312347	5q	scaffold-ma5	2.00E-119	378	45831798	45832379	
PPP1R7	BW999982	5q	scaffold-ma5	2.00E-92	228	56956062	56955736	
PDCD10	AU312342	5q	scaffold-ma5	4.00E-61	143	74805371	74805547	
TLOC1	AU312335	5q	scaffold-ma5	2.00E-45	101	76109988	76110125	
UCHL1	BW999983	6p	scaffold-ma7	4.00E-89	210	33298090	33298407	Did not map to predicted chromosome
GNAI2	BW999984	6p	scaffold-ma2	2.00E-106	126	49893686	49893841	Did not map to predicted chromosome
P4HB	BW999985	6p	scaffold-ma2	2.00E-69	100	97717890	97718012	Did not map to predicted chromosome
FLJ12571	AU312352	6q	scaffold-ma6	2.00E-46	117	46698606	46698752	
RANGAP1	AU312313	6q	scaffold-ma6	7.00E-71	95	47795604	47795500	
LDHB	BW999986	6q	scaffold-ma6	2.00E-60	117	69268248	69268418	
SEC3L1	AU312345	7p	scaffold-ma7	3.00E-58	125	55644074	55643916	
KIAA1109	AU312348	7q	scaffold-ma7	2.00E-60	124	30398905	30398711	
RAP1GDS1	AU312351	7q	scaffold-ma7	2.00E-91	112	12141068	12140931	

GAD2	BW999991	Zp	scaffold-Z	1.00E-109	136	17484512	17484336	
WAC	AU312355	Zp	scaffold-Z	3.00E-93	209	16303681	16303947	
KLF6	BW999992	Zp	scaffold-ma2	1.00E-99	366	47130305	47130796	Did not map to predicted chromosome
LOC90693	BW999993	Zp	scaffold-ma7	4.00E-127	301	34444161	34444577	Mapped to multiple chromosomes with high similarity
LOC90693	BW999993	Zp	scaffold-Z	1.00E-107	291	34827559	34827182	Mapped to multiple chromosomes with high similarity
TAX1BP1	AU312320	Zp	scaffold-Z	1.00E-86	141	36989995	36990174	
RAB5A	BW999994	Zp	scaffold-Z	9.00E-94	166	40227424	40227215	
CTNNB1	BW999995	Zcen	scaffold-Z	3.00E-129	275	49548885	49549226	
AMPH	BW999996	Zcen	scaffold-Z	1.00E-66	101	55612836	55612955	
TUBG1	BW999997	Zq	scaffold-Z	5.00E-89	116	17359265	17359113	
GH1	BW999998	Zq	scaffold-Z	2.00E-115	179	77397011	77396727	
MYST2	BW999999	Zq	scaffold-Z	6.00E-122	293	90785118	90784714	
NEF3	BW999987	micro	scaffold-mi1	1.00E-102	352	13833430	13832942	
ASB6	AU312340	micro	scaffold-mi7	1.00E-95	161	6270589	6270353	
RPL12	BW999988	micro	scaffold-mi7	6.00E-67	95.5	7974658	7974542	
FLJ25530	AU312336	micro	scaffold-mi1	4.00E-98	255	8157147	8156806	
HSPA8	BW999989	micro	scaffold-ma1	2.00E-124	236	20422342	20422662	Mapped to multiple chromosomes with high similarity
HSPA8	BW999989	micro	scaffold-mi1	3.00E-123	259	2089357	2089025	Mapped to multiple chromosomes with high similarity
GLCE	AU312330	micro	scaffold-mi10	1.00E-79	234	24861	24577	
POLG	AU312315	micro	scaffold-mi3	4.00E-97	116	10042696	10042845	
LOC283820	AU312323	micro	scaffold-mi5	8.00E-71	116	3659851	3659708	
PARN	AU312312	micro	scaffold-mi7	1.00E-66	73.9	12029447	12029361	
ATRX	BW999990	micro	scaffold-mi4	3.00E-63	102	1268001	1268126	

Table 5. Transcription factors significantly upregulated in the venom gland.

Gene ID	Rattlesnake Gene Detail
ATF6	augustus_masked-scaffold-ma3-processed-gene-300.3
ELF5	maker-scaffold-ma1-augustus-gene-235.5
FOXC2	augustus_masked-scaffold-mi6-processed-gene-2.1
CREB3L2	maker-scaffold-ma6-augustus-gene-195.2
HSP90B1	maker-scaffold-ma6-augustus-gene-185.14
GRHL1	maker-scaffold-ma1-augustus-gene-601.8
NCOA2	maker-scaffold-ma3-augustus-gene-89.6
NFIA	maker-scaffold-ma3-augustus-gene-414.2
NFIB	maker-scaffold-ma2-augustus-gene-569.3
NFIB	maker-scaffold-ma2-augustus-gene-569.2
NFIX	maker-scaffold-ma2-augustus-gene-473.3
NR4A2	maker-scaffold-ma1-augustus-gene-428.4
SREBF2	maker-scaffold-ma6-augustus-gene-158.15

Table 6. RNAseq libraries used in this study**.**

Sample ID	Tissue	Raw Reads	Quality Trimmed Reads
CroVirPan	pancreas	28,126,703	27,073,946
CroVirTon	tongue	24,451,116	23,561,349
CroVirVG1	venom gland	41,744,110	40,147,306
CroVirVG3	venom gland	29,216,664	28,035,353
Cvv01	liver	7,833,506	7,365,740
Cvv02	liver	7,451,792	7,064,234
Cvv11	liver	9,218,939	8,441,587
Cvv20	kidney	6,958,120	6,580,387
Cvv22	kidney	8,116,679	7,601,517
Cvv23	kidney	7,193,762	6,785,947

Cvv25	skin	7,849,895	7,303,441
Cvv26	pancreas	8,886,612	8,160,214
Cvv27	venom gland	3,098,151	2,928,974
Cvv28	lung	6,613,196	6,024,613
Cvv29	testes	5,055,189	4,745,375
Cvv30	accessory venom gland	3,261,326	3,053,142
Cvv31	shaker muscle	4,290,989	3,996,274
Cvv32	pancreas	4,836,715	4,566,165
Cvv33	brain	3,815,570	3,569,113
Cvv34	stomach	5,297,110	4,993,142
Cvv35	ovaries	3,737,870	3,528,104
Cvv36	rictal gland	6,654,626	6,070,883
Cvv37	spleen	7,776,020	6,975,210
Cvv38	blood	2,550,433	2,364,162

Table 7. Details of Illumina Nextera resequencing libraries used for comparative female/male read coverage across the rattlesnake genome.

Library Type	Read Length	Sample ID	Species	Sex	Number of Mapped Reads
Illumina Nextera	150 bp paired end	CV0007	*Crotalus viridis viridis*	Male	20,279,801
Illumina Nextera	150 bp paired end	CV0011	*Crotalus viridis viridis*	Female	4,975,491

Table 8. GC variation in windows of various sizes for 12 squamate species. Values for each species are measured as the standard deviation (SD) of GC content in all sampled windows of a given size. Information for 5, 20, and 80 Kb windows are also presented in Fig. 1c. Missing data (i.e., window sizes that were too large and contained greater than the threshold allowed missing data) are denoted with '-'.

Window Size (bp)	*Gekko japonicus*	*Eublepharis macularius*	*Ophisaurus gracilis*	*Shinisaurus crocodilurus*	*Pogona vitticeps*	*Anolis carolinensis*
5,000	0.039295606	0.037140406	0.037038224	0.03488877	0.03681681	0.032312269
20,000	0.028980944	0.027338004	0.029217483	0.027425317	0.030930264	0.021209
40,000	0.025219459	0.024838347	0.027141528	0.025322106	0.029367252	0.017608402
80,000	0.021385708	0.023326607	0.025558162	0.023843432	0.028238318	0.015121097
160,000	0.01811246	0.022646783	0.024536212	0.022632678	0.027330318	0.013089382
240,000	-	0.022203903	0.023356372	0.021943776	0.026943855	0.012088733
320,000	-	0.022121291	0.022899173	0.021312719	0.026617904	0.011287772

Window Size (bp)	*Boa constrictor*	*Python molurus*	*Ophiophagus hannah*	*Thamnophis sirtalis*	*Deinagkistrodon acutus*	*Crotalus viridis*
5,000	0.043942864	0.042024505	0.040098669	0.047076022	0.047062019	0.041210929
20,000	0.034934365	0.035837726	0.031894398	0.037865804	0.03882085	0.032232558
40,000	0.030576918	0.033337717	0.028952912	0.03429097	0.036517713	0.029884634
80,000	0.023292703	0.030197592	0.026685436	0.031202717	0.034964163	0.0281043
160,000	0.014736549	0.02736241	0.024597185	0.02894796	0.033486765	0.026806291
240,000	-	0.024725646	0.023968494	0.026250057	0.032562166	0.02616041

320,000	-	0.023707617	0.023468328	0.024606171	0.031784231	0.025840409

Table 9. Representative sequences for known snake venom gene families used to annotate venom genes in the rattlesnake genome.

Gene Family	Accession	Sequence Type	Species
5'Nucleotidase	AK291667.1	mRNA	*Homo sapiens*
Acetylcholinesterase	U54591.1	mRNA	*Bungarus fasciatus*
AVItoxin	EU195459.1	mRNA	*Varanus komodoensis*
C-type Lectin	JF895761.1	mRNA	*Crotalus oreganus helleri*
Cobra Venom Factor	U09969.2	mRNA	*Naja kaouthia*
CRISp (cysteine-rich secretory protein)	HQ414088.1	mRNA	*Crotalus adamanteus*
Cystatin	FJ411289.1	mRNA	*Naja kaouthia*
Extendin	EU790960.1	mRNA	*Heloderma suspectum*
Exonuclease	XM_015826835.1	mRNA	*Protobothrops mucrosquamatus*
Hyaluronidase	HQ414098.1	mRNA	*Crotalus adamanteus*
LAAO (L-amino acid oxidase)	HQ414099.1	mRNA	*Crotalus adamanteus*
SVMP I (class I snake venom metalloproteinase)	HM443635.1	mRNA	*Bothrops neuwiedi*
SVMP II (class II snake venom metalloproteinase)	HM443637.1	mRNA	*Bothrops neuwiedi*
SVMP III (class III snake venom metalloproteinase)	HM443632.1	mRNA	*Bothrops neuwiedi*
Nerve growth factor	AF306533.1	mRNA	*Crotalus durissus terrificus*
Phosphodiesterase	HQ414102.1	mRNA	*Crotalus adamanteus*
PLA2_I (vipers)	AF403134.1	mRNA	*Crotalus viridis viridis*
PLA2_II (elapids)	GU190815.1	mRNA	*Bungarus flaviceps*
Sarafotoxin	L07528.1	mRNA	*Atractaspis engaddensis*
Serine Proteinase	HQ414121.1	mRNA	*Crotalus adamanteus*
3FTX (Three-finger Toxin)	DQ273582.1	mRNA	*Ophiophagus hannah*
Veficolin	GU065323.1	mRNA	*Cerberus rynchops*
VEGF (Vascular Endothelial Growth Factor)	AB848141.1	mRNA	*Protobothrops mucrosquamatus*
Vespryn	EU401840.1	mRNA	*Oxyuranus scutellatus*
Waprin	EU401843.1	mRNA	*Oxyuranus scutellatus*

Kunitz (serine peptidase inhibitor, Kunitz type)	JU173666.1	mRNA	*Crotalus adamanteus*
Thrombin-like (thrombin-like venom gland enzyme)	AJ001209.1	mRNA	*Deinagkistrodon acutus*
Ficolin	GBUG01000048.1	mRNA	*Echis coloratus*
Disintegrin	AJ131345.1	mRNA	*Deinagkistrodon acutus*
FactorV (venom coagulation factor V)	XM_015815922.1	mRNA	*Protobothrops mucrosquamatus*
FactorX	XM_015819885.1	mRNA	*Protobothrops mucrosquamatus*
Prokineticin	XM_015822870.1	mRNA	*Protobothrops mucrosquamatus*
Ohanin (ohanin-like)	XM_015818414.1	mRNA	*Protobothrops mucrosquamatus*
Complement C3 (Cadam VF)	JU173742.1	mRNA	*Crotalus adamanteus*
Crotasin	AF250212.1	mRNA	*Crotalus durissus terrificus*
Endothelin	XM_015810852.1	mRNA	*Protobothrops mucrosquamatus*
Kallikrein	GALC01000005.1	mRNA	*Crotalus oreganus helleri*
Lynx1 (Ly6/neurotoxin 1)	XM_014066791.1	mRNA	*Thamnophis sirtalis*
Natriuretic Peptide (bradykinin potentiating peptide and C-type natriuretic peptide precursor isoform 2)	AF308594.2	mRNA	*Crotalus durissus terrificus*
sPla/ryanodine receptor	XM_015823102.1	mRNA	*Protobothrops mucrosquamatus*
WAP four-disulfide core domain protein 5 (Whey Acidic Protein/secretory leuki proteinase inhibitor)	XM_015822353.1	mRNA	*Protobothrops mucrosquamatus*
Myotoxin	HQ414100.1	mRNA	*Crotalus adamanteus*
PLA2	APD70899.1	protein	*Crotalus atrox*
SVMP	Q90282.1	protein	*Crotalus atrox*
Serine Proteinase	F8S114.1	protein	*Crotalus adamanteus*

Table 10. Annotated venom gene homologs in the prairie rattlesnake genome. Genes were annotated using materials detailed in Supplementary Table 9.

