scientific data

Article in Press

Check for updates

OPEN

DATA DESCRIPTOR

An improved genome and venom gene annotation for the Prairie Rattlesnake (*Crotalus viridis*)

Siddharth S. Gopalan[1], Kaas Ballard[1], Yannick Z. Francioli[1], Hannah D. Guss[1], Daren C. Card[2], Ivan Koludarov[3], Blair W. Perry[4], Drew R. Schield[5], Anthony J. Saviola[6], Richard H. Adams[7], Stephen P. Mackessy[8] & Todd A. Castoe[1]✉

The Prairie Rattlesnake, *Crotalus viridis*, is a widely distributed North American rattlesnake and has become a versatile model for studies of phenotypic evolution, genome biology, gene regulation, and hybridization in natural populations. While scaffolded to chromosomes, the previous assembly for this species produced a genome with thousands of contigs and was missing highly expressed venom genes of great medical and evolutionary importance. This made interpreting relationships among key venom loci and their regulatory architecture challenging or impossible. Here, we present a more complete and highly contiguous chromosome-level genome assembly for *C. viridis*, including a more complete and substantially expanded venom gene annotation. This assembly was produced using a combination of PacBio HiFi reads, Illumina short reads, Omni-C proximity ligation, followed by reference-guided scaffolding to existing rattlesnake assemblies. The final assembly spans 1.59 Gb across 17 autosomes, 2 sex chromosomes (W and Z), 23 unplaced scaffolds and a complete mitochondrial genome. With an average sequencing depth of ~56X, a scaffold N50 of ~212 Mb, a Tetrapoda BUSCO completeness of 94.6%, and >99.99% base pair accuracy, this highly contiguous assembly represents a much-improved high-quality resource for studies of venom evolution, gene regulation, population genomics, and comparative vertebrate genomics.

Background and Summary

The Prairie Rattlesnake (*Crotalus viridis*) is the widest ranging rattlesnake species, occurring from northern Mexico to southwestern Canada. Throughout this range, the species exhibits significant ecological and phenotypic variation, including extensive diversity in venom composition[1]. Venom from *C. viridis* was first studied biochemically, with early work focused on the isolation of individual peptides and a description of their pathological effects[2–4]. The release of a genome assembly for *C. viridis* in 2019[5] placed decades of laboratory investigation in a genomic context, and established a foundation for integrative studies of venom composition, venom gene regulation, and their geographic variation[1,6,7]. However, *C. viridis* has also become a valuable resource for other research areas including organ regeneration[8], the genetic and regulatory basis of phenotypic variation[9,10], recombination landscape evolution[11,12], genome structure and sex chromosome evolution[13–15], and the maintenance of adaptive variation through balancing selection[16]. Its propensity to hybridize with related *Crotalus* lineages throughout portions of its range has also made it a powerful model for studying introgression, genomic coupling, and the evolution of reproductive isolation in natural populations[17–21]. Considering these expanding bodies of work, *C. viridis* has emerged over the last decade as an unusually versatile squamate model whose

[1]Department of Biology, University of Texas at Arlington, Arlington, TX, 76019, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA. [3]School of Computation, Informatics and Technology, Technical University of Munich, 85748, Garching, Munich, Germany. [4]Department of Integrative Biology, Michigan State University, East Lansing, MI, 48824, USA. [5]Department of Biology, University of Virginia, Charlottesville, VA, 22903, USA. [6]Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, CO, 80045, USA. [7]Center for Agricultural Data Analytics, University of Arkansas, Fayetteville, AR, 72703, USA. [8]Department of Biological Sciences, University of Northern Colorado, Greeley, CO, 80639, USA. ✉e-mail: todd.castoe@uta.edu

biology has found application across diverse areas of genome biology, from gene regulation and chromosome evolution to adaptation, regeneration, and speciation.

Among the most challenging and biologically important regions of snake genomes are the loci encoding venom genes and their associated regulatory elements[9,10,22–24]. As studies of snake venoms increasingly move beyond toxin discovery toward understanding the genomic architecture of venom systems, highly contiguous genome assemblies have become essential resources. Consequently, complete genome assemblies are increasingly important not only for accurate venom gene annotation, but also for resolving toxin gene cluster organization, cis-regulatory landscapes, and structural variation underlying venom expression and phenotypic diversity[10,25]. However, venom loci are among the most challenging regions of snake genomes to assemble and annotate. Venom genes are often organized into large tandemly duplicated arrays containing highly similar paralogs embedded within repeat-rich genomic regions[22,23,26], features that can complicate assembly and annotation, particularly in genomes generated using short-read sequencing approaches.

Illustrating these challenges, the previous *C. viridis* assembly, which used mate-pair and short-read sequencing for the primary assembly, failed to resolve several biologically important venom loci, including the bradykinin-potentiating peptide (BPP) and crotamine/myotoxin[5,27]. Both of these toxins can comprise over half of the entire venom proteome in *C. viridis*[7,9], and contribute substantially to the lethal effects of envenomation, including severe hypotension, tetanic paralysis and cardiac arrest[28,29]. The absence of fully assembled loci limited interpretation of both venom gene content and the genomic architecture underlying medically and evolutionarily important venom phenotypes. Furthermore, it is likely that numerous other venom genes and gene families were incompletely assembled or annotated, as the previous annotation contains roughly half as many venom genes as recently published venomous snake genomes[25,30,31]. While some of this discrepancy reflects differences in annotation strategies and fluid definitions of venom genes[32,33], it likely also reflects an incomplete representation of the true venom complement and organization encoded in the genome. Therefore, a central objective of our current assembly effort was to generate a more complete representation of the venom gene content, genomic organization, and gene regulatory architecture of *C. viridis*, while bringing its venom annotation in line with other modern snake genomes[25,30,31,34]. We addressed this goal by combining new PacBio HiFi sequencing with existing genomic resources to generate a highly contiguous chromosome-level assembly, followed by a modified annotation strategy integrating venom gland PacBio Iso-Seq, Illumina mRNA-seq, and homology-based approaches. Together, these methods improved recovery of previously fragmented or missing toxin loci while enhancing representation and annotation of complex duplicated venom gene families and their genomic architecture.

A high-quality *C. viridis* genome provides an important foundation for studies of venom biology, genome regulation, structural variation, introgression, and chromosome evolution, while supporting comparative analyses across vertebrates. Here, we report an improved chromosome-level genome assembly and expanded venom gene annotation for the highly versatile model species *C. viridis*, providing a reference-quality genomic resource for diverse applications in vertebrate genome biology and evolutionary genomics.

Methods

Sample collection and sequencing. All animal collection, housing, and sampling was conducted in accordance with an approved IACUC protocol (2303D-SM-S-26; to S.P.M.) at the University of Northern Colorado. Animals were collected under an approved state permit (Colorado Parks and Wildlife Scientific Collection License #3084507943; to S.P.M.). A single adult female was collected from a mid-range population of *C. viridis* (Weld Co., Colorado; NCBI BioSample accession: SAMN49815070), temporarily housed at the University of Northern Colorado, and humanely sacrificed for tissue dissection. Liver tissue was collected and immediately flash frozen in liquid nitrogen, and then sent to Novogene (Sacramento, CA, USA) for DNA extraction, library preparation and sequencing of both PacBio HiFi and Illumina short-read libraries, as well as draft genome assembly. Briefly, high-molecular weight DNA was extracted from liver tissues using the Nucleobond HMW DNA Kit (Macherey-Nagel) and used for PacBio library preparation. The PacBio library was sequenced using circular consensus sequencing on a Sequel II system, using a single SMRT cell, resulting in 22.7 Gb of data. gDNA was extracted from liver tissues using the QIAamp DNA Mini Kit (Qiagen) and used for Illumina library preparation. Illumina libraries were sequenced on a NovaSeq 6000 using 150 bp paired-end reads, resulting in 118.4 Gb of data. HiFi reads were used as input into Hifiasm[35] to produce a primary set of contiguous sequences. This initial contig set was polished using short-reads to correct sequencing errors using NextPolish[36] resulting in a high-quality, error-corrected draft genome assembly. Blastn v2.12.0+[37] was then used to identify and remove contigs of non-eukaryotic origin prior to scaffolding (sidg13.github.io/bioinformatics/2024/06/04/blasting-genomes.html).

Scaffolding the draft assembly. Our scaffolding approach adopted two strategies in succession: a first round of scaffolding using Omni-C proximity-ligation data followed by further reference-based scaffolding to two existing assemblies (first to the prior *C. viridis* assembly, and finally to the *C. adamanteus* assembly). This hierarchical scaffolding approach, where scaffold inferences from successive steps build on fixed prior scaffolds, was designed to avoid conflicts introduced across assembly steps, and to minimize the potential that structural variation between *C. viridis* and *C. adamanteus* would introduce assembly errors. Venom gland Omni-C chromatin conformation data was sourced from a previous study[27] from an individual from the same population as the genome representative animal (NCBI BioSample accession: SAMN28204866). Omni-C data were used to perform a first-pass scaffolding of the draft assembly using YaHS v1.2.2[38], with default settings. This scaffolded assembly was then sequentially scaffolded to existing genome assemblies for *C. viridis*[5] and then *C. adamanteus*[25] using the 'scaffold' function from RagTag v2.1.0[39]. This two-step scaffolding procedure was used to improve confidence in scaffold-to-chromosome assignments and to fill remaining assembly gaps.

Metric	*Crotalus viridis* (this study)	*Crotalus viridis* (Schield *et al.*, 2019)	*Crotalus tigris* (Margres *et al.*, 2021)	*Crotalus adamanteus* (Hogan *et al.*, 2024)
Assembly size (Gb)	1.59 Gb	1.34 Gb	1.61 Gb	1.70 Gb
# Scaffolds	42	7,046	4,228	26
# Contigs	2,098	219,439	4,228	365
Longest scaffold length	363.204 Mb	311.713 Mbp	27.107 Mb	359.319 Mb
Longest contig length	34.028 Mb	0.208 Mb	27.107 Mb	172.052 Mb
Unplaced sequence (%)	6.67 Mb (0.42%)	43.2 Mb (3.22%)	—	1.86 Mb (0.11%)
% Ambiguous sequence (N)	0.013%	6.17%	0%	0.002%
Number of gaps	2,056	212,393	0	339
Scaffold N50	211.51 Mb	179.89 Mb	2.107 Mb	208.94 Mb
Contig N50	2.556 Mb	0.015 Mb	2.107 Mb	67.533 Mb
# Genes	20,261	18,585	18,240	21,841
BUSCO Complete (Tetrapoda, %)	94.6%	85.3%	96.2%	96.7%
# Venom genes	88	50	51	134

Table 1. Comparison of nuclear assembly metrics across high-quality, publicly available rattlesnake genomes. Note that statistics for *C. tigris* and *C. adamanteus* are calculated from their publicly available assemblies (NCBI GenBank accessions: GCA_016545835.1 and GCA_039797435.1). Note that the *Crotalus tigris* assembly is contig-level.

Pre-split chromosome name	RagTag scaffold ID	Total introduced gaps	Round 1 gaps (*C. viridis*)	Round 2 gaps (*C. adamanteus*)
scaffoldma1	CM077918.1_RagTag	294	203	91
scaffoldZ	CM077936.1_RagTag	259	197	62
scaffoldma2	CM077919.1_RagTag	248	180	68
scaffoldma3	CM077920.1_RagTag	216	147	69
scaffoldma4	CM077921.1_RagTag	164	111	53
scaffoldma7	CM077924.1_RagTag	150	106	44
scaffoldma6	CM077923.1_RagTag	136	89	47
scaffoldma5	CM077922.1_RagTag	121	81	40
scaffoldmi6	CM077931.1_RagTag	80	66	14
scaffoldmi1	CM077930.1_RagTag	62	60	2
scaffoldW	CM077935.1_RagTag	52	0	52
scaffoldmi8	CM077934.1_RagTag	50	43	7
scaffoldmi4	CM077928.1_RagTag	43	41	2
scaffoldmi7	CM077929.1_RagTag	38	38	0
scaffoldmi2	CM077926.1_RagTag	35	29	6
scaffoldmi3	CM077925.1_RagTag	33	32	1
scaffoldmi9	CM077933.1_RagTag	31	29	2
scaffoldmi5	CM077927.1_RagTag	26	26	0
scaffoldmi10	CM077932.1_RagTag	19	8	11

Table 2. Accounting of gaps introduced during scaffolding to prior *C. viridis* assembly and *C. adamanteus*. RagTag introduced gaps to the assembly at different stages of scaffolding. Note the W chromosome was entirely scaffolded to scaffolds from *C. adamanteus* as it was missing in the prior *C. viridis* assembly. Rows are arranged by descending order of total introduced gaps. Note that the scaffold and chromosome naming scheme follows outputs from RagTag scaffolding, and all values in this table represent the pre-microchromosome 1 split.

