Chromosome-level genome and the identification of sex chromosomes in Uloborus diversus.

Authors: Jeremiah Miller[1], Aleksey V Zimin[2,3], Andrew Gordus[1,4]

[1] Department of Biology, Johns Hopkins University, Baltimore, MD

[2] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

[3] Center for Computational Biology, Johns Hopkins University, Baltimore, MD

[4] Solomon H. Snyder Department of Neuroscience, Johns Hopkins University, Baltimore, MD

Abstract

The orb-web is a remarkable example of animal architecture that is observed in families of spiders that diverged over 200 million years ago. While several genomes exist for Araneid orb-weavers, none exist for other orb-weaving families, hampering efforts to investigate the genetic basis of this complex behavior. Here we present a chromosome-level genome assembly for the cribellate orb-weaving spider *Uloborus diversus*. The assembly reinforces evidence of an ancient arachnid genome duplication and identifies complete open reading frames for every class of spidroin gene, which encode the proteins that are the key structural components of spider silks. We identified the two X chromosomes for *U. diversus* and identify candidate sex-determining genes. This chromosome-level assembly will be a valuable resource for evolutionary research into the origins of orb-weaving, spidroin evolution, chromosomal rearrangement, and chromosomal sex-determination in spiders.

Introduction

Spiders are among the most successful and diverse terrestrial predators on Earth. Almost 400 million years of evolution has produced more than 50,000 extant spider species representing 128 families that are distributed over every continent except Antarctica(Gloor et al. 2017). The success of these animals is due in part to the diversity of behaviors that have evolved to capture prey in different environments(Vollrath and Selden 2007). Many spiders attack their prey by physically grabbing and immobilizing them with venom and use their silk exclusively for egg sacs. Others use silk to line their burrows, or construct webs of varying geometry and composition to detect or entrap prey. The diversity of web use correlates with a diversity of spidroin proteins that form silk, as well as the glands that produce these proteins(Vollrath and Selden 2007; Blackledge et al. 2009; Gatesy et al. 2001; Foelix 2011; Vollrath 1999). Spiders such as orb-weavers alternate between glands depending upon the web feature they are constructing. For example, load bearing parts of the web such as the radii are composed of major ampullate silk that has high tensile strength, whereas the anchors are made up of pyriform silk which is sticky and amorphous(Foelix 2011).

Remarkably, the orb web is not restricted to a single monophyletic group, but is observed in two lineages that diverged 250 million years ago, leading to considerable debate about its evolutionary origins(Blackledge et al. 2009; Fernández et al. 2018; Coddington et al. 2019; Kallal et al. 2021) (**Figure 1**). Araneoidea is the largest superfamily of orb weavers, which have evolved adhesive aggregate spidroins that are used in the capture spiral to adhere prey to the web(Gatesy et al. 2001; Sahni et al. 2010; Opell and Hendricks 2010; Hayashi and Lewis 2000, 1998). However, uloborids also build orb webs, but use a more ancient cribellate spidroin in their capture spiral to immobilize prey(Peters 1992, 1984; Blackledge and Hayashi 2006; Piorkowski et al. 2020). In addition to Uloboridae, other families such as Deinopidae, and Oecobiidae + Hersiliidae (UDOH grade in **Figure 1**) also build orb-webs, but with more derived behavioral and structural characteristics(Coddington 1986). Orb-weaving is an innate behavior, with discrete stages of web construction that are shared between araneoid and non-araneoid orb-weavers(Zschokke and Vollrath 1995). When exposed to neuroactive compounds, the behaviors within specific stages are altered, indicating that the neuronal targets of these compounds are more important for certain stages than for others(Witt and Reed 1965; Hesselberg and Vollrath 2004; Reed et al. 1982). This behavioral paradigm offers an excellent model for understanding not only how complex behaviors can be organized in a small brain(Corver et al. 2021), but also how such behaviors evolve.

However, a genetic understanding of the evolution of orb weaving behavior is hampered by a lack of sequenced genomes for non-araneoid families. Spider genomes are enormous with high repeat content, making them challenging to assemble(Sanggaard et al. 2014; Stellwagen and Renberg 2019; Ayoub et al. 2013). While 22 spider genomes have been assembled and made publicly available(Sanggaard et al. 2014; Babb et al. 2017; Kono et al. 2019; Sheffer et al. 2021; Sánchez-Herrero et al. 2019; Schwager et al. 2017; Fan et al. 2021; Yu et al. 2019; Liu et al. 2019; Escuer et al. 2022; Cerca et al. 2021; Zhu et al. 2022; Hendrickx et al. 2022; Kono et al. 2021a; Purcell and Pruitt 2019; Li et al. 2022; Kono et al. 2021b) (**Table S1**), most are highly fragmented, with 6 assembled to chromosome-scale(Sheffer et al. 2021; Fan et al. 2021; Escuer et al. 2022; Zhu et al. 2022; Hendrickx et al. 2022; Kono et al. 2021b). Of the 22 genomes, only 7 represent the Araneoidea(Kono et al. 2019; Sheffer et al. 2021; Fan et al.

2021; Kono et al. 2021a) or ecribellate orb-weavers, 2 of which have been assembled to chromosome-scale(Sheffer et al. 2021; Fan et al. 2021). There are no published genomes for the UDOH clade, and only 1 genome represents a member of the cribellate RTA clade (**Figure 1B**)(Garb et al. 2018). Chromosome-level assemblies are essential for understanding evolutionary divergence and identifying sites of chromosomal reorganization that play roles in adaptation and speciation.

Spiders have multiple sex chromosomes, with ♂X_1X_2/♀$X_1X_1X_2X_2$ being the most common sex determination observed. Sex chromosome dosage compensation has evolved multiple times, but all known genetic mechanisms are for single sex chromosome systems. A molecular understanding of dosage compensation in spiders is lacking, in part due to a paucity of sex-associated genetic loci.

Uloborus diversus (**Figure 1A**) of the UDOH clade, is a cribellate orb weaving spider native to the desert Southwest in the United States(Eberhard 1971) and an important model for understanding the evolution of spidroins and orb-weaving(Garb et al. 2018). The existence of a non-araneoid orb-weaving genome is crucial for addressing the evolutionary origins of orb-weaving. Recent work has demonstrated the utility of this species as a model system for understanding orb-weaving behavior(Corver et al. 2021). This, combined with the potential to compare both behaviors and their genetic underpinnings across divergent species of orb-weavers offers a rich opportunity to understand the underlying genetics that encode this behavior, and whether orb-weaving behaviors are conserved or convergently evolved. Here, we report a high-quality, chromosome-scale draft genome assembly of *Uloborus diversus* (NCBI: txid327109), as well as a complementary transcriptome assembly and gene annotations. This genome enabled the identification of full >10kb spidroin genes, as well as the identification of sex chromosomes for this species. This chromosome-level assembly will be a valuable resource for evolutionary research into the origins of orb-weaving, spidroin evolution, chromosomal rearrangement, and chromosomal sex-determination in spiders.

Results

Genome Sequencing

We sequenced and assembled a high-quality, chromosome-scale genome assembly for *Uloborus diversus* using a hybrid approach that leveraged the complementary benefits of multiple technologies. The genome of *U. diversus* contains long regions of low-complexity sequence, which hinders assembly using short-reads alone, as well as extremely long protein-coding genes, which makes long-reads necessary for a reference-quality assembly(Ayoub et al. 2013; Sanggaard et al. 2014; Babb et al. 2017; Kono et al. 2019; Stellwagen and Renberg 2019; Sheffer et al. 2021). Illumina sequencing provides high sequence fidelity but short read lengths, while ONT sequencing provides long read lengths, useful for scaffolding and spanning long, low complexity regions, but lower sequence fidelity(Giani et al. 2020). PacBio HiFi sequencing provides an excellent combination of long read lengths and high sequence fidelity, however, we were able to produce multiple megabase-long reads with ONT, which is not possible with PacBio. Each of these sequencing technologies provided unique advantages for improving the overall assembly.

To limit genetic variation, we used sequencing data from only 5 unmated female spiders in our assembly. We used Illumina to obtain high fidelity read data with from a single female spider, generating 795M 150 bp read pairs totaling 119.3 Gb. Because Illumina short reads are not sufficiently long to span long, highly repetitive regions encountered in spider genomes, we sequenced 3 ONT libraries, each from a single female, generating 14.7M reads totaling 98.4 Gb, with a read N50 of 6.7 kb. To obtain long sequencing reads with high sequence fidelity, we also sequenced a single adult female using PacBio HiFi, generating 35M subreads totaling 412.8 Gb with a read N50 of 12.8 kb, which yielded 2M consensus reads totaling 26 Gb with a read N50 of 13.0 kb. To investigate the sex determination system in *U. diversus*, we also generated an Illumina library from a single adult male, producing 937M 50 bp read pairs totaling 46.8 Gb. Sequencing library statistics are available in **Table 1**.

Genome Size, Heterozygosity, and Coverage Estimation

To assess the size and heterozygosity of the genome, we used Jellyfish(Marçais and Kingsford 2011) to count the frequency of canonical 21-mers in our adult female Illumina sequencing reads and used the 21-mer distribution as input to GenomeScope(Ranallo-Benavidez et al. 2020). The resulting model estimated a genome size of 1.98 Gb with a heterozygosity of 1.38% and 50.2% of the genome occurring as unique sequence (**Figure 2**), similar to other spider genomes (**Supplementary Table S1**)(Sanggaard et al. 2014; Babb et al. 2017; Kono et al. 2019; Sheffer et al. 2021; Sánchez-Herrero et al. 2019; Schwager et al. 2017; Fan et al. 2021; Yu et al. 2019; Liu et al. 2019). Given this genome size, our Illumina sequencing yielded 63x

coverage, our ONT sequencing yielded 68x coverage, and our PacBio HiFi yielded 208x in raw read coverage and 13x in consensus read coverage (**Table 1**).

Karyotype of *U. diversus*

To infer the expected number of pseudo-chromosomes in our final assembly, we determined the number of chromosomes in *U. diversus* using metaphase karyotyping. Mitotic chromosome spreads from developing embryos displayed two distinct patterns of chromosome number; either 18 or 20 (**Figure 3A**), consistent with ♂X_1X_2/♀$X_1X_1X_2X_2$ sex determination, which is the most common form of sex determination observed in spiders(Sember et al. 2020). Thus, *U. diversus* appears to have 8 autosomes and 2 sex chromosomes.

De novo Nuclear Genome Assembly

First, we used MaSuRCA(Zimin et al. 2013, 2017) to produce an initial assembly (*U. diversus* v.1.0) using Illumina short-read data scaffolded by ONT long-read data, consisting of 68,259 scaffolds spanning 3.22 Gb. The scaffold N50 was 98,014 bp and the scaffold L50 was 6,558 (**Table 2**), with 94.7% of complete BUSCOs (**Table 3**). The inferred redundancy accounts for the significant increase in the length of the assembly compared to the expected genome size. High heterozygosity leads to alternative haplotypes that can often be misassembled into their own contigs.

Next, we used Rascaf(Song et al. 2016) to improve continuity and ordering of scaffolds in the initial MaSuRCA assembly. Rascaf uses paired-end RNA-seq reads to improve the contiguity of gene models and scaffolds. We observed a modest improvement, reducing the number of scaffolds from 68,259 to 63,265 with the scaffold N50 increasing from 98,014 bp to 108,431 bp and the scaffold L50 decreasing from 6,885 to 5,994, with no change in the assembly span (**Table 2**). However, despite identifying 95.4% of BUSCOs, 22.5% were duplicated (**Table 3**)**.** This, combined with the large span of the genome, indicated a high degree of redundancy in the assembly.

To filter out redundant heterozygous contigs we used Pseudohaploid(Chen et al. 2019b). Pseudohaploid filters suspected homologous contigs and selects a single representative contig where high rates of heterozygosity prevent assemblers from appropriately identifying haplotypes. We reduced the number of scaffolds by 23.1%, with an increase in scaffold N50 and a decreased span from 3.22 Gb to 2.8 Gb (**Table 2**) and a drop in duplicated BUSCOs to 16.7% (**Table 3**). This suggests that Pseudohaploid was able to accurately collapse much of the redundant sequence attributable to alternative haplotypes.

To further improve our assembly, we sequenced using PacBio HiFi and assembled the resulting HiFi reads with PacBio's IPA (Improved Phased Assembly) pipeline, resulting in a substantial decrease in the new assembly span, from 2.8 Gbp to only 2.1 Gbp. The number of scaffolds in this assembly was only 9,734, a remarkable improvement. The scaffold N50 in the IPA assembly was 328,082 bp and the scaffold L50 in the IPA assembly was 1,789 (**Table 2**). The BUSCO score for the IPA assembly indicated that this assembly contained 92.3% of BUSCOs complete (83.7% in single copy, 9.3% duplicated) (**Table 3**).

We then used SAMBA(Zimin and Salzberg 2022) to merge the previous MaSuRCA assembly with the IPA assembly. SAMBA used scaffolds from the MaSuRCA assembly to merge the scaffolds in the IPA assembly. The new assembly had a length of 2.1 Gbp from only 7,197 scaffolds, with a scaffold N50 of 496,769 bp (**Table 2**), and 94% complete BUSCOs (**Table 3**).

To generate a chromosome-level assembly, we used HiRise to scaffold the IPA+MaSuRCA assembly with a Dovetail Hi-C library(Putnam et al. 2016). Scaffolding did not change the amount of sequencing in the assembly; however, the total number of scaffolds was reduced by 78% to only 1,586 scaffolds, with a remarkable improvement in scaffold N50, which increased to 185,519,777 bp in the Hi-C scaffolded assembly (**Table 2, Figure 3C**). Most importantly, 88% of the total assembly was represented by 10 large scaffolds that comprise 1.9 Gbp (**Figure 3B**), matching the expected number of chromosomes (**Figure 3A**). The BUSCO score for the final assembly showed 94.8% of the BUSCOs were complete (with 85.2% in single copy, 9.6% duplicated) (**Table 3**). Our final chromosome-level genome assembly statistics are consistent with previously published spider genomes, with 10 scaffolds representing the 10 chromosomes and high scaffold N50(Sanggaard et al. 2014; Babb et al. 2017; Kono et al. 2019; Sheffer et al. 2021; Sánchez-Herrero et al. 2019; Schwager et al. 2017; Fan et al. 2021; Yu et al. 2019; Liu et al. 2019) (**Supplementary Table S1**). We will refer to these 10 scaffolds as pseudochromosomes.

Repeat Annotation

To characterize repetitive sequences, we constructed a species-specific repeat library using RepeatModeler2(Flynn et al. 2020) This library was used in conjunction with the RepBase RepeatMasker Edition(Bao et al. 2015) database for masking the genome. RepeatMasker analysis of the combined *U. diversus* and RepBase repeats masked 66.6% of the final *U. diversus* genome assembly. Many (29.27%) of the repetitive regions were unclassified; however, DNA transposons accounted for a similar proportion (22.73%). Retroelements accounted for a much smaller proportion (7.7%). Total interspersed repeats account for 59.7% and simple repeats cover 2.61% of the genome (**Table 4**). The disparity between the GenomeScope estimation of 49.8% repetitive sequence and the RepeatMasker estimation of 66.6% suggests that the repeat content may be underestimated by GenomeScope. Therefore, the genome size may also be underestimated by GenomeScope. Because the length of the *U. diversus* genome is slightly less than the prediction by GenomeScope, this possibility suggests that the length of the assembly may more closely represent the true length of the *U. diversus*

genome than the GenomeScope prediction. The repeat content is typical of spider genomes (**Supplemental Table S2**).

Transcriptome Sequencing and Assembly

To identify protein-coding genes, we assembled a transcriptome. To capture a wide range of transcripts, we extracted RNA from spiders at multiple developmental stages and from male and female adults. We produced an Illumina short-read sequencing library from: a whole adult female, a whole adult male, the dissected prosoma (cephalothorax) from an adult female, the dissected opisthosoma (abdomen) from the same adult female, the dissected prosoma from an adult male, the dissected opisthosoma from the same adult male, the pooled legs from the dissected male and female, a single 4th instar female, and approximately 30 pooled 2nd instars. 302M read pairs were generated, totaling 45.4 Gbp.

We used Trinity(Haas et al. 2013) to assemble a genome-guided transcriptome. We then used TransDecoder(Haas et al. 2013) to find coding regions within our transcripts. We included homology searches to known proteins using both BLAST (Basic Local Alignment Search Tool)(Altschul et al. 1990; Altschul 1997) and Pfam(Mistry et al. 2021) searches. We assessed the BUSCO score of the long ORFs predicted by TransDecoder, finding that 90.9% of the BUSCOs were present and complete, with 54.8% single copy and 36.1% duplicated, with 1.5% present but fragmented and 7.6% missing. (**Table 3**).

Protein-Coding Gene Annotation

For protein coding gene annotations, we used BRAKER2(Hoff et al. 2016; Brůna et al. 2021) with our RNAseq data and homology evidence using a custom library of spider proteins obtained from NCBI (**Supplemental Table S3**). The number of predicted genes in the final *U. diversus* assembly was 44,408; with 40,466 models predicted on the 10 pseudochromosomes (**Table 5**), with 86.7% of complete BUSCOs. To functionally annotate these genes, we used Interproscan(Quevillon et al. 2005; Jones et al. 2014) to annotate the longest CDS for each gene. 30,911 models were assigned a domain or function from one or more of the databases used (**Table 6**).