Venom Gene Family	Rattlesnake Scaffold	Start Position (bp)	End Position (bp)
3-Finger toxin	scaffold-ma1	103004868	103021927
3-Finger toxin	scaffold-ma1	102999393	103000958
5' Nucleotidase	scaffold-ma5	46133017	46179118
5' Nucleotidase	scaffold-ma6	55711914	55732365
5' Nucleotidase	scaffold-mi1	18004217	18021456
5' Nucleotidase	scaffold-ma2	45090212	45121335
5' Nucleotidase	scaffold-ma2	134237148	134264183
Acetylcholinesterase	scaffold-ma2	4047955	4053281
Acetylcholinesterase	scaffold-ma2	3948506	3952373
Acetylcholinesterase	scaffold-ma2	4016363	4018146
Acetylcholinesterase	scaffold-ma2	4026170	4045822
Acetylcholinesterase	scaffold-ma5	73971094	73976212
Acetylcholinesterase	scaffold-ma5	74015346	74036663
Acetylcholinesterase	scaffold-un210	16032	17552
Bradykinin potentiating and natriuretic peptide	scaffold-un187	22386	23524
C-type lectin	scaffold-mi5	3276042	3284747
C-type lectin	scaffold-mi5	11650747	11653723
C-type lectin	scaffold-Z	21883578	21895509
C-type lectin	scaffold-Z	21706900	21776775
C-type lectin	scaffold-Z	21786524	21797211
C-type lectin	scaffold-Z	108214710	108236532
Cysteine-rich secretory protein	scaffold-ma1	169434958	169437996
Cysteine-rich secretory protein	scaffold-ma1	169423774	169434684
Cysteine-rich secretory protein	scaffold-ma3	25391938	25416947
Cysteine-rich secretory protein	scaffold-mi6	1021447	1040191
Exonuclease	scaffold-mi7	8097114	8103411
Exonuclease	scaffold-ma1	5804894	5842638
Exonuclease	scaffold-mi3	10271502	10274220
Exonuclease	scaffold-ma6	12590208	12591465
Factor V	scaffold-mi4	8493826	8518402
Factor V	scaffold-mi4	8479637	8493564
Factor V	scaffold-ma4	81074882	81113119
Glutaminyl cyclase	scaffold-ma1	256551622	256564040
Glutaminyl cyclase	scaffold-mi7	5091107	5094268
Hyaluronidase	scaffold-ma6	14952252	14955850
Hyaluronidase	scaffold-ma2	45901201	45920587
Hyaluronidase	scaffold-ma2	49137409	49145188
Hyaluronidase	scaffold-ma2	49106981	49118469
Kunitz peptide	scaffold-mi7	3590975	3597607

Kunitz peptide	scaffold-mi8	4992795	5002390
L-amino acid oxidase	scaffold-ma4	56914906	56948498
L-amino acid oxidase	scaffold-ma4	85461961	85468906
L-amino acid oxidase	scaffold-ma2	4658599	4661642
L-amino acid oxidase	scaffold-ma2	4654769	4658293
Myotoxin/crotamine	scaffold-ma1	289328153	289328605
Nerve growth factor	scaffold-Z	93342025	93347811
Nerve growth factor	scaffold-ma1	76711308	76727703
PLA2	scaffold-mi7	3019970	3021876
PLA2	scaffold-mi7	3027607	3029199
PLA2	scaffold-mi7	3031464	3033348
PLA2	scaffold-mi7	3037103	3038488
PLA2	scaffold-mi7	3042118	3043697
Serine Proteinase	scaffold-mi2	8569773	8575182
Serine Proteinase	scaffold-mi2	8588278	8593660
Serine Proteinase	scaffold-mi2	8628274	8636651
Serine Proteinase	scaffold-mi2	8664603	8670797
Serine Proteinase	scaffold-mi2	8739986	8745649
Serine Proteinase	scaffold-mi2	8752578	8759324
Serine Proteinase	scaffold-mi2	8864675	8879153
Serine Proteinase	scaffold-mi2	8937526	8947481
Serine Proteinase	scaffold-mi2	8960028	8980478
Snake venom metalloproteinase	scaffold-mi1	13901629	14014239
Snake venom metalloproteinase	scaffold-mi1	14022082	14075370
Snake venom metalloproteinase	scaffold-mi1	14091987	14112667
Snake venom metalloproteinase	scaffold-mi1	14147865	14170405
Snake venom metalloproteinase	scaffold-mi1	14174872	14190142
Snake venom metalloproteinase	scaffold-mi1	14211673	14242249
Snake venom metalloproteinase	scaffold-mi1	14248933	14272689
Snake venom metalloproteinase	scaffold-mi1	14281564	14300774
Snake venom metalloproteinase	scaffold-mi1	14368422	14393313
Snake venom metalloproteinase	scaffold-mi1	14401627	14424637
Snake venom metalloproteinase	scaffold-mi1	14310844	14338336
Veficolin/Ficolin	scaffold-mi7	5271880	5282014
Veficolin/Ficolin	scaffold-ma3	179788950	179790745
Veficolin/Ficolin	scaffold-ma1	232337083	232340714
Veficolin/Ficolin	scaffold-ma1	232312034	232335439
Vascular endothelial growth factor	scaffold-ma7	40288572	40327884
Vascular endothelial growth factor	scaffold-ma1	40733075	40747358
Vascular endothelial growth factor	scaffold-ma1	260248287	260272500
Venom Factor	scaffold-Z	79798672	79803249
Venom Factor	scaffold-Z	79749464	79761456

Venom Factor	scaffold-ma2	1573588	1616446
Venom Factor	scaffold-ma2	137559964	137560374
Venom Factor	scaffold-ma2	137553669	137558461
Venom Factor	scaffold-ma2	137623562	137648584
Venom Factor	scaffold-ma2	137651285	137653877
Venom Factor	scaffold-ma2	137710627	137728987
Venom Factor	scaffold-ma2	137753804	137775039
Venom Factor	scaffold-ma2	137735629	137741352
Vespryn/Ohanin	scaffold-ma2	4377779	4385668
Vespryn/Ohanin	scaffold-ma2	109834300	109838076
Waprin	scaffold-ma1	204655764	204666466

References

Agrawal AA, Laforsch C, Tollrian R (1999) Transgenerational induction of defenses in plants and animals. *Nature,* 401: 60-63.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics,* 25: 25-29.

Aubin-Horth N, Renn SPC (2009). Genomic reaction norms: using integrative biology to understand molecular mechanisms of phenotypic plasticity. *Molecular Ecology,* 18: 3763-3780.

Bashey F (2006). Cross-generational environmental effects and the evolution of offspring size in the Trinidadian guppy *Poecilia reticulata*. *Evolution*, 60: 348-361.

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

Bonduriansky R, Crean AJ, Day T (2012). The implications of nongenetic inheritance for evolution changing environments. *Evolutionary Applications,* 5: 192-201.

Bossdorf O, Richards CL, Pigliucci M (2008). Epigenetics for ecologists. *Ecology Letters*, 11: 106-115.

Boyko A, Belvins T, Yao Y, Golubov A, Bilichak A, Ilnytskyy Y, Hollander J, Meins F Jr., Kovalchuk I (2010). Transgenerational adaptation of Arabidopsis to stress requires DNA methylation and the function of Dicer-like proteins. *PLOS ONE,* 5.

Braunstein M, Rose, AB, Holmes, SG, Allis CD, Broach JR (1993). Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes and Development,* 7: 592-604.

Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, Bock C, Li C, Gu H, Zamore PD, Meissner A, Weng Z, Hofmann HA, Friedman N, Rando OJ (2010). Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell,* 143: 1084-1096.

Carpenter SR, Fisher SG, Grimm NB, Kitchell JF (1992). Global change and freshwater ecosystems. *Annual Review of Ecology and Systematics*, 23: 119-139.

Charmantier A, McCleery RH, Cole LR, Perrins C, E. L, Kruuk B, Sheldon BC (2008). Adaptive phenotypic plasticity in response to climate change in a wild bird population. *Science* 320: 800-803.

Roy Chowdhury P, Frisch D, Becker D, Lopez JA, Weider LJ, Colbourne JK, Jeyasingh PD (2015). Differential transcriptomic responses of ancient and modern *Daphnia* genotypes to phosphorus supply. *Molecular Ecology,* 24: 123-135.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Cáceres CE, Carmel L, Casola C, Choi J-H, Detter JC, Dong Q, Dusheyko S, Eads BD, Fröhlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kültz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL (2011). The Ecoresponsive genome of *Daphnia pulex*. *Science,* 331: 555-561.

Conesa A, Götz S (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008: 1-13.

Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21: 3674-3676.

Dall SRX, Giraldeau L-A, Olsson O, McNamara JM, Stephens DW (2005). Information and its use by animals in evolutionary ecology. *Trends in Ecology & Evolution,* 20: 187-193.

Day T, Bonduriansky R (2011). A unified approach to the evolutionary consequences of genetic and nongenetic inheritance. *American Naturalist*, 178: E18-E36.

Dixon P (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14: 927-930.

Donohue K, Schmitt J (1998). Maternal environmental effects in plants: adaptive plasticity? Oxford, UK: Oxford University Press.

Dyer AR, Brown CS, Espeland EK, McKay JK, Meimberg H, Rice KJ (2010). SYNTHESIS: The role of adaptive trans-generational palsticity in biological invasions of plants. *Evolutionary Applications,* 3: 179-192.

von Elert E, Agrawal MK, C. G, Jaensch H, Bauer U, Zitt A (2004). Protease activity in gut of *Daphnia magna*: evidence for trypsin and chymotrypsin enzymes. *Comparative Biochemistry and Physiology*, 137: 287-296.

Ezard THG, Prizak R, Hoyle RB (2014). The fitness costs of adaptation via phenotypic plasticity and maternal effects. *Functional Ecology*, 28: 693-701.

Fischer B, Taborsky B, Kokko H (2011). How to balance the offspring quality–quantity tradeoff when environmental cues are unreliable. Oikos: 258-270.

Fox CW, Mousseau TA (1998). Maternal effects as adaptations for transgenerational phenotypic plasticity in insects. Maternal Effects as Adaptations: 159-177.

Galloway LF (2005). Maternal effects provide phenotypic adaptation to local environmental conditions. *New Phytoligist,* 166: 93-100.

Galloway LF (2009). Plasticity to canopy shade in a monocarpic herb: within-and between-generation effects. *New Phytologist,* 182: 1003-1012.

Galloway LF, Etterson JR (2007). Transgenerational plasticity Is adaptive in the wild. *Science*, 318: 1134-1136.

Schröder T, Gilbert JJ (2004). Transgenerational plasticity for sexual reproduction and diapause in the life cycle of monogonont rotifers: intraclonal, intraspecific and interspecific variation in the response to crowding. *Functional Ecology*, 18: 458-466.

Götz S, Arnold R, Sebastián-León P, Martín-Rodríguez S, Tischler P, Jehl MA, Dopazo J, Rattei T, Conesa A (2011). B2G-FAR, a species centered GO annotation repository. *Bioinformatics*, 27: 919-924.