Despite our hierarchical assembly approach, reference-based scaffolding using distinct species could potentially introduce misassemblies, particularly in regions of structural divergence between the reference and target genomes. To evaluate the extent of reference-guided modification of our assembly, we quantified the joins introduced by RagTag at each scaffolding step. Of the 2,057 total joins introduced by RagTag, 1,486 were introduced during the initial scaffolding against *C. viridis*, with an additional 571 introduced during subsequent scaffolding against *C. adamanteus*. Collectively, gaps with these joins comprise only 0.013% of the final assembly, compared with 6.17% gap sequence in the prior *C. viridis* reference assembly (Table 1), indicating that reference-guided scaffolding introduced very little additional unresolved sequence. We further evaluated the behavior of the hierarchical scaffolding procedure using the W chromosome, which was absent from the prior *C. viridis* reference but present in the *C. adamanteus* reference. As expected, the W chromosome was not erroneously incorporated during scaffolding against *C. viridis* and was subsequently correctly scaffolded using only the *C. adamanteus* reference (Table 2). Together, the limited contribution of RagTag-introduced gaps and the expected recovery of

Final chromosome name	Total contacts	% intra	% inter
scaffoldma1	125,808,680	56.89%	43.11%
scaffoldma2	99,526,294	53.57%	46.43%
scaffoldma3	74,288,673	50.66%	49.34%
scaffoldma4	40,786,908	44.02%	55.98%
scaffoldma5	38,760,431	43.97%	56.03%
scaffoldma6	32,214,670	41.67%	58.33%
scaffoldma7	29,534,990	41.19%	58.81%
scaffoldZ	29,679,320	42.31%	57.69%
scaffoldW	1,307,874	29.51%	70.49%
scaffoldmi1	11,674,110	34.21%	65.79%
scaffoldmi2	11,547,918	35.40%	64.60%
scaffoldmi3	9,175,855	33.61%	66.39%
scaffoldmi4	8,761,459	33.71%	66.29%
scaffoldmi5	9,390,244	34.22%	65.78%
scaffoldmi6	7,757,635	32.18%	67.82%
scaffoldmi7	7,028,820	32.94%	67.06%
scaffoldmi8	6,813,900	33.61%	66.39%
scaffoldmi9	5,415,500	31.40%	68.60%
scaffoldmi10	3,795,159	31.00%	69.00%

Table 3. Intra- and inter-chromosomal contacts in *C. viridis* assembly. Number of contacts within and across chromosomes, calculated from Omni-C data.

the W chromosome provide support for the validity of our hierarchical scaffolding approach and suggest that reference-specific structural differences did not substantially compromise the final assembly.

Previous comparisons of independently assembled *C. viridis* and *C. adamanteus* genomes revealed rearrangements among microchromosomes, whereas macrochromosomes largely retained their structure[25]. These findings raise the possibility that scaffolding the *C. viridis* genome against the *C. adamanteus* reference could introduce errors by imposing microchromosome arrangements that differ between species. A second concern is that elevated interchromosomal contact frequencies among microchromosomes can complicate contact-based scaffolding, potentially leading to erroneous joins requiring manual correction[5]. Our Omni-C data confirmed relatively higher interchromosomal contact frequencies among microchromosomes than among macrochromosomes (Table 3), consistent with observations in other vertebrates[40]. We therefore closely examined microchromosome assembly to assess potential errors arising from both reference-guided scaffolding and elevated contacts between distinct chromosomes. To identify potential erroneous chromosome fusions or splits, we aligned the previous and new *C. viridis* assemblies using minimap2 v2.31[41] and visually inspected the resulting alignment. This comparison indicated that microchromosomes 6 and 8 in the previous *C. viridis* assembly had likely been artificially joined to form microchromosome 1 in the new assembly during scaffolding against *C. adamanteus* (Fig. 1). Omni-C data supported this interpretation, showing an abrupt discontinuity in the 10-kb-resolution contact map at the inferred junction (Fig. 1). We therefore split the scaffold at this junction in the final assembly.

The final nuclear genome assembly contains 42 scaffolds and has a scaffold N50 of 211.51 Mb, placing it among other highly contiguous squamate genomes[19,31,34,42–44] (Table 1). All 7 macrochromosomes and 10 microchromosomes known from karyotyping[45] were assigned, as well as the Z and W chromosomes. 23 scaffolds were unplaced, representing 6,673,954 bp, or ~0.42% of the total assembly. Omni-C reads were re-mapped to the scaffolded assembly using Juicer Tools v1.19.02[46] and visualized using Juicebox v2.17.00[47] to confirm the accurate assembly of the genome into chromosomes (Fig. 2).

Mitochondrial assembly and annotation. The mitochondrial genome was assembled using MITObim v1.9.1[48] from the same Illumina-based short-reads used for the genome assembly. MITObim was run using the reference mitochondrial genome of *C. oreganus helleri* to capture matching reads for assembly[49]. Mitochondrial genes were automatically annotated using the MITOS2 webserver (Galaxy Version 2.1.10 + galaxy0)[50], which identified 37 genes including 22 tRNAs, 2 rRNAs and 13 protein coding genes, matching the gene number and composition of mitochondrial genomes from other snakes[51].

Non-venom gene annotation transfer. Gene annotations were transferred from the prior annotation of *C. viridis*[5] using LiftOff v1.6.3[52] by transferring transcript, exon and CDS features. Single-copy non-venom orthologs from *C. atrox* (NCBI BioSample accession: SAMN56617639) that were missing in *C. viridis* were also transferred, resulting in a total of 20,135 non-venom genes being transferred. Non-venom genes were given human ortholog identifiers by translating the longest isoform per gene using the AGAT tools v1.7.0 (github.com/NBISweden/AGAT) script "agat_sp_keep_longest_isoform.pl", then by comparing isoform orthology to the Ensembl human proteome (Homo_sapiens.GRCh38.pep.all.fa) using Orthofinder v2.5.5[53].

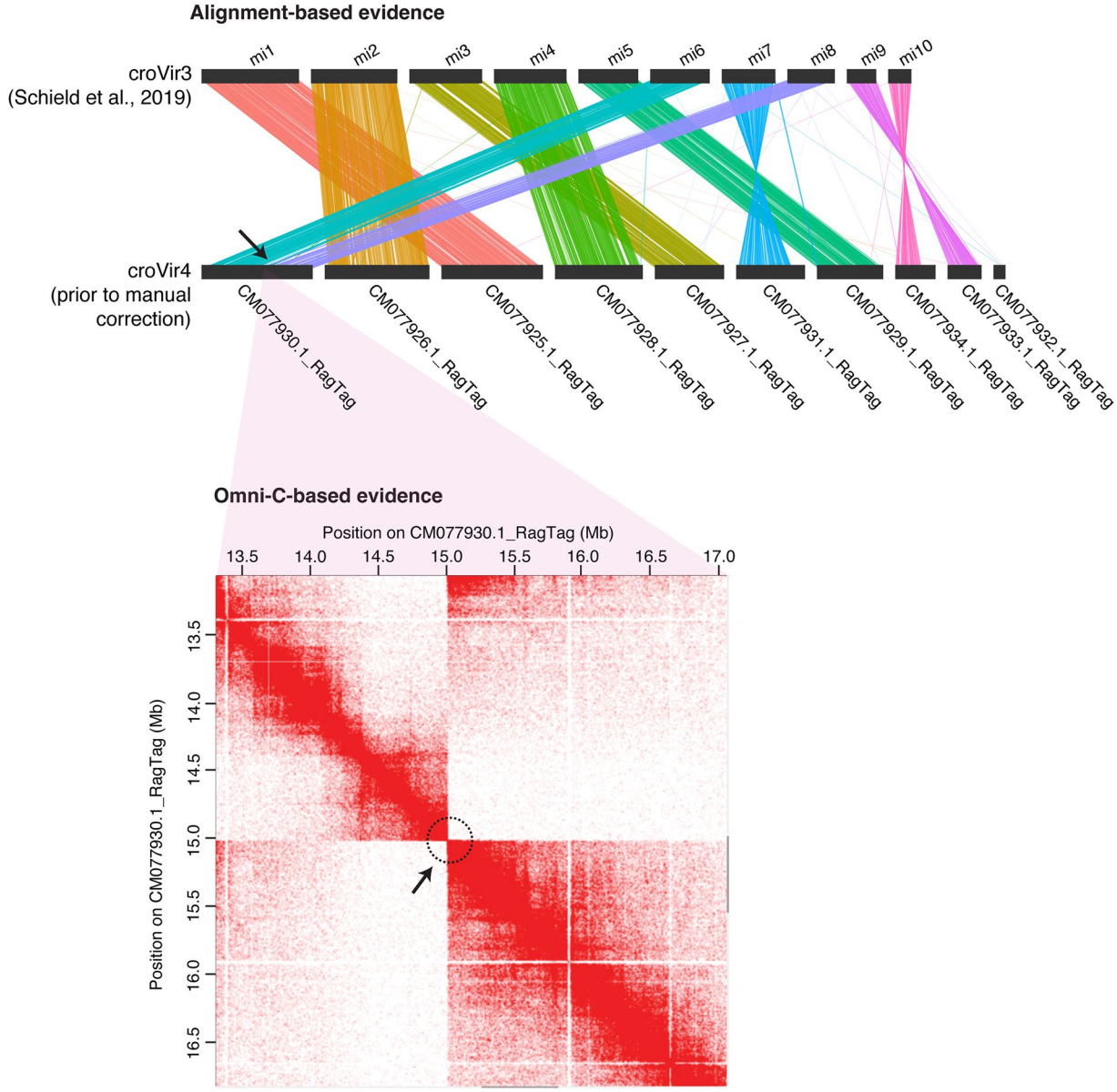


Fig. 1 Genome alignment and Omni-C data support an erroneous microchromosomal join introduced during scaffolding against the *C. adamanteus* reference. The upper panel shows alignments between microchromosomes in the previous *C. viridis* assembly (croVir3; Schield *et al.*[5]) and the new assembly before manual correction (croVir4). During reference-guided scaffolding against *C. adamanteus*, sequences corresponding to microchromosomes 6 and 8 in croVir3 were joined into a single scaffold, designated CM077930.1_RagTag in the pre-correction croVir4 assembly. The arrow indicates the inferred junction. The lower panel shows the Omni-C contact map for this scaffold at 10-kb resolution. An abrupt discontinuity in contacts across the inferred junction, approximately 15 Mb along the scaffold, supports an erroneous join (arrow and dashed circle). We split this scaffold at the inferred junction for the final assembly. The croVir4 scaffold naming scheme refers to the pre-chromosome assignments taken from RagTag scaffolding outputs. Alignment revealed that the smallest croVir4 scaffold shown in the top panel (CM077932.1_RagTag) was truncated and had few alignments to croVir3 microchromosomes and so was downgraded to an unplaced scaffold to match the *C. viridis* karyotype.

