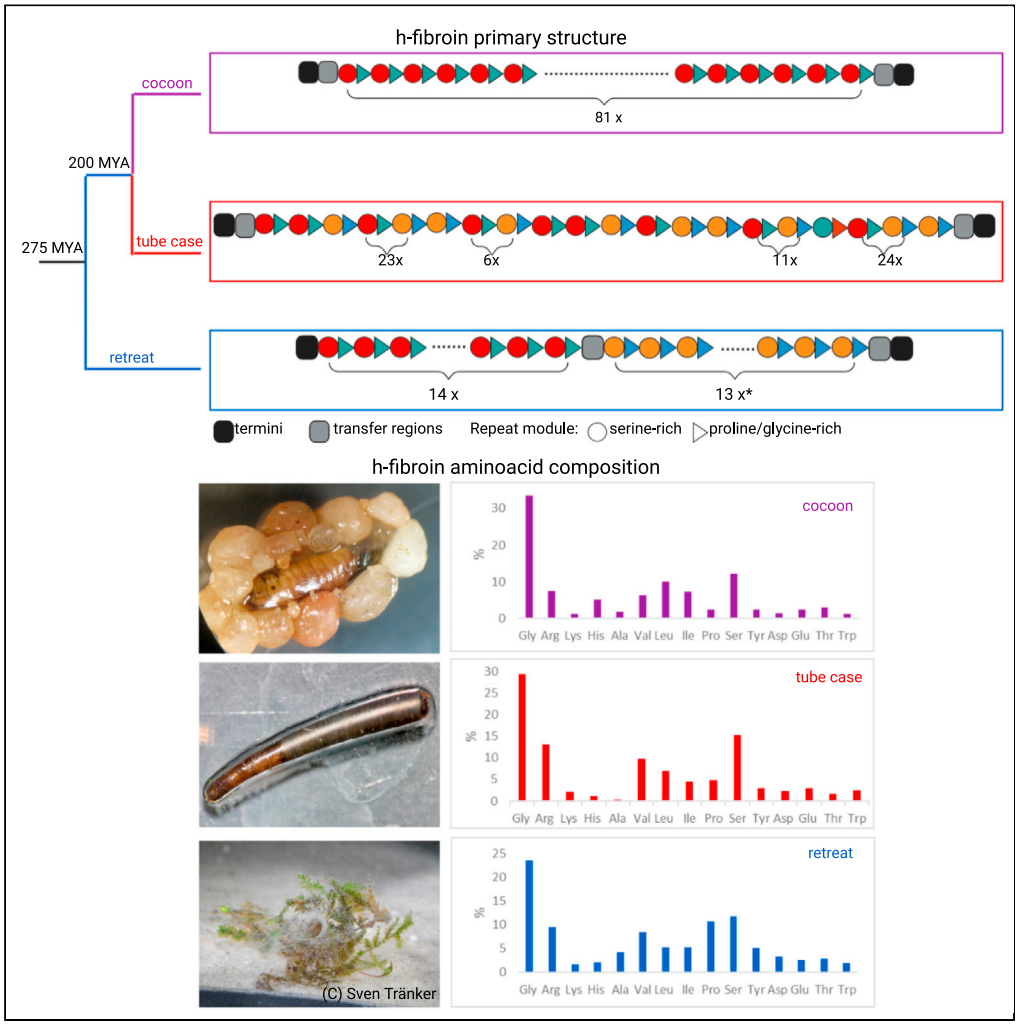
iScience

CellPress
OPEN ACCESS

Article

Characterization of the primary structure of the major silk gene, *h-fibroin*, across caddisfly (Trichoptera) suborders



Jacqueline Heckenhauer, Russell J. Stewart, Blanca Ríos-Touma, Ashlyn Powell, Tshering Dorji, Paul B. Frandsen, Steffen U. Pauls

jacqueline.heckenhauer@senckenberg.de

Highlights

New sequences of the major caddisfly silk protein, h-fibroin, are reconstructed

Repetitive modules vary widely across clades with different silk use

H-fibroin of retreat/capture net makers have higher proportion of proline

Amino acid differences may be linked to mechanical properties of silk

Heckenhauer et al., iScience 26, 107253
August 18, 2023
https://doi.org/10.1016/j.isci.2023.107253

iScience

CellPress
OPEN ACCESS

Article

Characterization of the primary structure of the major silk gene, *h-fibroin*, across caddisfly (Trichoptera) suborders

Jacqueline Heckenhauer,[1,2,9,*] Russell J. Stewart,[3] Blanca Ríos-Touma,[4] Ashlyn Powell,[5] Tshering Dorji,[6] Paul B. Frandsen,[1,5,7] and Steffen U. Pauls[1,2,8]

SUMMARY

Larvae of caddisflies (Trichoptera) produce silk to build various underwater structures allowing them to exploit a wide range of aquatic environments. The silk adheres to various substrates underwater and has high tensile strength, extensibility, and toughness and is of interest as a model for biomimetic adhesives. As a step toward understanding how the properties of underwater silk evolved in Trichoptera, we used genomic data to identify full-length sequences and characterize the primary structure of the major silk protein, h-fibroin, across the order. The h-fibroins have conserved termini and basic motif structure with high variation in repeating modules and variation in the percentage of amino acids, mainly proline. This finding might be linked to differences in mechanical properties related to the different silk usage and sets a starting point for future studies to screen and correlate amino acid motifs and other sequence features with quantifiable silk properties.