Harris KDM, Bartlet NJ, Lloyd VK (2012). *Daphnia* as an emerging epigenetic model organism. *Genetics Research International* 2012.

Herman JJ, Spencer HG, Donohue K, Sultan SE (2014). How stable 'should' epigenetic modifications be? Insights from adaptive plasticity and bet hedging. *Evolution,* 68: 632-643.

Herman JJ, Sultan SE (2011). Adaptive transgenerational plasticity in plants: case studies, mechanisms, and implications for natural populations. *Frontiers in Plant Science,* 6: 1-10.

Herrera CM, Bazaga P (2010). Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet *Viola cazorlensis*. *New Phytologist*, 187: 867-876.

Herrera CM, Pozo MI, Bazaga P (2012). Jack of all nectars, master of most: DNA methylation and the epigenetic basis of niche width in a flower-living yeast. *Molecular Ecology*, 21: 2602-2616.

Hoyle RB, Ezard THG (2012). The benefits of maternal effects in novel and in stable environments. *Journal of the Royal Society Interface*, 9: 2403-2413.

Jablonka E, Lachmann M, Lamb MJ (1992). Evidence, mechanisms and models for the inheritance of acquired characters. J*ournal of Theoretical Biology*, 158: 245-268.

Jablonka E, Lachmann M, Lamb MJ (1989). The inheritance of acquired epigenetic variations. *Journal of Theoretical Biology,* 1989: 69-83.

Jablonka E, Oborny B, Molnar I, Kisdi E, Hofbauer J, Czaran T (1995). The adaptive advantage of phenotypic memory in changing environments. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 350: 133-141.

Jablonka E, Raz G (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology*, 84: 131-176.

Jaenisch R, Bird A (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics* 33: 245-254.

Kalisz S, Purugganan MD (2004). Epialleles via DNA methylation: consequences for plant evolution. *Trends in Ecology and Evololution,* 19: 309-314.

Kilham SS, Kreeger DA, Lynn SG, Goulden CE, Herrera L (1998). COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia,* 377: 147-159.

Kinsella R, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P (2011). Ensembl BioMarts: a hub for data retrieval across the taxonomic space. *Database,* 2011.

Kooke R, Johannes F, Wardenaar R, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJ (2015). Epigenetic basis for morphological variation and phenotypic plasticity is *Arabidopsis thaliana*. *Plant Cell*, 27: 337-348.

Kuijper B, Hoyle RB (2015). When to rely on maternal effects and when on phenotypic plasticity? *Evolution,* 69: 950-968.

Kuijper B, Johnstone RA, Townley S (2014). The evolution of multivariate maternal effects. *PLoS Computational Biology*, 10: e1003550.

Laforsch C, Beccara L, Tollrian R (2006). Inducible defenses: The relevance of chemical alarm cues in *Daphnia*. *Limnology and Oceanography*, 51: 1466-1472.

Law JA, Jacobsen SE (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature,* 11: 204-220.

Leimar O, McNamara JM (2015). The evolution of transgenerational integration of information in heterogeneous environments. *American Naturalist*, 185: E55-E69.

Levins R (1968). Evolution in changing environments. Princeton, NJ: Princeton University Press.

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 (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25: 2078-2079.

Lin SM, Galloway LF (2010). Environmental context determines within-and potential between-generation consequences of herbivory. *Oecologia,* 163: 911-920.

Marshall DJ (2008). Transgenerational plasticity in the sea: context-dependent maternal effects across the life history. *Ecology,* 89: 418-427.

Merzendorfer H, Zimoch, L (2003). Chitin metabolism in insects: structure, function and regulation of chitin synthases and chitinases. *The Journal of Experimental Biology*, 206: 4393-4412.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD (2016). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research,* 41.

Miner BE, De Meester L, Pfrender ME, Lampert W, Hairston NG (2012). Linking genes to communities and ecosystems: *Daphnia* as an ecogenomic model. *Proceedings of the Royal Society of London B: Biological Sciences,* 279: 1873-1882.

Miyakawa H, Imai M, Sugimoto N, Ishikawa Y, Ishikawa A, Ishigaki H, Okada Y, Miyazaki S, Koshikawa S, Cornette R, Miura T (2010). Gene up-regulation in response to predator kairomones in the water flea, *Daphnia pulex*. *BMC Developmental Biololgy*, 10: 45.

Molinier J, Ries G, Zipfel C, Hohn B (2006). Transgeneration memory of stress in plants. *Nature,* 442: 1046-1049.

Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11: R25.

Otte KA, Schrank I, Frohlich T, Arnold GJ, Laforsch C (2015). Interclonal proteomic responses to predator exposure in *Daphnia magna* may depend on predator composition of habitats. *Molecular Ecology,* 24: 3901-3917.

Post DM, Palkovacs EP, Schielke EG, Dodson SI (2008). Intraspecific variation in a predator affects community structure and cascading trophic interactions. *Ecology,* 89: 2019-2032.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Astralia.

Riessen HP (1999). Predator-induced life history shifts in *Daphnia:* a synthesis of studies using meta-analysis. *Canadian Journal of Fisheries and Acquatic Sciences*, 56: 2487-2494.

Robichaud NF, Sassine J, Beaton MJ, Lloyd VK (2012). The epigenetic repertoire of *Daphnia magna* includes modified histones. *Genetics Research International,* 2012.

Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26: 139-140.

Rozenberg A, Parida M, Leese F, Weiss LC, Tollrian R, Manak JR (2015). Transcriptional profiling of predator-induced phenotypic plasticity in *Daphnia pulex*. *Frontiers in Zoology*, 12.

Salinas S, Munch SB (2012). Thermal legacies: transgenerational effects of temperature on growth in a vertebrate. *Ecology Letters*, 15: 159-163.

Schield DR, Walsh MR, Card DC, Andrew AL, Adams RH, Castoe TA (2016). EpiRADseq: scalable analysis of genomewide patterns of methylation using next-generation sequencing. *Methods in Ecology and Evolution*, 7: 60-69.

Schmitt J, Niles J, Wulff RD (1992). Norms of reaction of seed traits to maternal environments in *Plantago lanceolata*. *American Naturalist*, 139: 451-466.

Schwarzenberger A, Courts C, von Elert E (2009). Target gene approaches: gene expression in *Daphnia magna* exposed to predator-borne kairomones or to microcystin-producing and microcystin-free *Microcysts aeruginosa*. *BioMed Central Genomics*, 10: 527-541.

Schwerin S, Zeis B, Lamkemeyer T, Paul RJ, Koch M, Madlung J, Fladerer C, Pirow R (2009). Acclimatory responses of the *Daphnia pulex* proteome to environmental changes. II. Chronic exposure to different temperatures (10 and 20°C) mainly affects protein metabolism. *BMC Physiology*, 9: 8.

Shea N, Pen I, Uller T (2011). Three epigenetic information channels and their different roles in evolution. *Journal of Evolutionary Biology,* 24: 1178-1187.

Simons AM (2011). Adaptive transgenerational plasticity in plants: case studies, mechanisms, and implications for natural populations. *Frontiers Plant Science,* 2: 1-10.

Soetaert A, Vandenbrouck T, van der Ven K, Maras M, van Remortel P, Blust R, De Coena WM (2007). Molecular responses during cadmium-induced stress in *Daphnia magna*:Integration of differential gene expression with higher-level effects. *Aquatic Toxicology*, 83: 212-222.

Stibor H (1992). Predator induced life-history shifts in freshwater cladoceran. *Oecologia*, 92: 162-165.

Stollewerk A (2010). The water flea *Daphnia*- a ‘new’ model system for ecology and evolution? *Journal of Biology,* 9.

Sultan SE, Barton K, Wilczek AM (2009). Contrasting patterns of transgenerational plasticity in ecologically distinct congeners. *Ecology,* 90: 1831-1839.

Sun J, Li C, Wang S (2015). The up-regulation of ribosomal proteins further regulates protein expression profile in female *Schistosoma japonicum* after pairing. *PLoS,* 10: e0129626.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011). REVIGO summarizes and visualizes long lists of Gene Ontology Terms. *PLoS,* 6.

Tollrian R, Harvell DC (1999). The Ecology and Evolution of Inducible Defenses: Princeton University Press.

Tollrian R, Leese F (2010). Ecological genomics: steps towards unraveling the genetic basis of inducible defenses in *Daphnia*. *BMC Biology*, 5.

Turck F, Coupland G (2014). Natural variation in epigenetic gene regulation and its effects on plant developmental traits. *Evolution,* 68: 620-631.

Uller T (2008). Developmental plasticity and the evolution of parental effects. *Trends Ecology Evololution,* 23: 432-438.

Uller T, English S, Pen I (2015)a. The evolution of transgenerational integration of information in heterogeneous environments. *Evolutionary Applications,* 185: 179-192.

Uller T, English S, Pen I (2015)b. When is incomplete epigenetic resetting in germ cells favoured by natural selection? *Proceedings of the Royal Society of London B: Biological Sciences,* 282: 20150682.

Vandegehuchte MB, Coninck DD, Vandenbrouck T, Coen WMD, Janssen CR (2010). Gene transcription profiles, global DNA methylation and potential transgenerational epigenetic effects related to Zn exposure history in *Daphnia magna*. *Environmental Pollution*, 158: 3323-3329.

Vandegehuchte MB, Janssen CR (2011). Epigenetics and its implications for ecotoxicology. *Ecotoxicology,* 20: 607-624.

Walsh MR, Castoe TA, Holmes J, Packer M, Biles K, Walsh MJ, Munch SB, Post DM (2016). Local adaptation in transgenerational responses to prey. *Proceedings of the Royal Society of London B: Biological Sciences,* 283: 20152271.

Walsh MR, Cooley F, Biles K, Munch SB (2015). Predator-induced phenotypic plasticity within- and across-generations: a challenge for theory? *Proceedings of the Royal Society B-Biological Sciences*, 282: 20142205.

Walsh MR, Post DM (2012). The impact of intraspecific variation in a fish predator on the evolution of phenotypic plasticity and investment in sex in *Daphnia ambigua*. *Journal of Evolutionary Biology*, 25: 80-89.

Walsh MR, Post DM (2011). Interpopulation variation in a fish predator drives evolutionary divergence in prey in lakes. *Proceedings of the Royal Society B-Biological Sciences*, 278: 2628-2637.

Wang J, Lan P, Gao H, Zheng L, Li W, Schmidt W (2013). Expression changes of ribosomal proteins in phosphate- and iron-deficient *Arabidopsis* roots predict stress-specific alterations in ribosome composition. *BMC Genomics*, 14: 783.

Zhang Y, Fischer M, Colot V, Bossdorf O (2013). Epigentic variation creates potential for evolution of plant phenotypic plasticity. *New Phytologist*, 197: 314-322.

Zhou X, Liao W-J, Liao J-M, Liao P, Lu H (2015). Ribosomal proteins: functions beyond the ribosome. *Journal of Molecular Cell Biology*, 7: 92-104.

Alfoldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477(7366): 587-591.

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(4): 485-495.

Baker RJ, Bull JJ, Mengden GA. 1972. Karyotypic Studies of 38 Species of North-American Snakes. *Copeia* (2): 257-&.

Beckstette M, Homann R, Giegerich R, Kurtz S. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* 7(1): 389.

Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat Genet* 49(3): 387-394.

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(4): e1003537.

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J et al. 2015. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* 48: 427-437.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2(1): 10.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18(1): 188-196.

Casewell NR, Harrison RA, Wüster W, Wagstaff SC. 2009. Comparative venom gland transcriptome surveys of the saw-scaled vipers (Viperidae: *Echis*) reveal substantial intra-family gene diversity and novel venom transcripts. *BMC Genomics* 10(1): 564.

Casewell NR, Huttley GA, Wuster W. 2012. Dynamic evolution of venom proteins in squamate reptiles. *Nat Commun* 3: 1066.

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. *P Natl Acad Sci USA* 110(51): 20645-20650.

Cohn MJ, Tickle C. 1999. Developmental basis of limblessness and axial patterning in snakes. *Nature* 399(6735): 474-479.

Darrow EM, Huntley MH, Dudchenko O, Stamenova EK, Durand NC, Sun Z, Huang S-C, Sanborn AL, Machol I, Shamim M. 2016. Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. *P Natl Acad Sci USA*: 201609643.