Venom gene annotation and manual curation. Venom gene annotation was performed manually by combining lines of evidence from multiple sources. All venom gene annotations that were initially transferred by LiftOff were removed and manually re-annotated using a combination of short-read and long-read venom gland mRNA-seq data. Both datasets were generated for a previous study[27] (NCBI SRA accessions: SRR19221439 and SRR19221440) sampled from an individual from the same population as the representative genome animal (NCBI BioSample accession: SAMN28204866). Short-reads were trimmed using Trim Galore![54] and mapped to the *C. viridis* assembly using STAR v2.7.11b[55]. Polymerase reads were processed into full-length isoforms using

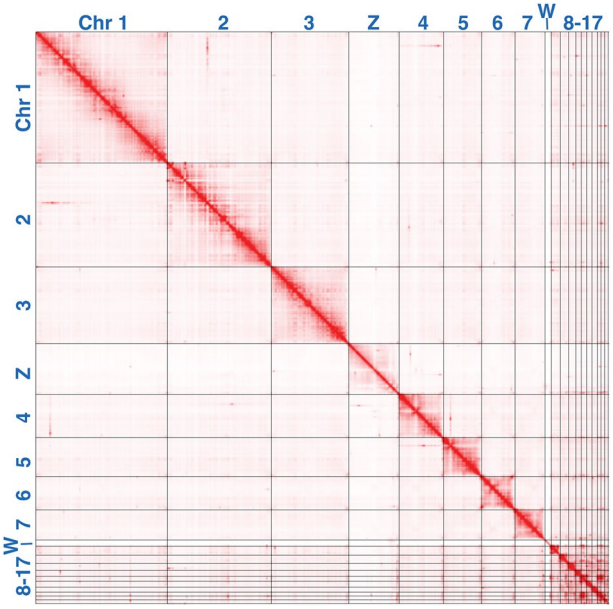


Fig. 2 Omni-C re-mapping to scaffolded assembly confirms chromosome-level assembly. After scaffolding and fusion correction, Omni-C reads were mapped back to the assembly to visually assess the degree of self-association (interchromosomal vs. intrachromosomal contact frequency) of chromosomes. Note that chromosomes 8–17 are considered 'microchromosomes', which have a higher baseline level of interchromosomal association than macrochromosomes[40].

the PacBio command line workflow for primer removal, read refinement, clustering, mapping, collapsing, and transcript classification (isoseq.how). The short-read mRNA-seq was then mapped to the isoform-level transcriptome using Kallisto v0.51.0[56] to quantify isoform expression.

For each toxin family we compiled a reference set of 557 full-length toxin precursors from UniProt/ToxProt (Serpentes), retrieved by recommended protein name and restricted to precursors ≥ 50 aa. Candidate venom loci in the genome were identified by protein-to-genome alignment of these references against the assembly using miniprot v0.18-r281[57] ('–gff', one query per invocation) and by ToxCodAn-Genome[58] to perform homology-based identification of assembled transcripts from venom gland short-read mRNA-seq data ('-r') against a curated viper-specific database. Gene models were built and verified manually in Geneious Prime 2025 (Biomatters, Auckland, New Zealand). For each locus, the candidate genomic region was visualized alongside the aligned reference set with the homologous segments overlain, exon boundaries were placed by eye against the alignment, and each model was translated and read directly in the browser. A model was accepted when every intron began with GT and ended with AG, the translation ran from an initiator methionine through a complete signal peptide to the stop codon with no internal stop or frameshift, exon count and precursor length matched the family, the closest match of the translated protein was the venom clade of its family rather than a non-venom paralog family, and the gene lay at the family's venom locus[26]. Candidates failing the family-membership or architecture criteria were reclassified as non-venom paralogs. A candidate was called a pseudogene when a premature stop, frameshifting indel or splice-site substitution was visible in the genomic sequence within a core exon of an otherwise recognizable family member at the venom locus, though such candidates were not removed.

We then performed several secondary validations of the predicted venom gene models. The alignment of Iso-Seq to gene models was first manually validated using IGV v2.19.7[59]. In cases where multiple isoforms were detected for a single gene, the highest expressed isoform (based on Kallisto TPM estimates) was chosen as the canonical gene annotation. Intergenic regions were then screened for missing genes using the Augustus webserver (bioinf.uni-greifswald.de/augustus/submission.php) with pre-trained *Gallus gallus* gene models[60]. All candidate Augustus predictions were aligned to the ToxCodAn-Genome Viperidae database to validate homology. Predicted coding sequences were finally translation-validated using the 'agat_sp_extract_sequences.pl' from AGAT tools. Additionally, protein sequences were checked for the presence of a signal peptide using SignalP v6.0[61]. Transcript support was not strictly required for a gene to be annotated, though the alignment of short read mRNA-seq was noted, and occasionally used to further manually correct some intron-exon junctions.

The final *C. viridis* venom annotation contained 88 manually curated venom genes; an increase of 38 venom genes compared to the previous annotation (Fig. 3, Table 4). This included additional genes for the SVSP, SVMP and the PLA2 loci: three phenotypically dominant venom families in viperid snakes (Fig. 4), as well as the identification of previously fragmented or missing loci encoding BPP and crotamine/myotoxin. We used DIAMOND[62] ('-blastp') to perform reciprocal-best-blast (RBB) analysis to assess if novel venom loci were identified in the updated version of the assembly relative to the prior version. RBB analysis supported the novel identification of crotamine/myotoxin and several 3-finger toxins, which were known to be missing in the prior assembly (Table 5).

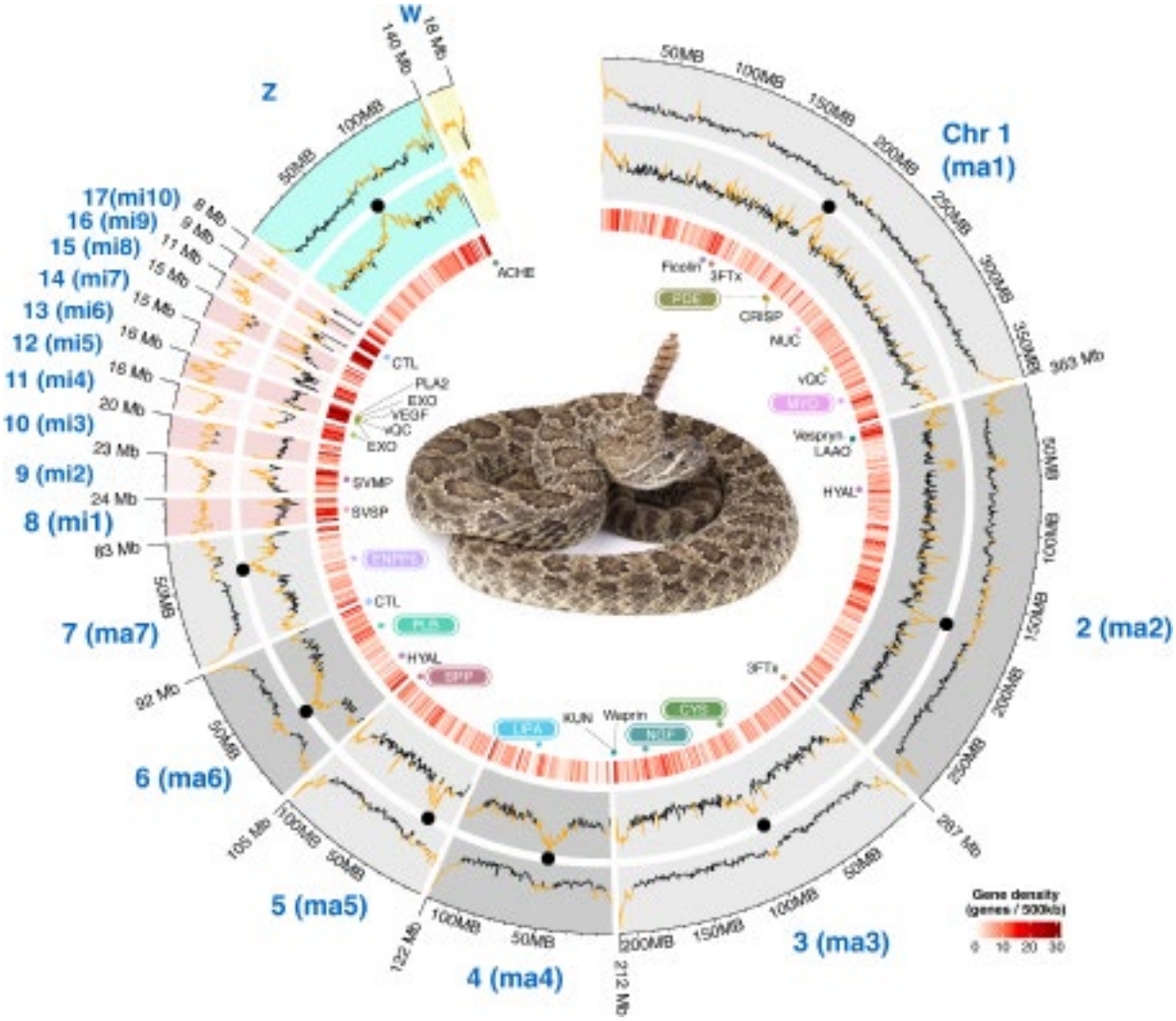


Fig. 3 Genomic landscape and venom annotations of the updated *C. viridis* assembly. The circos plot displays features of the genomic landscape in concentric rings around the genome. From innermost to outermost: gene density in genes per 500 kb, repeat (%) in 500 kb bins and GC (%) in 500 kb bins. Repeat and GC content values (ranging from 0 to 1) above the genome-wide 75th percentile are highlighted in orange. The approximate positions of centromeres are labelled with a black circle, shown between the repeat and GC content tracks. Chromosomes 8 to 17 are classified as microchromosomes (mi1 to mi10) and so are shown in a different colour to the sex chromosomes (Z and W) and other autosomes (macrochromosomes; ma1 to ma7). Venom family locations are labelled along the inner circle, with any novel families, relative to the prior annotation, shown in a coloured oval with white text. Photo credit (adult male *C. viridis* from Weld Co. Colorado, USA): Y.Z.F.

Repeat annotation. Repetitive elements in the *C. viridis* assembly were annotated using *de novo* and homology-based approaches. *De novo* repeat libraries were first generated with RepeatModeler v2.0.3[63]. Elements initially classified as unknown were iteratively reclassified using 'repclassifier' v1.1 (github.com/darencard/GenomeAnnotation/blob/master/repclassifier). This pipeline uses the Tetrapoda library from Repbase v20181026[64] and a curated library of consensus repeat sequences from 18 snake genomes[5,65]. The use of 'repclassifier' achieved assignment of 89% of previously unclassified repeats after 10 rounds of classification. The resulting *de novo* and reference libraries were then combined for homology-based masking. RepeatMasker v4.1.7-p1[66], with Dfam v3.9[67] and additional repeat libraries, were applied iteratively. Masking began with simple repeats, followed by a custom BovB/CR1 LINE library[13], Tetrapoda RepBase elements, and finally snake-specific repeats. Masking outputs were processed using 'ProcessRepeats' from RepeatMasker to characterize the repeat landscape (Table 6). To enable quantitative repeat content comparisons between the prior[5] and new genome assemblies, we re-masked the prior genome assembly with the same protocol and repeat libraries used for the new genome assembly and find a substantial relative increase in repeat element content (~10%) in the new genome assembly. Among repeat classes, the relative increases in inferred repeat element content were highest for LINEs, Ty3/Gypsy, hAT elements, and simple sequence repeats in the new assembly (Table 6, Fig. 5).