INTRODUCTION

Insects use silk for a variety of purposes[1,2] and research has focused on terrestrial Lepidoptera (moths and butterflies), especially on the commercially important silkworm *Bombyx mori*.[3–11] Less well studied are aquatic insects that use silk. These include the most speciose primary aquatic insect order, Trichoptera (caddisflies), which exhibit diverse silk usage strategies.[12] Similar to their sister order Lepidoptera, several caddisfly species construct cocoons for metamorphosis in the final larval instar.[13] In earlier larval stages, they produce a diverse array of underwater structures, which is reflected in their phylogeny.[13,14] Trichoptera is divided into two suborders, which are distinguished by differences in morphology, habitat, and use of silk[14]: Integripalpia (cocoon and tube case makers) and Annulipalpia (fixed retreat makers). Further differentiation in silk use also occurs at the subordinal and superfamily level (reviewed in[13]). Basal Integripalpia, which show diverse case-making behaviors (free-living, tortoise case-making, purse case-making), are referred to as cocoon makers because they pupate in a silken pupal cocoon,[15] whereas larvae of tube case-making Integripalpia build portable, tubular cases made purely from silk or from diverse materials encountered in their habitats, such as small stones or plant materials that are "taped" together with silk. The cases provide protection and channel oxygenated water past the body allowing them to obtain oxygen in lentic environments.[15] They are of various shapes and materials and often species-specific, i.e., distinctive among congeneric species (i.e., *Micrasema longulum*: silk case, *Micrasema setiferum*: sand-stone material, *Micrasema wataga*: plant material). Larvae of Annulipalpia create stationary shelters of silk, often with mineral particles or plant material, which are fixed to the substrate (e.g., stones or aquatic plants). These retreats serve primarily as physical protection against predation. In addition, members of Annulipalpia construct various kinds of nets to filter suspended organic particles from flowing water or to catch small invertebrates.[13]

Caddisfly silk is a tough adhesive multi-network fiber.[16,17] Studies on the mechanical properties of single silk fibers revealed that they are viscoelastic, have an extensibility of >100%,[18] display large strain cycle hysteresis, and self-recover 99% of their initial stiffness and strength.[16] Caddisfly silk is of practical interest as a model for the development of biomimetic materials with applications in aqueous environments,[17,19] such as the human body (e.g., in medicine for tissue engineering, as tough hydrogels,[20] surgical structures or bio-bandages[21]).

[1]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Hesse 60325, Germany

[2]Department of Terrestrial Zoology, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt, Hesse 60325, Germany

[3]Department of Biomedical Engineering, University of Utah, Salt Lake City, UT 84112, USA

[4]Facultad de Ingenierías y Ciencias Aplicadas, Ingeniería Ambiental, Grupo de Investigación en Biodiversidad, Medio Ambiente y Salud (BIOMAS), Universidad de Las Américas, Quito, EC 170124, Ecuador

[5]Department of Plant and Wildlife Science, Brigham Young University, Provo, UT 84602, USA

[6]Department of Environment and Climate Studies, Royal University of Bhutan, Punakha 13001, Bhutan

[7]Data Science Lab, Smithsonian Institution, Washington, DC 20560, USA

[8]Institute for Insect Biotechnology, Justus-Liebig-University, Gießen, Hesse 35392; Germany

[9]Lead contact

*Correspondence: jacqueline.heckenhauer@senckenberg.de

https://doi.org/10.1016/j.isci.2023.107253

Their diverse net- and case-making behavior allow caddisflies to exploit a range of ecological niches, which raises the question of how the properties of underwater silk evolved in Trichoptera. To begin to answer this question, and to gauge the potential of caddisfly silk in material sciences, a comprehensive study of the primary molecular structure of the major silk protein, h-fibroin, is necessary.