Dixon JR, Gorkin DU, Ren B. 2016. Chromatin domains: the unit of chromosome organization. *Mol Cell* 62(5): 668-680.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3(1): 95-98.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Ann Rev Genomics Hum Genet* 10: 285-311.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5): 1792-1797.

Fane M, Harris L, Smith AG, Piper M. 2017. Nuclear factor one transcription factors as epigenetic regulators in cancer. *Int J Cancer* 140(12): 2634-2641.

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

Gamble T, Castoe TA, Nielsen SV, Banks JL, Card DC, Schield DR, Schuett GW, Booth W. 2017. The Discovery of XY Sex Chromosomes in a Boa and Python. *Curr Biol : CB* 27(14): 2148-2153 e2144.

Geffeney S, Brodie ED, Ruben PC. 2002. Mechanisms of adaptation in a predator-prey arms race: TTX-resistant sodium channels. *Science* 297(5585): 1336-1339.

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(7): 644.

Graves JA. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. *Nat Rev Genet* 17(1): 33-46.

Gronostajski RM. 2000. Roles of the *NFI*/*CTF* gene family in transcription and development. *Gene* 249(1): 31-45.

Harris RS. 2007. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.

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(7018): 695-716.

Ikeda N, Chijiwa T, Matsubara K, Oda-Ueda N, Hattori S, Matsuda Y, Ohno M. 2010. Unique structural characteristics and evolution of a cluster of venom phospholipase A 2 isozyme genes of *Protobothrops flavoviridis* snake. *Gene* 461(1): 15-25.

Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schütz F, Daish T, Grützner F, Kaessmann H. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biology* 10(5): e1001328.

Kerchove CM, Luna MSA, Zablith MB, Lazari MFM, Smaili SS, Yamanouye N. 2008. α1-adrenoceptors trigger the snake venom production cycle in secretory cells by activating phosphatidylinositol 4, 5-bisphosphate hydrolysis and ERK signaling pathway. *Comp Biochem Physiol A Mol Integr Physiol* 150(4): 431-437.

Kim M, McGinnis W. 2011. Phosphorylation of Grainy head by ERK is essential for wound-dependent regeneration but not for development of an epidermal barrier. *P Natl Acad Sci USA* 108(2): 650-655.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14): 1754-1760.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950): 289-293.

Mackessy SP. 2008. Venom composition in rattlesnakes: trends and biological significance. In *The Biology of Rattlesnakes*, (ed. WK Hayes, KR Beaman, MD Cardwell, SP Bush). Loma Linda University Press, Loma Linda, CA.

Mackessy SP. 2010. The field of reptile toxinology. In *Handbook of venoms and toxins of reptiles*, (ed. SP Mackessy). CRC Press, New York, NY.

Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Brüning T, Mössinger K. 2017. Convergent origination of a *Drosophila*-like dosage compensation mechanism in a reptile lineage. *Genome Res* 27(12): 1974-1987.

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. *P Natl Acad Sci USA* 103(48): 18190-18195.

O'Connor RE, Romanov MN, Kiazim LG, Barrett PM, Farré M, Damas J, Furguson-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.

Olmo E. 2005. Rate of chromosome changes and speciation in reptiles. *Genetica* 125(3): 185-203.

Organ CL, Godínez Moreno R, Edwards SV. 2008. Three tiers of genome evolution in reptiles. *Integr Comp Biol* 48(4): 494-504.

Pasquesi GI, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, Reyes-Velasco J, Ruggiero RP, Vandewege MW, Shortt JA et al. In Press. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat Commun* 9: 2774.

Perry BW, Card DC, McGlothlin JW, Pasquesi GI, Hales NR, Corbin AB, Adams RH, Schield DR, Fujita MK, Demuth JP et al. In Review. Molecular adaptations for sensing and securing prey, and insight into amniote genome diversity, revealed by the garter snake genome.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 26: 1-9.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* 9(1): 189.

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(7): 1665-1680.

Rice ES, Kohno S, John JS, Pham S, Howard J, Lareau LF, O'Connell BL, Hickey G, Armstrong J, Deran A et al. 2017. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Res* 27(5): 686-696.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11(3): R25.

Rokyta DR, Lemmon AR, Margres MJ, Aronow K. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics* 13: 312.

Saviola AJ, Pla D, Sanz L, Castoe TA, Calvete JJ, Mackessy SP. 2015. Comparative venomics of the Prairie Rattlesnake (*Crotalus viridis viridis*) from Colorado: Identification of a novel pattern of ontogenetic changes in venom composition and assessment of the immunoreactivity of the commercial antivenom CroFab®. *J Proteomics* 121: 28-43.

Secor S, Diamond J. 1998. A vertebrate model of extreme physiological regulation. *Nature* 395: 659-662.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19): 3210-3212.

Smeds L, Kawakami T, Burri R, Bolivar P, Husby A, Qvarnstrom A, Uebbing S, Ellegren H. 2014. Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. *Nat Commun* 5: 5448.

Smit AF, Hubley R. 2015. RepeatModeler Open 1.0.

Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. *Institute for Systems Biology http://repeatmasker org*.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* 7(1): S10.

Srikulnath K, Nishida C, Matsubara K, Uno Y, Thongpan A, Suputtitada S, Apisitwanich S, Matsuda Y. 2009. Karyotypic evolution in squamate reptiles: comparative gene mapping revealed highly conserved linkage homology between the butterfly lizard (*Leiolepis reevesii rubritaeniata*, Agamidae, Lacertilia) and the Japanese four-striped rat snake (*Elaphe quadrivirgata*, Colubridae, Serpentes). *Chromosome Res* 17(8): 975-986.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33(suppl_2): W465-W467.

Ting SB, Caddy J, Hislop N, Wilanowski T, Auden A, Zhao L-l, Ellis S, Kaur P, Uchida Y, Holleran WM. 2005. A homolog of *Drosophila* grainy head is essential for epidermal integrity in mice. *Science* 308(5720): 411-413.

Vicoso B, Emerson JJ, Zektser Y, Mahajan S, Bachtrog D. 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biology* 11(8): e1001643.

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. *P Natl Acad Sci USA* 110(51): 20651-20656.

Voss SR, Kump DK, Putta S, Pauly N, Reynolds A, Henry R, 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.

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(7289): 757-762.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol* 15(12): 549.

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 J. 2016. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4,162 species. *Mol Phylogenetics Evol* 94: 537-547.

Chapter 4

Transmission patterns, relatedness, and genetic diversity inferred from whole genome resequencing of archival blood fluke miracidia (*Schistosoma japonicum*).

Zachary L. Nikolakis[1,*], Nicole R. Hales[1,*], Drew R. Schield[1], Blair W. Perry[1], Laura E. Timm[4],Andrea Buchwald[2], Elise Grover[2], Yang Liu[3], Bo Zhong[3], Laura E. Timm[4], Elizabeth J. Carlton[2], and David D. Pollock[4], and Todd A. Castoe[1, §]

[1]Department of Biology, 501 S. Nedderman Drive, University of Texas at Arlington, Arlington, TX 76019 USA

[2]Department of Environmental and Occupational Health, University of Colorado, Colorado School of Public Health, Aurora, CO 80045, USA

[3]Institute of Parasitic Disease, Sichuan Center for Disease Control and Prevention, Chengdu, The People's Republic of China

[4]Department of Biochemistry & Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

Abstract

Schistosomiasis is a neglected tropical disease caused by helminths within the genus *Schistosoma* that affects an estimated 200 million people worldwide. Genomic approaches hold great promise for understanding detailed patterns of schistosome transmission and regional persistence that could inform strategic control measures. However, the cost of collecting genomic data together with the challenges associated with obtaining sufficient DNA from individual schistosome miricidia – the only readily available life stage of schistosomes – have limited the application of population genomic data for studying schistosome disease transmission. Here we leverage the decreasing costs of genome resequencing together with whole genome amplification to demonstrate the feasibility and utility of using whole genome sequencing (WGS) methods to study detailed patterns of schistosome infection and transmission. We performed WGS on 22 miracidial samples from 10 infected hosts across 2 villages in Sichuan, China. From these data, we analyzed patterns of genetic diversity and relatedness among samples that provide new insight into patterns of transmission, parasite diversity, and potential drivers of regional persistence. Our results broadly demonstrate that true population-level genomic analyses of schistosomes is now feasible and holds great potential for providing informative and actionable insight into transmission that can inform control measures.

Introduction

Human infection by parasitic helminthiases affect an estimated 1 billion people globally (Hampton, 2012; Hotez, Fenwick, Savioli, & Molyneux, 2009; WHO, 2012), and the prevalence of these diseases are concentrated in low and middle-income countries, especially within regions of sub-Sahara Africa and southeast Asia (Barry, Bezek, Serpa, Hotez, & Woc-Colburn, 2012; Hotez et al., 2009). Schistosomiasis is a neglected tropical disease caused by helminths within the genus *Schistosoma* and alone affects an estimated 200 million people globally, causing fibrosis of the liver and bladder, anemia and, in the case of *S. haematobium*, cancer (Bouvard et al., 2009; Friedman, Kanzaria, & McGarvey, 2005; Gryseels, Polman, Clerinx, & Kestens, 2006). Adult schistosome worms live in mammalian hosts, mate, and shed eggs that are excreted in stool (*S. japonicum* and *S. mansoni*) or urine (*S. haematobium*)(Gryseels et al., 2006). Excreted eggs then hatch into miracidia in fresh water and must

infect a snail host before maturing into cercariae, the clonal, larval life stage infectious to mammalian hosts. The World Health Organization (WHO) has recently made efforts to improve drug distribution for the millions of people who lack access to antihelminthic treatment (Hotez, Engels, Fenwick, & Savioli, 2010; Schistosomiasis, 2015) with the goal of moving towards regional elimination of schistosomiasis (Boatin et al., 2012; Bockarie, Kelly-Hope, Rebollo, & Molyneux, 2013; L. Wang, Utzinger, & Zhou, 2008).

Among regions impacted by schistosomiasis, China has emerged as a global example of success for reducing schistosomiasis prevalence. Indeed, schistosomiasis in China is illustrative of both the opportunities and challenges in controlling human helminthiases globally. Control programs for this disease in China were initiated in the 1950s, when there were approximately 12 million documented cases (Xu et al., 2016). Since then, Chinese control efforts to reduce schistosomiasis prevalence and transmissions through a multi-pronged approach including antihelmenthic treatment, reduction of snail populations, and improvements to water and sanitation infrastructure have been incredibly effective (Liang et al., 2014; L.-D. Wang et al., 2009), reducing schistosomiasis prevalence by 99% (Lei et al., 2015). Despite these highly effective control efforts, regional pockets of transmission persist, and the disease has reemerged in previously controlled areas. An example of this pattern is Sichuan Province, China, where over the past decade researchers have documented the reemergence and persistence of schistosomiasis infections despite ongoing aggressive disease control programs targeting these and other regions of the country (Carlton, Bates, Zhong, Seto, & Spear, 2011; Carlton, Hubbard, Wang, & Spear, 2013; Liang, Yang, Zhong, & Qiu, 2006). These patterns of persistence and reemergence suggest that while aggressive control measures were highly effective at reducing disease prevalence, they are insufficient for accomplishing regional eradication of schistosomiasis.

Key steps towards understanding patterns of schistosomiasis persistence and reemergence that could inform elimination strategies include the development of detailed information on patterns of recent and ongoing parasite transmission, and an understanding of how parasites might be responding to control efforts. Previous studies have used small-scale genotyping approaches (e.g., microsatellite genotyping) to study parasite population structure and transmission (e.g., Prugnolle et al., 2005; Rudge et al., 2009; Barbosa et al., 2013; Gower et al., 2013), although the low numbers of genetic markers available for such

studies limit resolution of transmission patterns. More recent studies have used higher-resolution reduced-representation genomic methods (i.e., double-digest RAD-seq; Peterson et al., 2012) to infer parasite population genetic structure and transmission patterns (Shortt et al., 2017), although even these higher-resolution data have limitations for inferring patterns of relatedness, they fail to provide information on functional genomic diversity (e.g., protein-coding variation) that may be important for understanding parasite responses to control-driven selection.