Non-Coding RNA Annotation

We used tRNAscan-SE(Chan and Lowe 2019; Chan et al. 2021) to annotate transfer RNAs. We found 3,084 tRNAs coding for the standard 20 amino acids and 14 tRNAs coding for selenocysteine (TCA) tRNAs. We found 21 tRNAs with undetermined or unknown isotypes, 537 tRNAs with mismatched isotypes, and 57,824 putative tRNA pseudogenes. We identified no putative suppressor tRNAs. We used Barrnap(Seeman 2018) to annotate ribosomal rRNAs. We found 114 rRNAs, of which 100 were located on the 10 pseudo-chromosomes. These included: 6 copies of the 18S subunit, with 4 on pseudochromosomes; 83 copies of the 5S subunit, with

81 on pseudochromosomes; 6 copies of the 5.8S subunit, with 2 on pseudochromosomes; and 19 copies of the 28S subunit, with 13 pseudochromosomes.

Mitogenome Assembly

Animal mitochondrial genomes comprise 37 genes: 13 protein coding genes, 22 tRNAs, 2 rRNAs, and at least one control region(Boore 1999). We assembled the mitochondrial genome sequence with NOVOplasty(Dierckxsens et al. 2016) using the adult female Illumina DNA seq data and each of the mitochondrial genome sequences listed in **Supplemental Table S4** as sources for seed sequences(Wang et al. 2019b; Zhu and Zhang 2017; Wang et al. 2014; Liu et al. 2015; Fang et al. 2016; Masta and Boore 2008; Kumar et al. 2020; Li et al. 2016; Qiu et al. 2005; Pan et al. 2014; Kim et al. 2014; Pan et al. 2016; Tian et al. 2016; Wang et al. 2016a). Each run produced the same single, circularized 14,737 bp mitochondrial sequence, consistent with the expected size for an arachnid mitochondrial sequence(Boore 1999). We annotated the sequence with the MITOS2 web server(Bernt et al. 2013) and found all 13 of the expected protein coding genes, 20 of 22 tRNAs, 2 rRNAs, and the control region (**Fig. 3E**). All identified tRNAs were truncated and lacked T-arms, which is unique to spiders and has been observed in other species(Masta and Boore 2008; Wang et al. 2016b; Li et al. 2016; Pons et al. 2019).

Identification and Analysis of Spidroins

Spidroins are a unique class of proteins that are the primary components of spider silk. While all spiders produce silk, spidroins have evolved for different uses in web-making. Orb-weavers in particular evolved several silk glands that each produce a different repertoire of spidroins to make different silks with varying utility. Several ecribellate Araneidae spidroins have been sequenced, and many of these spidroins are also made by cribellate orb-weavers such as *U. diversus*. However, ecribellate spiders evolved a unique type of hydrated flageliform silk for their capture spiral, whereas cribellate spiders such as *U. diversus* use a dry cribellate silk in their capture spirals.

Spider dragline silk has the strongest stress and strain capabilities of any known substance. Interest in silk properties extends beyond their evolved use, as silk has many potential human applications in both industry and medicine(Kumari et al. 2020; Xu et al. 2019; Öksüz et al. 2021; Choi and Choy 2020; Mayank et al. 2022; Liu et al. 2020; Lewis 2006; Teulé et al. 2007). However, the genetic characterization of spidroins is often challenging due to their exceptional length (coding regions >5 kb) and high repeat content(Stellwagen and Renberg 2019). The annotation of these genes is difficult and often fragmented because reads rarely span the entire length of these genes. With the exceptional contiguity and read depth of our assembly, due to the diversity of sequencing technologies employed, we identified the entire open-read frames of all major spidroins in the *U. diversus* genome.

We found 10 full-length candidate sequences, including at least one candidate for each of the seven types of spidroin used by cribellate orb-weavers. There were no gaps in the assembly interrupting any of our candidate spidroin sequences. We performed read mapping to validate

the continuity of each full-length sequence and ensure that the predicted sequences were not chimeric. In 8 cases, including 3 minor spidroin (MiSp) candidates, a major spidroin (MaSp) MaSp-1 candidate, 2 MaSp-2 candidates, a tubuliform spidroin (TuSp) candidate, and an aciniform spidroin (AcSp) candidate, the full length of the predicted genomic region was spanned entirely by at least one HiFi consensus read. In the remaining 3 cases, consisting of a pyriform spidroin (PySp) candidate, a cribellate spidroin CrSp candidate, and a pseudoflagelliform spidroin (Pflag) candidate, no more than 2 HiFi reads were necessary to span the entirety of the predicted genomic region, and in each of these cases there was sufficient depth and overlap in the reads to call the region with high confidence (see **Supplemental Figure S1**). The length of the coding regions for the spidroins ranged from 5.5 kb to 20 kb. This is consistent with expectations of full-length sequences found in other spiders(Babb et al. 2017; Kono et al. 2019, 2020). Only in the case of MaSp-1 were we unable to call a complete, full-length sequence. The finalized spidroin annotations included only two spidroins with multiple exons: CrSp and MaSp-1; however, the structure of MaSp1 remains unclear. All other spidroins were found to be single exon sequences. While most single-exon genes tend to be small highly expressed proteins such as histones, spidroins are a rare exception. The single exon structure of spidroin genes has been noted in other species(Ayoub et al. 2013; Kono et al. 2020; Garb and Hayashi 2005; Motriuk-Smith et al. 2005; Ayoub and Hayashi 2008; Liu et al. 2022; Wen et al. 2020; Wang et al. 2019a; Wen et al. 2017).

Aciniform spidroin (AcSp)

Aciniform silk is one of the toughest spider silks, and is typically used for wrapping prey(Tremblay et al. 2015). A single exon for AcSp was identified on Chromosome 7 (**Figure 4A** and **Table 7**), consistent with the sequences in the Araneid orb-weaving spiders *Araneus ventricosus*(Wen et al. 2018) and *Argiope agentata*(Chaw et al. 2014), as well as the cobweb spider *Latrodectus hesperus*(Ayoub et al. 2013). Our confidence in this sequence is high since the complete sequence was spanned entirely by multiple PacBio HiFi reads. In the repetitive region, we found 13 iterated repeats of a 357 amino acid motif, and a 14th partial repeat (**Supplemental Figure S1**). As with previous reports on the structure of AcSp(Ayoub et al. 2013; Hayashi 2004; Chaw et al. 2014; Wen et al. 2018), we also found that the repeats are remarkably well-homogenized. After removal of the signal peptide between Ser-23 and Arg-24, the remaining N-terminal domain secondary structure includes 5 alpha-helices, and a C-Terminal domain consisting of 4 alpha helices which is consistent with the structure found in other AcSps(Wen et al. 2018).

Pseudoflagelliform spidroin (Pflag)

The capture spiral of an orb-web is a composite of two types of silk(Tarakanova and Buehler 2012). For cribellate spiders such as *U. diversus*, the core fiber is the pseudoflagelliform silk made up of pseudoflagelliform spidroin (Pflag). When produced, this core fiber is coated with finely brushed cribellate silk which provides adhesive properties to the capture silk (see Cribellate spidroin).

We found a single candidate for Pflag on Chromosome 7 (**Figure 4A** and **Table 7**). We determined that the internal structure consists of 91 repeats, each of which range between 39 and 70 aa in length and are composed of two parts: a glycine-poor spacer, which is usually either 7 or 12 aa, followed by a glycine-rich repeat with variations on the motif *PSSGGXGG*. The final repeat motif in each module always terminates in a proline.

Cribellate spidroin (CrSp)

Cribellate silk is produced by numerous silk glands with hundreds to thousands of spigots in the cribellum. These numerous fibers are combined into a single silk which is combed into a “wooly” silk by calamistra located on the posterior legs. This wooly silk soaks into the waxy cuticle of insects by means of van der Waals interactions and hygroscopic forces and is used as the capture silk by cribellate orb-weavers(Hawthorn and Opell 2003). No full-length sequence for CrSp has been reported to date, although partial spidroin sequences have been reported for some CrSps in *Tengella perfuga*(Correa-Garhwal et al. 2018) and several *Octonoba* species(Kono et al. 2020). The whole genomic length of the *CrSp* locus on Chromosome 10 is 20,195 nt (**Figure 4A** and **Table 7**). The *U. diversus CrSp* gene was predicted to be a 2-exon gene, with a single 83 nt intron, which is consistent with what we found by manual inspection. While the entire *CrSp* locus was not spanned by single HiFi reads, we are still confident in the sequence produced, since no more than 2 reads were required to span the entire sequence. The N-terminal region of the predicted protein product consists of 874 aa and the C-terminal region consists of 296 aa (**Table 7**). The long N-terminal domain is consistent with that found in *Octonoba* spp, which were also found to have coding regions more than 2 kbp(Kono et al. 2020). We found an internal region that consists of variations on 4 repetitive motifs, with the first half of the sequence made up of motifs 1 and 2 alone, and the second half of the sequence including all 4 motifs. Motifs 1 and 3 are similar, whereas motifs 2 and 4 are distinct from each other motif (**Supplemental Figure S1)**.

Major ampullate spidroin (MaSp)

Draglines are produced by the major ampullate gland which produces two major ampullate spidroins (MaSp1 & MaSp2). This silk has extremely high tensile strength and elasticity and is commonly used for the primary load-bearing parts of the web such as the frame and radii. It is also the primary silk produced by spiders when they are navigating their environment(Foelix 2011). We found three candidates for major ampullate spidroin (MaSp). Based on previous work that identified multiple distinct classes of MaSps, we were able to assign one of our candidate sequences to the MaSp-1 class and the other two candidates to the MaSp-2 class. All three MaSp candidates are on Chromosome 6, although the MaSp-1 locus was located distantly from the two MaSp-2 loci (**Figure 4A**).

The MaSp-1 candidate is the single spidroin sequence we were not able to call as a complete sequence. In our annotation, the sequence appears as a two-exon gene, with the 5’ sequence

and 3' sequence found in different reading frames; however, a close inspection of the data suggests that this is not likely to be correct. We found instead that there is a large region to which the PacBio HiFi reads mapped poorly. There is consensus between the reads that indicates sequence found in the reference assembly that is not found in the reads. However, it is not clear from inspection exactly where the boundaries should be called for this region. This is likely an artifact of the assembly, since the reference was assembled from polymerase-based sequencing which is susceptible to polymerase-slippage.

The first MaSp-2 candidate, MaSp-2a, is a single exon sequence. We found that there were two distinct regions. Interestingly, the first repetitive region, which is 958 aa in length, contains mostly *GPGPQ* motifs reminiscent of the *GPGPX* motifs found in the MaSp-4 sequence recently reported in *Caerosris darwini*, but not elsewhere in the known catalog of spidroins (Garb, *et al*., 2019). The second repetitive region, which is 1,215 aa long, contains runs of poly-A and *GPX*, although *GPGPQ* repeats are also found less frequently in this region.

The second MaSp-2 candidate, MaSp-2b is also a single exon sequence. In the repetitive region, GPGPQ occurs in a few instances, but is relatively rare compared to MaSp-2a. Alternating runs of polyalanine and variations on the motif *GSGPGQQGPGQQGPGGYGPG* characterize the repetitive region. Unlike the case of the first MaSp-2 candidate, MaSp-2b does not have two distinct repetitive regions.

Minor ampullate spidroin (MiSp)

Minor ampullate silk has lower strength, but greater extensibility, and is composed of spidroin made by the minor ampullate gland. While it is commonly used for the construction of the auxiliary spiral in orb-weavers, it is used for prey wrapping by cob-weavers(Vienneau-Hathaway et al. 2017). We found three candidates for minor ampullate spidroin (MiSp). All three MiSp loci were located near one another on Chromosome 1 (**Figure 4A**).

The first candidate, MiSp-1 is a single exon sequence (**Table 7**). There are three repetitive regions in MiSp-1, separated by short spacers. Previous work in *Araneus ventricosus* and the cobweb weaving spiders *Latrodectus hesperus*, *L. tredecimguttatus*, *L. geometricus*, *Steatoda grossa*, and *Parasteatoda tepidariorum* has suggested that MiSp length and sequence are conserved(Chen et al. 2012; Vienneau-Hathaway et al. 2017); however, while the spacers we observed shared some sequence similarities, such as the presence of serine, threonine, and valine residues, the lengths of the spacers observed in *U. diversus* are much shorter.

The second and third candidates, MiSp-2a and MiSp-2b, shared a nearly identical amino acid composition, which was slightly different from that of MiSp-1. Both are single exon sequences (**Table** 7). Their hydrophobicity profiles are also slightly different from MiSp-1.

436

437 *Pyriform spidroin (PySp)*

438

439 Pyriform silk serves as an adhesive compound used to adhere silk lines to one-another, or to
440 substrate that holds the web(Foelix 2011). We found one single exon candidate pyriform
441 spidroin (PySp) on Chromosome 1 (**Figure 4A**), which is consistent in size with a prior PySp
442 sequence reported from *Araneus ventricosus*(Wang et al. 2019a). We found that the internal
443 repetitive region was preceded by a Q-rich N-linker region. 19 tandem repeat motifs, ranging
444 from 188 - 196 aa were found.

445

446 *Tubuliform spidroin (TuSp)*

447

448 Tubuliform silk is used to encase the egg sac and is spun from tubuliform glands. We found a
449 single candidate for tubuliform spidroin (TuSp) on Chromosome 2 (**Figure 4A** and **Table 7**)
450 which is a single exon We found an internal region that was composed of 10 repeats, ranging
451 from 262 aa - 302 aa in length. This is consistent with other reported TuSp repeats, which have
452 been observed between 176 and 375 residues, although it seems that the typical TuSp module
453 is repeated 15 to 20 times(Wen et al. 2017). The N-terminal region has 3 Cys residues: Cys-21,
454 Cys-52, and Cys-132. Other TuSp N-terminal sequences have been reported with 2 Cys
455 residues(Wen et al. 2017), however Cys-21 is expected to be removed during signal peptide
456 cleavage. After cleavage, the N-terminal domain contains five predicted alpha helices. Cys-52
457 and Cys-132 are found in alpha helix 1 and alpha helix 4 after cleavage, which is also where the
458 AcSp Cys residues are found. This conservation suggests functionality.

459

460 **Whole Genome Duplication**

461

462 Gene duplication acts as a primary mode of evolutionary diversification by providing new
463 genetic material that serves as a reservoir for subfunctionalization and neofunctionalization
464 under selective pressure(Ohno 2013; Sémon and Wolfe 2007; Zhang 2003). In previous studies
465 in spiders, evidence of whole genome duplication has been reported, including the presence of
466 multiple copies of *Hox* genes(Sheffer et al. 2021; Cerca et al. 2021; Clarke et al. 2015, 2014)
467 and the expansion of silk genes and chemosensory genes(Cerca et al. 2021). We analyzed the
468 *U. diversus* genome to identify signatures of whole genome duplication.

469

470 Evidence for whole genome duplication can be ascribed by assessing the number of *Hox* gene
471 clusters(Garcia-Fernàndez and Holland 1994). We used published *Hox* gene sequences from
472 the spider *Parasteatoda tepidariorum* as query sequences for BLAST searches against the *U.*
473 *diversus* genome, identifying two *Hox* gene clusters on Chromosome 5 and Chromosome 10
474 (**Figure 4A**), which were found to retain the expected order of *Hox* genes(Pace et al. 2016).
475 Each cluster was missing a single *Hox* gene; however, the specific missing gene was different

for each of the clusters. Cluster A on Chromosome 5 is missing the *fushi tarazu* (*ftz*) gene sequence, while Cluster B on Chromosome 10 is missing the *labial* gene sequence. Our findings are consistent with the discovery of two *Hox* gene clusters in *P. tepidariorum*(Schwager et al. 2017). In the genome of *T. antipodiana*, two *Hox* clusters were found(Fan et al. 2021), including one complete cluster on chromosome 12 which included a copy of all 10 expected *Hox* genes, and a second cluster on chromosome 8 which was found to be missing *abdominal-A*, *Hox3*, *ftz*, and *Ultrabithorax*.The presence of multiple *Hox* clusters in the *U. diversus* genome adds further support to an ancient, ancestral whole genome duplication.

We also searched for evidence of synteny between pseudochromosomes using AnchorWave(Song et al. 2022). The initial identification of syntenic blocks was quite ubiquitous across all 10 pseudochromosomes, but with large gaps between mRNAs and/or with very large interanchor distances. To constrain these results, we chose to include only mRNAs where the number of missing mRNAs between anchors was 4 or less. We made this allowance to account for the fact that we expect there to be significant loss of copies of duplicated genes(Ohno 2013; Sémon and Wolfe 2007; Hakes et al. 2007). This resulted in retaining 196 blocks of at least 2 mRNAs (**Figure 5**). Considering the conservative nature of our analysis, this line of evidence provides further support for an ancient duplication event. However, considerable reorganization has occurred since the duplication event, an observation also made in mammalian genomes, and potentially associated with their successful adaptation to diverse environments(Waters et al. 2021; White 1978).

Sex Chromosomes

The most common and likely ancestral system of chromosomal sex determination in spiders is ♂X_1X_2/♀$X_1X_1X_2X_2$.(Sember et al. 2020) However, spiders exhibit a diversity of sex determining systems, with some including Y chromosomes, and others possessing up to 13 X chromosomes(Král et al. 2019). Usually, these sex determining systems are determined by karyotyping(Sember et al. 2020; Král et al. 2019) (**Figure 3**). The genetic basis of sex determination and chromosomal dosage compensation are unknown for spiders. Determining the genetic identity of X chromosomes has been challenging, in part due to significant levels of shared synteny between sex chromosomes and autosomes(Sember et al. 2020) (**Figure 5**), as well as a paucity of spider genomes with chromosome-level scaffolds. X chromosome scaffolds from more fragmented genomes have been identified by quantifying the relative difference in read depth from sperm with or without the X chromosomes(Bechsgaard et al. 2019). Recently, the X chromosomes for *Argiope bruennichi* (also ♂X_1X_2/♀$X_1X_1X_2X_2$) were identified through disparities in read coverage of X chromosomes between males and females(Sember et al. 2020). In principle, because males have only one copy of each X chromosome, the average read depth for scaffolds from these chromosomes should be half that of autosomes. To determine the sex chromosome in *U. diversus*, we assembled an Illumina short-read library from a single male spider and mapped the reads onto the 10 assembled pseudochromosomes (**Figure 6**).

