Allelic resolution of insect and spider silk genes reveals hidden genetic diversity

Paul B. Frandsen[1,2,*], Scott Hotaling, Ashlyn Powell[1], Jacqueline Heckenhauer[2,4], Akito Y. Kawahara[5], Richard H. Baker[6], Cheryl Y. Hayashi[6], Blanca Ríos-Touma[7], Ralph Holzenthal[8], Steffen U. Pauls[2,4,9], Russell J. Stewart[10]

Affiliations:
[1] Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, USA
[2] LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany
[3] Department of Watershed Sciences, Utah State University, Logan, UT, USA
[4] Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt, Germany
[5] McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA
[6] Division of Invertebrate Zoology and Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA
[7] Grupo de Investigación en Biodiversidad, Medio Ambiente y Salud (BIOMAS), Universidad de las Américas, Quito, Ecuador
[8] Department of Entomology, University of Minnesota, St. Paul, Minnesota, USA
[9] Institute for Insect Biotechnology, Justus-Liebig-University, Gießen, Germany
[10] Department of Biomedical Engineering, University of Utah, Salt Lake City, UT 84112, USA
*To whom correspondence should be addressed: paul_frandsen@byu.edu

Abstract:
Arthropod silk is vital to the evolutionary success of hundreds of thousands of species. The primary proteins in silks are often encoded by long, repetitive gene sequences. Until recently, sequencing and assembling these complex gene sequences has proven intractable given their repetitive structure. Here, using high-quality long-read sequencing, we show that there is extensive variation—both in terms of length and repeat motif order—between alleles of silk genes within individual arthropods. Further, this variation exists across two deep, independent origins of silk which diverged more than 500 million years ago: (1) the insect clade containing caddisflies and butterflies and (2) spiders. This remarkable convergence in previously overlooked patterns of allelic variation across multiple origins of silk suggests common mechanisms for the generation and maintenance of structural protein-coding genes. Future genomic efforts to connect genotypes to phenotypes should account for such allelic variation.

Intro:
Silk is fundamental to the life histories of hundreds of thousands of arthropods (1). The genes that encode for silk proteins are often long and repetitive, and their variation is directly tied to silk phenotypes. Despite major headway in resolving difficult-to-assemble regions in *de novo* genome assembly [e.g., the human telomere-to-telomere consortium (2)], assemblies of biodiverse organisms vary widely in quality and contiguity (3, 4) and long repetitive regions pose a particularly difficult challenge (5). Arthropod silk genes are difficult to assemble due to their repetitive internal region, which forms the semi-crystalline protein structure underlying the unique properties of silk (1). In the sister orders Lepidoptera (butterflies and moths) and Trichoptera (caddisflies), the gene that encodes the primary protein component of silk is *heavy chain fibroin* (*H-fibroin*), which originated in their common ancestor. However, structural silk protein-coding genes have independently evolved multiple times across the arthropod tree of life. For example, in spiders, there are multiple repetitive silk genes that encode a suite of proteins collectively known as spidroins (6).

The first full length *H-fibroin* and spidroin genes were published in 2000 and 2007, respectively through sequencing of bacterial artificial chromosome (BAC) and fosmid libraries (7, 8). Subsequent attempts using high throuput sequencing to resolve full-length *H-fibroin* and spidroin sequences were unsuccessful due to the difficulty in resolving long repetitive regions with short-read sequencing (9). This lack of full-length silk gene sequences has hindered the analysis of the variation present in these large (>20kbp), repetitive proteins, leaving a significant gap in our understanding of their evolution and structure. Recent advances in long-read sequencing enabled recovery of these regions (10–13). However, only single consensus sequences were recovered, leaving allelic variation hidden. New sequencing technologies, e.g., PacBio HiFi, generate accurate long-reads that can be assembled into haploid-resolved genomes, even for regions of the genome that were previously intractable to assemble (2, 5). Here we present the first examination of allelic structure in insect and spider silk genes and show that there is substantial diversity within individuals, demonstrating a wealth of genomic variation that was previously overlooked. Ultimately, characterizing and understanding this allelic variation is essential to uncovering the molecular mechanisms that shaped these highly modular structural proteins that are central to the success of hundreds of thousands of animals.