As genome sequencing costs continue to decrease, the increasing feasibility of whole genome sequencing (WGS) of schistosomes provides an opportunity to estimate genetic diversity, relatedness, and population structure at a much higher resolution and with greater certainty. Furthermore, WGS data also enables analyses of functional regions of the genome and can identify patterns associated with control-driven selection on parasite populations, and potentially detect genomic evidence of emerging drug resistance. To date, however, population level whole genome sequencing efforts have been limited to analyses of pooled *S. japonicum* samples to understand broad regional patterns of *S. japonicum* genomic diversity. While such pooled-sample studies lack the ability to provide much needed insight into detailed transmission patterns, the pooling of samples was previously required due to prior higher costs of WGS, combined with the limited DNA yield of individual schistosome miracidia – the only readily available life stage of schistosomes that can be harvested from host stool samples (Shortt et al., 2017). However, recent studies have demonstrated that individual miracidia can be effectively used for large-scale genotyping through the use of whole genome amplification (WGA), thereby removing the limitation imposed by the low yield of single miracidial samples (Shortt et al., 2017).

In this study we demonstrate the feasibility and utility of using WGA and WGS of single archived *S. japonicum* miracidia and apply this approach to study a set of samples collected from human hosts in Sichuan, China. Using our WGS dataset, we discern genetic structure and patterns of relatedness within and among hosts and adjacent villages. We then apply these results to address key questions about parasite biology and epidemiology relevant to understanding transmission patterns, and next-step priorities for control measures. We also use these data to estimate how future larger-scale studies may feasibly sample greater numbers of individuals in economical ways to fully leverage full population-scale

genomic investigations of schistosome parasites, which could provide valuable insight into patterns of transmission and control-driven selection hold that would transform the efficacy of current and future control and elimination efforts.

Materials and Methods

Sample Collection

A total of 22 archived field-collected samples of *S. japonicum* miracidia were obtained from infected humans in Suchuan Province, China in 2016 using methods described elsewhere (Carlton et al., 2013; Xiao et al., 2013). Individual miracidia were collected from 2 villages, denoted as Village A & B, from 9 individual human hosts in village A and one host in village B. Briefly, participants were tested for infection using the miracidia hatching test and miracids were collected from positive hatching tests. Miracidia were collected from the top of the hatching test flask, isolated using a hematocrit tube or Pasteur pipette drawn to a narrow bore with a flame, washed three times with autoclaved deionized water and placed on a Whatman FTA indicating card (GE) for long-term storage. After drying, cards were stored in a desiccator at room temperature.

Whole Genome Amplification and Sequencing

Due to the limited about of DNA available from a single miracidia, whole genome amplification (WGA) was used, similar to previous studies that have used this approach to conduct reduced genome representation libraries (Shortt et al., 2017). Individual miracidia were extracted from Whatman cards using a Whatman Harris 2mm micro-core punch (Whatman; cat. WB100029). Following excision, punches underwent five consecutive 5-minute washes with 200uL TE buffer. After the final wash, punches were left to dry for at least one hour at room temperature. DNA from miracids was amplified directly from the punch using the Illustra Ready-To-Go GenomiPhi V3 DNA Amplification Kit (GE Healthcare; cat. 25-6601-96) following the manufacturer's recommended protocol for amplification with minor adjustments made to accommodate amplification from a 2mm disk. Specifically, dried disks were transferred to an amplification tube containing 20uL of 1x denaturation buffer. Tubes were incubated for 95°C for 3 minutes and then immediately placed on ice. Liquid from the tube was then added to individual

amplification pellets provided in the kit and allowed to dissolve the pellet for 10min on ice. After gentle mixing, the liquid was transferred back to its original tube with the 2mm disk still present, and each amplification tube was then subjected to 90 minutes of amplification at 30°C, followed by enzymatic heat kill at 65°C for 10 minutes, and ended with a hold at 4°C. WGA samples were stored at -20°C until ready for library preparation. 200ng of WGA DNA was used to construct libraries using the KAPA HyperPlus kit (KAPA Biosystems; cat. KK8514). Each sample was prepared using half-reactions and were uniquely indexed using IDT for Illumina TruSeq UD Indexes (Illumina; cat. 20022370). All 22 individual libraries were multiplexed into a single, pooled library and sequenced on a single Illumina NovaSeq lane using 150bp paired-end sequencing.

Sequencing Read Processing, Mapping, and Variant Calling

Whole genome sequencing libraries were demultiplexed using the FASTQ Generation application available on the Ilumina BaseSpace Sequence Hub (basespace.illimina.com) and paired reads were quality trimmed using Trimmomatic v0.36 (Bolger, Lohse, & Usadel, 2014) with the following options:LEADING:20 TRAILING:20 MINLEN:75 AVGQUAL:20. Trimmed reads were mapped to the *S. japonicum* reference genome (ASM636876v1; downloaded from https://www.ncbi.nlm.nih.gov/) using default parameters in BWA (Li & Durbin, 2009) and reads were sorted using SAMtools for downstream analysis (Li et al., 2009). We called variants using Bcftools (Li, 2011) and recorded all raw variant calls for all individuals with a read depth of less than five as missing data. We filtered variants to include only biallelic SNPs with a minor allele frequency (MAF) greater than 0.05 and that contained data from at least 80% of samples between the two villages. We further subdivided SNPs according to genomic regions (i.e., introns, exons, & intergenic.

Population genomic analyses, rare allele sharing and posterior estimates of relatedness

We performed a principle components analysis (PCA) using the R package 'adegenet' to visualize the spatial distribution of genetic variance between all individuals for exonic SNPs. Using the

same dataset we also inferred relatedness among all samples by constructing a neighbor-joining tree using the average pairwise distance in the R package 'ape'(Paradis & Schliep, 2018).

To infer close order familial relationships, we estimated pairwise rare allele sharing (minor allele frequency (MAF) <0.1) between all pairs of miracidia using a custom perl script in which we randomly sampled a subset of 2,000 variants for 50 generations from our exonic dataset. We estimated the posterior probability of each degree of relatedness between every pair of miracidia using posterior probability distributions estimated in Shortt et al. (in review). In brief, the posterior probability distribution for each degree of relatedness was estimated by taking the average level of allele sharing between the most geographically distant individuals, $\hat{\mu}_{unrelated}$, and the mean and variance of allele sharing from clusters of 3 or more miracidia, $\hat{\mu}_{sibs}$ & $\hat{\sigma}_{sibs}$, from the same host with the proportion of rare allele shared being >= 0.30. For intermediate degrees of relatedness, means and variances were estimated by halving the distance from sibs to unrelated. Posterior probabilities were calculated assuming even prior probabilities for each degree of relatedness from siblings to 5th degree relatives assuming that allele sharing probabilities were distributed normally, i.e, $\sim N\left(\hat{\mu}_{degree}, \hat{\sigma}_{degree}\right)$.

Analysis of Coverage Across Individuals

To assess the distribution of mapped read coverage across individuals, we examined coverage within exonic regions using samtools depth on all sites, including sites with zero coverage, and pulled out exonic regions using bedtools intersect (Quinlan & Hall, 2010). Distributions of exonic coverage were plotted from each individual. We specifically focused on analyses of coverage within exon regions because *S. japonicum* genome contains a high fraction of repetitive DNA (~45%) that may lead to inaccurate inferences of read mapping and coverage, and we expect exonic sequences to be mostly single or low-copy sequences that should suffer the least amount of such mapping error.

To further examine how potential differences in coverage may obfuscate measures of genetic variation and heterozygosity, we analyzed a subset four individual and down sampled reads to estimate the impact sampling lower levels of coverage (35x, 30x, 20x, 15x, 10x, and 5x) may have on the accuracy

of variant calls and population genetic inferences. We used the methods described above to to call and filter variants for all genomic regions and calculated heterozygosity based upon matching coverage estimates using VCFTools (Danecek et al., 2011). We performed an ANOVA and used the Tukey Honest method in the statistical program (R Core Team, 2017)to test for significant differences between heterozygosity estimates inferred based on various levels of coverage.

Results

Genomic sequencing, mapping, and coverage

We sequenced a total of 22 whole genome amplified miracidia collected in 2016 that spanned two villages in Sichuan, China. On average, we recovered an average of 263M reads per individual with Q-scores > 20 (a Q-score of 20 indicates a 1 in 100 probability of an incorrect base call). The average number of mapped reads across all individuals was 228M, with 20 out of 22 individuals having > 90% mapped reads, and two individual samples with low coverage (<20%; **Table 1**). These two individuals with low coverage were subsequently excluded from all further downstream analyses, due to the expectation that they may contain a substantial amount of contaminated reads (e.g., host DNA) that may negatively impact our inferences. To assess read coverage distributions across the 20 remaining samples, we surveyed coverage within exonic regions (**Supp. Fig. 1**) and found the mean exon coverage across these individuals was 271x (see **Table 1** for additional information and summary statistics related to coverage across samples).

Analyses of genetic variation among samples

Using our exonic variant dataset of 439,722 SNPs, our broad estimate of relatedness among samples based on neighbor-joining clustering suggests that samples from each of the two villages are distantly related, and represent distinct genetic clusters (**Fig. 1**). This analysis also indicates that the relative genetic similarity of miracidia within hosts is variable. Within Village A (where we have samples from multiple hosts), we find that miracidia from the same host vary in relatedness from forming distinct host-specific clusters (e.g., host 9), to representing fairly divergent lineages within the larger Village A

cluster (e.g., host 2; **Fig. 1B**). The principle component analysis (PCA) of these same data indicates similar patterns that include the clear genetic differentiation between samples from each of the two villages – this between-village distinction underlies much of the variation in PC1, which represented 15% of the variation among samples. The second principle component (PC2; explaining 10% of the variation) separated samples within villages, and like the neighbor-joining tree, highlights distinctions among miracidia within some hosts (e.g., host 2) and the similarity among miracidia in other hosts (**Fig. 1C**).

Estimates of relatedness among miracidial samples

To infer degrees of relatedness among miracidial samples we first estimated patterns of allele sharing among all pairs of miracids sampled, based on the same exonic SNP dataset used above (**Fig. 2A**). To provide context to the distributions of allele sharing, we also labeled the X-axis of allele sharing with posterior probability distributions for various degrees of relatedness (**Fig. 2B**). Distributions of allele sharing suggests that relationships between miracidia within hosts vary widely, between 2nd degree (siblings) and 5th degree relatives (second cousins). Within Village A, where we sampled miracidia from multiple hosts, most relationships are 4th degree (first cousin-level) and 5th degree. Within Village B allele sharing, which represents comparisons within a single sampled host, we find only close 2nd and 3rd (avuncular/pibling) degree relationships. The distribution of allele sharing between villages indicates that more distant relationships link samples from the two adjacent villages, with most being 5th degree or greater (less related) (Fig. 2)

To better understand detailed patterns of allele sharing in the context of relatedness, we estimated and plotted posterior probabilities of discrete degrees of relatedness among all miracidial pairs sampled (**Fig. 3**). Estimates of relatedness based on posterior probabilities from both villages suggest that there are multiple sibling pairs, but all are restricted to within individual hosts, and no 2nd (or 3rd) degree relationships were inferred with high probability among hosts. In miracidia sampled from Village A, we find that within-host comparisons between miracids suggests that miracids are predominantly 4th or 5th degree relatives, suggesting surprisingly distant relatedness among miracidia produced by a single

host – this is in contrast to the high degree of relatedness observed in the host sampled from Village B. We also estimate that only 5th degree or more distant relationships likely link miracids between villages.

Estimation of the relationship between SNP calling accuracy and coverage

We assessed the feasibility and utility of sequencing individual miracidia whole genomes at relatively low-coverage and calling heterozygous sites by down sampling raw reads to mimic a range of different coverages (**Fig 4**). We find that measures of heterozygosity were mainly consistent with an expected trend of decreased called heterozygous sites as coverage was reduced. Differences between mimicked coverages did significantly differ until coverages fell between 5-10x as even samples at full coverage showed similar distributions as multiple other coverage ranges (**Table S2)**. Our Tukey multiple comparison test showed that samples at 15x coverage were not significantly different from any other coverage with the exception of our 5x dataset (p-value < 0.05) (**Table S2**) and that the distribution of heterozygosity does not differ between coverages at 15x or greater (p-value > 0.05). These results indicate that the ability to confidently call heterozygous SNPs rapidly declines between 10-15x coverage and as such, individuals sampled at a population scale would need to be at least 15x or greater to accurately infer genetic variation.