Gene name	Chromosome	Start	End	Strand	Number of exons	Toxin family
Venom_ACHE-1	scaffoldZ	139328552	139332915	−	3	ACHE
Venom_Ficolin-1	scaffoldma1	112533915	112546999	+	8	Ficolin
Venom_Ficolin-2	scaffoldma1	112552453	112561495	+	8	Ficolin
Venom_3FTx-10	scaffoldma1	123076475	123084635	−	3	3FTx
Venom_PDE-1	scaffoldma1	192497090	192510987	−	4	PDE
Venom_CRISP-1	scaffoldma1	195769383	195775142	+	7	CRISP
Venom_CRISP-2	scaffoldma1	195868699	195879720	+	7	CRISP
Venom_CRISP-3	scaffoldma1	195935444	195953487	+	7	CRISP
Venom_CRISP-4	scaffoldma1	196006965	196015765	+	7	CRISP
Venom_NUC-1	scaffoldma1	242850748	242886853	+	9	NUC
Venom_NUC-2	scaffoldma1	244066342	244101133	+	9	NUC
Venom_vQC-1	scaffoldma1	296181799	296194338	−	7	vQC
Venom_MYO-1	scaffoldma1	334525514	334526886	+	3	MYO
Venom_Vespryn-1	scaffoldma2	10080283	10088204	+	4	Vespryn
Venom_LAAO-1	scaffoldma2	10410331	10434241	−	7	LAAO
Venom_LAAO-2	scaffoldma2	10567550	10586286	+	6	LAAO
Venom_HYAL-1	scaffoldma2	66113284	66124565	+	5	HYAL
Venom_3FTx-1	scaffoldma3	2532620	2535715	−	4	3FTx
Venom_3FTx-2	scaffoldma3	2547560	2549633	−	3	3FTx
Venom_3FTx-3	scaffoldma3	2567867	2570436	+	3	3FTx
Venom_3FTx-4	scaffoldma3	2603233	2606706	−	3	3FTx
Venom_3FTx-5	scaffoldma3	2615546	2619118	+	3	3FTx
Venom_3FTx-6	scaffoldma3	2636125	2641927	−	3	3FTx
Venom_3FTx-7	scaffoldma3	2643605	2645175	−	3	3FTx
Venom_3FTx-8	scaffoldma3	2654166	2655648	−	3	3FTx
Venom_3FTx-9	scaffoldma3	2666769	2670804	−	3	3FTx
Venom_CYS-1	scaffoldma3	89410523	89418906	+	3	CYS
Venom_NGF-1	scaffoldma3	175655947	175747776	−	3	NGF
Venom_KUN-1	scaffoldma3	210477940	210478612	−	2	KUN
Venom_KUN-2	scaffoldma3	210507937	210509330	−	2	KUN
Venom_KUN-3	scaffoldma3	210522193	210523604	−	2	KUN
Venom_Waprin-1	scaffoldma3	210584682	210594624	+	3	Waprin
Venom_pgKUN-4	scaffoldma3	210601005	210606697	−	3	pgKUN
Venom_KUN-5	scaffoldma3	210661771	210662905	+	2	KUN
Venom_LIPA-1	scaffoldma4	76867583	76887046	+	10	LIPA
Venom_BPP-1	scaffoldma5	101174289	101187372	+	4	BPP
Venom_HYAL-2	scaffoldma6	22343098	22347718	−	5	HYAL
Venom_HYAL-3	scaffoldma6	22391576	22422893	−	4	HYAL
Venom_PLB-1	scaffoldma6	62425504	62468877	−	11	PLB
Venom_CTL-1	scaffoldma6	91813982	91817482	+	3	CTL
Venom_CTL-2	scaffoldma6	91824073	91828801	−	6	CTL
Venom_CTL-3	scaffoldma6	91838715	91844430	+	5	CTL
Venom_CTL-4	scaffoldma6	91854057	91860581	−	6	CTL
Venom_ENPP6-1	scaffoldma7	44804748	44845284	+	8	ENPP6
Venom_SVSP-1	scaffoldmi1	11031030	11037097	+	6	SVSP
Venom_SVSP-2	scaffoldmi1	11058642	11069759	+	5	SVSP
Venom_SVSP-3	scaffoldmi1	11134193	11140411	+	6	SVSP
Venom_SVSP-4	scaffoldmi1	11147287	11152699	+	5	SVSP
Venom_SVSP-5	scaffoldmi1	11192892	11205803	+	7	SVSP
Venom_SVSP-6	scaffoldmi1	11216333	11229088	+	6	SVSP
Venom_SVSP-7	scaffoldmi1	11236306	11248585	+	5	SVSP
Venom_SVSP-8	scaffoldmi1	11278476	11284602	+	6	SVSP
Venom_SVSP-9	scaffoldmi1	11301521	11311029	+	6	SVSP
Venom_SVSP-10	scaffoldmi1	11354154	11368228	+	6	SVSP
Venom_SVSP-11	scaffoldmi1	11407183	11423432	+	6	SVSP
Venom_SVSP-12	scaffoldmi1	11465615	11473959	+	6	SVSP
Venom_SVSP-13	scaffoldmi1	11538167	11555150	+	6	SVSP

Continued

Gene name	Chromosome	Start	End	Strand	Number of exons	Toxin family
Venom_SVMP-1	scaffoldmi2	14730996	14758803	–	18	SVMP
Venom_SVMP-2	scaffoldmi2	14783402	14822472	–	17	SVMP
Venom_SVMP-3	scaffoldmi2	14835493	14870227	–	15	SVMP
Venom_SVMP-4	scaffoldmi2	14883386	14900651	–	12	SVMP
Venom_SVMP-5	scaffoldmi2	14913823	14934120	–	12	SVMP
Venom_SVMP-6	scaffoldmi2	14940408	14968781	+	17	SVMP
Venom_SVMP-7	scaffoldmi2	14980781	15007230	–	17	SVMP
Venom_SVMP-8	scaffoldmi2	15014583	15036269	–	17	SVMP
Venom_SVMP-9	scaffoldmi2	15045312	15068813	–	17	SVMP
Venom_SVMP-10	scaffoldmi2	15089527	15115091	–	16	SVMP
Venom_SVMP-11	scaffoldmi2	15134556	15164382	–	17	SVMP
Venom_SVMP-12	scaffoldmi2	15175462	15193331	–	15	SVMP
Venom_SVMP-13	scaffoldmi2	15200555	15219666	–	15	SVMP
Venom_SVMP-14	scaffoldmi2	15229002	15255987	–	17	SVMP
Venom_SVMP-15	scaffoldmi2	15258726	15277887	–	15	SVMP
Venom_SVMP-16	scaffoldmi2	15290908	15315652	–	17	SVMP
Venom_SVMP-17	scaffoldmi2	15323564	15346833	–	17	SVMP
Venom_ADAM28-1	scaffoldmi2	15361435	15402020	–	25	ADAM28
Venom_EXO-1	scaffoldmi4	10675240	10677230	–	3	EXO
Venom_EXO-2	scaffoldmi5	4667010	4673525	–	8	EXO
Venom_vQC-2	scaffoldmi5	7376421	7382841	–	7	vQC
Venom_VEGF-1	scaffoldmi5	8798361	8801448	+	5	VEGF
Venom_EXO-3	scaffoldmi5	9705955	9708797	–	4	EXO
Venom_PLA2-F	scaffoldmi5	9907624	9908983	–	4	PLA2
Venom_PLA2-A1	scaffoldmi5	9911828	9913706	–	5	PLA2
Venom_PLA2-C1	scaffoldmi5	9917143	9918528	+	4	PLA2
Venom_PLA2-K	scaffoldmi5	9921688	9926845	+	4	PLA2
Venom_PLA2-B1	scaffoldmi5	9930485	9932380	+	4	PLA2
Venom_PLA2-gIIe	scaffoldmi5	9938490	9939852	+	2	PLA2
Venom_CTL-5	scaffoldmi9	6398903	6404869	–	5	CTL
Venom_CTL-6	scaffoldmi9	6414835	6421184	+	5	CTL

Gene name	Protein length (aa)	Has internal stop codon	SignalP prediction	SignalP probability	Has signal peptide	Number of overlapping IsoSeq transcripts
Venom_ACHE-1	610	FALSE	SP(Sec/SPI)	0.985545	TRUE	0
Venom_Ficolin-1	337	FALSE	SP(Sec/SPI)	0.999524	TRUE	0
Venom_Ficolin-2	329	FALSE	SP(Sec/SPI)	0.993881	TRUE	0
Venom_3FTx-10	94	FALSE	SP(Sec/SPI)	0.999865	TRUE	0
Venom_PDE-1	453	FALSE	SP(Sec/SPI)	0.936345	TRUE	3
Venom_CRISP-1	242	FALSE	SP(Sec/SPI)	0.94728	TRUE	0
Venom_CRISP-2	242	FALSE	SP(Sec/SPI)	0.945413	TRUE	0
Venom_CRISP-3	241	FALSE	SP(Sec/SPI)	0.97873	TRUE	22
Venom_CRISP-4	241	FALSE	SP(Sec/SPI)	0.978273	TRUE	0
Venom_NUC-1	590	FALSE	SP(Sec/SPI)	0.775839	TRUE	16
Venom_NUC-2	590	FALSE	SP(Sec/SPI)	0.775839	TRUE	10
Venom_vQC-1	369	FALSE	SP(Sec/SPI)	0.877921	TRUE	10
Venom_MYO-1	66	FALSE	SP(Sec/SPI)	0.999866	TRUE	3
Venom_Vespryn-1	219	FALSE	OTHER	0.055635	FALSE	1
Venom_LAAO-1	517	FALSE	SP(Sec/SPI)	0.994829	TRUE	95
Venom_LAAO-2	508	FALSE	SP(Sec/SPI)	0.753806	TRUE	0
Venom_HYAL-1	475	FALSE	SP(Sec/SPI)	0.971142	TRUE	1
Venom_3FTx-1	111	TRUE	SP(Sec/SPI)	0.991125	TRUE	0
Venom_3FTx-2	86	FALSE	SP(Sec/SPI)	0.999597	TRUE	0
Venom_3FTx-3	100	FALSE	SP(Sec/SPI)	0.999559	TRUE	0
Venom_3FTx-4	85	FALSE	SP(Sec/SPI)	0.999789	TRUE	0
Venom_3FTx-5	82	FALSE	SP(Sec/SPI)	0.999452	TRUE	0
Venom_3FTx-6	76	FALSE	SP(Sec/SPI)	0.999847	TRUE	0
Venom_3FTx-7	83	FALSE	SP(Sec/SPI)	0.999766	TRUE	0