Results and Discussion:
We obtained high-quality reference genomes from single individuals of a butterfly (*Vanessa cardui*), three caddisfly species across a diversity of silk use (*Hesperophylax magnus*, a case-making caddisfly; *Atopsyche davidsoni*, a cocoon-maker; and *Arctopsyche grandis*, a retreat-maker), and a spider (*Argiope argentata*). Two of these were newly generated and the primary assembles of the others were previously published (Table 1).

We recovered full-length, fully resolved sequences for both alleles of *H-fibroin* across all three caddisflies, unveiling substantial, and previously hidden, heterozygosity within each individual (Fig. 1). The variation between alleles can largely be ascribed to indels resulting in allele sequences with considerable differences in length (Table 1). Because the origin of *H-fibroin* can be traced back to the common, silk-spinning ancestor of Trichoptera and Lepidoptera more than 290 million years ago, we also investigated the *H-fibroin* sequence in the butterfly *V. cardui* to determine whether such patterns of heterozygosity were consistent across the evolutionary history of the gene. As with the caddisflies, we recovered two distinct *H-fibroin* alleles. Across all samples of Trichoptera and Lepidoptera, the structure of the *H-fibroin* gene was conspicuously conserved. For example, each *H-fibroin* sequence included conserved termini with a repetitive internal region consisting of modular repetitive units. Each gene was structured by a short initial exon followed by an intron and a long terminal exon that contained the entire repetitive region in a single open reading frame. Nevertheless, within each individual, allelic variation was dramatic, resulting from apparent deletions and insertions of repetitive modules (Fig. 1). However, despite the similarity in *H-fibroin* gene structure, the sequences and number of repeats in the internal