Discussion

From genetic markers to genomes

The application of single or multi-locus genetic markers in the context of epidemiology studies of *S. japonicum* has facilitated a broad understanding of transmission dynamics and patterns of overall parasite relatedness within regional areas (Shrivastava, Qian, Mcvean, & Webster, 2005). These studies have provided invaluable information regarding population level structuring across wide geographical regions. However, the recent use of larger genomic scale datasets has furthered our understanding of detailed population structuring of these parasites and provided some of the first insight into how schistosome populations may respond to selection pressures (Young et al., 2015). While these previous studies have been valuable for understanding relatively broad-scale patterns of genetic variation and selection, details of local transmission patterns, infection routes, and patterns of local and regional

persistence have not been addressed. This knowledge gap is largely the result of previous studies which have examined whole genome variation within *S. japonicum* doing so by utilized pooling methods (pooling multiple individuals), which reduces the ability to discern individual variation and limits inferences to relatively large geographic scales (i.e., provinces).

In addition to the costs of WGS, which continue to decline, a major barrier preventing individual whole genome sequencing of schistosomes that justified pooled-genome sequencing was the challenge of obtaining sufficient DNA for WGS from single schistosome individuals. Indeed, because the only readily-available life stage – miracidia, obtained from host stool samples – are very small and yield low quantities of DNA, PCR-based genetic markers (e.g., microsatellites) or pooled-WGS sampling were the only viable options. Recently, we demonstrated the feasibility of using whole genome amplification (WGA) on single archival miracidia, preserved on Whatman cards, to generate reduced representation genomic libraries of individual miracidia (Shortt et al., 2017). In this study, we leveraged this WGA approach, and the decreasing costs of sequencing, to demonstrate the feasibility of whole genome resequencing of individual archival miracidia and to illustrate the resolution that these data provide. This in turn provides valuable information regarding the potential for scaling up such individual schistosome WGS sampling to accomplish true population genomic analyses of schistosome parasites.

As a first step towards understanding the potentials, and the challenges, of large-scale population genomic sampling of schistosomes, in this study we sampled 22 individual miracidia from geographically close localities (~12 km). We also sampled each individual at a very high level of genomic coverage (mean exonic coverage >250x) – far higher than is typical for WGS genotyping studies (e.g., 10-40x). This design was chosen to illustrate the ability of WGS data to discern fine-scale patterns of relatedness and differentiation, and to enable us to down-sample our large datasets to infer how greater numbers of individuals could be sequenced at lower costs by reducing the amount of data per individual, enabling collection of larger sample sizes per cost.

When considering how much data is required per individual, it is important to appreciate that different types of inferences based on WGS data may depend more or less on the accuracy of genotyping inferences, and thus on coverage obtained per individual. As a simple indication of genotyping accuracy,

we compared inferred heterozygosity across individual samples based on full-coverage, which we assume is the approximate best case degree of accuracy, to heterozygosity estimates derived from much lower coverage. We also note that, while there is substantial empirical and theoretical literature available that predicts the relationships between genotype uncertainty and coverage and sequence quality(Davey et al., 2011; Nielsen, Paul, Albrechtsen, & Song, 2011), we conducted our own empirical estimates here because none of these previous studies incorporated WGA which has the potential to introduce additional error. We find that our estimates suggest that even at 15x there is minimal (and non-significant) reduction in the accuracy of estimation of heterozygosity – these results closely match estimates from the literature, and suggest that the additional WGA step in our approach does not introduce substantial levels of error. These estimates also suggest that per individual coverage as low as 15x yields highly accurate per-base genotype estimates. Furthermore, with whole genome data, even lower per-sample coverage may be far more than sufficient for many relevant types of inferences, including those of relatedness, population structure and genomic scans, which tend to be less sensitive to per-base genotyping errors (Martin et al., 2013). Collectively, our results highlight the potential to sequence individual whole genomes from miracidia on a population-level scale at low-moderate coverage at a cost that enables the inclusion of many samples without compromising the accuracy of a wide breadth of epidemiologically- and biologically-relevant inferences.

Whole genome comparisons highlight contrasting infection patterns at fine scales

Developing and understanding of patterns of schistomiasis infection in Sichuan, China is particularly motivating and valuable because this region represents a model for both success and the persistent challenges associated with end-game regional elimination of the disease. Control measures and success in this region are therefore far ahead of most other schistosome-impacted regions globally, and lessons learned about end-came control and elimination efforts will likely be relevant to other regions after they advance to this stage of control Schistosomiasis in this region has persisted despite ongoing disease control measures, and infection prevalence and intensity, as well as morbidity, has declined markedly in Sichuan and nationwide following the introduction of praziquantel in the 1990s and

complementary activities to control snail populations and promote schistosomiasis health education (Collins, Xu, & Tang, 2012; L. Wang et al., 2008; Xu et al., 2016). By 2004, prevalence was estimated to be <1% in most endemic counties in the province, placing them on a course for presumed interruption of transmission (Liang et al., 2006; Zhou et al., 2007). However, in 2004, reemergence was documented in 8 counties that had previously met transmission control targets (Liang et al., 2006), and by 2016 *S. japonicum* infection prevalence was as high as 27% (Carlton et al., 2013). Here we analyzed 20 miricidia, sampling multiple individual parasites from 9 human hosts from a single village, and an additional host from a second nearby village. This sampling was designed to illustrate patterns of parasite relatedness, and underlying patterns of transmission and diversity, both within and among human hosts, and assess how these patterns compared to a host from a different village.

Despite the restricted scale of our sampling in this study, our results highlight key features of schistosome diversity at fine scales that provide new insight into regional patterns of infection. Comparing patterns of diversity and relatedness of individual miricidia within hosts, we find substantial differences between hosts from different villages that suggest remarkably different infection dynamics occurring between villages. Miracidial samples within hosts from village A exhibit a pattern of elevated within-host diversity, with only a single host containing a sibling pair (shared-parentals), and two other hosts with individual miracidia inferred to be 3rd degree relatives (half-sibs) (**Fig 3b**); the remaining relationships among miracidia within hosts from village A showed 4th and 5th degree relationships (**Fig 3c-d**). In stark contrast, miracidia from the host from village B are all either 2nd or 3rd degree relatives. These results suggest that the number of infection events and overall schistosome population driving infection in village A is much higher, whereas a very small number of infection events can account for occurences in village B. While village B is only represented by a single host, the overall pattern of low degrees of relatedness within this single individual is in stark contrast compared to hosts in village A with similar parasite proportions. These higher degrees of relatedness (4th and 5th) within village A are likely a reflection relative to the overall number of infection opportunities pertaining to each locality, with village A in particular having a higher number of sources.

Our results also highlight substantial differences in the number of adult parental worms infecting a single host compared to all hosts between villages. For hosts from village A, sibling relationships among miricids from the same host were rarely seen, indicating that almost all miricids sampled within a host were derived from distinct parents, suggesting multiple adult worm infections per host. Consistent with findings discussed above that suggest a diverse population of infections in village A, these results further suggest that most hosts in village A harbor multiple mating pairs. Such contrasting patterns of parasite diversity observed between hosts from different villages could potentially be driven by corresponding differences in intermediate hosts (i.e., snails) between villages, or by other factors (such as parasite import and within-village infection rate) that may lead to differences in the number and diversity of schistosomes in each village. While the scope of this study is not sufficient to differentiate these factors, these findings do illustrate how such insight may lead to an understanding of how control measures may be directed differently for different areas that demonstrate contrasting profiles of infection.

Within village A, we find multiple hosts that contained more than one individual miracid that are more closely related to other individuals outside of that host, but within the same village. For instance, a single host from Village A contained four miracidia but only one pair were found to be 3rd degree relatives (half-sibs) and all others were estimated to have 5th degree levels of relatedness. This also indicates that this particular host harbors individuals that share the same level of relatedness to other individuals not only outside of the host but also to miracidia from the adjacent village. While most of the relationships within village A are 4th and 5th degrees, the host from village B had a high number of 3rd degree relationships suggesting that this host contained either multiple mating pairs or that these individuals were double first cousins potentially due to the local source infection being highly inbred in addition to low number of adult worms contributing to this source. The relationships found within village A, indicate that the local re-infection sources are likely relevant, and that these infected hosts may have contributed to the transmission or re-infection to other host individuals. These findings highlight the differences in infection dynamics underlying the observed patterns for hosts of different villages.

Patterns of relatedness and diversity between villages also indicate that infection sources tend to be local (i.e., village specific), rather than imported from adjacent villages. Our measures of genetic

diversity and relatedness suggest that, despite the close proximity of the two villages (~12 km), miracidia are more closely related to other individual parasites within that same village than they are to parasites in another village. Overall the inferred patterns of relatedness, and genetic diversity provide insight to transmission and source dynamics between geographically proximate villages and highlight the level/scale of resolution that these data provide.

Future directions for schistosome population genomics

Understanding patterns of persistence and reemergence of schistosomiasis despite ongoing control efforts, and the potential impacts that control efforts have had on the biology of the parasite, are important epidemiological problems that could have major global health relevance. Population genomic approaches to studying schistomiasis have the potential to answer such previously inaccessible questions about the persistence of this disease and how it has responded to control efforts, and thereby provide actionable information for reforming control efforts to be more effective. China in particular has pursued one of the most aggressive and long standing schistosomiasis control programs in the world, starting in the 1950s, and re-invigorated most recently through a multi-pronged campaign to eliminate schistosomiasis nationwide (L.-D. Wang et al., 2009; Xu et al., 2016). In our study region, for example, control activities have included over a decade of mass and targeted chemotherapy in humans and bovines, as well as snail control and other environmental modifications, strategies that have been pursued throughout Sichuan(Liu et al., 2016).

Because of the relatively small number of samples in this exploratory study, we focused on inferring patterns of relatedness and parasite diversity at fine scales, and our results highlight the variation parasite diversity across villages and hosts, and suggest a high degree of genetic structure among parasite populations at even very small (between-village) scales. However, our results support the practical and economic feasibility of far larger-scale population genomic studies on schistosomes, and further suggest that such studies have the potential to incorporate archival material to analyze genomic composition and diversity through time (Shortt et al., 2017). Future studies with larger sample sizes therefore have the potential to incorporate more detailed analysis of genomic diversity and inferences of

selection, with (both temporal and spatial dimensions that may be particularly informative for interpreting patterns of past and ongoing transmission, as well as shifts in parasite biology linked to selection-driven genomic changes in parasite populations.

Evolutionary theory suggests that the parasite will respond to strong control-induced selective pressures by increasing the frequency of genomic variants that promote responses would alter fundamental transmission parameters, with important implications for control activities. Selection pressures from control measures leading to drug resistance is a particular concern, as praziquantel is the only drug in widespread use for schistosomiasis control (Albonico et al., 2015). Praziquantel resistance has been demonstrated in the laboratory (e.g., Cioli et al., 2004), but to date has not been documented at the population level(Albonico et al., 2015; Greenberg, 2013; W. Wang, Wang, & Liang, 2012). However, *S. mansoni* evolved resistance to the antischistosomal, oxamniquine, in Brazil, leading to its eventual discontinuation as a first line treatment and raising concerns about resistance to praziquantel (Valentim et al., 2013). The potential for using population genomic approaches to assess the impacts of control-driven selection, including drug resistance, represents an exciting and finally plausible application for a new generation of population genomic studies of schistosomes and other infectious diseases.

Acknowledgments

Support for this work was provided by a National Institute for Health (NIH) grant 1RO1AI134673-01A1 to EJC, DDP and TAC.