While 8 of the 10 pseudochromsomes had a median read depth of 40 ± 2, pseudochromosomes 3 and 10 were outliers, with read depths of 36 and 33, respectively. If these pseudochromosomes were exclusively unique X chromosomes, the expected read depth would have been ~20. However, as observed in other species(Sember et al. 2020) and our own (**Figure 5**), orthologous autosomal regions should decrease the expected depth disparity. The higher than expected read depth could also be due to mis-assembly of these

pseudochromosomes, however very little linkage was observed between pseudochromosomes 3 and 10 in the Hi-C data (**Figure 3**). Despite these caveats, the lower median read depth in males for pseudochromosomes 3 and 10 is a strong indicator these likely represent the two X chromosomes for *U. diversus*.

Prior work with *Stegodyphus mimosarum* (also ♂X_1X_2/♀$X_1X_1X_2X_2$) identified sex-linked scaffolds based on lower read depth of sperm lacking X chromosomes(Bechsgaard et al. 2019). When genes identified on these X-linked *S. mimosarum* scaffolds were mapped on to the *U. diversus* pseudochromosomes (**Table 5**), 62% of these genes mapped onto pseudochromosomes 3 and 10 (**Figure 6B**). This large fraction of predicted X-linked genes between two distantly related species of spiders is a strong indicator that not only are pseudochromosomes 3 and 10 likely to be the X chromosomes, but that the genetic composition of these chromosomes has remained fairly stable amongst spiders. Since two X chromosomes were recently identified in *A. bruennichi*, we compared the genetic composition (**Supplemental Table S6**) and synteny between the X chromosomes identified in both species (**Figure 6**). In addition to shared X-linked genes (**Supplemental Table S6, Figure 6B**), *A. bruennichi* scaffolds 10 and 9 appear to share considerable synteny with *U. diversus* pseudochromosomes 10 and 3, respectively. One of these syntenic blocks is the *Hox* gene cluster located on chromosome 10 for both species. The presence of a *hox* cluster on a sex chromosome was surprising since these genes play critical roles in development. Therefore, either dosage compensation is needed in males, or dosage disparity between males and females plays a role in developmental sexual dimorphism.

In insects, the primary sex chromosome dosage sensor is *sex lethal* (*sxl*), which then triggers a cascade of sex-defining signaling events leading to sexually dimorphic expression of genes and/or splice variants. While no *sxl* homologue has been found in spider genomes (including *U. diversus*), other genes involved in sexual dimorphism, such as *doublesex* (*dsx*) are present. Thus, the mechanism spiders use for sensing X:autosome ratio differences remains unknown, but relevant genes are likely shared between *U. diversus*, *A. bruennichi*, and *S. mimosarum*. Of the 534 shared sex-linked genes in these three species, 14 are predicted to be DNA/RNA-binding, and may play a role in sex-determination. The X-linked genes shared between these three species (**Supplemental Table S6**) will be a resource for comparative analysis to identify conserved genes that serve as sex-specifying triggers for spiders. Uncovering how spiders perform sex-linked dosage compensation can not only illuminate how arthropods evolved different sex-determining systems, but also how dosage compensation has evolved independently in numerous animals.

Conclusions

Here, we present a high-quality chromosome-level genome and complementary transcriptome assembly of the hackled orb-weaver *Uloborus diversus*. The 2.15 Gbp draft genome assembly comprises 1,586 scaffolds, including 10 pseudochromosomes that contain 1.9 Gbp (88%) of the total assembly, comparable to the estimated genome size (1.98 Gbp) predicted by GenomeScope2 and contains the vast majority of highly conserved orthologs (94.1% complete, with 88.6% complete and in single copy) as estimated by BUSCO. We predicted a total of 44,408 protein-coding gene models with a BUSCO completeness of 86.7%. Despite the aforementioned technical hurdles, the contiguity and completeness of this assembly, along with the recovery of a complete catalog of full-length spidroin gene sequences, demonstrates the utility of using multiple complementary sequencing technologies for large, repetitive, and highly heterozygous genomes.

The repetitive nature and length of spidroin genes have posed a technical challenge for identifying and reporting full-length sequences. However, it is exactly these qualities that lend spidroins their unique mechanical properties(Malay et al. 2017; Rising et al. 2005; Li et al. 2017); underscoring the need for accurate assemblies. Recent studies leveraging single molecule, long read sequencing technology have predicted longer spidroin sequences than those using PCR approaches(Kono et al. 2019). Here, we used ONT and PacBio HiFi reads to achieve a complete catalog of full-length spidroin sequences for *Uloborus diversus*. The ability to recover full-length sequences for this family of genes is an indication of the high quality of the assembly.

All current models of chromosomal dosage compensation are based on single-sex chromosome animals, however multiple sex-chromosome systems exist in both vertebrates and invertebrates(Yoshido et al. 2020; Rens et al. 2007). Spiders exhibit considerable morphological and behavioral sexual dimorphism that is based on a multiple-sex chromosome system. Understanding the genetic underpinnings of spider sexual development will contribute to a fuller understanding of how chromosomal sex determination can evolve independently in different species. Here we provide evidence for the identities of sex chromosomes in *U. diversus* and leverage this information to identify 14 candidate DNA-binding genes that are shared between three divergent species of spiders.

Our genome will facilitate comparative studies and meets a specific need in the field for a greater representation of genomes from the UDOH+RTA clade that represent nearly half of all known spider species (**Figure 1**)(Garb et al. 2018). We expect that the highly contiguous draft genome and transcriptome datasets we produced for *U. diversus* will serve as a valuable resource for continuing research into the evolution, development, and physiology of spiders, as well as a vital tool to study the genetic basis of orb-weaving behavior. While a handful of spider genomes have been published, all orb-weaving genomes have been from ecribellate Araneid spiders, with no representative genomes from the cribellate families Uloboridae, Deinopidae, Oecobiidae, or Hersiliidae. Improved knowledge of genomes from these families, combined with behavioral and cellular analyses of orb-weaving behavior, will offer a crucial foundation for understanding how and when orb-weaving evolved.

603

604

Materials and Methods

Sample Collection and Husbandry

We collected spiders of the species *Uloborus diversus* from the ancestral lands of the Ramaytush, in Half Moon Bay, California, USA. We collected colony founders from a single greenhouse during several trips between 2016 and 2019 and transported them to custom-fabricated habitats in an on-campus greenhouse at Johns Hopkins University. We later transferred experimental animals from the on-site greenhouse to custom-fabricated habitats in the laboratory until required for experiments. We fed all animals alternately *Drosophila melanogaster* or *Drosophila virilis* once per week.

Karyotyping

We soaked embryos soaked in Grace's insect medium (Gibco) containing 0.1% colchicine for 2 hours. We then added an equal volume of hypotonic solution. After 15 min, we transferred the embryos to a 3:1 ethanol:acetic acid solution for 1 hour. After fixing, we transferred embryos to gelatin-coated microscope slides and dissociated them in a drop of 45% acetic acid. We used siliconized coverslips to squash the dissociated tissue and briefly froze them in liquid nitrogen. After removing the slides from LN2, we immediately removed the coverslips with a razor blade and transferred the slides as quickly as possible to 95% ethanol. We then performed a step-down series from 95% ethanol to 70%, 35%, and finally to Grace's insect medium to return the tissue to an aqueous solution. We then transferred the slides to a 1ug/mL DAPI solution. After a 10-minute incubation, we transferred the slides to de-ionized water to rinse and mounted coverslips with a drop of Vectamount (Vector Laboratories, Burlingame, CA, USA).

RNA Extraction, Library Preparation, and Sequencing

We extracted total RNA from multiple samples: a whole adult female, a whole adult male, adult female prosoma and opisthosoma, adult male prosoma and opisthosoma, pooled legs from both the adult female and adult male dissections, a 4th instar female, and approximately 30 pooled 2nd instars. We used the Qiagen RNeasy Mini Kit (Qiagen, Hilden, Germany) to extract total RNA, following the manufacturer's protocol. We estimated the quality and quantity of total RNA using a NanoDrop One Microvolume UV-Vis Spectrophotometer (ThermoFisher Scientific, Waltham, MA, USA). Before library preparation, we also measured the quality, quantity, and fragment length of our total RNA using a TapeStation 4200 System with RNA ScreenTape and reagents (Agilent, Santa Clara, CA, USA). We prepared barcoded, directional, paired-end RNA-seq libraries with the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina using the NEBNext Poly(A) mRNA Magnetic Isolation Module. We submitted the resulting libraries to the Johns Hopkins Genomics Core Resources Facility to be sequenced on an Illumina HiSeq 2500 Sequencing System with 150 bp paired-end chemistry.

Genomic DNA Extraction, Library Preparation, and Sequencing

Prior to extraction of DNA, we withdrew food for 3 days to minimize the potential contribution of contaminating DNA from dietary sources. We extracted high molecular weight (HMW) DNA using the QIAgen MagAttract HMW DNA kit. Prior to HMW purification, we followed the manufacturer's protocol for disruption/lysis of tissue. We avoided fast pipetting and prolonged vortexing to minimize shearing of DNA. We flash froze adult spiders in liquid N_2 and crushed them with a pellet pestle (Fisher, 12-141-364) in a Protein LoBind tube (Eppendorf, 022431081) containing 220 uL of Buffer ATL. We then added 20 uL Proteinase K and briefly vortexed the sample. We next incubated the sample overnight at 56C with 900 rpm shaking on a ThermoMixer C (Eppendorf, 5382000023). After the overnight incubation, we then briefly centrifuged the sample to spin down condensate on the tube. We next transferred 200 uL of lysate to a fresh 2 mL sample tube and followed the manufacturer's protocol for manual purification of HMW DNA from fresh or frozen tissue. We estimated DNA quality using a NanoDrop One Microvolume UV-Vis Spectrophotometer and quantified DNA using a Qubit 4 Fluorometer (ThermoFisher) with a Quant-iT dsDNA HS Assay Kit. We also measured DNA quality, quantity, and fragment length distributions using the Agilent TapeStation 4200 System with Genomic DNA ScreenTape and reagents before proceeding to library prepration. A typical preparation from a 20 mg spider yielded 8.5ug of DNA.

Illumina Sequencing

For Illumina sequencing, we extracted genomic DNA from a single, whole, unmated penultimate stage female to minimize the potential contribution of extraneous haplotypes from stored sperm after mating events. We submitted the HMW gDNA to the Johns Hopkins GCRF, where they prepared a PCR-free library of approximately 400 bp DNA insert size using the Illumina TruSeq PCR-Free High Throughput Library Prep Kit (San Diego, CA, USA), according to the manufacturer's protocol. They then sequenced the prepared library on an Illumina NovaSeq 6000 Sequencing System with 150bp paired-end chemistry.

ONT Sequencing

For ONT sequencing, we extracted HMW genomic DNA from 3 adult females. We prepared sequencing libraries using the Ligation Sequencing Kit (SQK-LSK109) (Oxford Nanopore Technologies, UK), according to the manufacturer's protocols. Third party reagents we used during library preparation included: New England Biolabs (New England Biolabs, Ipswitch, MA, USA) NEBNext End Repair/dA-Tailing Module (E7546), NEBNext FFPE DNA Repair Mix (M6630), and NEB Quick Ligation Module (E6056). We then sequenced the libraries, using ONT R.9.4.1 flowcells (FLO-PRO002) on an ONT PromethION sequencing platform. We then used

ONT's Albacore basecalling software v.2.0.1 (RRID:SCR_015897) to basecall the raw fast5 data.

PacBio HiFi Sequencing

For PacBio sequencing, HMW DNA was extracted from a single adult female spider provided to Circulomics (Baltimore, MD, USA). They extracted DNA using a modified protocol with the Nanobind Tissue Kit (Circulomics, #NB-900-701-01). Briefly, they froze and crushed a single, adult female spider with a pellet pestle (Fisher, #12-141-364) in a Protein LoBind tube (Eppendorf, #022431081) containing 200 uL of Buffer CT. The crushed spider was centrifuged at 16,000 x g at 4 C for 2 min. The supernatant was discarded, and the pellet was resuspended in 500 ul Buffer CT and the mixture was transferred to a 2.0 mL Protein LoBind tube (Eppendorf # 022431102). The suspension was spun again at 16,000 x g at 4 C for 2 min and the supernatant discarded. The spider tissue pellet was combined with 20 ul Proteinase K and 150 ul Buffer PL1 and resuspended by pipetting with a P200 wide bore pipette tip. The tissue was incubated on a ThermoMixer at 55 C with 900 rpm mixing for 1 hour. After lysis, 20 ul RNaseA was added, and the lysate was mixed by pipetting with a P200 wide bore pipette tip. The lysate was incubated at RT for 3 min. After RNaseA incubation, 25 ul Buffer SB was added, the lysate was vortexed 5 x 1 sec pulses, and then centrifuged at 16,000 x g at 4 C for 5 min. The supernatant (~200 ul) was transferred to a 70uM filter (Fisher # NC1444112) set in a new 1.5 mL Protein LoBind tube (Eppendorf # 022431081). The tube with the 70 uM filter was spun on a mini-centrifuge (Ohaus # FC5306) for 1 sec and then the filter was dicarded. 50 ul Buffer BL3 was added to the cleared lysate and the tube was inversion mixed 10X. The tube was then incubated on a ThermoMixer at 55 C with 900 rpm mixing for 5 min. After incubation, the tube was allowed to come to RT, which took about 2 min. The tube was spun for 1 sec on a mini-centrifuge to spin down condensate from the lid. One 5 mm Nanobind disk was added to the tube followed by 250 uL isopropanol and then the tube was inversion mixed 5X. The tube was then rocked on a platform rocker (ThermoScientific # M48725Q) at RT and max speed for 30 min. The DNA-bound Nanobind disk was washed according to handbook directions with one 500 ul CW1 wash and one 500 ul CW2 wash. The tube with the disk was tap spun for 2 x 1 sec to dry the disk. The DNA was eluted with 50 ul Buffer EB and incubated at RT overnight. The next day, the eluate was pipette mixed with a standard bore pipette tip 5x and then quantitated with Nanodrop and Qubit dsDNA BR assay and then sized by pulsed-field gel electrophoresis.

We then submitted the DNA sample to the University of Maryland School of Medicine Genomics Core Facility for PacBio HiFi sequencing. There, they size selected the DNA using a Safe Science BluePippin with a 9kb high-pass cutoff. They prepared the sequencing library using the Express v2 kit, according to the standard protocol for preparing HiFi sequencing libraries. They then sequenced the library on a PacBio Sequel II 8M SMRT Cell using a 30 hour HiFi run mode and processed using SMRT Link v.9.0 software.

Dovetail Chicago and Dovetail Hi-C Sequencing

To further improve the *U. diversus* genome assembly, we used proximity ligation-based sequencing techniques to scaffold intermediate versions of our assembly. We provided 19 spider specimens to Dovetail Genomics (Scotts Valley, CA, USA) for Chicago and Hi-C library preparation as previously described(Putnam et al. 2016). They prepared a Chicago library using 15 pooled adult females and a Hi-C library using 4 pooled adult females. They sequenced both the prepared Chicago and Dovetail Hi-C libraries on an Illumina HiSeq X sequencing platform on 1 flowcell.

DNA-seq and RNA-seq QA/QC

For Illumina, we examined read quality using FastQC(Andrews 2010) v.0.11.9 (RRID:SCR_014583). For DNA-seq data, we determined that, due to high quality of reads and absence of adapter sequences, no further processing would be required and proceeded to assembly with raw read data.For RNA-seq data, we used TrimGalore(Krueger et al. 2021) v.0.4.2 (RRID:SCR_011847) to apply quality filtering and remove adapter sequences from the FASTQ files. We performed additional filtering for quality with Trimmomatic(Bolger et al. 2014) v.0.33 (RRID:SCR_011848) . For ONT, reads shorter than 3 kbp were discarded. The length-filtered ONT long reads were used in downstream assembly.

Genome Size, Heterozygosity, and Unique Sequence Estimation

Prior to assembly, we used Jellyfish(Marçais and Kingsford 2011) v.2.2.4 (RRID:SCR_005491) to count the frequency of canonical 21-mers in our Illumina sequencing data. We used the resulting sorted *k*-mer frequencies vs counts histogram as input to GenomeScope(Ranallo-Benavidez et al. 2020; Vurture et al. 2017) v.2.0 to estimate genome size, heterozygosity, and repetitiveness.

Recovery of Mitogenome

We used Novoplasty(Dierckxsens et al. 2016) v.4.2 (RRID:SCR_017335) to generate a complete circularized mitochondrial sequence using raw Illumina read data. The mitochondrial sequences of several spider species were used to provide seed sequences (**Supplemental Table S3**). The resulting mitogenome sequences assembled by Novoplasty were compared for consensus. The consensus mitogenome was uploaded to the MITOS 2 web server(Bernt et al. 2013) for annotation. The CGView web server(Stothard and Wishart 2005) (RRID:SCR_011779) was used to visualize the annotated mitogenome.