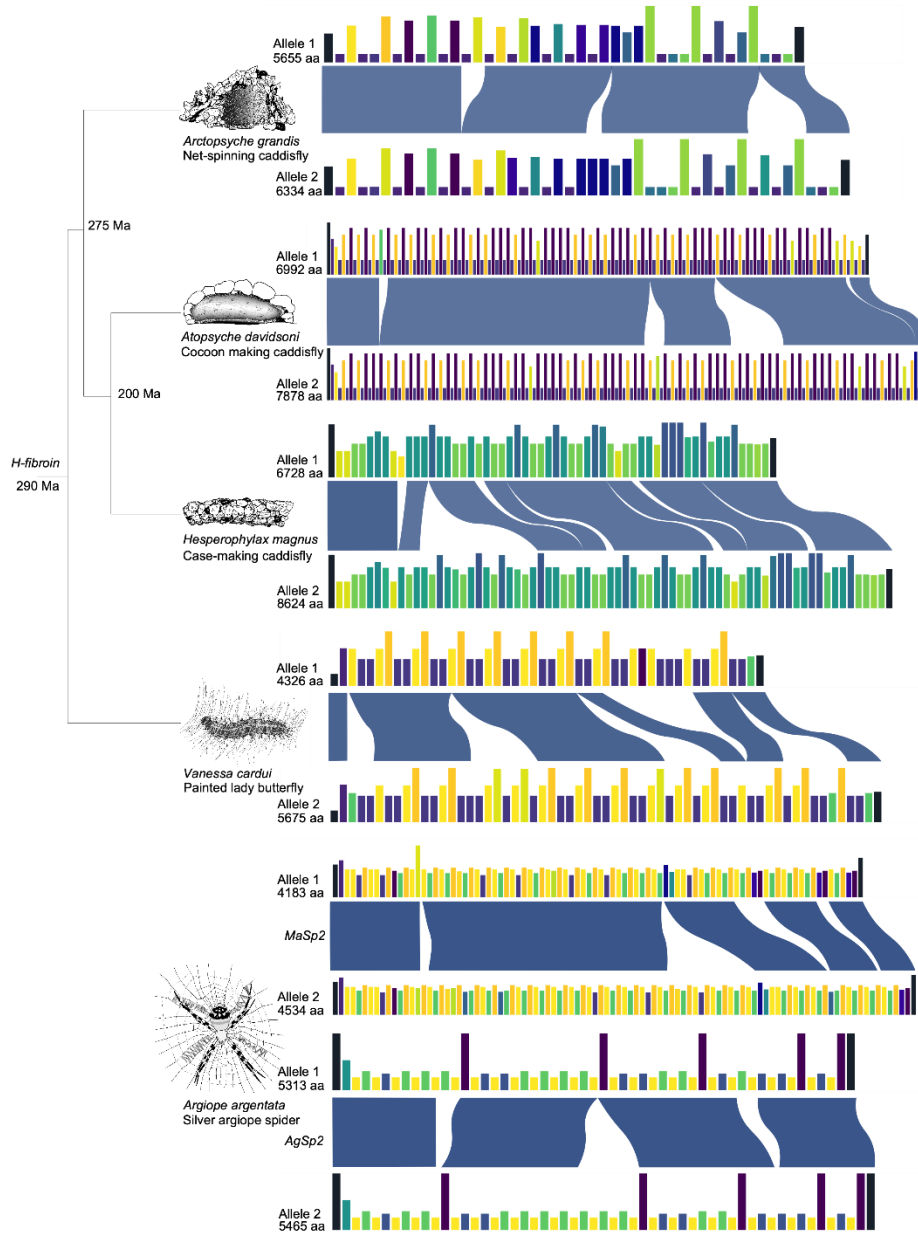regions varied widely across species and orders.

Figure 1 legend. Allelic variation in silk protein sequences of caddisflies, a butterfly, and a spider. Illustrations depict variation in silk use across these distantly related organisms. Bars of the same color represent the same repetitive motif within an individual, but not among individuals, and blue ribbons represent aligned regions highlighting insertions and deletions between alleles. The height of the bars indicates proportional unit length of each repeat.

While the level of variation that we observed between alleles of *H-fibroin* within individual butterflies and caddisflies was striking, it is perhaps unsurprising that a functionally important gene with a common evolutionary origin would share similar patterns of variation. To determine if our findings extended to silk with a different evolutionary origin, we analyzed two spidroin silk genes from the spider *Argiope argentata*: *major ampullate spidroin-2* (*MaSp2*) and *aggregate spidroin-2* (*AgSp2*). For both genes, we observed patterns of allelic variation that were strikingly

3

convergent with *H-fibroin*, including modular heterozygosity between alleles (Fig. 1). Indeed, across all samples, "complete repeat indels," defined as indels that encompassed one or more repeat unit(s), accounted for the majority of the variation between alleles (Table 1).

Table 1. Comparison of genome and gene sequences from organisms used in this study. BUSCO scores reflect the insecta_odb10 dataset for the Trichoptera and Lepidoptera and the arachnida_odb10 dataset for the spider. Percent of complete repeat indels (%CRI) indicates the percentage of total allelic variation that results from complete repeat indels between alleles.