Figures

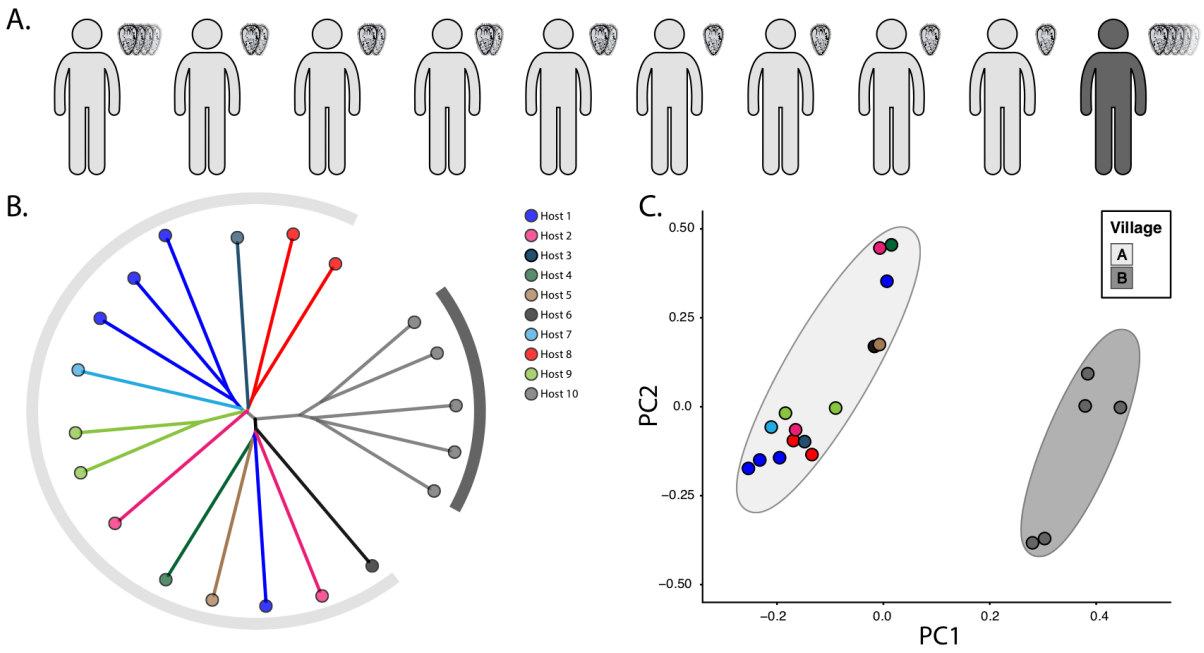

Figure 1. Overview of sampling and genetic separation between villages. A) Number of samples from each host with individual host colored by village. B) NJ tree using all exonic variants from all miracids excluding the two lowest mapped. C) PCA of all exonic variants with the first principal component representing 15% of the variation and the second representing 10%.

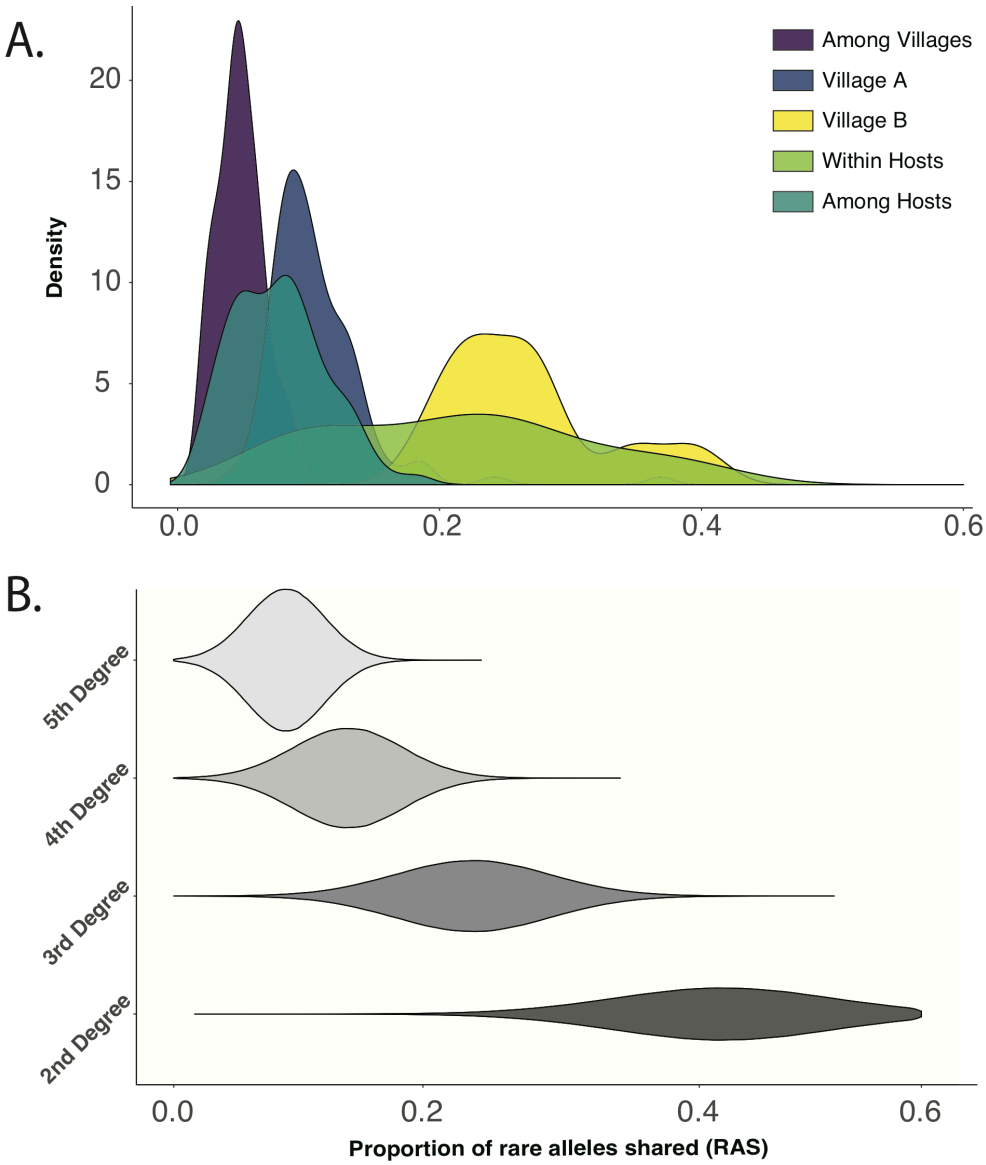

Figure 2. Rare allele sharing between all pairs of miracids sampled. A) Proportions of rare alleles shared within and among villages and hosts. B) Posterior estimates of relatedness.

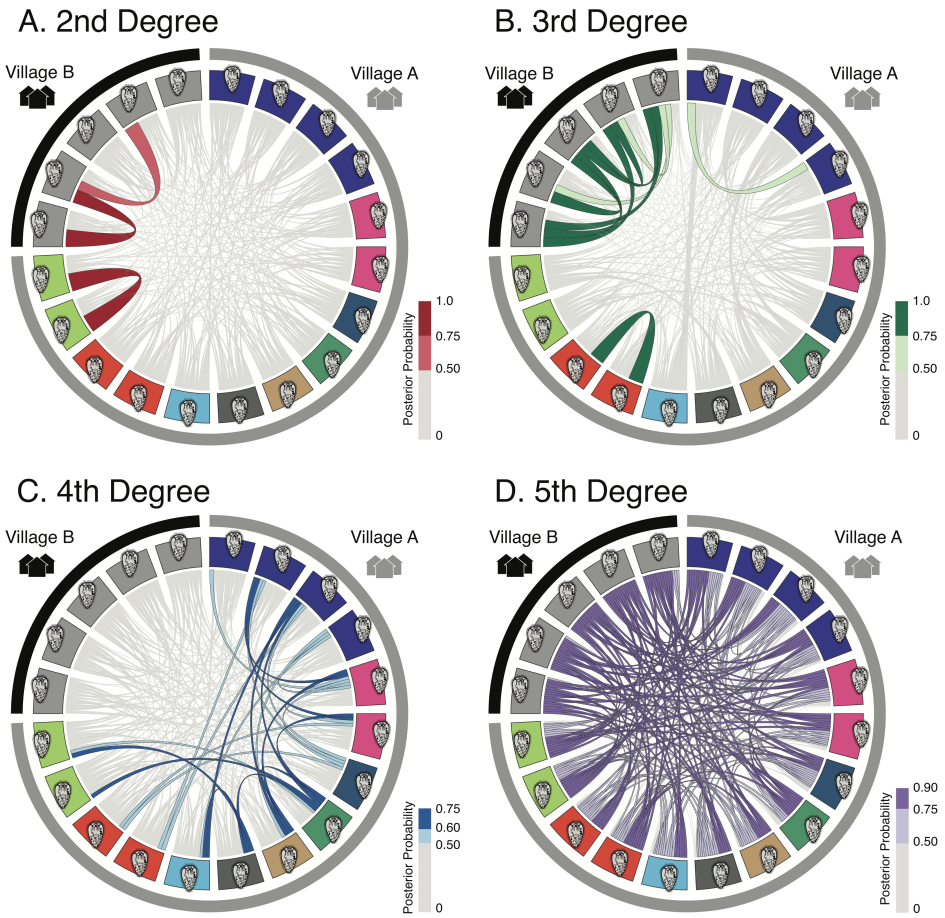

Figure 3. Posterior probability estimates of relatedness among miracidial sampled based on rare allele sharing. Bands represent villages, and miracids are colored by host. All posterior probability estimates are shown in light gray, while colored ribbons represent relationships that have a posterior probability relating to the panel. A) Red ribbons representing 2nd degree relationships among miracids that have a posterior probability estimate of of 0.50 or higher – full-sibling relationships. B) Green ribbons representing 3rd degree relationships among miracids that have a posterior probability estimate of of 0.50 or higher – half-sibling relationships. C) Blue ribbons representing 4th degree relationships among miracids that have a posterior probability estimate of of 0.50 or higher D) Purple ribbons representing 5th degree relationship among miracids.

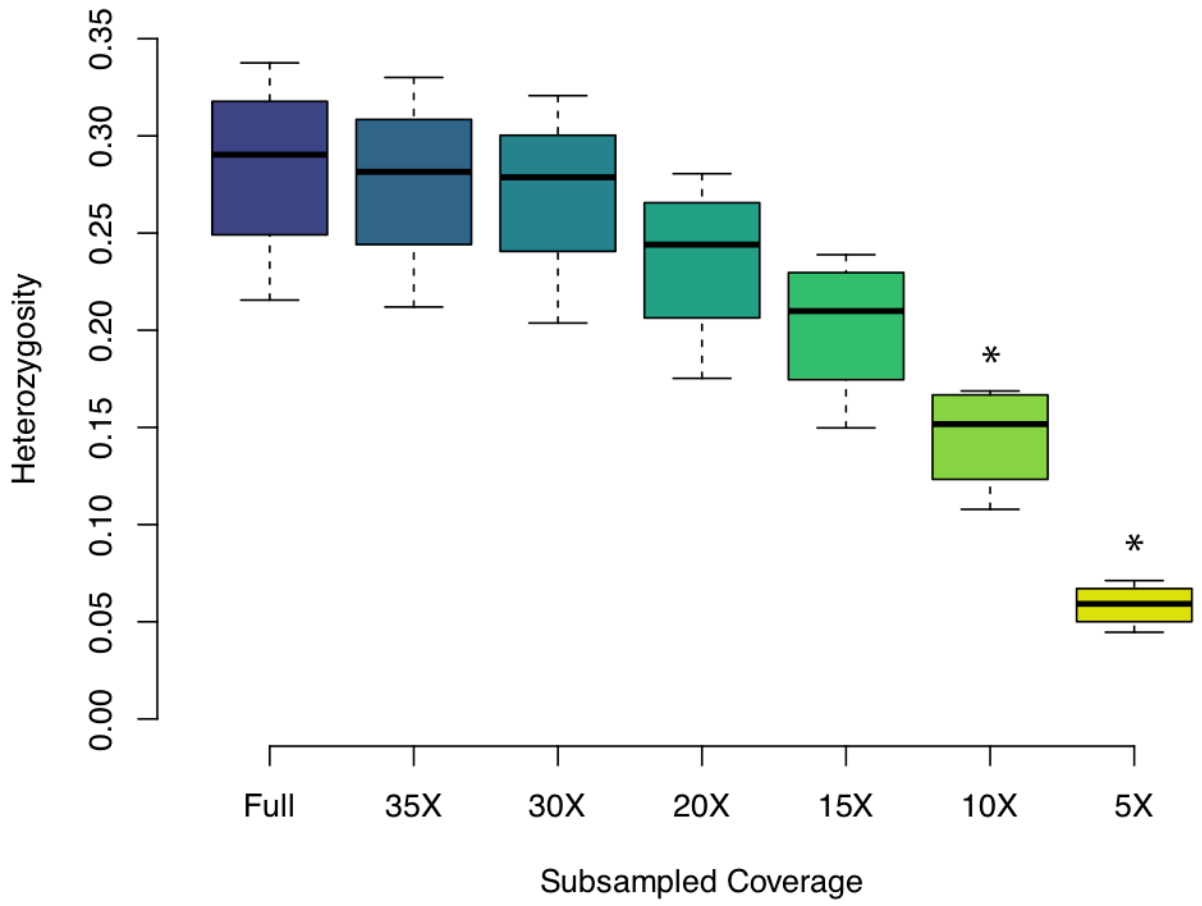

Figure 4. Effects of downsampling overall genomic coverage on heterozygosity estimates. Boxplots represent distribution of heterozygosity estimates among individuals at various degrees of randomly downsampling data, such that estimates were determined from subsets of data where total genomic coverage is either 35, 30, 30, 15, 10 or 5x.