Nuclear Genome Assembly

De novo Nuclear Genome Assembly with MaSuRCA

Illumina reads were assembled into contigs and the resulting contigs were scaffolded with ONT long reads using the MaSuRCA assembly pipeline(Zimin et al. 2013, 2017) v.3.4.2 (RRI:010691). We used default settings, including the default CABOG contigging module in lieu of the Flye assembler. The resulting genome assembly is referred to as *U. diversus* v.1.0.

To improve the assembly, we used *Rascaf*(Song et al. 2016) v.2016-11-29 to scaffold with Illumina RNA-seq read data. The resulting genome assembly is referred to as *U. diversus* v.1.1. To reduce redundancy in the assembly due to the presence of alternative haplotigs, we used Pseudohaploid with default settings. The resulting genome assembly is referred to as *U. diversus* v.1.2

De novo Nuclear Genome Assembly with PB-IPA

We used PacBio's Improved Phased Assembly (IPA) HiFi Genome Assembler with default settings, specifying a genome size of 1.9 Gbp, to assemble the HiFi reads. The resulting genome assembly is referred to as *U. diversus* v.2.0.

Merging MaSuRCA and PB-IPA Assemblies with SAMBA

We used the SAMBA tool distributed with MaSuRCA to merge the MaSuRCA assembly, *U. diversus* v.1.2, and the PB-IPA assembly, *U. diversus* v.2.0. The resulting genome assembly is referred to as *U. diversus* v.3.0.

Scaffolding Assemblies with *HiRise*

The initial *U. diversus* v.1.2 draft assembly obtained using a combination of MaSuRCA, Rascaf, and Pseudohaploid was provided to Dovetail Genomics in FASTA format. The resulting genome assembly is referred to as *U. diversus* v.1.3.

The merged MaSuRCA and PB-IPA assembly, *U. diversus* v.3.0, was provided to Dovetail Genomics in FASTA format. The resulting genome assembly is referred to as *U. diversus* v.3.1.

Genome Assembly Metrics and Assessments

For each assembly, completeness was estimated with Benchmarking Universal Single-Copy Orthologs (BUSCO)(Seppey et al. 2019; Simão et al. 2015; Waterhouse et al. 2018) v.5.2.1 (RRID:SCR_015008) using the arachnida_odb10 database(Kriventseva et al. 2019). Contiguity of each assembly was evaluated for comparison using Quast(Gurevich et al. 2013) v.5.0.2 (RRID:SCR_001228).

Genome-Guided Transcriptome Assembly

Cleaned and trimmed Illumina RNA-seq reads were aligned to the genome using *HISAT2*(Kim et al. 2019) v.2.2.1. We then used the Trinity assembler[43] v.2.12.0 to produce a genome-guided transcriptome assembly [--CPU 60 –max_memory 200G –genome_guided_max_intron 20000 –SS_lib_type RF –include_supertranscripts –verbose]. We used TransDecoder(Haas et al. 2013) v.5.5.0 with default settings, including homology searches using both BlastP(Altschul et al. 1990; Altschul 1997) against a SwissProt UniProt database(The UniProt Consortium 2019) as well as the Pfam database(Mistry et al. 2021) v.32, as ORF retention criteria.

Repeat Annotations

To characterize the repeat elements in the *U. diversus* genome, we generated a custom *de novo* repeat library using RepeatModeler(Flynn et al. 2020) v.2.0.2 with default parameters. We used RepeatMasker(Tarailo Graovac and Chen 2009) v.4.1.2 to screen and mask repeat and low-complexity regions of the genome with the Dfam consensus(Storer et al. 2021) v.3.4 and RepBase RepeatMasker Edition(Bao et al. 2015) v.2018-10-26 repeat libraries.

Annotation of Protein Coding Genes

We performed gene annotation using the BRAKER 2 pipeline(Hoff et al. 2016; Brůna et al. 2021; Lomsadze 2005; Lomsadze et al. 2014; Stanke et al. 2006, 2008; Gotoh 2008; Li et al. 2009; Barnett et al. 2011; Iwata and Gotoh 2012; Buchfink et al. 2015; Hoff et al. 2019; Brůna et al. 2020) v.2.1.6 with RNA-seq evidence and protein homology evidence based on a custom library of spider sequences obtained from NCBI. BRAKER2 uses RNA-seq data to produce intron hints for training the *ab initio* gene prediction program AUGUSTUS(Stanke et al. 2006, 2008) on a species-specific model. This species-specific model is then used in conjunction with RNA-seq data to predict protein coding genes. The bam file previously generated in transcriptome assembly and analysis was passed to BRAKER2, which was run with default settings.

Annotation of Non-Coding RNAs

We used tRNAscan-SE(Chan and Lowe 2019; Chan et al. 2021) v.2.0.7 with default settings to predict tRNAs. We then used Barrnap(Seeman 2018) v.0.9 (RRID:SCR_015995) with default settings to predict rRNAs.

Functional Annotation

We started the annotation of predicted genes used the BLAST+ blastp algorithm. First, we obtained the longest coding sequence for each gene predicted by BRAKER2. We then used the EMBOSS(Rice et al. 2000) v.6.6.0.0 Transeq tool to translate and trim the coding sequences. Once translated and trimmed, we used the BLAST+ v.2.10.1+ Blastp tool to search against the UniProt SwissProt database with an e-value cutoff of 1e-10. We used InterProScan(Jones et al. 2014; Quevillon et al. 2005) (RRID:SCR_005829) to predict motifs, domains, and gene ontology (GO)(Ashburner et al. 2000; The Gene Ontology Consortium et al. 2021) terms (RRID:SCR_002811), as well as MetaCyc(Caspi et al. 2016, 2018) and Reactome(Gillespie et al. 2022; Jassal et al. 2019) pathways,using the following analyses: CDD(Lu et al. 2020) v.3.18, Coils v.2.2.1, Gene3D(Lewis et al. 2018) v.4.3.0, Hamap(Pedruzzi et al. 2015) v.2020-05, MobiDBLite(Necci et al. 2017) v.2.0, PANTHER(Mi et al. 2019) v.15.0, Pfam(Mistry et al. 2021) v.34.0, PIRSF(Wu 2004) v.3.10, PIRSR(Chen et al. 2019a) v.2021-02, PRINTS(Attwood 2003) v.42.0, ProSitePatterns(Sigrist 2002; Sigrist et al. 2012) v.2021-01, ProSiteProfiles(Sigrist 2002; Sigrist et al. 2012) v.2021-01, SFLD(Akiva et al. 2014) v.4, SMART(Letunic and Bork 2018; Letunic et al. 2021) v.7.1, SUPERFAMILY(Pandurangan et al. 2019; Gough et al. 2001) v.1.75, and TIGRFAM(Haft et al. 2012; Selengut et al. 2007; Haft 2003, 2001) v.15.0.

Spidroins

Identification of Spidroin Candidate Sequences

We identified *Uloborus diversus* by conducting BLAST(Altschul et al. 1990; Altschul 1997) searches using the list of spidroin sequences included in **Supplemental Table S5** as queries against the assembled genome, transcriptome, and gene models predicted by BRAKER2. We looked for matches to both N- and C-terminal sequences from members of each type of spidroin, as well as to available repetitive motifs. After cross-referencing genomic coordinates with gene models and transcripts, we used JBrowse(Buels et al. 2016) to visualize mapping of Illumina RNAseq data and PacBio HiFi reads to the assembled genome. RNAseq reads were mapped to the genome with HISAT2(Kim et al. 2019), while minimap2(Li 2018, 2021) was used to map PacBio HiFi reads. Samtools(Li et al. 2009) was used to convert the resulting SAM files to BAM files, as well as to sort and index the BAM files. For each spidroin candidate, the entire sequence from start codon to stop codon, ignoring any predicted splicing, with an additional 5

kb of sequence on both the 5' and 3' end, was translated in all six frames using the ExPASy Translate Tool via the ExPASy web server(Gasteiger 2003) and inspected for ORFs as well as the presence of repetitive motifs characteristic of spidroins. Predicted splice sites were compared with RNAseq data. Unsupported splice sites, either by lack of evidence in the mapping of RNAseq reads or by the obvious presence of spidroin repeat motifs within the predicted intronic region, were removed from the annotations. Spidroins sequences were called based upon the preponderance of available evidence, which in some cases conflicted with the structure predicted by BRAKER2.

Spidroin Sequence Analysis

We used the ExPASy web server tool ProtScale to find the amino acid composition of each sequence, as well as to estimate the hydrophobicity using the the Kyte-Doolittle method(Kyte and Doolittle 1982; Gasteiger 2003). We used the PSIPRED v.4.0 tool in the UCL Bioinformatics Group's PSIPRED Protein Analysis Workbench(Buchan et al. 2013) to predict the secondary structure of each sequence. The sequences were often too long and necessitated judicious segmentation into reasonable sequences that were short enough for analysis. In such cases, we selected natural breaks in the sequence structure, such as separating the N-terminal region from the repetitive regions, etc. We used SignalP(Teufel et al. 2022) v.6.0 to predict the presence signal peptides and signal peptidase cleavage sites in the N-terminal regions.

Data Availability

The raw sequencing data and assembled genome presented in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRNA846873.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

J.M. acknowledges funding from the NSF Graduate Research Fellowship Program (DGE-1746891). A.G. acknowledges funding from NIH (R35GM124883). A.V.Z. acknowledges funding from the USDA National Institute of Food and Agriculture (2018-67015-28199), NSF (IOS-1744309), and NIH (R01-HG006677 and R35-GM130151).

Author Contributions

J.M., A.Z. and A.G. designed the research study. J.M. performed DNA purification and sample preparation for Illumina and Oxford Nanopore sequencing. J.M. performed all computational analyses, except for HiRise scaffolding (performed by Dovetail), MaSuRCA and SAMBA. A.Z. performed MaSuRCA assembly and merging with SAMBA. J.M. and A.G. analyzed the data and wrote the paper.

Acknowledgements

We thank Circulomics Inc., particularly Kelvin Liu and Michelle Kim, for assistance in DNA extraction for PacBio HiFi sequencing. We thank the Johns Hopkins University Genomics Core, and David Mohr in particular, for Illumina sequencing and consultation. We additionally thank the University of Maryland Genomic Resource Center, and Luke Tallon specifically, for PacBio HiFi sequencing. We thank Dovetail Genomics, particularly Mark Daly and Tom Swale, for Chicago and Dovetail Hi-C library preparation and sequencing, as well as HiRise assembly scaffolding and consultation. We thank the members of the Timp Lab, in particular Winston Timp, Norah Sadowski, and Rachael Workman, for training and graciously permitting the use of their ONT PromethION and TapeStation. We thank Gordus lab members, James Taylor, Michael Schatz, Bob Johnston, Rajiv McCoy, Prashant Sharma, and Ben Matthews for helpful discussions and comments on the manuscript.

Figure Legends

Figure 1: Spider Phylogeny

A) A female *U. diversus.*

B) Phylogeny of spiders. Orb weaver families are highlighted in orange. Species with sequenced genomes are highlighted in blue. *U. diversus* is highlighted in red. Example webs from Rooney(Roberson et al. 2016), Glatz(Glatz 1967), Coddington(Coddington 1986). UDOH = Uloboridae, Deinopidae, Oecobiidae, Hersiliidae. RTA = Retrolateral tibial apophysis clade. O = Ordovician, S = Silurian, D = Devonian, C = Carboniferous, P = Permian, T = Triassic, J = Jurassic, K = Cretaceous, Pg. = Paleogene, Ng. = Neogene, Mya = millions of years ago

Figure 2: GenomeScope Plot from Illumina Data

A) Kmer spectra for Illumina reads from a single, virgin female. The diploid and haploid peaks are at 70x and 35x coverage, respectively.

Figure 3: Chromosome Scale Genome Assembly

A) Karyotype of female and male embryos. Female and male diploid sizes of 20 and 18, respectively, indicate a ♂X1X2/♀X1X1X2X2 sex-determination system, with 8 autosomes.

B) Hi-C linkage map of assembled scaffolds. The 10 largest scaffolds are annotated.

C) Comparison of HiRise and MaSuRCA assemblies. The majority of the HiRise assembly is captured by the first 10 scaffolds.

D) Circos plot of 10 largest nuclear scaffolds, highlighting GC content, repeat content, and gene content across the scaffolds.

E) Circos plot of mitochondrial scaffold, highlighting tRNA-coding loci, protein-coding loci, and GC content.

Figure 4: Gene Annotations

A) Gene loci for spidroins and *hox* gene clusters.

B) Domain composition of identified spidoins. (Repeat region annotations are condensed for clarity.)

Figure 5: Synteny and Chromosomal Rearrangements

A) Inter-anchor mRNA ID difference distribution of syntenic blocks identified by AnchorWave analysis. Each syntenic block is defined by ORF or inter-ORF anchors. All ORFS are numerically annotated in consecutive order from scaffold 1 through scaffold 10. Inter-anchor mRNA ID difference is defined as the difference in these numerical

ORF IDs between consecutive ORF anchors. If the distance equals 1, it means the two anchors are consecutive ORFs within the block. Asterisk indicates syntenic blocks used in **E**.

B) Inter-anchor Mbp difference distribution of syntenic blocks identified with AnchorWave analysis. Inter-anchor difference was calculated as the base-pair distance between consecutive ORF anchors within a syntenic block. Asterisk indicates syntenic blocks used in **E**.

C) Ribbon plot of all AnchorWave-defined syntenic blocks shared between chromosomal scaffolds.

D) Ribbon plot of filtered AnchorWave-defined syntenic blocks shared between chromsomal scaffolds. Only blocks consisting of consecutive ORF anchors < 4 mRNA IDs apart are plotted.

Figure 6: Sex Chromosomes

A) Read depth of Illumina reads from a male spider aligned to the chromosomal scaffolds. Scaffolds 3 and 10 (asterisks) exhibited lower read depth than other scaffolds.

B) Venn diagram of shared sex-associated genes identified in *U. diversus*, *S. mimosarum*, and *A. bruennichi*.

C) Ribbon plot of shared synteny between predicted X chromosomes from *U. diversus* and *A. bruennichi*.

Table 1 - Summary of Library Statistics.

Table 2 - Summary of Genome Assembly Statistics.

Table 3 - Summary of Genome Assembly BUSCO Scores.

Table 4 - Summary of Repeat Content.

Table 5 - Summary of Annotation Statistics.

Table 6 - Summary of InterproScan Results.

Table 7 - Summary of Spidroin Gene Features.

Table S1 - Comparison of Genome Statistics for Published Genomes.

Table S2 - Summary of Spider Genome Repeat Content.

Table S3 – Library of Annotated Genes from Spider Genomes.

Table S4 - Mitogenome Sequences Used for NOVOplasty Seeds.

Table S5 - Spidroin Protein Sequences Used in BLAST Searches.

Table S6 - Summary of Common Sex-Linked Annotations.

1008 **Literature Cited**

1009

1010 Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, Huang CC, Lauck F,
1011 Mashiyama ST, Meng EC, et al. 2014. The Structure–Function Linkage Database. *Nucl*
1012 *Acids Res* **42**: D521–D530.

1013 Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search
1014 programs. *Nucleic Acids Research* **25**: 3389–3402.

1015 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
1016 *Journal of Molecular Biology* **215**: 403–410.

1017 Andrews S. 2010. FastQC: A quality control tool for high throughput sequencing.
1018 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

1019 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight
1020 SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet*
1021 **25**: 25–29.

1022 Attwood TK. 2003. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*
1023 **31**: 400–402.

1024 Ayoub NA, Garb JE, Kuelbs A, Hayashi CY. 2013. Ancient Properties of Spider Silks Revealed
1025 by the Complete Gene Sequence of the Prey-Wrapping Silk Protein (AcSp1). *Molecular*
1026 *Biology and Evolution* **30**: 589–601.

1027 Ayoub NA, Hayashi CY. 2008. Multiple Recombining Loci Encode MaSp1, the Primary
1028 Constituent of Dragline Silk, in Widow Spiders (Latrodectus: Theridiidae). *Molecular*
1029 *Biology and Evolution* **25**: 277–286.

1030 Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, Hogenesch JB, Kuntner M,
1031 Higgins L, Hayashi CY, Agnarsson I, et al. 2017. The Nephila clavipes genome
1032 highlights the diversity of spider silk genes and their complex expression. *Nat Genet* **49**:
1033 895–903.

1034 Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in
1035 eukaryotic genomes. *Mobile DNA* **6**: 11.

1036 Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. BamTools: a C++ API
1037 and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**: 1691–1692.

1038 Bechsgaard J, Schou MF, Vanthournout B, Hendrickx F, Knudsen B, Settepani V, Schierup MH,
1039 Bilde T. 2019. Evidence for Faster X Chromosome Evolution in Spiders ed. A.
1040 Larracuente. *Molecular Biology and Evolution* **36**: 1281–1293.

1041 Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M,
1042 Stadler PF. 2013. MITOS: Improved de novo metazoan mitochondrial genome
1043 annotation. *Molecular Phylogenetics and Evolution* **69**: 313–319.

1044 Blackledge TA, Hayashi CY. 2006. Unraveling the mechanical properties of composite silk
1045 threads spun by cribellate orb-weaving spiders. *Journal of Experimental Biology* **209**:
1046 3131–3140.

1047 Blackledge TA, Scharff N, Coddington JA, Szüts T, Wenzel JW, Hayashi CY, Agnarsson I.
1048 2009. Reconstructing web evolution and spider diversification in the molecular era. *Proc*
1049 *Natl Acad Sci USA* **106**: 5229–5234.