Continued

Gene name	Protein length (aa)	Has internal stop codon	SignalP prediction	SignalP probability	Has signal peptide	Number of overlapping IsoSeq transcripts
Venom_3FTx-8	83	FALSE	SP(Sec/SPI)	0.999766	TRUE	0
Venom_3FTx-9	81	FALSE	SP(Sec/SPI)	0.99983	TRUE	0
Venom_CYS-1	137	FALSE	SP(Sec/SPI)	0.978449	TRUE	0
Venom_NGF-1	242	FALSE	SP(Sec/SPI)	0.986175	TRUE	81
Venom_KUN-1	81	FALSE	SP(Sec/SPI)	0.999732	TRUE	0
Venom_KUN-2	85	FALSE	SP(Sec/SPI)	0.999552	TRUE	0
Venom_KUN-3	85	FALSE	SP(Sec/SPI)	0.999587	TRUE	0
Venom_Waprin-1	135	FALSE	SP(Sec/SPI)	0.997235	TRUE	0
Venom_pgKUN-4	75	TRUE	SP(Sec/SPI)	0.999722	TRUE	0
Venom_KUN-5	84	FALSE	SP(Sec/SPI)	0.999788	TRUE	0
Venom_LIPA-1	476	FALSE	OTHER	0.311016	FALSE	0
Venom_BPP-1	229	FALSE	SP(Sec/SPI)	0.987793	TRUE	244
Venom_HYAL-2	542	FALSE	OTHER	0.071174	FALSE	0
Venom_HYAL-3	450	FALSE	OTHER	0.378102	FALSE	12
Venom_PLB-1	554	FALSE	SP(Sec/SPI)	0.84321	TRUE	27
Venom_CTL-1	109	FALSE	OTHER	0.001829	FALSE	0
Venom_CTL-2	143	FALSE	SP(Sec/SPI)	0.999424	TRUE	2
Venom_CTL-3	123	FALSE	SP(Sec/SPI)	0.999311	TRUE	3
Venom_CTL-4	149	FALSE	SP(Sec/SPI)	0.995072	TRUE	23
Venom_ENPP6-1	439	FALSE	SP(Sec/SPI)	0.97013	TRUE	17
Venom_SVSP-1	263	FALSE	SP(Sec/SPI)	0.826463	TRUE	14
Venom_SVSP-2	263	FALSE	SP(Sec/SPI)	0.950998	TRUE	351
Venom_SVSP-3	264	FALSE	SP(Sec/SPI)	0.953527	TRUE	99
Venom_SVSP-4	264	TRUE	SP(Sec/SPI)	0.970434	TRUE	29
Venom_SVSP-5	259	FALSE	SP(Sec/SPI)	0.955384	TRUE	120
Venom_SVSP-6	261	FALSE	SP(Sec/SPI)	0.943653	TRUE	166
Venom_SVSP-7	259	FALSE	SP(Sec/SPI)	0.994067	TRUE	115
Venom_SVSP-8	259	FALSE	SP(Sec/SPI)	0.951859	TRUE	130
Venom_SVSP-9	268	FALSE	SP(Sec/SPI)	0.869918	TRUE	50
Venom_SVSP-10	117	FALSE	SP(Sec/SPI)	0.963045	TRUE	26
Venom_SVSP-11	260	FALSE	SP(Sec/SPI)	0.9685	TRUE	190
Venom_SVSP-12	258	FALSE	SP(Sec/SPI)	0.982827	TRUE	54
Venom_SVSP-13	258	FALSE	SP(Sec/SPI)	0.969505	TRUE	222
Venom_SVMP-1	613	FALSE	SP(Sec/SPI)	0.907952	TRUE	44
Venom_SVMP-2	611	FALSE	SP(Sec/SPI)	0.866849	TRUE	5
Venom_SVMP-3	488	FALSE	SP(Sec/SPI)	0.907952	TRUE	14
Venom_SVMP-4	447	FALSE	SP(Sec/SPI)	0.946535	TRUE	0
Venom_SVMP-5	423	FALSE	SP(Sec/SPI)	0.873871	TRUE	27
Venom_SVMP-6	602	FALSE	SP(Sec/SPI)	0.767284	TRUE	49
Venom_SVMP-7	611	FALSE	SP(Sec/SPI)	0.758355	TRUE	4
Venom_SVMP-8	603	FALSE	SP(Sec/SPI)	0.833783	TRUE	25
Venom_SVMP-9	603	FALSE	SP(Sec/SPI)	0.77066	TRUE	19
Venom_SVMP-10	593	FALSE	SP(Sec/SPI)	0.876864	TRUE	116
Venom_SVMP-11	610	FALSE	SP(Sec/SPI)	0.844424	TRUE	18
Venom_SVMP-12	478	FALSE	SP(Sec/SPI)	0.929608	TRUE	46
Venom_SVMP-13	489	FALSE	SP(Sec/SPI)	0.872466	TRUE	11
Venom_SVMP-14	609	FALSE	SP(Sec/SPI)	0.860351	TRUE	9
Venom_SVMP-15	485	FALSE	SP(Sec/SPI)	0.877545	TRUE	66
Venom_SVMP-16	612	FALSE	SP(Sec/SPI)	0.880319	TRUE	80
Venom_SVMP-17	619	FALSE	SP(Sec/SPI)	0.941729	TRUE	0
Venom_ADAM28-1	827	FALSE	SP(Sec/SPI)	0.941126	TRUE	6
Venom_EXO-1	212	FALSE	OTHER	0.001176	FALSE	3
Venom_EXO-2	442	FALSE	OTHER	0.006237	FALSE	16
Venom_vQC-2	382	FALSE	OTHER	0.148693	FALSE	1
Venom_VEGF-1	133	FALSE	SP(Sec/SPI)	0.991983	TRUE	20
Venom_EXO-3	351	FALSE	OTHER	0.000227	FALSE	1

Continued

Gene name	Protein length (aa)	Has internal stop codon	SignalP prediction	SignalP probability	Has signal peptide	Number of overlapping IsoSeq transcripts
Venom_PLA2-F	146	FALSE	SP(Sec/SPI)	0.592745	TRUE	1
Venom_PLA2-A1	139	FALSE	SP(Sec/SPI)	0.999755	TRUE	14
Venom_PLA2-C1	137	FALSE	SP(Sec/SPI)	0.999394	TRUE	0
Venom_PLA2-K	138	FALSE	SP(Sec/SPI)	0.999594	TRUE	0
Venom_PLA2-B1	116	FALSE	OTHER	0.033473	FALSE	88
Venom_PLA2-gIIe	92	FALSE	SP(Sec/SPI)	0.95938	TRUE	50
Venom_CTL-5	167	FALSE	SP(Sec/SPI)	0.999025	TRUE	0
Venom_CTL-6	159	FALSE	SP(Sec/SPI)	0.999043	TRUE	0

Gene name	Supported by IsoSeq	Novel relative to prior genome	% of introns supported by splice junctions
Venom_ACHE-1	FALSE	No	50
Venom_Ficolin-1	FALSE	No	71.42857
Venom_Ficolin-2	FALSE	No	0
Venom_3FTx-10	FALSE	No	100
Venom_PDE-1	TRUE	No	100
Venom_CRISP-1	FALSE	No	0
Venom_CRISP-2	FALSE	No	0
Venom_CRISP-3	TRUE	No	100
Venom_CRISP-4	FALSE	No	83.33333
Venom_NUC-1	TRUE	No	75
Venom_NUC-2	TRUE	No	87.5
Venom_vQC-1	TRUE	No	100
Venom_MYO-1	TRUE	Yes	50
Venom_ Vespryn-1	TRUE	No	100
Venom_LAAO-1	TRUE	No	100
Venom_LAAO-2	FALSE	No	40
Venom_HYAL-1	TRUE	No	75
Venom_3FTx-1	FALSE	No	0
Venom_3FTx-2	FALSE	No	0
Venom_3FTx-3	FALSE	No	0
Venom_3FTx-4	FALSE	Yes	0
Venom_3FTx-5	FALSE	Yes	0
Venom_3FTx-6	FALSE	Yes	0
Venom_3FTx-7	FALSE	Yes	0
Venom_3FTx-8	FALSE	Yes	0
Venom_3FTx-9	FALSE	Yes	0
Venom_CYS-1	FALSE	No	100
Venom_NGF-1	TRUE	No	100
Venom_KUN-1	FALSE	No	0
Venom_KUN-2	FALSE	No	0
Venom_KUN-3	FALSE	No	0
Venom_Waprin-1	FALSE	No	100
Venom_pgKUN-4	FALSE	No	50
Venom_KUN-5	FALSE	No	0
Venom_LIPA-1	FALSE	No	0
Venom_BPP-1	TRUE	No	66.66667
Venom_HYAL-2	FALSE	No	0
Venom_HYAL-3	TRUE	No	100
Venom_PLB-1	TRUE	No	100
Venom_CTL-1	FALSE	Yes	100
Venom_CTL-2	TRUE	No	80
Venom_CTL-3	TRUE	Yes	75
Venom_CTL-4	TRUE	No	100
Venom_ENPP6-1	TRUE	No	100
Venom_SVSP-1	TRUE	No	100
Venom_SVSP-2	TRUE	No	100

Continued

Gene name	Supported by IsoSeq	Novel relative to prior genome	% of introns supported by splice junctions
Venom_SVSP-3	TRUE	No	100
Venom_SVSP-4	TRUE	No	100
Venom_SVSP-5	TRUE	No	100
Venom_SVSP-6	TRUE	No	100
Venom_SVSP-7	TRUE	No	100
Venom_SVSP-8	TRUE	No	100
Venom_SVSP-9	TRUE	No	100
Venom_SVSP-10	TRUE	No	40
Venom_SVSP-11	TRUE	No	100
Venom_SVSP-12	TRUE	No	100
Venom_SVSP-13	TRUE	No	100
Venom_SVMP-1	TRUE	No	94.11765
Venom_SVMP-2	TRUE	No	31.25
Venom_SVMP-3	TRUE	No	57.14286
Venom_SVMP-4	FALSE	No	54.54545
Venom_SVMP-5	TRUE	No	81.81818
Venom_SVMP-6	TRUE	No	93.75
Venom_SVMP-7	TRUE	No	87.5
Venom_SVMP-8	TRUE	No	100
Venom_SVMP-9	TRUE	No	93.75
Venom_SVMP-10	TRUE	No	93.33333
Venom_SVMP-11	TRUE	No	100
Venom_SVMP-12	TRUE	No	100
Venom_SVMP-13	TRUE	No	100
Venom_SVMP-14	TRUE	No	100
Venom_SVMP-15	TRUE	No	100
Venom_SVMP-16	TRUE	No	100
Venom_SVMP-17	FALSE	No	12.5
Venom_ ADAM28-1	TRUE	No	8.333333
Venom_EXO-1	TRUE	No	0
Venom_EXO-2	TRUE	No	57.14286
Venom_vQC-2	TRUE	No	83.33333
Venom_VEGF-1	TRUE	No	75
Venom_EXO-3	TRUE	No	100
Venom_PLA2-F	TRUE	No	66.66667
Venom_PLA2-A1	TRUE	No	100
Venom_PLA2-C1	FALSE	No	100
Venom_PLA2-K	FALSE	No	66.66667
Venom_PLA2-B1	TRUE	No	100
Venom_PLA2-gIIe	TRUE	No	100
Venom_CTL-5	FALSE	No	0
Venom_CTL-6	FALSE	No	0

Table 4. Venom gene annotation and supporting evidence details. Details of chromosomal location, gene structure, family assignment and lines of supporting evidence for each of the venom genes annotated.

Data Record

We have deposited data and scripts associated with this study in NCBI, Zenodo and GitHub datastores. The NCBI datastore (NCBI BioProject accession: PRJNA1454988) contains the genome (NCBI GenBank accession: GCA_059756665.2[68]) as well as raw assembly data[69]. This genome's nucleotide sequences were deposited at GenBank under the accession JBXUFQ000000000, and the version described in this paper is version JBXUFQ020000000. To increase the reusability of this resource, we have deposited custom datafiles, which were used to perform analyses and make the figures presented here, in a Zenodo datastore available at https://doi.org/10.5281/zenodo.21683217[70]. These files include the custom genome (FASTA file), the custom gene annotation (GTF file), repeat element libraries (FASTA files) used to mask genomic repeat elements (including classified and unclassified *de novo* repeats modelled from the assembly and existing repeat libraries used for masking), the final genomic repeat annotation (GFF3 file), the soft-masked genome (FASTA file), and the Omni-C contact matrix