Order	Species	Contig N50 (Mbp)	BUSCO % complete	Gene	Number of amino acids allele 1	Number of amino acids allele 2	% complete repeat indels (CRI)	Study
Trichoptera	*Arctopsyche grandis*	9.4	98.9	*H-fibroin*	6375	5696	97.6	Present study
Trichoptera	*Atopsyche davidsoni*	14.1	98.8	*H-fibroin*	7878	6992	94.1	Ríos-Touma et al. 2021
Trichoptera	*Hesperophylax magnus*	11.2	95.6	*H-fibroin*	8624	6728	91.7	Hotaling et al. 2022
Lepidoptera	*Vanessa cardui*	7.0	98.8	*H-fibroin*	5675	4326	95.3	Lohse et al. 2021
Araneae	*Argiope argentata*	32.3	98.4	*MaSp2*	4534	4183	94.7	Present study
Araneae	*Argiope argentata*	32.3	98.4	*AgSp2*	5465	5313	69.1	Present study

In this study, we uncovered remarkable convergence in the structure and variation between alleles of silk protein-coding genes across deeply divergent arthropod orders, including two independent origins of silk. Historically, this allelic variation has been overlooked because short-read sequences are not well-suited to resolving long, highly repetitive regions. It is clear from these comparisons that common methods for assessing genomic variation, such as the analysis of single nucleotide variants, cannot fully capture the extent of variation in these repetitive regions. The patterns we observed lend insight into shared mechanisms driving silk gene evolution across two divergent groups of arthropods (Fig. 1). For example, the allelic variation of large indels is consistent with unequal crossing over which could play a role in driving allelic variation. As such, selection may tolerate the presence of diverse alleles that differ in the organization of repeat units while maintaining nearly identical sequence within each repeat. Furthermore, since sequence length is linked to the properties of silk fibers (12), variation among alleles may be critical to the function of silk fibers and represent the product of selection. However, fully uncovering the evolutionary mechanisms underlying variation in silk genes will require population-level sampling. Ultimately, to effectively connect genotype to phenotype, we must account for the full suite of allelic variation that exists within structural silk protein-coding genes. Our findings echo a broader trend; advances in long-read sequencing are leading to the

discovery of substantial, and previously unobserved, genomic variation across the tree of life (14).

Materials and Methods:
We extracted and sequenced genomic DNA from four of the five species, two of which were available previously, and retrieved the other species from publicly available data (Table 1). For each species, we generated *de novo* assemblies using Hifiasm v.0.13-r307 (15) and used a custom pipeline to retrieve, annotate, and align the *H-fibroin* and spidroin sequences (Supporting Information). We then used a custom script to visualize and compare the alleles of fibroin and spidroin genes within each species.

Data availability:
Data is available in the following GenBank accessions: XXXX-XXXX. The script used to visualize the repeats is available at https://doi.org/10.5281/zenodo.7783469.

References:

1. T. D. Sutherland, J. H. Young, S. Weisman, C. Y. Hayashi, D. J. Merritt, Insect Silk: One Name, Many Materials. *Annual Review of Entomology* **55**, 171–188 (2010).

2. S. Nurk, *et al.*, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

3. S. Hotaling, J. L. Kelley, P. B. Frandsen, Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences* **118**, e2109019118 (2021).

4. R. A. Marks, S. Hotaling, P. B. Frandsen, R. VanBuren, Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* **7**, 1571–1578 (2021).

5. J. Wagner, *et al.*, Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**, 672–680 (2022).

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

7. C.-Z. Zhou, *et al.*, Fine organization of Bombyx mori fibroin heavy chain gene. *Nucleic Acids Research* **28**, 2413–2419 (2000).

8. N. A. Ayoub, J. E. Garb, R. M. Tinghitella, M. A. Collin, C. Y. Hayashi, Blueprint for a High-Performance Biomaterial: Full-Length Spider Dragline Silk Genes. *PLOS ONE* **2**, e514 (2007).

9. N. Yonemura, K. Mita, T. Tamura, F. Sehnal, Conservation of Silk Genes in Trichoptera and Lepidoptera. *J Mol Evol* **68**, 641–653 (2009).

10. P. B. Frandsen, *et al.*, Exploring the underwater silken architectures of caddisworms: comparative silkomics across two caddisfly suborders. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20190206 (2019).