1050 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
1051 data. *Bioinformatics* **30**: 2114–2120.

1052 Boore JL. 1999. Animal mitochondrial genomes. *Nucleic Acids Research* **27**: 1767–1780.

1053 Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic
1054 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a
1055 protein database. *NAR Genomics and Bioinformatics* **3**: lqaa108.

1056 Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with
1057 self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics* **2**:
1058 lqaa026.

1059 Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT. 2013. Scalable web services for
1060 the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research* **41**: W349–W357.

1061 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat*
1062 *Methods* **12**: 59–60.

1063 Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG,
1064 Lewis SE, Stein L, et al. 2016. JBrowse: a dynamic web platform for genome
1065 visualization and analysis. *Genome Biol* **17**: 66.

1066 Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker
1067 M, Latendresse M, Mueller LA, et al. 2016. The MetaCyc database of metabolic
1068 pathways and enzymes and the BioCyc collection of pathway/genome databases.
1069 *Nucleic Acids Res* **44**: D471–D480.

1070 Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M,
1071 Midford PE, Ong Q, Ong WK, et al. 2018. The MetaCyc database of metabolic pathways
1072 and enzymes. *Nucleic Acids Research* **46**: D633–D639.

1073 Cerca J, Armstrong EE, Vizueta J, Fernández R, Dimitrov D, Petersen B, Prost S, Rozas J,
1074 Petrov D, Gillespie RG. 2021. The *Tetragnatha kauaiensis* Genome Sheds Light on the
1075 Origins of Genomic Novelty in Spiders ed. T. Gossmann. *Genome Biology and Evolution*
1076 **13**: evab262.

1077 Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and
1078 functional classification of transfer RNA genes. *Nucleic Acids Research* **49**: 9077–9096.

1079 Chan PP, Lowe TM. 2019. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences.
1080 In *Gene Prediction* (ed. M. Kollmar), Vol. 1962 of *Methods in Molecular Biology*, pp. 1–

14, Springer New York, New York, NY http://link.springer.com/10.1007/978-1-4939-9173-0_1 (Accessed March 8, 2022).

Chaw R, Zhao Y, Wei J, Ayoub NA, Allen R, Atrushi K, Hayashi CY. 2014. Intragenic homogenization and multiple copies of prey-wrapping silk genes in Argiope garden spiders. *BMC Evol Biol* **14**: 31.

Chen C, Wang Q, Huang H, Vinayaka CR, Garavelli JS, Arighi CN, Natale DA, Wu CH. 2019a. PIRSitePredict for protein functional site prediction using position-specific rules. *Database* **2019**. https://academic.oup.com/database/article/doi/10.1093/database/baz026/5363830 (Accessed March 5, 2022).

Chen G, Liu X, Zhang Y, Lin S, Yang Z, Johansson J, Rising A, Meng Q. 2012. Full-Length Minor Ampullate Spidroin Gene Sequence ed. V.N. Uversky. *PLoS ONE* **7**: e52293.

Chen L-Y, VanBuren R, Paris M, Zhou H, Zhang X, Wai CM, Yan H, Chen S, Alonge M, Ramakrishnan S, et al. 2019b. The bracteatus pineapple genome and domestication of clonally propagated crops. *Nat Genet* **51**: 1549–1558.

Choi D, Choy KL. 2020. Spider silk binder for Si-based anode in lithium-ion batteries. *Materials & Design* **191**: 108669.

Clarke TH, Garb JE, Hayashi CY, Arensburger P, Ayoub NA. 2015. Spider Transcriptomes Identify Ancient Large-Scale Gene Duplication Event Potentially Important in Silk Gland Evolution. *Genome Biol Evol* **7**: 1856–1870.

Clarke TH, Garb JE, Hayashi CY, Haney RA, Lancaster AK, Corbett S, Ayoub NA. 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. *BMC Genomics* **15**: 365.

Coddington JA. 1986. ORB WEBS IN “NON-ORB WEAVING” OGRE-FACED SPIDERS (ARANEAE: DINOPIDAE): A QUESTION OF GENEALOGY. *Cladistics* **2**: 53–67.

Coddington JA, Agnarsson I, Hamilton CA, Bond JE. 2019. Spiders did not repeatedly gain, but repeatedly lost, foraging webs. *PeerJ* **7**: e6703.

Correa-Garhwal SM, Chaw RC, Clarke TH, Alaniz LG, Chan FS, Alfaro RE, Hayashi CY. 2018. Silk genes and silk gene expression in the spider Tengella perfuga (Zoropsidae), including a potential cribellar spidroin (CrSp) ed. P. Heneberg. *PLoS ONE* **13**: e0203563.

Corver A, Wilkerson N, Miller J, Gordus A. 2021. Distinct movement patterns generate stages of spider web building. *Current Biology* **31**: 4983-4997.e5.

Dierckxsens N, Mardulyn P, Smits G. 2016. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res* gkw955.

Eberhard WG. 1971. The ecology of the web ofUloborus diversus (Araneae: Uloboridae). *Oecologia* **6**: 328–342.

1118 Escuer P, Pisarenco VA, Fernández-Ruiz AA, Vizueta J, Sánchez-Herrero JF, Arnedo MA,
1119 Sánchez-Gracia A, Rozas J. 2022. The chromosome-scale assembly of the Canary
1120 Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin
1121 and genome structure of chemoreceptor gene families in chelicerates. *Molecular*
1122 *Ecology Resources* **22**: 375–390.

1123 Fan Z, Yuan T, Liu P, Wang L-Y, Jin J-F, Zhang F, Zhang Z-S. 2021. A chromosome-level
1124 genome of the spider *Trichonephila antipodiana* reveals the genetic basis of its
1125 polyphagy and evidence of an ancient whole-genome duplication event. *GigaScience*
1126 **10**: giab016.

1127 Fang W-Y, Wang Z-L, Li C, Yang X-Q, Yu X-P. 2016. The complete mitogenome of a jumping
1128 spider Carrhotus xanthogramma (Araneae: Salticidae) and comparative analysis in four
1129 salticid mitogenomes. *Genetica* **144**: 699–709.

1130 Fernández R, Kallal RJ, Dimitrov D, Ballesteros JA, Arnedo MA, Giribet G, Hormiga G. 2018.
1131 Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the
1132 Spider Tree of Life. *Current Biology* **28**: 1489-1497.e5.

1133 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
1134 RepeatModeler2 for automated genomic discovery of transposable element families.
1135 *Proc Natl Acad Sci USA* **117**: 9451–9457.

1136 Foelix RF. 2011. *Biology of spiders*. 3rd ed. Oxford University Press, Oxford□; New York.

1137 Garb JE, Hayashi CY. 2005. Modular evolution of egg case silk genes across orb-weaving
1138 spider superfamilies. *Proc Natl Acad Sci USA* **102**: 11379–11384.

1139 Garb JE, Sharma PP, Ayoub NA. 2018. Recent progress and prospects for advancing arachnid
1140 genomics. *Current Opinion in Insect Science* **25**: 51–57.

1141 Garcia-Fernàndez J, Holland PWH. 1994. Archetypal organization of the amphioxus Hox gene
1142 cluster. *Nature* **370**: 563–566.

1143 Gasteiger E. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis.
1144 *Nucleic Acids Research* **31**: 3784–3788.

1145 Gatesy J, Hayashi C, Motriuk D, Woods J, Lewis R. 2001. Extreme Diversity, Conservation, and
1146 Convergence of Spider Silk Fibroin Sequences. *Science* **291**: 2603–2605.

1147 Giani AM, Gallo GR, Gianfranceschi L, Formenti G. 2020. Long walk to genomics: History and
1148 current approaches to genome sequencing and assembly. *Computational and Structural*
1149 *Biotechnology Journal* **18**: 9–19.

1150 Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C,
1151 Matthews L, Gong C, et al. 2022. The reactome pathway knowledgebase 2022. *Nucleic*
1152 *Acids Research* **50**: D687–D692.

1153 Glatz L. 1967. Zur biologie und morphologie von Oecobius annulipes lucas (Araneae,
1154 Oecobiidae). *Z Morph Tiere* **61**: 185–214.

1155 Gloor D, Nentwig W, Blick T, Kropf C. 2017. World Spider Catalog. http://wsc.nmbe.ch
1156 (Accessed February 4, 2022).

1157 Gotoh O. 2008. A space-efficient and accurate method for mapping and aligning cDNA
1158 sequences onto genomic sequence. *Nucleic Acids Research* **36**: 2630–2638.

1159 Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome
1160 sequences using a library of hidden Markov models that represent all proteins of known
1161 structure. *Journal of Molecular Biology* **313**: 903–919.

1162 Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome
1163 assemblies. *Bioinformatics* **29**: 1072–1075.

1164 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D,
1165 Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq
1166 using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–
1167 1512.

1168 Haft DH. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Research* **31**: 371–
1169 373.

1170 Haft DH. 2001. TIGRFAMs: a protein family resource for the functional identification of proteins.
1171 *Nucleic Acids Research* **29**: 41–43.

1172 Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2012. TIGRFAMs and
1173 Genome Properties in 2013. *Nucleic Acids Research* **41**: D387–D395.

1174 Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal:
1175 the difference between small-scale and genome duplication. *Genome Biol* **8**: R209.

1176 Hawthorn AC, Opell BD. 2003. van der Waals and hygroscopic forces of adhesion generated by
1177 spider capture threads. *Journal of Experimental Biology* **206**: 3905–3911.

1178 Hayashi CY. 2004. Molecular and Mechanical Characterization of Aciniform Silk: Uniformity of
1179 Iterated Sequence Modules in a Novel Member of the Spider Silk Fibroin Gene Family.
1180 *Molecular Biology and Evolution* **21**: 1950–1959.

1181 Hayashi CY, Lewis RV. 1998. Evidence from flagelliform silk cDNA for the structural basis of
1182 elasticity and modular nature of spider silks 1 1Edited by M. F. Moody. *Journal of*
1183 *Molecular Biology* **275**: 773–784.

1184 Hayashi CY, Lewis RV. 2000. Molecular Architecture and Evolution of a Modular Spider Silk
1185 Protein Gene. *Science* **287**: 1477–1479.

1186 Hendrickx F, De Corte Z, Sonet G, Van Belleghem SM, Köstlbacher S, Vangestel C. 2022. A
1187 masculinizing supergene underlies an exaggerated male reproductive morph in a spider.
1188 *Nat Ecol Evol* **6**: 195–206.

1189 Hesselberg T, Vollrath F. 2004. The effects of neurotoxins on web-geometry and web-building
1190 behaviour in Araneus diadematus Cl. *Physiology & Behavior* **82**: 519–529.

1191 Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised
1192 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.
1193 *Bioinformatics* **32**: 767–769.

1194 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with
1195 BRAKER. In *Gene Prediction* (ed. M. Kollmar), Vol. 1962 of *Methods in Molecular*
1196 *Biology*, pp. 65–95, Springer New York, New York, NY
1197 http://link.springer.com/10.1007/978-1-4939-9173-0_5 (Accessed January 29, 2022).

1198 Iwata H, Gotoh O. 2012. Benchmarking spliced alignment programs including Spaln2, an
1199 extended version of Spaln that incorporates additional species-specific features. *Nucleic*
1200 *Acids Research* **40**: e161–e161.

1201 Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J,
1202 Gillespie M, Haw R, et al. 2019. The reactome pathway knowledgebase. *Nucleic Acids*
1203 *Research* gkz1031.

1204 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A,
1205 Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification.
1206 *Bioinformatics* **30**: 1236–1240.

1207 Kallal RJ, Kulkarni SS, Dimitrov D, Benavides LR, Arnedo MA, Giribet G, Hormiga G. 2021.
1208 Converging on the orb: denser taxon sampling elucidates spider phylogeny and new
1209 analytical methods support repeated evolution of the orb web. *Cladistics* **37**: 298–316.

1210 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and
1211 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

1212 Kim JY, Yoo JS, Park YC. 2014. The complete mitochondrial genome of the green crab spider
1213 *Oxytate striatipes* (Araneae: Thomisidae). *Mitochondrial DNA* 1–2.

1214 Kono N, Nakamura H, Mori M, Tomita M, Arakawa K. 2020. Spidroin profiling of cribellate
1215 spiders provides insight into the evolution of spider prey capture strategies. *Sci Rep* **10**:
1216 15721.

1217 Kono N, Nakamura H, Mori M, Yoshida Y, Ohtoshi R, Malay AD, Pedrazzoli Moran DA, Tomita
1218 M, Numata K, Arakawa K. 2021a. Multicomponent nature underlies the extraordinary
1219 mechanical properties of spider dragline silk. *Proc Natl Acad Sci USA* **118**:
1220 e2107065118.

1221 Kono N, Nakamura H, Ohtoshi R, Moran DAP, Shinohara A, Yoshida Y, Fujiwara M, Mori M,
1222 Tomita M, Arakawa K. 2019. Orb-weaving spider Araneus ventricosus genome
1223 elucidates the spidroin gene catalogue. *Sci Rep* **9**: 8380.

1224 Kono N, Ohtoshi R, Malay AD, Mori M, Masunaga H, Yoshida Y, Nakamura H, Numata K,
1225 Arakawa K. 2021b. Darwin's bark spider shares a spidroin repertoire with *Caerostris*
1226 *extrusa* but achieves extraordinary silk toughness through gene expression. *Open Biol*
1227 **11**: 210242.

1228 Král J, Forman M, Kořínková T, Lerma ACR, Haddad CR, Musilová J, Řezáč M, Herrera IMÁ,
1229 Thakur S, Dippenaar-Schoeman AS, et al. 2019. Insights into the karyotype and genome

1230 evolution of haplogyne spiders indicate a polyploid origin of lineage with holokinetic
1231 chromosomes. *Sci Rep* **9**: 3001.

1232 Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019.
1233 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral
1234 genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids*
1235 *Research* **47**: D807–D811.

1236 Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B. 2021.
1237 *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*. Zenodo
1238 https://zenodo.org/record/5127899 (Accessed February 4, 2022).

1239 Kumar V, Tyagi K, Chakraborty R, Prasad P, Kundu S, Tyagi I, Chandra K. 2020. The Complete
1240 Mitochondrial Genome of endemic giant tarantula, Lyrognathus crotalus (Araneae:
1241 Theraphosidae) and comparative analysis. *Sci Rep* **10**: 74.

1242 Kumari S, Lang G, DeSimone E, Spengler C, Trossmann VT, Lücker S, Hudel M, Jacobs K,
1243 Krämer N, Scheibel T. 2020. Engineered spider silk-based 2D and 3D materials prevent
1244 microbial infestation. *Materials Today* **41**: 21–33.

1245 Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a
1246 protein. *Journal of Molecular Biology* **157**: 105–132.

1247 Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. *Nucleic*
1248 *Acids Research* **46**: D493–D496.

1249 Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in
1250 2020. *Nucleic Acids Research* **49**: D458–D460.

1251 Lewis RV. 2006. Spider Silk: Ancient Ideas for New Biomaterials. *Chem Rev* **106**: 3762–3774.

1252 Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, Orengo C, Lees J. 2018. Gene3D:
1253 Extensive prediction of globular domains in proteins. *Nucleic Acids Research* **46**: D435–
1254 D439.

1255 Li C, Wang Z-L, Fang W-Y, Yu X-P. 2016. The complete mitochondrial genome of the orb-
1256 weaving spider *Neoscona theisi* (Walckenaer) (Araneae: Araneidae). *Mitochondrial DNA*
1257 *Part A* **27**: 4035–4036.

1258 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences ed. I. Birol. *Bioinformatics*
1259 **34**: 3094–3100.

1260 Li H. 2021. New strategies to improve minimap2 alignment accuracy. *arXiv:210803515 [q-bio]*.
1261 http://arxiv.org/abs/2108.03515 (Accessed February 18, 2022).

1262 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
1263 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
1264 format and SAMtools. *Bioinformatics* **25**: 2078–2079.

1265 Li X, Shi C-H, Tang C-L, Cai Y-M, Meng Q. 2017. The correlation between the length of
1266 repetitive domain and mechanical properties of the recombinant flagelliform spidroin.
1267 *Biology Open* bio.022665.

1268 Li Y-Y, Tsai J-M, Wu C-Y, Chiu Y-F, Li H-Y, Warrit N, Wan Y-C, Lin Y-P, Cheng R-C, Su Y-C.
1269 2022. In Silico Assessment of Probe-Capturing Strategies and Effectiveness in the
1270 Spider Sub-Lineage Araneoidea (Order: Araneae). *Diversity* **14**: 184.

1271 Liu FYC, Liu JYX, Yao X, Wang B. 2022. Hybrid sequencing reveals the full-length Nephila
1272 pilipes pyriform spidroin 1 (PySp1). *International Journal of Biological Macromolecules*
1273 **200**: 362–369.

1274 Liu M, Zhang Z, Peng Z. 2015. The mitochondrial genome of the water spider *Argyroneta*
1275 *aquatica* (Araneae: Cybaeidae). *Zool Scr* **44**: 179–190.

1276 Liu S, Aagaard A, Bechsgaard J, Bilde T. 2019. DNA Methylation Patterns in the Social Spider,
1277 Stegodyphus dumicola. *Genes* **10**: 137.

1278 Liu Z, Zhang M, Zhang Y, Xu Y, Zhang Y, Yang X, Zhang J, Yang J, Yuan L. 2020. Spider silk-
1279 based tapered optical fiber for humidity sensing based on multimode interference.
1280 *Sensors and Actuators A: Physical* **313**: 112179.