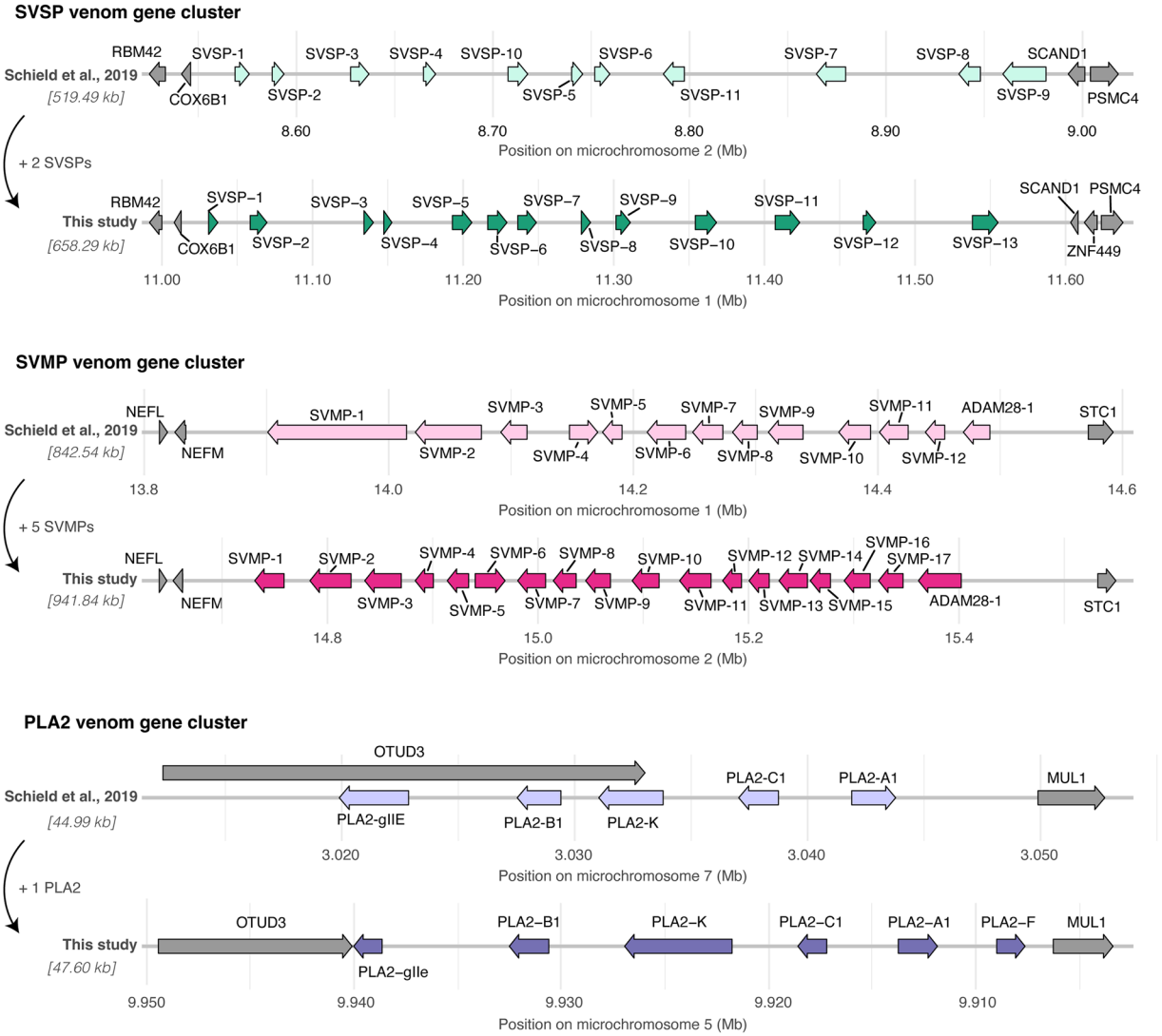


Fig. 4 All major viperid venom gene clusters see improved annotations. SVSPs, SVMPs and PLA2s are major contributors to the secreted venoms of most viperid snakes, and we highlight their annotations here to visualize the overall improvement in venom annotations. In each venom gene cluster diagram, the top row represents the prior annotation[5], and the bottom row represents this study's annotation. Venom gene colours also represent the prior annotation (lighter shades) and this study's annotation (darker shades). The conserved, non-venom flanking genes used to anchor syntenic regions at each cluster are shown in gray. To the left of each cluster diagram, the increase in the number of annotated paralogs is shown beside the arrow. Note that the PLA2 array is oriented in the reverse complement direction in the current assembly and is shown here in the forward orientation for clarity.

(.hic file). The GitHub repository (github.com/SidG13/croVir4) contains sequencing reports, scripts and smaller datafiles used to recreate major analyses, and scripts to reproduce key figures and perform several major data audits.

Technical Validation

We evaluated assembly quality on the basis of annotation completeness, chromosomal scaffolding consistency, and contamination screening. Annotation quality was assessed using BUSCO v6.0.0[71] against the Tetrapoda_odb12 dataset, yielding 94.6% complete single-copy orthologs, indicating a high recovery of conserved tetrapod genes. All 19 nuclear chromosomes (7 macrochromosomes, 10 microchromosomes, and 2 sex chromosomes) were recovered, with only ~0.42% of genome sequence unplaced. We also used the reference-free completeness approach from Merqury v1.3[72] to assess assembly consensus quality value (QV), resulting in a score of 41.8, surpassing 99.99% consensus sequence accuracy (QV > 40), reflecting an error rate of $6.6 \cdot 10^{-5}$ (1 error every ~15 kb). Assembly metrics were uniformly compared to other high-quality *Crotalus* genomes using assembly-stats v1.0.1 (github.com/sanger-pathogens/assembly-stats) and 'stats.sh' from bbmap v40.00 (github.com/BioInfoTools/bbmap). The mitochondrial genome assembled as a single complete circular contig with conserved gene order and full annotation support. Contamination screening ensured no non-eukaryotic sequences were present in the final assembly. Omni-C re-mapping was used to assess the accuracy of assembly into

Gene name	Best hit in prior genome	pident	qlen	slen	aln_length	query_coverage
Venom_ACHE-1	crovir-protein-17118	100	610	506	506	0.829508
Venom_Ficolin-1	crovir-protein-9182	47.4	337	669	662	0.997033
Venom_Ficolin-2	crovir-protein-9182	77.1	329	669	340	0.896657
Venom_3FTx-10	crovir-protein-8023	92.3	94	101	78	0.829787
Venom_PDE-1	crovir-protein-8584	99.8	453	471	452	0.997792
Venom_CRISP-1	crovir-protein-8572	96.3	242	113	107	0.442149
Venom_CRISP-2	crovir-protein-8572	97.2	242	113	107	0.442149
Venom_CRISP-3	crovir-protein-8573	93.2	241	196	88	0.365145
Venom_CRISP-4	crovir-protein-8573	86.4	241	196	88	0.365145
Venom_NUC-1	crovir-protein-8937	98	590	683	589	0.998305
Venom_NUC-2	crovir-protein-8937	97.8	590	683	589	0.998305
Venom_vQC-1	crovir-protein-9447	99.7	369	368	368	0.99729
Venom_MYO-1	NA	NA	NA	NA	NA	NA
Venom_Vespryn-1	crovir-protein-13215	100	219	220	218	0.995434
Venom_LAAO-1	crovir-protein-3893	92.5	517	445	441	0.852998
Venom_LAAO-2	crovir-protein-3893	87.3	508	445	441	0.822835
Venom_HYAL-1	crovir-protein-13660	93.7	475	444	474	0.997895
Venom_SVSP-1	crovir-protein-SVSP_1	98.1	263	262	262	0.996198
Venom_SVSP-2	crovir-protein-SVSP_1	80.8	263	262	261	0.992395
Venom_SVSP-3	crovir-protein-SVSP_5	96.1	264	260	259	0.981061
Venom_SVSP-4	crovir-protein-SVSP_5	97.7	264	260	260	0.984848
Venom_SVSP-5	crovir-protein-SVSP_8	80.5	259	258	257	0.992278
Venom_SVSP-6	crovir-protein-SVSP_9	94.1	261	288	254	0.931034
Venom_SVSP-7	crovir-protein-SVSP_8	96.1	259	258	258	0.996139
Venom_SVSP-8	crovir-protein-SVSP_5	74.5	259	260	259	0.992278
Venom_SVSP-9	crovir-protein-SVSP_8	78.9	268	258	266	0.992537
Venom_SVSP-10	crovir-protein-SVSP_6	77.3	117	242	66	0.547009
Venom_SVSP-11	crovir-protein-SVSP_4	80.7	260	260	259	0.992308
Venom_SVSP-12	crovir-protein-SVSP_5	75.2	258	260	258	0.988372
Venom_SVSP-13	crovir-protein-SVSP_7	90.7	258	258	258	0.996124
Venom_SVMP-1	crovir-protein-SVMP_2	71	613	599	613	0.998369
Venom_SVMP-2	crovir-protein-SVMP_2	68.7	611	599	613	0.998363
Venom_SVMP-3	crovir-protein-SVMP_2	69.7	488	599	488	0.995902
Venom_SVMP-4	crovir-protein-SVMP_4	72.3	447	611	611	0.997763
Venom_SVMP-5	crovir-protein-SVMP_2	72.6	423	599	441	0.992908
Venom_SVMP-6	crovir-protein-SVMP_2	86.9	602	599	602	0.998339
Venom_SVMP-7	crovir-protein-SVMP_2	71.4	611	599	611	0.998363
Venom_SVMP-8	crovir-protein-SVMP_2	88.5	603	599	602	0.998342
Venom_SVMP-9	crovir-protein-SVMP_2	95.5	603	599	602	0.998342
Venom_SVMP-10	crovir-protein-SVMP_2	80.4	593	599	602	0.983137
Venom_SVMP-11	crovir-protein-SVMP_6	89.4	610	595	625	0.998361
Venom_SVMP-12	crovir-protein-SVMP_7	97.4	478	536	427	0.872385
Venom_SVMP-13	crovir-protein-SVMP_8	94.7	489	458	451	0.881391
Venom_SVMP-14	crovir-protein-SVMP_9	92.9	609	567	608	0.998358
Venom_SVMP-15	crovir-protein-SVMP_7	86.8	485	536	433	0.870103
Venom_SVMP-16	crovir-protein-SVMP_10	80.7	612	496	611	0.998366
Venom_SVMP-17	crovir-protein-SVMP_11	99.5	619	618	618	0.998384
Venom_ADAM28-1	crovir-protein-SVMP_11	62.8	827	618	629	0.758162
Venom_3FTx-1	crovir-protein-17052	90	111	110	110	0.990991
Venom_3FTx-2	crovir-protein-17052	46	86	110	50	0.430233
Venom_3FTx-3	crovir-protein-17052	87	100	110	23	0.23
Venom_3FTx-4	NA	NA	NA	NA	NA	NA
Venom_3FTx-5	NA	NA	NA	NA	NA	NA
Venom_3FTx-6	NA	NA	NA	NA	NA	NA
Venom_3FTx-7	NA	NA	NA	NA	NA	NA
Venom_3FTx-8	NA	NA	NA	NA	NA	NA
Venom_3FTx-9	NA	NA	NA	NA	NA	NA

Continued

Gene name	Best hit in prior genome	pident	qlen	slen	aln_length	query_coverage
Venom_CYS-1	crovir-protein-5488	99.3	137	136	136	0.992701
Venom_NGF-1	crovir-protein-2707	45.7	242	221	151	0.599174
Venom_KUN-1	crovir-protein-10196	50	81	409	62	0.765432
Venom_KUN-2	crovir-protein-14273	58.8	85	576	51	0.6
Venom_KUN-3	crovir-protein-14273	56.9	85	576	51	0.6
Venom_Waprin-1	crovir-protein-8898	100	135	126	100	0.740741
Venom_pgKUN-4	crovir-protein-8900	68	75	306	50	0.626667
Venom_KUN-5	crovir-protein-8348	47.4	84	358	57	0.678571
Venom_LIPA-1	crovir-protein-3745	88.4	476	420	475	0.997899
Venom_BPP-1	crovir-protein-17319	87.3	229	231	245	0.995633
Venom_EXO-1	crovir-protein-11292	82.1	212	308	173	0.683962
Venom_HYAL-2	crovir-protein-11737	99.1	542	450	435	0.802583
Venom_HYAL-3	crovir-protein-13659	43.2	450	451	433	0.962222
Venom_PLB-1	crovir-protein-12042	89.6	554	502	556	0.998195
Venom_CTL-1	NA	NA	NA	NA	NA	NA
Venom_CTL-2	crovir-protein-564	67	143	103	103	0.692308
Venom_CTL-3	NA	NA	NA	NA	NA	NA
Venom_CTL-4	crovir-protein-564	87.4	149	103	103	0.691275
Venom_EXO-2	crovir-protein-1466	99	442	388	388	0.868778
Venom_vQC-2	crovir-protein-1322	83.6	382	240	268	0.701571
Venom_VEGF-1	crovir-protein-7188	40	133	133	105	0.75188
Venom_EXO-3	crovir-protein-11720	100	351	297	297	0.846154
Venom_PLA2-F	crovir-protein-PLA2_B1	42.9	146	138	126	0.863014
Venom_PLA2-A1	crovir-protein-PLA2_A1	100	139	138	138	0.992806
Venom_PLA2-C1	crovir-protein-PLA2_C1	100	137	136	136	0.992701
Venom_PLA2-K	crovir-protein-PLA2_K	97.6	138	154	126	0.913043
Venom_PLA2-B1	crovir-protein-PLA2_B1	100	116	138	115	0.991379
Venom_PLA2-gIIe	crovir-protein-PLA2_gIIE	98.8	92	156	85	0.923913
Venom_ENPP6-1	crovir-protein-4587	90	439	406	439	0.997722
Venom_CTL-5	crovir-protein-5477	89.8	167	169	167	0.994012
Venom_CTL-6	crovir-protein-17168	100	159	216	43	0.27044