11. A. Y. Kawahara, *et al.*, Long-read HiFi sequencing correctly assembles repetitive heavy fibroin silk genes in new moth and caddisfly genomes. *Gigabyte* **2022**, 1–14 (2022).

12. K. Arakawa, *et al.*, 1000 spider silkomes: Linking sequences to silk physical properties. *Science Advances* **8**, eabo6043 (2022).

13. R. H. Baker, A. Corvelo, C. Y. Hayashi, Rapid molecular diversification and homogenization of clustered major ampullate silk genes in Argiope garden spiders. *PLOS Genetics* **18**, e1010537 (2022).

14. R. E. Mukamel, *et al.*, Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).

15. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).

Supporting Information for "Allelic resolution of insect and spider silk genes reveals hidden genetic diversity"

We sequenced and assembled four of the five species represented here, two of which were available previously (Table 1). The other, *Vanessa cardui*, was publicly available (1). For each species, we extracted DNA from head and thorax (silk glands in the case of *Argiope argentata*) of a single individual using a Qiagen Genomic-tip extraction kit. After visualization of high molecular weight DNA using pulse-field gel electrophoresis, we sheared the DNA to 15 kbp fragments with the Diagenode Megaruptor. We generated PacBio HiFi libraries for each DNA extraction using the SMRTbell Express Template Prep Kit 2.0. Depending on genome size, each library was run on 30-hour 8M PacBio SMRT cells on the Sequel II instrument at the Brigham Young University DNA Sequencing Center.

We computed HiFi reads from the raw data (reads with quality above Q20) using PacBio SMRTlink software. For *Arctopsyche grandis*, we used DeepConsensus v.1.0.0 (2) to generate additional HiFi reads from subreads. We then assembled the raw reads into assemblies using HiCanu and Hifiasm v.0.13-r307 (3, 4). To extract and analyze the *H-fibroin* sequences, we used BLAST (5) to search the assembled genomes using the N-terminus and C-terminus sequences from previously published data. We then verified that both BLAST hits were isolated to the same contig in the genome assembly. We then extracted the region encompassed between the BLAST hits along with 20 kbp of flanking sequence on each end. We generated a gene annotation of the sequence using Augustus v.3.3.3 (6). The first species that we sequenced was *Hesperophylax magnus*. To generate assemblies for both *H-fibroin* alleles, we mapped HiFi reads back to the gene sequence from the primary assembly. We then separated reads into two bins by counting the number of mismatches in each read. We then assembled each bin of reads separately using HiCanu v.1.9 (4). Both assemblies resulted in full-length fully resolved *H-fibroin* sequences. We then verified that the allelic sequences matched those in the primary and alternate assemblies generated by Hifiasm (3). Both assemblers yielded identical sequences. For all other species, we simply extracted the alleles from the primary and alternate assemblies generated by Hifiasm (3) as described above.

We then aligned the amino acid sequences of the genes with the MAFFT-linsi algorithm to identify regions of shared alignment (7). We analyzed each alignment and classified the variation between alleles into three classes: (1) single site substitutions, (2) small indels, and (3) complete repeat indels (CRI), defined as those indels that encompassed one or more complete repeat units. To visualize the differences in repeat structure among species we generated a custom visualization script that splits the gene into repetitive motifs and then generates a bar plot, in which each unique motif is assigned a unique color (Fig. 1). To compare the nucleotide identity across motif classes, we wrote an additional script to extract the corresponding nucleotide sequences and then aligned them with MAFFT (7).

References:
1.   K. Lohse, *et al.*, The genome sequence of the painted lady, *Vanessa cardui* Linnaeus 1758 (2021) https:/doi.org/10.12688/wellcomeopenres.17358.1 (November 19, 2022).

2. G. Baid, *et al.*, DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2023).

3. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

4. S. Nurk, *et al.*, HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.*, gr.263566.120 (2020).

5. S. Altschul, Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403–410 (1990).

6. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).

7. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).