1281 Lomsadze A. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm.
1282 *Nucleic Acids Research* **33**: 6494–6506.

1283 Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into
1284 automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* **42**:
1285 e119–e119.

1286 Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI,
1287 Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in
1288 2020. *Nucleic Acids Research* **48**: D265–D268.

1289 Malay AD, Arakawa K, Numata K. 2017. Analysis of repetitive amino acid motifs reveals the
1290 essential features of spider dragline silk proteins ed. Q. Zou. *PLoS ONE* **12**: e0183397.

1291 Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of
1292 occurrences of k-mers. *Bioinformatics* **27**: 764–770.

1293 Masta SE, Boore JL. 2008. Parallel Evolution of Truncated Transfer RNA Genes in Arachnid
1294 Mitochondrial Genomes. *Molecular Biology and Evolution* **25**: 949–959.

1295 Mayank, Bardenhagen A, Sethi V, Gudwani H. 2022. Spider-silk composite material for
1296 aerospace application. *Acta Astronautica* **193**: 704–709.

1297 Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD. 2019. Protocol Update for
1298 large-scale genome and gene function analysis with the PANTHER classification system
1299 (v.14.0). *Nat Protoc* **14**: 703–721.

1300 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE,
1301 Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in
1302 2021. *Nucleic Acids Research* **49**: D412–D419.

1303 Motriuk-Smith D, Smith A, Hayashi CY, Lewis RV. 2005. Analysis of the Conserved N-Terminal
1304 Domains in Major Ampullate Spider Silk Proteins. *Biomacromolecules* **6**: 3152–3159.

1305 Necci M, Piovesan D, Dosztányi Z, Tosatto SCE. 2017. MobiDB-lite: Fast and highly specific
1306 consensus prediction of intrinsic disorder in proteins. *Bioinformatics* btx015.

1307 Ohno S. 2013. *Evolution by Gene Duplication.* Springer Berlin / Heidelberg, Berlin, Heidelberg
1308 https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5585971
1309 (Accessed March 7, 2022).

1310 Öksüz KE, Özkaya NK, İnan ZDŞ, Özer A. 2021. Novel natural spider silk embedded
1311 electrospun nanofiber mats for wound healing. *Materials Today Communications* **26**:
1312 101942.

1313 Opell BD, Hendricks ML. 2010. The role of granules within viscous capture threads of orb-
1314 weaving spiders. *Journal of Experimental Biology* **213**: 339–346.

1315 Pace RM, Grbić M, Nagy LM. 2016. Composition and genomic organization of arthropod Hox
1316 clusters. *EvoDevo* **7**: 11.

1317 Pan W-J, Fang H-Y, Zhang P, Pan H-C. 2016. The complete mitochondrial genome of flat
1318 spider *Selenops bursarius* (Araneae: Selenopidae). *Mitochondrial DNA* **27**: 1488–1489.

1319 Pan W-J, Fang H-Y, Zhang P, Pan H-C. 2014. The complete mitochondrial genome of striped
1320 lynx spider *Oxyopes sertatus* (Araneae: Oxyopidae). *Mitochondrial DNA* 1–2.

1321 Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J. 2019. The SUPERFAMILY
1322 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids*
1323 *Research* **47**: D490–D494.

1324 Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA,
1325 Bougueleret L, Poux S, et al. 2015. HAMAP in 2015: updates to the protein family
1326 classification and annotation system. *Nucleic Acids Research* **43**: D1064–D1070.

1327 Peters HM. 1992. On the spinning apparatus and the structure of the capture threads
1328 ofDeinopis subrufus (Araneae, Deinopidae). *Zoomorphology* **112**: 27–37.

1329 Peters HM. 1984. The spinning apparatus of Uloboridae in relation to the structure and
1330 construction of capture threads (Arachnida, Araneida). *Zoomorphology* **104**: 96–104.

1331 Piorkowski D, Blackledge TA, Liao C-P, Joel A-C, Weissbach M, Wu C-L, Tso I-M. 2020.
1332 Uncoiling springs promote mechanical functionality of spider cribellate silk. *Journal of*
1333 *Experimental Biology* jeb.215269.

1334 Pons J, Bover P, Bidegaray-Batista L, Arnedo MA. 2019. Arm-less mitochondrial tRNAs
1335 conserved for over 30 millions of years in spiders. *BMC Genomics* **20**: 665.

1336 Purcell J, Pruitt JN. 2019. Are personalities genetically determined? Inferences from subsocial
1337 spiders. *BMC Genomics* **20**: 867.

1338 Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley
1339 PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro
1340 method for long-range linkage. *Genome Res* **26**: 342–350.

1341 Qiu Y, Song D, Zhou K, Sun H. 2005. The Mitochondrial Sequences of Heptathela
1342 hangzhouensis and Ornithoctonus huwena Reveal Unique Gene Arrangements and
1343 Atypical tRNAs. *J Mol Evol* **60**: 57–71.

1344 Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005.
1345 InterProScan: protein domains identifier. *Nucleic Acids Research* **33**: W116–W120.

1346 Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for
1347 reference-free profiling of polyploid genomes. *Nat Commun* **11**: 1432.

1348 Reed CF, Witt PN, Scarboro MB. 1982. Maturation andd-amphetamine-induced changes in web
1349 building. *Dev Psychobiol* **15**: 61–70.

1350 Rens W, O'Brien PC, Grutzner F, Clarke O, Graphodatskaya D, Tsend-Ayush E, Trifonov VA,
1351 Skelton H, Wallis MC, Johnston S, et al. 2007. The multiple sex chromosomes of
1352 platypus and echidna are not completely identical and several share homology with the
1353 avian Z. *Genome Biol* **8**: R243.

1354 Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open
1355 Software Suite. *Trends in Genetics* **16**: 276–277.

1356 Rising A, Nimmervoll H, Grip S, Fernandez-Arias A, Storckenfeldt E, Knight DP, Vollrath F,
1357 Engström W. 2005. Spider Silk Proteins – Mechanical Property and Gene Sequence.
1358 *Zoological Science* **22**: 273–281.

1359 Roberson EJ, Chips MJ, Carson WP, Rooney TP. 2016. Deer herbivory reduces web-building
1360 spider abundance by simplifying forest vegetation structure. *PeerJ* **4**: e2538.

1361 Sahni V, Blackledge TA, Dhinojwala A. 2010. Viscoelastic solids explain spider web stickiness.
1362 *Nat Commun* **1**: 19.

1363 Sánchez-Herrero JF, Frías-López C, Escuer P, Hinojosa-Alvarez S, Arnedo MA, Sánchez-
1364 Gracia A, Rozas J. 2019. The draft genome sequence of the spider Dysdera silvatica
1365 (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic
1366 studies in chelicerates. *GigaScience* **8**: giz099.

1367 Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, Jiang X, Cheng L, Fan
1368 D, Feng Y, et al. 2014. Spider genomes provide insight into composition and evolution of
1369 venom and silk. *Nat Commun* **5**: 3765.

1370 Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y,
1371 Esposito L, Bechsgaard J, Bilde T, et al. 2017. The house spider genome reveals an
1372 ancient whole-genome duplication during arachnid evolution. *BMC Biol* **15**: 62.

Seeman T. 2018. *Barrnap: BAsic Rapbid Ribosomal RNA Predictor*. github.com/tseemann/barrnap.

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O. 2007. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Research* **35**: D260–D264.

Sember A, Pappová M, Forman M, Nguyen P, Marec F, Dalíková M, Divišová K, Doležálková-Kaštánková M, Zrzavá M, Sadílek D, et al. 2020. Patterns of Sex Chromosome Differentiation in Spiders: Insights from Comparative Genomic Hybridisation. *Genes* **11**: 849.

Sémon M, Wolfe KH. 2007. Consequences of genome duplication. *Current Opinion in Genetics & Development* **17**: 505–512.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. In *Gene Prediction* (ed. M. Kollmar), Vol. 1962 of *Methods in Molecular Biology*, pp. 227–245, Springer New York, New York, NY http://link.springer.com/10.1007/978-1-4939-9173-0_14 (Accessed February 4, 2022).

Sheffer MM, Hoppe A, Krehenwinkel H, Uhl G, Kuss AW, Jensen L, Jensen C, Gillespie RG, Hoff KJ, Prost S. 2021. Chromosome-level reference genome of the European wasp spider *Argiope bruennichi*: a resource for studies on range expansion and evolutionary adaptation. *GigaScience* **10**: giaa148.

Sigrist CJA. 2002. PROSITE: A documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics* **3**: 265–274.

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2012. New and continuing developments at PROSITE. *Nucleic Acids Research* **41**: D344–D347.

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

Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. 2022. AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc Natl Acad Sci USA* **119**: e2113075119.

Song L, Shankar DS, Florea L. 2016. Rascaf: Improving Genome Assembly with RNA Sequencing Data. *Plant Genome* **9**. https://onlinelibrary.wiley.com/doi/10.3835/plantgenome2016.03.0027 (Accessed February 4, 2022).

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637–644.

1411 Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a
1412 generalized hidden Markov model that uses hints from external sources. *BMC*
1413 *Bioinformatics* **7**: 62.

1414 Stellwagen SD, Renberg RL. 2019. Toward Spider Glue: Long Read Scaffolding for Extreme
1415 Length and Repetitious Silk Family Genes AgSp1 and AgSp2 with Insights into
1416 Functional Adaptation. *G3 Genes|Genomes|Genetics* **9**: 1909–1919.

1417 Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of
1418 transposable element families, sequence models, and genome annotations. *Mobile DNA*
1419 **12**: 2.

1420 Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView.
1421 *Bioinformatics* **21**: 537–539.

1422 Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to Identify Repetitive Elements in
1423 Genomic Sequences. *Current Protocols in Bioinformatics* **25**.
1424 https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0410s25 (Accessed February
1425 4, 2022).

1426 Tarakanova A, Buehler MJ. 2012. The role of capture spiral silk properties in the diversification
1427 of orb webs. *J R Soc Interface* **9**: 3240–3248.

1428 Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O,
1429 Brunak S, von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal
1430 peptides using protein language models. *Nat Biotechnol*.
1431 https://www.nature.com/articles/s41587-021-01156-3 (Accessed February 4, 2022).

1432 Teulé F, Furin WA, Cooper AR, Duncan JR, Lewis RV. 2007. Modifications of spider silk
1433 sequences in an attempt to control the mechanical properties of the synthetic fibers. *J*
1434 *Mater Sci* **42**: 8974–8985.

1435 The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall
1436 CJ, Basu S, Chisholm RL, Dodson RJ, et al. 2021. The Gene Ontology resource:
1437 enriching a GOld mine. *Nucleic Acids Research* **49**: D325–D334.

1438 The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids*
1439 *Research* **47**: D506–D515.

1440 Tian X-X, Pan W-J, Chen L-L, Xu Y-Y, Pan H-C. 2016. The complete mitochondrial genome of
1441 stretch spider *Tetragnatha maxillosa* (Araneae: Tetragnathidae). *Mitochondrial DNA Part*
1442 *A* **27**: 3469–3470.

1443 Tremblay M-L, Xu L, Lefèvre T, Sarker M, Orrell KE, Leclerc J, Meng Q, Pézolet M, Auger M,
1444 Liu X-Q, et al. 2015. Spider wrapping silk fibre architecture arising from its modular
1445 soluble protein precursor. *Sci Rep* **5**: 11502.

1446 Vienneau-Hathaway JM, Brassfield ER, Lane AK, Collin MA, Correa-Garhwal SM, Clarke TH,
1447 Schwager EE, Garb JE, Hayashi CY, Ayoub NA. 2017. Duplication and concerted
1448 evolution of MiSp-encoding genes underlie the material properties of minor ampullate
1449 silks of cobweb weaving spiders. *BMC Evol Biol* **17**: 78.

1450 Vollrath F. 1999. Biology of spider silk. *International Journal of Biological Macromolecules* **24**:
1451 81–88.

1452 Vollrath F, Selden P. 2007. The Role of Behavior in the Evolution of Spiders, Silks, and Webs.
1453 *Annu Rev Ecol Evol Syst* **38**: 819–846.

1454 Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC.
1455 2017. GenomeScope: fast reference-free genome profiling from short reads ed. B.
1456 Berger. *Bioinformatics* **33**: 2202–2204.

1457 Wang K, Wen R, Jia Q, Liu X, Xiao J, Meng Q. 2019a. Analysis of the Full-Length Pyriform
1458 Spidroin Gene Sequence. *Genes* **10**: 425.

1459 Wang Z-L, Li C, Fang W-Y, Yu X-P. 2014. The complete mitochondrial genome of orb-weaving
1460 spider *Araneus ventricosus* (Araneae: Araneidae). *Mitochondrial DNA* 1–2.

1461 Wang Z-L, Li C, Fang W-Y, Yu X-P. 2016a. The complete mitochondrial genome of the wolf
1462 spider *Wadicosa fidelis* (Araneae: Lycosidae). *Mitochondrial DNA Part A* **27**: 3909–3910.

1463 Wang Z-L, Li C, Fang W-Y, Yu X-P. 2016b. The Complete Mitochondrial Genome of two
1464 *Tetragnatha* Spiders (Araneae: Tetragnathidae): Severe Truncation of tRNAs and Novel
1465 Gene Rearrangements in Araneae. *Int J Biol Sci* **12**: 109–119.

1466 Wang Z-L, Wang Z-Y, Huang J, Yu X-P. 2019b. The complete mitochondrial genome of an orb-
1467 weaver spider *Araneus angulatus* (Araneae: Araneidae). *Mitochondrial DNA Part B* **4**:
1468 3870–3871.

1469 Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV,
1470 Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction
1471 and Phylogenomics. *Molecular Biology and Evolution* **35**: 543–548.

1472 Waters PD, Patel HR, Ruiz-Herrera A, Álvarez-González L, Lister NC, Simakov O, Ezaz T, Kaur
1473 P, Frere C, Grützner F, et al. 2021. Microchromosomes are building blocks of bird,
1474 reptile, and mammal chromosomes. *Proc Natl Acad Sci USA* **118**: e2112494118.

1475 Wen R, Liu X, Meng Q. 2017. Characterization of full-length tubuliform spidroin gene from
1476 Araneus ventricosus. *International Journal of Biological Macromolecules* **105**: 702–710.

1477 Wen R, Wang K, Liu X, Li X, Mi J, Meng Q. 2018. Molecular cloning and analysis of the full-
1478 length aciniform spidroin gene from Araneus ventricosus. *International Journal of*
1479 *Biological Macromolecules* **117**: 1352–1360.

1480 Wen R, Wang K, Meng Q. 2020. The three novel complete aciniform spidroin variants from
1481 Araneus ventricosus reveal diversity of gene sequences within specific spidroin type.
1482 *International Journal of Biological Macromolecules* **157**: 60–66.

1483 White MJD. 1978. *Modes of speciation*. W. H. Freeman, San Francisco.

1484 Witt PN, Reed CF. 1965. Spider-Web Building: Measurement of web geometry identifies
1485 components in a complex invertebrate behavior pattern. *Science* **149**: 1190–1197.

1486 Wu CH. 2004. PIRSF: family classification system at the Protein Information Resource. *Nucleic*
1487 *Acids Research* **32**: 112D – 114.

1488 Xu M, Jiang Y, Pradhan S, Yadavalli VK. 2019. Use of Silk Proteins to Form Organic, Flexible,
1489 Degradable Biosensors for Metabolite Monitoring. *Front Mater* **6**: 331.

1490 Yoshido A, Šíchová J, Pospíšilová K, Nguyen P, Voleníková A, Šafář J, Provazník J, Vila R,
1491 Marec F. 2020. Evolution of multiple sex-chromosomes associated with dynamic
1492 genome reshuffling in Leptidea wood-white butterflies. *Heredity* **125**: 138–154.

1493 Yu N, Li J, Liu M, Huang L, Bao H, Yang Z, Zhang Y, Gao H, Wang Z, Yang Y, et al. 2019.
1494 *Genome sequencing and neurotoxin diversity of a wandering spider* Pardosa
1495 pseudoannulata *(pond wolf spider)*. Genomics
1496 http://biorxiv.org/lookup/doi/10.1101/747147 (Accessed February 4, 2022).

1497 Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**:
1498 292–298.

1499 Zhu B, Jin P, Hou Z, Li J, Wei S, Li S. 2022. Chromosomal-level genome of a sheet-web spider
1500 provides insight into the composition and evolution of venom. *Molecular Ecology*
1501 *Resources* 1755–0998.13601.

1502 Zhu X-L, Zhang Z-S. 2017. The complete mitochondrial genome of *Agelena silvatica* (Araneae:
1503 Agelenidae). *Mitochondrial DNA Part B* **2**: 58–59.

1504 Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome
1505 assembler. *Bioinformatics* **29**: 2669–2677.

1506 Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL.
1507 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii* , a
1508 progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* **27**:
1509 787–792.

1510 Zimin AV, Salzberg SL. 2022. The SAMBA tool uses long reads to improve the contiguity of
1511 genome assemblies ed. M. Shao. *PLoS Comput Biol* **18**: e1009860.

1512 Zschokke S, Vollrath F. 1995. Web construction patterns in a range of orb weaving spiders
1513 (Araneae). *European Journal of Entomology* **92**: 523–541.