Tables

Table 1. Mapping and coverage statistics for each 22 WGS sample. Total reads, proportion of mapped reads and central tendency statistics on exon coverage reported.

Miracid ID	Village alt ID	Host ID	Total Reads (Q>20)	Total Mapped Reads	% Mapped Reads	Mean Exon Coverage	Median Exon Coverage	Mode Exon Coverage
17090B	A	1	242964690	224059064	92.22	237.7698	45	18
17090A	A	1	121503915	112818813	92.85	103.3254	28	12
17048B	A	2	256689097	230890466	89.95	342.9313	75	39
17048C	A	2	202070081	183539245	90.83	176.3338	43	21
17052A	A	2	271323109	58201626	21.45	76.39758	0	0
17052B	A	3	234545104	223500519	95.29	286.8104	34	11
17053B	A	4	234200156	214819846	91.72	256.398	93	62
17082C	A	5	317925609	299109366	94.08	354.9228	49	27
17082B	A	5	203315847	183415003	90.21	258.8655	45	21
17064A	A	5	194083616	185010305	95.33	101.0228	13	6
17112A	A	6	227823406	218704118	96	82.84688	15	5
17059D	A	7	513724448	493482501	96.06	632.2784	57	21
17103B	A	8	830095814	773700617	93.21	901.8963	174	107
17103A	A	8	128019168	120061639	93.78	107.8554	36	14
17088A	A	9	173672554	161720975	93.12	248.9284	44	29
17088B	A	9	52399022	48480107	92.52	68.3051	12	6
17148C	B	10	289746165	273320135	94.33	372.1544	116	64
17141B	B	10	230826536	215001306	93.14	340.5262	81	46
17148B	B	10	410543401	338471019	82.44	277.2063	80	34
17128C	B	10	289249117	271422649	93.84	173.9289	34	17
17131C	B	10	108726702	99523184	91.54	101.7835	31	14
17128B	B	10	463499246	95771466	20.66	34.76015	0	0

Table S1. Reported p-values for Tukey Multiple Comparison on heterozygosity estimates determined from subsets of total genomic data, where data was randomly down sampled to 35, 20, 15, 10 or 5x coverage. Significant pairwise comparisons are in bold.

Downsampling	p-value
35X	0.9999758
30X	0.999226
20X	0.6588906
15X	0.1191918
10X	**0.0016341**
5X	**0.0000024**

References

Albonico, M., Levecke, B., LoVerde, P. T., Montresor, A., Prichard, R., Vercruysse, J., & Webster, J. P. (2015). Monitoring the efficacy of drugs for neglected tropical diseases controlled by preventive chemotherapy. *Journal of Global Antimicrobial Resistance*, *3*(4), 229–236.

Barbosa, L. M., Silva, L. K., Reis, E. A., Azevedo, T. M., Costa, J. M., Blank, W. A., … Blanton, R. E. (2013). Characteristics of the human host have little influence on which local Schistosoma mansoni populations are acquired. *PLoS Neglected Tropical Diseases*, *7*(12), e2572.

Barry, M. A., Bezek, S., Serpa, J. A., Hotez, P. J., & Woc-Colburn, L. (2012). Neglected infections of poverty in Texas and the rest of the United States: management and treatment options. *Clinical Pharmacology & Therapeutics*, *92*(2), 170–181.

Boatin, B. A., Basáñez, M.-G., Prichard, R. K., Awadzi, K., Barakat, R. M., García, H. H., … N'Goran, E. K. (2012). A research agenda for helminth diseases of humans: towards control and elimination. *PLoS Neglected Tropical Diseases*, *6*(4), e1547.

Bockarie, M. J., Kelly-Hope, L. A., Rebollo, M., & Molyneux, D. H. (2013). Preventive chemotherapy as a strategy for elimination of neglected tropical parasitic diseases: endgame challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1623), 20120144.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., El Ghissassi, F., … Galichet, L. (2009). A review of human carcinogens—Part B: biological agents. *The Lancet Oncology*, *10*(4), 321–322.

Carlton, E. J., Bates, M. N., Zhong, B., Seto, E. Y. W., & Spear, R. C. (2011). Evaluation of mammalian and intermediate host surveillance methods for detecting schistosomiasis reemergence in southwest China. *PLoS Neglected Tropical Diseases*, *5*(3), e987.

Carlton, E. J., Hubbard, A., Wang, S., & Spear, R. C. (2013). Repeated Schistosoma japonicum infection following treatment in two cohorts: evidence for host susceptibility to helminthiasis? *PLoS Neglected Tropical Diseases*, *7*(3), e2098.

Cioli, D., Botros, S. S., Wheatcroft-Francklow, K., Mbaye, A., Southgate, V., Tchuenté, L.-A. T., … Sabra, A.-N. A. (2004). Determination of ED50 values for praziquantel in praziquantel-resistant and-susceptible Schistosoma mansoni isolates. *International Journal for Parasitology*, *34*(8), 979–987.

Collins, C., Xu, J., & Tang, S. (2012). Schistosomiasis control and the health system in PR China. *Infectious Diseases of Poverty*, *1*(1), 8.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510.

Friedman, J. F., Kanzaria, H. K., & McGarvey, S. T. (2005). Human schistosomiasis and anemia: the relationship and potential mechanisms. *Trends in Parasitology*, *21*(8), 386–392.

Gower, C. M., Gouvras, A. N., Lamberton, P. H. L., Deol, A., Shrivastava, J., Mutombo, P. N., … Stothard, J. R. (2013). Population genetic structure of Schistosoma mansoni and Schistosoma haematobium from across six sub-Saharan African countries: Implications for epidemiology, evolution and control. *Acta Tropica*, *128*(2), 261–274.

Greenberg, R. M. (2013). New approaches for understanding mechanisms of drug resistance in schistosomes. *Parasitology*, *140*(12), 1534–1546.

Gryseels, B., Polman, K., Clerinx, J., & Kestens, L. (2006). Human schistosomiasis. *The Lancet*, *368*(9541), 1106–1118.

Hampton, T. (2012). Collaborative effort targets 17 tropical diseases for control, elimination. *JAMA*, *307*(8), 772.

Hotez, P. J., Engels, D., Fenwick, A., & Savioli, L. (2010). Africa is desperate for praziquantel. *The Lancet*, *376*(9740), 496–498.

Hotez, P. J., Fenwick, A., Savioli, L., & Molyneux, D. H. (2009). Rescuing the bottom billion through control of neglected tropical diseases. *The Lancet*, *373*(9674), 1570–1575.

Lei, Z. L., Zhang, L. J., Xu, Z. M., Dang, H., Xu, J., Lv, S., … Zhou, X. N. (2015). Endemic status of schistosomiasis in People's Republic of China in 2014. *Zhongguo Xue Xi Chong Bing Fang Zhi Za Zhi= Chinese Journal of Schistosomiasis Control*, *27*(6), 563–569.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Liang, S., Yang, C., Zhong, B., Guo, J., Li, H., Carlton, E. J., … Remais, J. V. (2014). Surveillance systems for neglected tropical diseases: global lessons from China's evolving schistosomiasis reporting systems, 1949–2014. *Emerging Themes in Epidemiology*, *11*(1), 19.

Liang, S., Yang, C., Zhong, B., & Qiu, D. (2006). Re-emerging schistosomiasis in hilly and mountainous areas of Sichuan, China. *Bulletin of the World Health Organization*, *84*, 139–144.

Liu, Y., Zhou, Y.-B., Li, R.-Z., Wan, J.-J., Yang, Y., Qiu, D.-C., & Zhong, B. (2016). Epidemiological features and effectiveness of schistosomiasis control programme in mountainous and hilly region of the People's Republic of China. In *Advances in parasitology* (Vol. 92, pp. 73–95). Elsevier.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., … Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, *23*(11), 1817–1828.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443.

Paradis, E., & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One*, *7*(5), e37135.

Prugnolle, F., Théron, A., Pointier, J. P., Jabbour-Zahab, R., Jarne, P., Durand, P., & Meeûs, T. de. (2005). Dispersal in a parasitic worm and its two hosts: consequence for local adaptation. *Evolution*, *59*(2), 296–303.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Rudge, J. W., LU, D., FANG, G., WANG, T., BASÁÑEZ, M., & Webster, J. P. (2009). Parasite genetic differentiation by habitat type and host species: molecular epidemiology of Schistosoma japonicum in hilly and marshland areas of Anhui Province, China. *Molecular Ecology*, *18*(10), 2134–2147.

Schistosomiasis, W. H. O. (2015). number of people treated worldwide in 2013. *Wkly Epidemiol Rec*, *90*(5), 25–32.

Shortt, J. A., Card, D. C., Schield, D. R., Liu, Y., Zhong, B., Castoe, T. A., … Pollock, D. D. (2017). Whole genome amplification and reduced-representation genome sequencing of Schistosoma japonicum miracidia. *PLoS Neglected Tropical Diseases*, *11*(1), e0005292.

Shrivastava, J., Qian, B. Z., Mcvean, G., & Webster, J. P. (2005). An insight into the genetic variation of Schistosoma japonicum in mainland China using DNA microsatellite markers. *Molecular Ecology*, *14*(3), 839–849.

Team, R. C. (2017). *R: A language and environment for statistical computing*.

Valentim, C. L. L., Cioli, D., Chevalier, F. D., Cao, X., Taylor, A. B., Holloway, S. P., … Tsai, I. J. (2013). Genetic and molecular basis of drug resistance and species-specific drug action in schistosome parasites. *Science*, *342*(6164), 1385–1389.

Wang, L.-D., Chen, H.-G., Guo, J.-G., Zeng, X.-J., Hong, X.-L., Xiong, J.-J., … Xia, G. (2009). A strategy to control transmission of Schistosoma japonicum in China. *New England Journal of Medicine*, *360*(2), 121–128.

Wang, L., Utzinger, J., & Zhou, X.-N. (2008). Schistosomiasis control: experiences and lessons from China. *The Lancet*, *372*(9652), 1793–1795.

Wang, W., Wang, L., & Liang, Y.-S. (2012). Susceptibility or resistance of praziquantel in human schistosomiasis: a review. *Parasitology Research*, *111*(5), 1871–1877.

WHO. (2012). *Accelerating work to overcome the global impact of neglected tropical diseases – A roadmap for implementation*. Geneva: WHO.

Xiao, N., Remais, J. V, Brindley, P. J., Qiu, D.-C., Carlton, E. J., Li, R.-Z., … Blair, D. (2013). Approaches to genotyping individual miracidia of Schistosoma japonicum. *Parasitology Research*, *112*(12), 3991–3999.

Xu, J., Steinman, P., Maybe, D., Zhou, X.-N., Lv, S., Li, S.-Z., & Peeling, R. (2016). Evolution of the national schistosomiasis control programmes in the People's Republic of China. In *Advances in parasitology* (Vol. 92, pp. 1–38). Elsevier.

Young, N. D., Chan, K.-G., Korhonen, P. K., Chong, T. M., Ee, R., Mohandas, N., … Jex, A. R. (2015). Exploring molecular variation in Schistosoma japonicum in China. *Scientific Reports*, *5*, 17345.

Zhou, X.-N., Guo, J.-G., Wu, X.-H., Jiang, Q.-W., Zheng, J., Dang, H., … Wu, G.-L. (2007). Epidemiology of schistosomiasis in the People's Republic of China, 2004. *Emerging Infectious Diseases*, *13*(10), 1470.