Gene name	subject_coverage	Reciprocal best hit	evalue	bitscore	Novel relative to prior genome
Venom_ACHE-1	1	TRUE	0	1042	No
Venom_Ficolin-1	0.971599	FALSE	3.76E-139	409	No
Venom_Ficolin-2	0.502242	TRUE	3.61E-148	431	No
Venom_3FTx-10	0.772277	TRUE	2.06E-56	168	No
Venom_PDE-1	0.95966	TRUE	0	924	No
Venom_CRISP-1	0.946903	FALSE	1.01E-80	236	No
Venom_CRISP-2	0.946903	TRUE	2.50E-81	238	No
Venom_CRISP-3	0.44898	FALSE	7.72E-60	186	No
Venom_CRISP-4	0.44898	FALSE	1.49E-54	173	No
Venom_NUC-1	0.844802	TRUE	0	1152	No
Venom_NUC-2	0.844802	FALSE	0	1150	No
Venom_vQC-1	1	TRUE	3.33E-282	761	No
Venom_MYO-1	NA	NA	NA	NA	Yes
Venom_Vespryn-1	0.990909	TRUE	2.03E-165	454	No
Venom_LAAO-1	0.986517	TRUE	8.08E-304	826	No
Venom_LAAO-2	0.986517	FALSE	2.65E-279	763	No
Venom_HYAL-1	1	TRUE	0	918	No
Venom_SVSP-1	1	TRUE	1.18E-199	544	No
Venom_SVSP-2	0.996183	FALSE	3.19E-163	452	No
Venom_SVSP-3	0.996154	FALSE	5.61E-191	522	No
Venom_SVSP-4	1	TRUE	1.75E-194	531	No
Venom_SVSP-5	0.996124	FALSE	1.04E-150	420	No
Venom_SVSP-6	0.881944	TRUE	1.59E-180	496	No
Venom_SVSP-7	1	TRUE	1.42E-189	518	No

Continued

Gene name	subject_coverage	Reciprocal best hit	evalue	bitscore	Novel relative to prior genome
Venom_SVSP-8	0.996154	FALSE	5.92E-139	390	No
Venom_SVSP-9	0.996124	FALSE	5.75E-148	413	No
Venom_SVSP-10	0.272727	FALSE	1.50E-28	103	No
Venom_SVSP-11	0.996154	FALSE	9.35E-157	435	No
Venom_SVSP-12	0.992308	FALSE	3.61E-142	398	No
Venom_SVSP-13	1	TRUE	6.88E-174	478	No
Venom_SVMP-1	1	FALSE	0	884	No
Venom_SVMP-2	1	FALSE	0.00E+00	850	No
Venom_SVMP-3	0.806344	FALSE	1.02E-245	684	No
Venom_SVMP-4	1	TRUE	0.00E+00	842	No
Venom_SVMP-5	0.717863	FALSE	2.06E-218	612	No
Venom_SVMP-6	1	FALSE	0	1096	No
Venom_SVMP-7	1	FALSE	0	885	No
Venom_SVMP-8	1	FALSE	0	1122	No
Venom_SVMP-9	1	TRUE	0	1209	No
Venom_SVMP-10	1	FALSE	0	1011	No
Venom_SVMP-11	1	TRUE	0	1133	No
Venom_SVMP-12	0.796642	TRUE	1.28000000000001e-310	845	No
Venom_SVMP-13	0.984716	TRUE	2.2299999999738e-313	849	No
Venom_SVMP-14	1	TRUE	0	1162	No
Venom_SVMP-15	0.796642	FALSE	7.39E-277	760	No
Venom_SVMP-16	1	TRUE	0	984	No
Venom_SVMP-17	1	TRUE	0	1292	No
Venom_ADAM28-1	0.998382	FALSE	1.34E-294	822	No
Venom_3FTx-1	1	TRUE	8.51E-69	201	No
Venom_3FTx-2	0.454545	FALSE	9.14E-05	37.4	No
Venom_3FTx-3	0.209091	FALSE	6.58E-06	40.8	No
Venom_3FTx-4	NA	NA	NA	NA	Yes
Venom_3FTx-5	NA	NA	NA	NA	Yes
Venom_3FTx-6	NA	NA	NA	NA	Yes
Venom_3FTx-7	NA	NA	NA	NA	Yes
Venom_3FTx-8	NA	NA	NA	NA	Yes
Venom_3FTx-9	NA	NA	NA	NA	Yes
Venom_CYS-1	1	TRUE	8.66E-97	273	No
Venom_NGF-1	0.683258	TRUE	2.05E-37	130	No
Venom_KUN-1	0.139364	TRUE	4.51E-15	67.4	No
Venom_KUN-2	0.088542	TRUE	1.32E-16	72	No
Venom_KUN-3	0.088542	FALSE	4.59E-16	70.5	No
Venom_Waprin-1	0.793651	TRUE	1.01E-49	154	No
Venom_pgKUN-4	0.153595	TRUE	2.73E-17	72.8	No
Venom_KUN-5	0.159218	TRUE	4.06E-18	75.9	No
Venom_LIPA-1	1	TRUE	2.75E-308	834	No
Venom_BPP-1	1	TRUE	4.71E-105	302	No
Venom_EXO-1	0.561688	TRUE	6.38E-87	258	No
Venom_HYAL-2	0.966667	TRUE	0	899	No
Venom_HYAL-3	0.94235	TRUE	1.42E-137	402	No
Venom_PLB-1	1	TRUE	0	1003	No
Venom_CTL-1	NA	NA	NA	NA	Yes
Venom_CTL-2	1	FALSE	2.72E-49	152	No
Venom_CTL-3	NA	NA	NA	NA	Yes
Venom_CTL-4	1	TRUE	9.38E-70	204	No
Venom_EXO-2	1	TRUE	4.48E-256	699	No
Venom_vQC-2	0.983333	TRUE	1.18E-153	432	No
Venom_VEGF-1	0.781955	TRUE	2.71E-23	87.4	No
Venom_EXO-3	1	TRUE	1.26E-214	587	No
Venom_PLA2-F	0.884058	FALSE	5.74E-35	117	No
Venom_PLA2-A1	1	TRUE	1.80E-108	303	No

Continued

Gene name	subject_coverage	Reciprocal best hit	evalue	bitscore	Novel relative to prior genome
Venom_PLA2-C1	1	TRUE	1.88E-109	305	No
Venom_PLA2-K	0.818182	TRUE	1.00E-97	276	No
Venom_PLA2-B1	0.833333	TRUE	1.25E-93	265	No
Venom_PLA2-gIIe	0.544872	TRUE	1.84E-63	188	No
Venom_ENPP6-1	1	TRUE	5.62E-288	781	No
Venom_CTL-5	0.893491	TRUE	2.19E-112	315	No
Venom_CTL-6	0.199074	TRUE	8.75E-26	97.1	No

Table 5. Results of reciprocal best blast between prior and new *C. viridis* assemblies. Reciprocal best blast was used to assess the novelty of venom gene annotations relative to the prior genome assembly. Note that 3-finger toxins and myotoxin in particular are novel to the new genome assembly.

Class	Length (bp)	Masked (%)	% change
Class I: LTR-RTs	—	—	—
Ty1/Copia	9,754,417	0.62	+0.05
Ty3/Gypsy	78,048,565	4.92	+1.38
Other LTR	55,972,078	3.53	+0.81
Non-LTR-RTs	—	—	—
LINE	405,101,349	25.54	+6.29
SINE	52,622,108	3.32	+0.40
Class II: DNA Transposon	—	—	—
CACTA	983,355	0.06	0.00
Helitron	3,995,850	0.25	+0.08
MULE	1,266,847	0.08	−0.01
PIF-Harbinger	9,965,618	0.63	+0.06
Polinton	1,039,267	0.07	+0.04
Tc1 Mariner	62,880,185	3.97	−0.32
hAT	108,082,822	6.82	+0.56
Other	—	—	—
Low complexity	8,340,281	0.53	+0.17
RNA	1,903,285	0.12	+0.04
Satellite	4,634,531	0.29	+0.08
Simple repeat	59,794,164	3.77	+1.45
Other	26,282,296	1.66	−0.05
Unknown	20,460,261	1.29	+0.11
Total	828,453,241	52.24	+10.31

Table 6. Description of repeat elements annotated in the updated chromosome-level *C. viridis* assembly. Percent change represents the change in masked percent from the prior annotation[5] to the current annotation. Because of some overlaps in repeat annotations, values listed in columns do not sum to totals shown.

chromosomes. Finally, integration of short-read mRNA-seq and Iso-Seq data confirmed expression-supported annotation of venom genes, including dozens of loci absent, or fragmented in the previous assembly. 62.5% of venom genes were supported by at least 1 Iso-Seq transcript, 43.2% had intron structures fully supported by RNA-seq splice junctions (71.9% of all introns individually supported) and 88.6% had validated signal peptides (Table 4). To perform an audit of the annotation file, which included counting genes and GTF validation, we used the AGAT tools scripts "agat_sp_statistics.pl" and "agat_convert_sp_gxf2gxf.pl" which returned no errors. Additionally, protein RBB suggests that several genes previously fragmented or missing in the prior assembly are correctly assembled and annotated in the current assembly. Results of these technical validation steps, as well as figure and other key analysis scripts can be found on GitHub at github.com/SidG13/croVir4. Overall, these results demonstrate a greatly improved, contiguous, and biologically complete genome assembly (Table 1) suitable for downstream comparative and functional genomic analyses.

Usage Notes

The custom gene annotation file, in GTF format, includes the annotation of all 88 manually curated venom genes, which are prepended with the string "Venom" in the gene name column. The gene name column contains standard Ensembl identifiers for non-venom genes, if available. Human-rattlesnake paralogs (many-to-one assignments due to gene duplications between human and rattlesnake) are denoted using a '.#' suffix to the gene name.

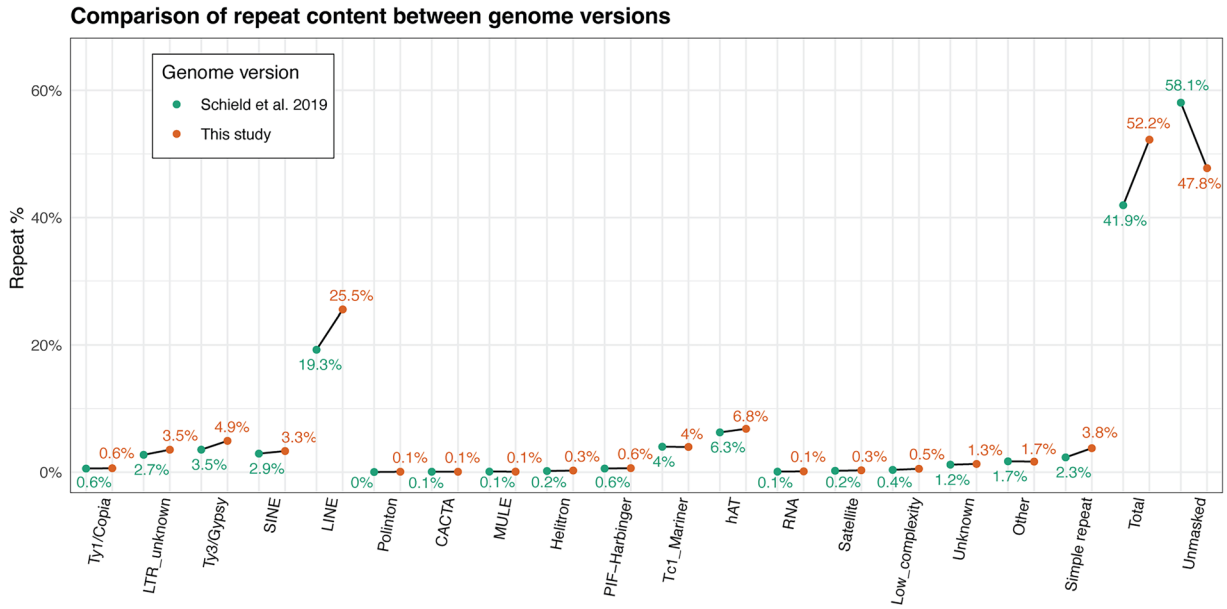


Fig. 5 Head-to-head comparison of repeat annotations demonstrates marked increase in annotated repeat element content in the updated *C. viridis* genome. The increase in repeat content is primarily driven by LINEs, Ty3/Gypsy, simple repeats and hAT elements. Each point represents the fraction of a repeat family masked in the Schield *et al.*, 2019 assembly[5] (green) and this study's assembly (orange). The percentage of the genome identified as a repeat has increased by 10.31%, supporting a more complete genome assembly and annotation of repeat elements.