1514

Figure 1

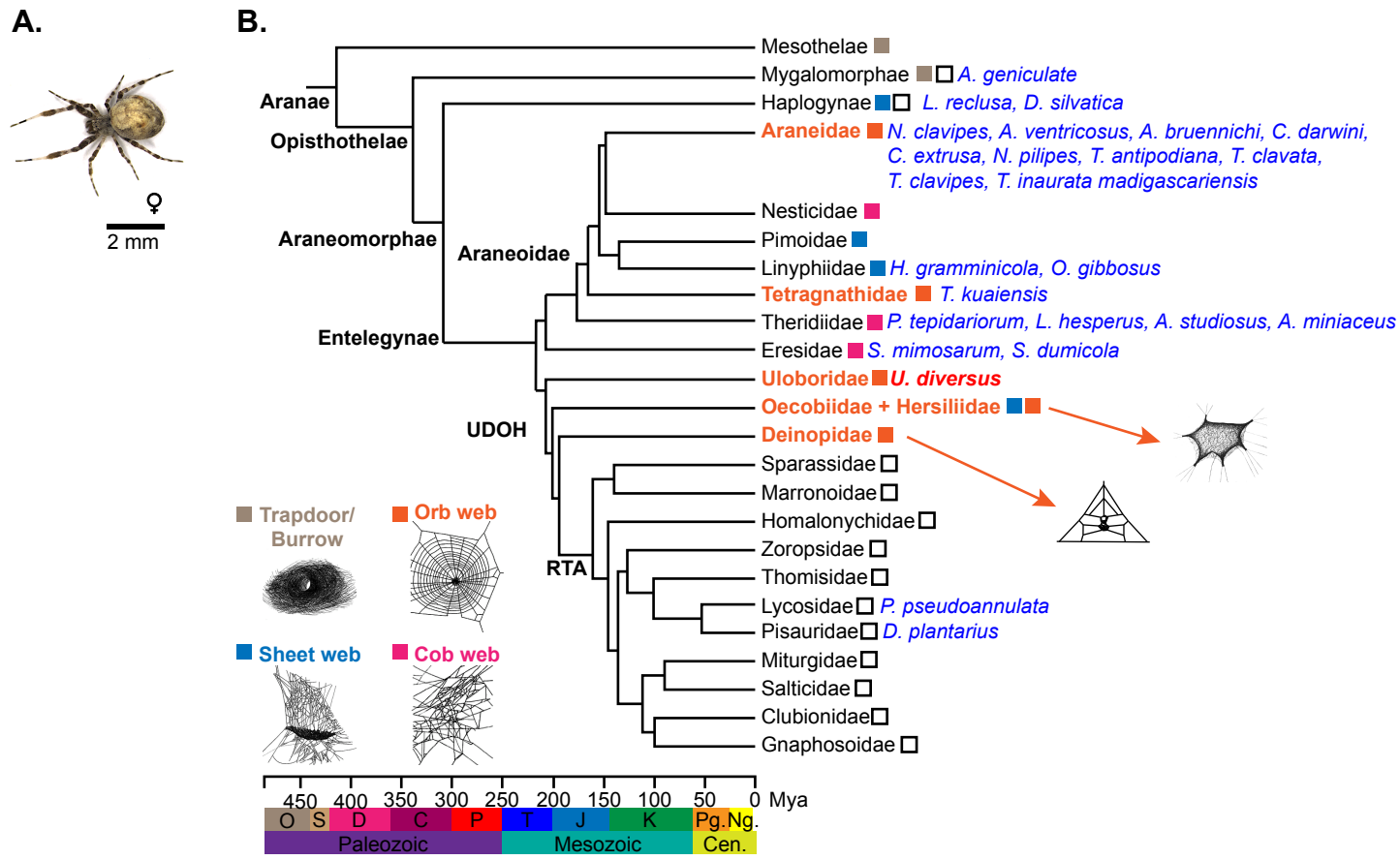


Figure 2

A.

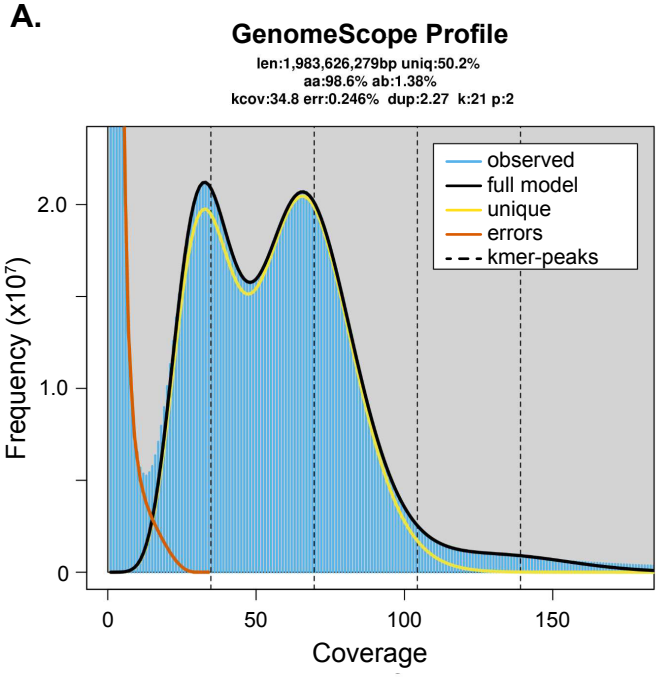


Figure 3

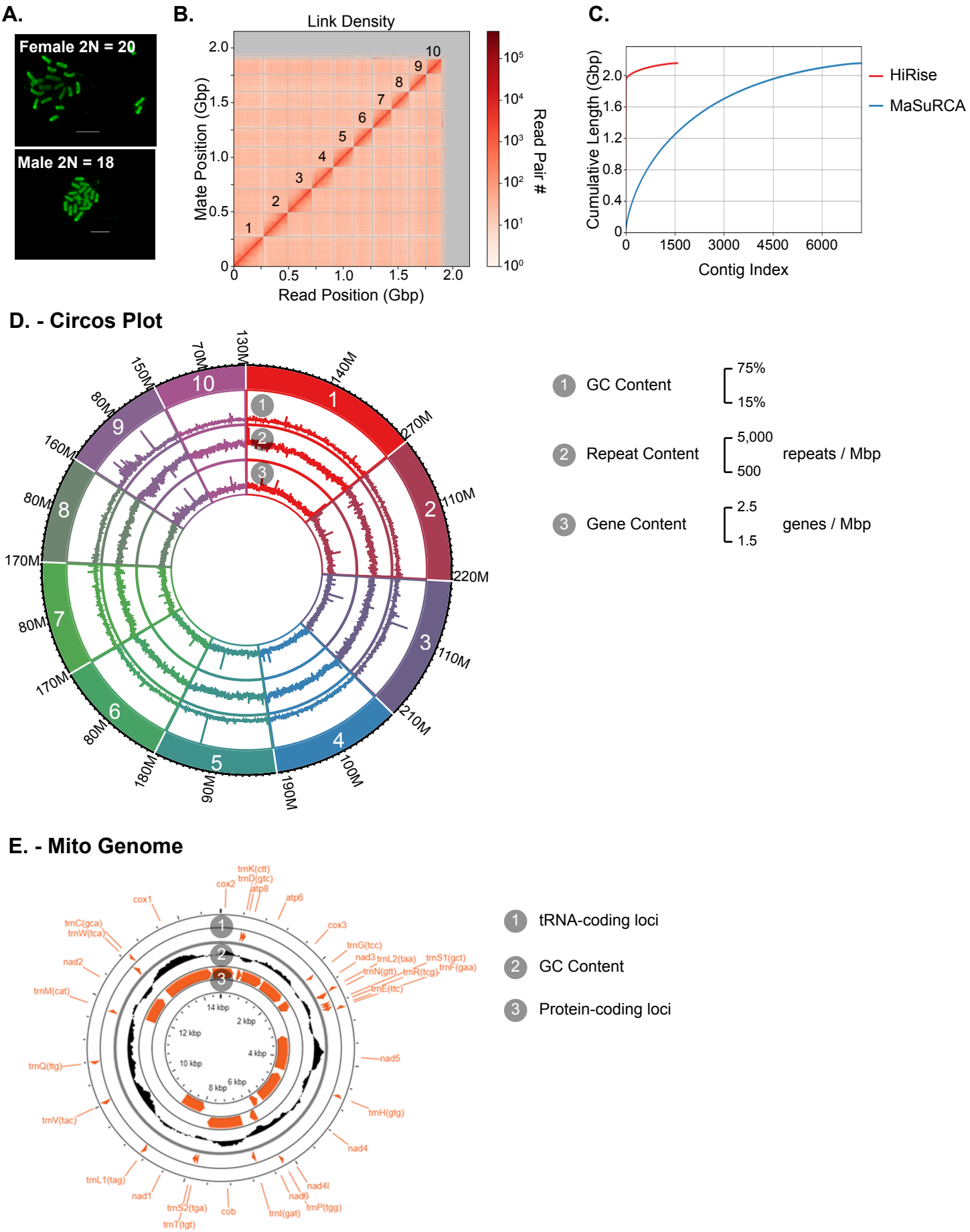
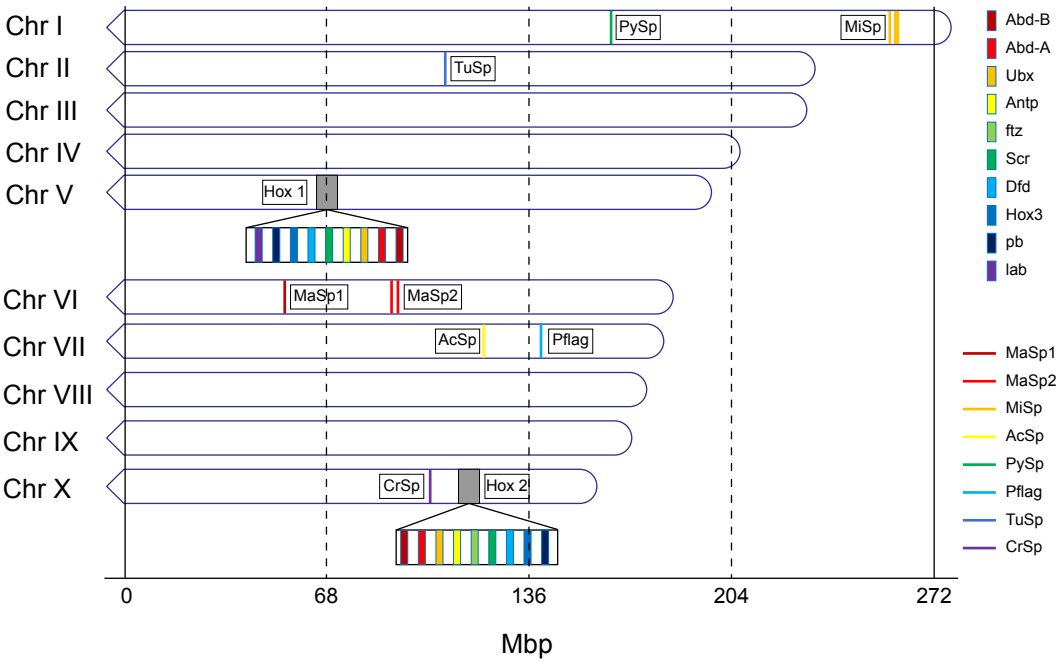


Figure 4

A.



B.

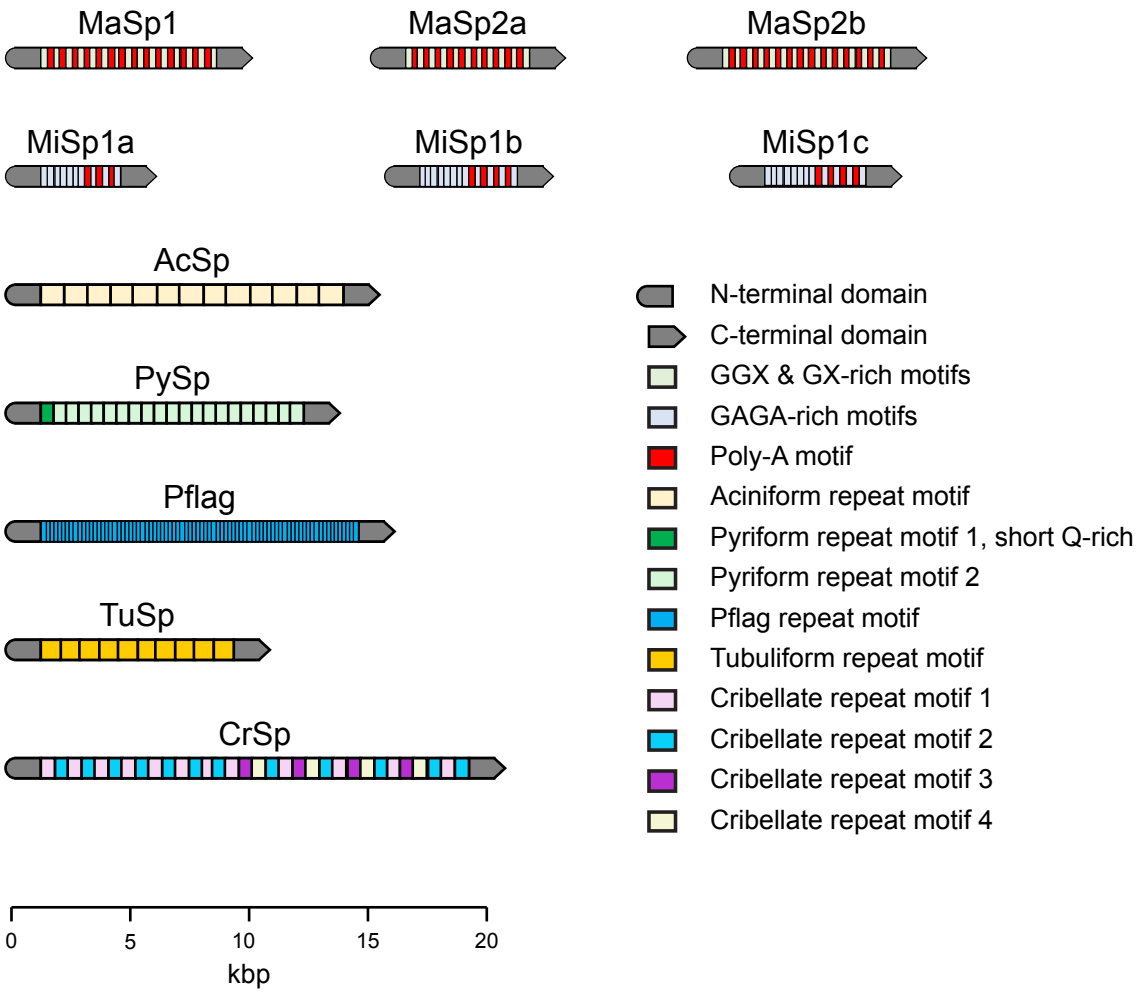


Figure 5

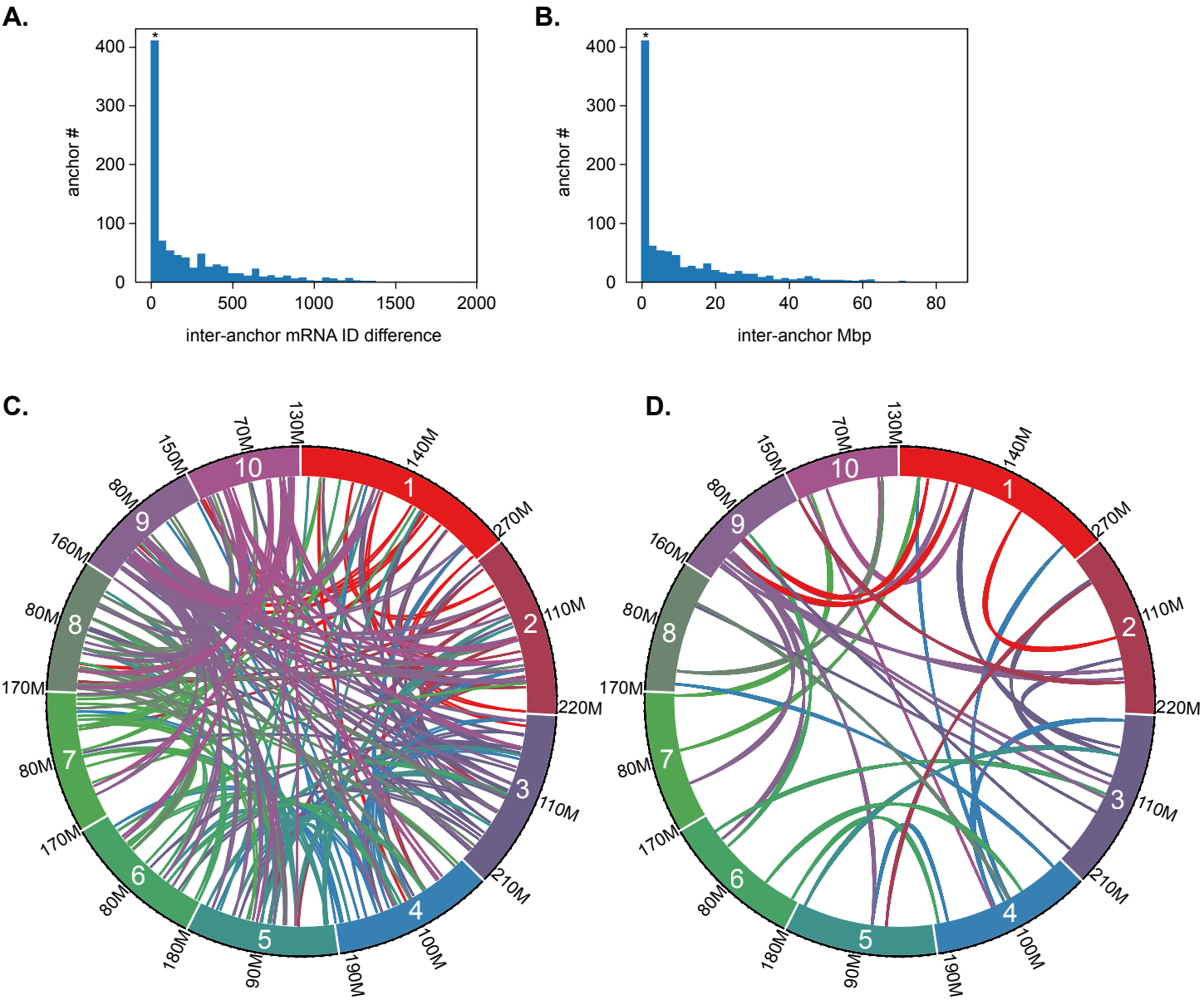
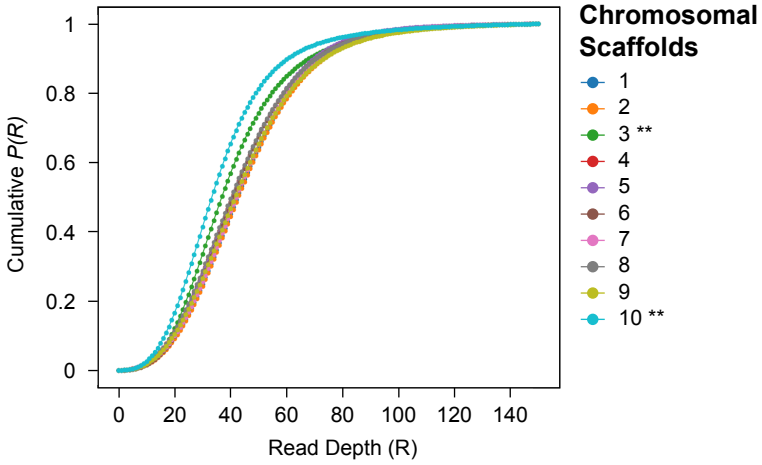
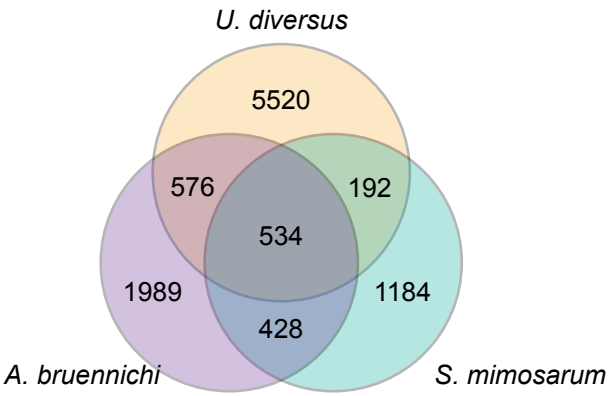


Figure 6

A.



B.



C.

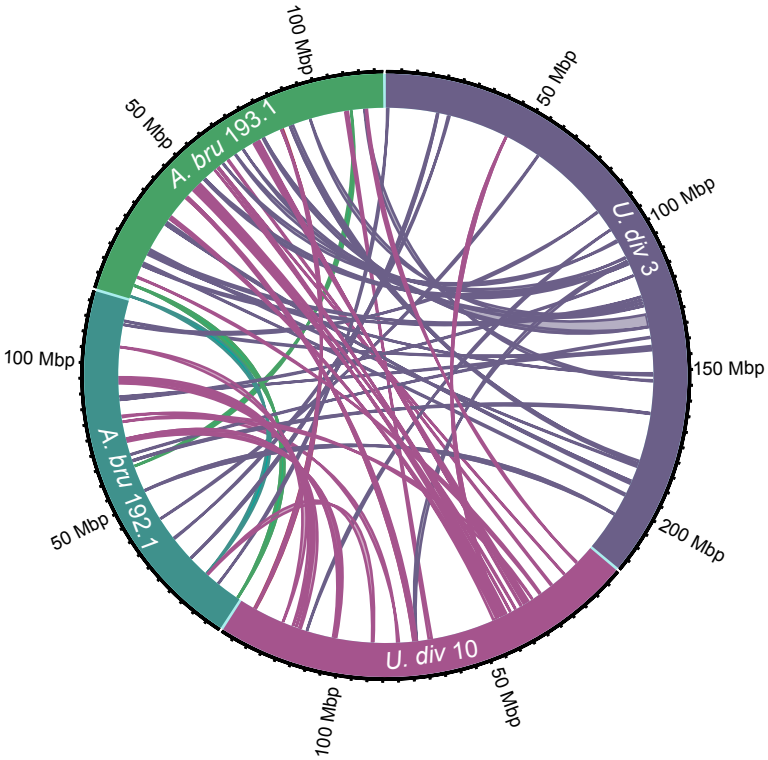


Table 1. Summary of Sequencing Library Statistics

Library Type	Instrument	Mean Read Length	Number of Reads or Read Pairs	Bases Sequenced	Coverage (X)[a]
Illumina	NovaSeq 6000	2 x 150 bp	795 million (female) 937 million (male)	119.3 Gb (female) 46.8 Gb (male)	60 (female) 24 (male)
ONT	PromethION	6.7 kb	14.7 million	98.4 Gb	50
PacBio	Sequel II	12.8 kb (subreads) 13.0 kb (consensus)	34.9 million (subreads) 2.0 million (consensus)	412.8 Gb (subreads) 26.2 Gb (consensus)	208 (subreads) 13.1 (consensus)
Chicago Hi-C	HiSeq X	2 x 150 bp	286 million	85.8 Gb	21b 43c
Dovetail Hi-C	HiSeq X	2 x 150 bp	537 million	161.1 Gb	1,009b 81c

[a] Based on in silico genome size estimate of 1.98 Gb by *k*-mer analysis with GenomeScope 2.0

[b] Physical coverage, defined as the number of read pairs that span a base pair

[c] Sequence coverage, defined as number of times a base pair is directly observed in sequencing data

Table 2. Summary of *Uloborus diversus* Draft Genome Assembly Statistics

	U. div v.1.0		*U. div v.1.1*		*U. div v.1.2*		*U. div v.1.3*		*U. div v.2.0*		*U. div v.3.0*		*U. div v.3.1*	
	Contigs	**Scaffolds**	**Contigs**	**Scaffolds**	**Contigs**	**Scaffolds**	**Contigs**	**Scaffolds**	**Contigs**	**Scaffolds**	**Contigs**	**Scaffolds**	**Contigs**	**Scaffolds**
Total Length	3,218,774,062		3,218,805,127		2,798,889,957		2,802,061,857		2,122,354,655		2,151,304,433		2,151,890,133	
Number of Contigs / Scaffolds	68,581	68,259	68,581	66,417	50,738	48,632	55,142	21,317	9,997	9,734	7,467	7,197	7,713	1,586
Largest Contig / Scaffold	3,987,394	3,987,394	3,987,394	6,446,995	3,987,394	6,446,995	3,987,394	367,325,988	4,617,239	4,617,239	5,877,357	5,877,357	5,877,357	272,330,431
Mean Contig / Scaffold Length	46,933	47,155	46,933	48,463	55,162	57,552	50,756	131,447	212,298	218,035	288,107	298,916	278,918	1,356,803
Median Contig / Scaffold Length	21,841	21,923	21,841	21,546	23,184	22,648	24,516	12,994	137,429	218,035	166,676	175,879	168,537	94,191
Smallest Contig / Scaffold	1,106	1,667	1,106	1,667	1,106	1,667	10	312	10,045	10,045	10,045	10,045	380	10,220
N10 / L10	558,595 / 374	564,850 / 373	558,595 / 374	686,213 / 308	613,610 / 302	744,166 / 249	432,729 / 4,437	367,325,988 /1	1,100,324 / 132	1,100,324 / 132	1,705,719 / 86	1,715,742 / 84	1,564,267 / 91	272,330,431 / 1
N20 / L 20	330,632 / 1,135	332,714 / 1,131	330,632 / 1,135	395,570 / 943	375,534 / 894	452,633 / 744	271,565 / 1,278	289,477,137 / 2	711,683 / 383	713,631 / 382	1,147,222 / 244	1,163,865 / 241	1,032,642 / 237	221,850,441 / 2
N30 / L30	214,753 / 2,358	215,989 / 2,348	214,753 / 2,358	252,183 / 1,983	252,860 / 1,816	296,524 / 1,521	192,451 / 2,518	264,078,423 / 3	516,155 / 741	518,859 / 736	826,976 / 467	844,177 / 461	763,983 / 509	218,885,758 / 3
N40 / L40	114,360 / 4,189	145,056 / 4,171	114,360 / 4,189	166,010 / 3,572	178,542 / 3,143	205,473 / 2,661	139,664 / 4,234	243,450,885 / 4	401,815 / 1,210	404,285 / 1,205	636,610 / 763	648,167 / 753	587,015 / 831	185,519,777 / 4
N50 / L50	97,524 / 6,913	98,015 / 6,885	97,524 / 6,913	108,431 / 5,994	126,236 / 5,019	142,390 / 4,302	103,046 / 6,850	241,033,576 / 5	326,127 / 1,798	328,082 / 1,789	487,746 / 1,150	496,769 / 1,133	452,781 / 1,250	185,519,777 / 5
N60 / L60	64,898 / 10,992	65,255 / 10,942	64,898 / 10,992	69,185 / 9,724	88,433 / 7,678	97,899 / 6,679	75,059 /9,775	217,901,867 / 7	259,595 / 2,527	261,845 / 2,511	380,517 / 1,651	387,963 / 1,626	359,368 /1,784	172,099,698 / 7
N70 / L70	42,949 / 17,131	43,310 / 17,041	42,949 / 17,131	44,720 / 15,580	60,079 / 11,529	64,377 / 10,226	53,001 / 14,221	213,666,783 / 8	201,138 / 3,460	204,029 / 3,433	292,644 / 2,296	298,089 / 2,258	280,097 / 2,463	161,310,338 / 8
N80 / L80	28,594 / 26,353	28,760 / 26,208	28,594 / 26,353	29,173 / 24,561	38,159 / 17,400	39,689 / 15,782	35,344 / 20,692	183,225,947 / 9	148,105 / 4,686	151,467 / 4,638	206,980 / 3,169	211,349 / 3,116	200,858 / 3,369	159,483,530 / 9
N90 / L90	18,127 / 40,511	18,198 / 40,294	18,127 / 40,511	18,295 / 38,527	21,027 / 27,290	21,374 / 25,391	20,312 / 31,107	37,840 / 1,399	97,979 / 6,436	100,002 / 6,354	127,987 / 4,484	131,815 / 4,398	126,330 / 4,713	9,355,840 / 14
N100 / L100	1,106 / 68,581	1,667 / 68,259	1,106 / 68,581	1,667 / 66,417	1,106 / 50,738	1,667 / 48,632	10 / 55142	312 / 21,317	10,045 / 9,997	10,045 / 9,734	10,045 / 7,467	10,045 / 7,197	380 / 7,713	10,220 / 1,586
Gaps	322		2,164				33,825		263		270		6,127	
Ns	32,200		63,265		57,963		3,229,863		6,049		6,749		592,449	
GC Content (%)	33.78		33.78		33.72		33.72		33.83		33.82		33.82	

U. div v.1.0 is the MaSuRCA assembly
U. div v.1.1 is the MaSuRCA assembly with further scaffolding using Rascaf
U. div v.1.2 is the MaSuRCA assembly with Rascaf and with reduction of redundancy using Pseudohaploid
U. div v.1.3 is the MaSuRCA assembly with Rascaf and Pseudohaploid further scaffolded using Chicago and Dovetail Hi-C
U. div v.2.0 is the PacBio IPA assembly
U. div v.3.0 is the MaSuRCA assembly and PacBio IPA assembly merged using MaSuRCA
U. div v.3.1 is the merged MaSuRCA/IPA assembly further scaffolded using Dovetail Hi-C

Table 3. Summary of Uloborus diversus Draft Genome Assembly BUSCO Scores

	U. div v.1.0	*U. div v.1.1*	*U. div v.1.2*	*U. div v.1.3*	*U. div v.2.0*	*U. div v.3.0*	*U. div v.3.1*
Complete	94.7	95.4	95.1	97.0	92.4	94.1	94.8
Single Copy	70.9	72.9	78.4	84.6	82.1	82.9	85.2
Duplicated	23.8	22.5	16.7	12.4	10.3	11.2	9.6
Fragmented	2.3	1.9	2.0	1.2	1.7	1.3	1.0
Missing	3.0	2.7	2.9	1.8	5.9	4.6	4.2

Total BUSCOs 2,934

U. div v.1.0 is the MaSuRCA assembly

U. div v.1.1 is the MaSuRCA assembly with further scaffolding using Rascaf

U. div v.1.2 is the MaSuRCA assembly with Rascaf and with reduction of redundancy using Pseudohaploid

U. div v.1.3 is the MaSuRCA assembly with Rascaf and Pseudohaploid further scaffolded using Chicago and Dovetail Hi-C

U. div v.2.0 is the PacBio IPA assembly

U. div v.3.0 is the MaSuRCA assembly and PacBio IPA assembly merged using MaSuRCA

U. div v.3.1 is the merged MaSuRCA/IPA assembly further scaffolded using Dovetail Hi-C

Table 4. Summary of the Repeat Content of the Uloborus diversus Draft Genome Assembly

Type of Element	Number of Elements	Total Length	Percent of Assembly
Retroelements	311,947	165,713,554	7.70
SINEs	87,037	46,791,087	2.17
Penelope	29,911	11,105,124	0.52
LINEs	151,953	66,189,386	3.08
CRE/SLAC	0	0	0.00
L2 / CR1 / Rex	31,294	18,030,662	0.84
R1 / LOA / Jockey	17,662	9,587,620	0.45
R2 / R4 / NeSL	0	0	0.00
RTE / Bov-B	40,080	15,062,698	0.70
L1 / CIN4	21,465	5,682,132	0.26
LTR elements	72,957	52,733,081	2.45
BEL / Pao	10,210	8,094,671	0.38
Ty1 / Copia	24,043	10,565,700	0.49
Gypsy / DIRS1	27,837	29,214,338	1.36
Retroviral	10,867	4,858,372	0.23
DNA transposons	1,510,089	489,113,225	22.73
hobo-Activator	545,025	164,842,101	7.66
Tc1-IS630-Pogo	428,233	150,644,883	7.00
En-Spm	0	0	0.00
MuDR-IS905	0	0	0.00
PiggyBac	11,411	4,281,311	0.20
Tourist / Harbinger	6,573	2,466,806	0.11
Other (Mirage, P-element, Transib)	2,942	1,822,173	0.08
Rolling-circles	229,864	88,460,474	4.11
Unclassified	2,482,847	629,886,344	29.27
Total Interspersed Repeats		1,284,713,123	59.70
Small RNA	20,797	5,420,377	0.25
Satellites	0	0	0.00
Simple repeats	529,329	56,254,893	2.61
Low complexity	67,980	3,208,338	0.15
		Total:	66.58

Table 5. Summary of Annotation Statistics for the Uloborus diversus Draft Genome Assembly

Number of Gene Models	45,762
Minimum Gene Model Length (bp)	60
Maximum Gene Model Length (bp)	408,348
Average Gene Model Length (bp)	16,764
Number of Exons	222,483
Average Number of Exons per Gene Model	5
Average Exon Length (bp)	237
Number of Transcripts	47,540
Average Number of Transcripts per Gene Model	1
Number of Gene Models < 200 bp	37

Table 6. Summary of Interproscan Results

Database	Total Hits	Individual mRNAs with Hits
CDD	11047	6560
Coils	9254	5947
Gene3D	34377	16424
Hamap	270	260
MobiDBLite	44028	14046
PANTHER	46497	20941
Pfam	35843	19491
PIRSF	916	736
PRINTS	16380	3134
ProSitePatterns	8222	3898
ProSiteProfiles	23252	9922
SFLD	120	62
SMART	25638	7561
SUPERFAMILY	26766	15559
TIGRFAM	821	736

Table 7. Summary of Spidroins in the Uloborus diversus Draft Genome Assembly

Spidroin	Gene Length (bp)	CDS Length (bp)	Protein Length (aa)	N-terminal Length (aa)	C-terminal Length (aa)	Signal Peptide Start	Signal Peptide Stop	Signal Peptide Probability	Signal Peptidase Cleavage Site Probability
Aciniform	14,907	14,907	4,968	148	109	Met-1	Ser-23	0.9955	0.9056
Pseudoflagelliform	15,546	15,546	5,181	195	96	Met-1	Gly-29	0.9991	0.9778
Cribellate	20,198	20,115	6,704	874	296	Met-1	Gly-23	0.9997	0.9805
Major Ampullate 2a	7,317	7,317	2,438	165	100	Met-1	Gly-25	0.9997	0.9662
Major Ampullate 2b	9,141	9,141	3,046	195	109	Met-1	Gly-25	0.9687	0.9992
Minor Ampullate 1a	5,523	5,523	1,840	242	95	Met-1	Gly-23	0.9998	0.9643
Minor Ampullate 2a	6,255	6,255	2,084	254	99	Met-1	Gly-23	0.9998	0.9670
Minor Ampullate 2b	6,432	6,432	2,143	254	98	Met-1	Gly-23	0.9998	0.9670
Pyriform	13,179	13,179	4,392	168	266	Met-1	Gly-23	0.9997	0.9768
Tubuliform	10,146	10,146	3,381	179	346	Met-1	Ala-25	0.9993	0.9808