Data availability

The genome assembly (NCBI GenBank accession: GCA_059756665.2) and raw data were submitted to the NCBI under BioProject accession PRJNA1454988 and the SRA accession SRR38315645. The custom genome, gene annotation including manually annotated venom genes, and files associated with repeat annotation are available on Zenodo at https://doi.org/10.5281/zenodo.21683217. Scripts and files to reproduce technical validation are found on GitHub at https://github.com/SidG13/croVir4. All other data requests should be sent to the Corresponding Author.

Code availability

Custom code associated with this manuscript can be found on GitHub at https://github.com/SidG13/croVir4.

Received: 16 June 2026; Accepted: 22 October 2026;
Published: xx xx xxxx

References

1. Mackessy, S. P. Evolutionary trends in venom composition in the Western Rattlesnakes (Crotalus viridis sensu lato): Toxicity vs. tenderizers. *Toxicon* **55**, 1463–1474 (2010).
2. Cameron, D. L. & Tu, A. T. Characterization of myotoxin a from the venom of prairie rattlesnake (Crotalus viridis viridis). *Biochemistry* **16**, 2546–2553 (1977).
3. Gleason, M. L., Odell, G. V. & Ownby, C. L. Isolation and Biological Activity of Viriditoxin and a Viriditoxin Variant from Crotalus Viridis Viridis Venoms. *J. Toxicol. Toxin Rev.* **2**, 235–265 (1983).
4. Ownby, C. L., Cameron, D. & Tu, A. T. Isolation of myotoxic component from rattlesnake (Crotalus viridis viridis) venom. Electron microscopic analysis of muscle damage. *Am. J. Pathol.* **85**, 149–166 (1976).
5. Schield, D. R. *et al.* The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res.* **29**, 590–601 (2019).
6. Smith, C. F. *et al.* Snakes on a plain: biotic and abiotic factors determine venom compositional variation in a wide-ranging generalist rattlesnake. *BMC Biol.* **21**, 136 (2023).
7. Saviola, A. J. *et al.* 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 (2015).
8. Perry, B. W. *et al.* Multi-species comparisons of snakes identify coordinated signalling networks underlying post-feeding intestinal regeneration. *Proc. R. Soc. B Biol. Sci.* **286**, 20190910 (2019).
9. Gopalan, S. S. *et al.* Diverse Gene Regulatory Mechanisms Alter Rattlesnake Venom Gene Expression at Fine Evolutionary Scales. *Genome Biol. Evol.* **16**, evae110 (2024).
10. Perry, B. W. *et al.* Snake venom gene expression is coordinated by novel regulatory architecture and the integration of multiple co-opted vertebrate pathways. *Genome Res.* (2022).
11. Schield, D. R. *et al.* Snake Recombination Landscapes Are Concentrated in Functional Regions despite PRDM9. *Mol. Biol. Evol.* **37**, 1272–1294 (2020).
12. Hoge, C. *et al.* Patterns of recombination in snakes reveal a tug-of-war between PRDM9 and promoter-like features. *Science* **383**, eadj7026 (2024).

13. Pasquesi, G. I. *et al*. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 1–11 (2018).
14. Schield, D. R. *et al*. The Rattlesnake W Chromosome: A GC-Rich Retroelement Refugium with Retained Gene Function Across Ancient Evolutionary Strata. *Genome Biol. Evol.* **14**, evac116 (2022).
15. Waters, P. D. *et al*. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proc. Natl. Acad. Sci.* **118**, e2112494118 (2021).
16. Schield, D. R. *et al*. The roles of balancing selection and recombination in the evolution of rattlesnake venom. *Nat. Ecol. Evol.* **6**, 1367–1380 (2022).
17. Francioli, Y. Z. *et al*. Estimation of genome-wide coupling in rattlesnake hybrids provides insight into the process of speciation and its progress. *Nat. Commun.* **16**, 10242 (2025).
18. Schield, D. R. *et al*. Allopatric divergence and secondary contact with gene flow: a recurring theme in rattlesnake speciation. *Biol. J. Linn. Soc.* **128**, 149–169 (2019).
19. Farleigh, K. *et al*. Evolution of genome-wide barriers to gene flow during complex speciation in rattlesnakes. *Proc. Natl. Acad. Sci.* **123**, e2609058123 (2026).
20. Nikolakis, Z. L. *et al*. Evidence that genomic incompatibilities and other multilocus processes impact hybrid fitness in a rattlesnake hybrid zone. *Evolution* **76**, 2513–2530 (2022).
21. Smith, C. F. *et al*. The best of both worlds? Rattlesnake hybrid zones generate complex combinations of divergent venom phenotypes that retain high toxicity. *Biochimie* **213**, 176–189 (2023).
22. Hargreaves, A. D., Swain, M. T., Hegarty, M. J., Logan, D. W. & Mulley, J. F. Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. *Genome Biol. Evol.* **6**, 2088–2095 (2014).
23. Dowell, N. L. *et al*. The Deep Origin and Recent Loss of Venom Toxin Genes in Rattlesnakes. *Curr. Biol.* **26**, 2434–2445 (2016).
24. Casewell, N. R., Wüster, W., Vonk, F. J., Harrison, R. A. & Fry, B. G. Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol. Evol.* **28**, 219–229 (2013).
25. Hogan, M. P. *et al*. The genetic regulatory architecture and epigenomic basis for age-related changes in rattlesnake venom. *Proc. Natl. Acad. Sci.* **121**, e2313440121 (2024).
26. Almeida, D. D. *et al*. Tracking the recruitment and evolution of snake toxins using the evolutionary context provided by the Bothrops jararaca genome. *Proc. Natl. Acad. Sci.* **118**, e2015159118 (2021).
27. Gopalan, S. S. *et al*. Origins, genomic structure and copy number variation of snake venom myotoxins. *Toxicon* **216**, 92–106 (2022).
28. Linz, W., Wiemer, G., Gohlke, P., Unger, T. & Schölkens, B. A. Contribution of kinins to the cardiovascular actions of angiotensin-converting enzyme inhibitors. *Pharmacol. Rev.* **47**, 25–49 (1995).
29. Mebs, D. & Ownby, C. L. Myotoxic components of snake venoms: Their biochemical and biological activities. *Pharmacol. Ther.* **48**, 223–236 (1990).
30. Margres, M. J. *et al*. The Tiger Rattlesnake genome reveals a complex genotype underlying a simple venom phenotype. *Proc. Natl. Acad. Sci.* **118** (2021).
31. Suryamohan, K. *et al*. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nat. Genet.* **52**, 106–117 (2020).
32. Castoe, T. A. *et al*. A Vision for VenomsBase: An Integrated Knowledgebase for the Study of Venoms and Their Applications. *Integr. Org. Biol.* **7**, obaf026 (2025).
33. Jenner, R. A., Casewell, N. R. & Undheim, E. A. B. What is animal venom? Rethinking a manipulative weapon. *Trends Ecol. Evol.* **40**, 852–861 (2025).
34. Mochales-Riaño, G. *et al*. Chromosome-level reference genome for the medically important Arabian horned viper (Cerastes gasperettii). *GigaScience* **14**, giaf030 (2025).
35. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
36. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
37. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
38. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
39. Alonge, M. *et al*. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
40. Perry, B. W., Schield, D. R., Adams, R. H. & Castoe, T. A. Microchromosomes Exhibit Distinct Features of Vertebrate Chromosome Structure and Function with Underappreciated Ramifications for Genome Evolution. *Mol. Biol. Evol.* **38**, 904–910 (2021).
41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
42. Roseman, M. A. *et al*. Insights from the timber rattlesnake (Crotalus horridus) genome for MHC gene architecture and evolution in threatened rattlesnakes. *J. Hered.* **116**, 591–602 (2025).
43. Peng, C. *et al*. Large-scale snake genome analyses provide insights into vertebrate development. *Cell* **186**, 2959–2976.e22 (2023).
44. Geneva, A. J. *et al*. Chromosome-scale genome assembly of the brown anole (Anolis sagrei), an emerging model species. *Commun. Biol.* **5**, 1126 (2022).
45. Baker, R. J., Mengden, G. A. & Bull, J. J. Karyotypic Studies of Thirty-Eight Species of North American Snakes. *Copeia* **1972**, 257–265 (1972).
46. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
47. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst.* **6**, 256–258.e1 (2018).
48. Hahn, C., Bachmann, L. & Chevreux, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* **41**, e129 (2013).
49. Westeen, E. P. *et al*. A genome assembly for the southern Pacific rattlesnake, Crotalus oreganus helleri, in the western rattlesnake species complex. *J. Hered.* **114**, 681–689 (2023).
50. Bernt, M. *et al*. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* **69**, 313–319 (2013).
51. Yan, J., Li, H. & Zhou, K. Evolution of the mitochondrial genome in snakes: Gene rearrangements and phylogenetic relationships. *BMC Genomics* **9**, 569 (2008).
52. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
53. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
54. Krueger, F. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. *Babraham Inst.* https://cir.nii.ac.jp/crid/1370294643762929691 (2015).
55. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
56. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
57. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
58. Nachtigall, P. G., Durham, A. M., Rokyta, D. R. & Junqueira-de-Azevedo, I. L. M. ToxCodAn-Genome: an automated pipeline for toxin-gene annotation in genome assembly of venomous lineages. *GigaScience* **13**, giad116 (2024).
59. Robinson, J. T. Integrative genomics viewer (2011).

60. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
61. Teufel, F. *et al*. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
62. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
63. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
64. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
65. Castoe, T. A. *et al*. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc. Natl. Acad. Sci.* **110**, 20645–20650 (2013).
66. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **5**, 4.10.1–4.10.14 (2004).
67. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
68. Gopalan, S. S. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_059756665.2 (2026).
69. *NCBI Sequence Read Archive* http://identifiers.org/insdc.sra:SRP695941 (2026).
70. Gopalan, S. croVir4 custom genome and annotation. *Zenodo* https://doi.org/10.5281/zenodo.21683217 (2026).
71. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
72. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

Author contributions

S.S.G. and T.A.C. conceived of the study. Y.Z.F., H.D.G., B.W.P., D.R.S., A.J.S, R.H.A, S.S.G. and S.P.M. collected samples and generated data. S.S.G., Y.Z.F., D.C.C., K.B. and I.K. performed computational analyses and formal investigation. S.S.G. and T.A.C. wrote the manuscript. All authors contributed to editing the final manuscript.

Funding

This work was funded by National Science Foundation grant IOS-2307044 awarded to T.A.C, S.P.M., R.H.A., and A.J.S., and a UTA RISE-100 Postdoctoral Fellowship to S.S.G.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.