The silk fiber in Trichoptera consists of two filaments derived from a pair of labial glands (see also[22]). As in most Lepidoptera, the fiber core is assembled from a large (200–500 kDa) heavy-chain fibroin (h-fibroin) and the smaller (~25 kDa) light-chain fibroin (l-fibroin) protein.[23] In contrast to the silk of the silkworm *B. mori*, a homolog of glycoprotein P25 (fibrohexamerin) has not been detected in Trichoptera[23] and, instead of being surrounded by a sericin layer, the central fiber core is covered with a thinner, poorly characterized peripheral layer.[24] Additional silk fiber components were identified by transcriptome and proteome analysis of the silk gland (e.g., peroxinectin, a novel structural component with sequence similarity to the elastic PEVK region of the muscle protein titin, and mucins).[25,26] The h-fibroin is the major silk protein by size and mass. On a macroscale level, the structural organization of the trichopteran h-fibroin is similar to Lepidoptera.[23] It consists of non-repetitive amino(n)- and carboxyl(c)-terminal domains flanking a central region, composed of repeated structural modules.[17,27] However, the central regions exhibit no sequence conservation between orders.[28] In Trichoptera, these consist of repeating $(SX)_nE$ motifs in which the S (serine) is often phosphorylated[19,29,30] and where X is primarily an amino acid with hydrophobic or aromatic side chains and sometimes arginine, E is glutamic acid, and n is 3–5. The $(SX)_n$ E motifs are separated by glycine-rich regions of variable length.[1,19,27] The identification of *h-fibroin* gene sequences has been difficult because of their length (>20 kilobase pairs [kbp]) and their highly repetitive regions.[31–33] However, partial *h-fibroin* sequences of six species of Trichoptera were derived from sequencing the ends of cDNAs[17,23,26,27,34] and the combination of long- and short-read sequencing approaches resulted in the assembly of two full-length *h-fibroin* sequences.[31,35] However, the *h-fibroin* could not be assembled in more than 20 species with these sequencing techniques. The lack of high-quality, full-length *h-fibroin* sequences has hindered the characterization and comparison of the primary molecular structure across trichopteran clades with different silk usage. Sequencing and assembling the entire repetitive central region of the h-fibroin is a crucial step toward understanding how phenotypes are encoded genetically,[32] because the repetitive regions are responsible for the strength and elasticity properties of silk fibers. Recently, new genomic long-read sequencing techniques with low sequencing error rates (e.g., PacBio HiFi) and new genome assembly tools (e.g., hifiasm[36]) allowed for full-length assembly of the *h-fibroin* gene sequences of four Trichoptera species (two case makers[32,33]; one retreat and one cocoon maker[37]).

In this study, we increased the number of high-quality full-length Trichoptera *h-fibroin* sequences from four to eleven. Specifically, we identified complete protein-coding *h-fibroin* gene sequences from seven high-quality genomes, including two newly assembled genomes and five publicly available genome assemblies. To increase the number of h-fibroin sequences of retreat and cocoon makers, we generated two *de novo* genomes and identified the h-fibroin sequences in these as well. Our final taxon sampling covers all three major silk usage strategies (six case makers, three retreat makers and two cocoon makers). We characterized and compared the primary structure of these h-fibroins and compared their amino acid composition to investigate differences between species with different silk usage. To examine differences between terrestrial and aquatic silk usage, we compared the amino acid composition of the h-fibroin to that of various terrestrial Lepidoptera.

RESULTS

Expansion of genomic resources and heavy-chain fibroin sequences

Aquatic insects have been neglected concerning genome sequencing efforts[38] and the lack of well-resolved genome assemblies has hindered progress in understanding the genomic basis of aquatic insect traits, such as silk. Here, we used PacBio HiFi sequencing to generated genome assemblies for two species of Trichoptera: *Leptonema lineaticorne* (retreat-/capture net maker, 525,771 ccs-sequences with a total of 6,127,219,457 bp, ~24.5× sequencing coverage) and *Himalopsyche tibetana* (cocoon maker, 1,391,005 ccs-sequences with a total of 14,730,644,101 bp, ~23.3 × sequencing coverage). Genome size estimates derived from different methods were consistent. For *L. lineaticorne*, Genomescope2 revealed a genome size of 249,938,531 bp (http://qb.cshl.edu/genomescope/genomescope2.0/analysis.php?code=Tn1oE0FzenfoB8gl5Y0L). The back-mapping approach revealed a genome size of 265.85Mb (Figure S1). *For H. tibetana*, *K-mer* analysis estimated the genome size to

Table 1. Comparison of currently published high-quality long-read-based genome assemblies of caddisflies

Species	Suborder (silk usage)	Accession number	Assembly length (bp)	Contig N50 (contig/scaffold)	No. of contigs/ scaffolds)	% BUSCO (n = 2124)
Atopsyche davidsoni[40]	Basal Intergripalpia (cocoon)	GCA_022113835.1	370,818,532	14,095,054/n.a.	80/n.a	C:96.9%[S:96.1%,D:0.8%], F:1.8%,M:1.3%
Himalopsyche tibetana[a]	Basal Intergripalpia (cocoon)	JAPJYX000000000	691,323,649	28,889,006/n.a.	282/n.a.	C:96.5%[S:95.6%,D:0.9%], F:2.4%,M:1.1%
Eubasilissa regina[32]	Integripalpia (tube case)	GCA_022840565.1	917,621,729	32,427,664/n.a.	123/n.a.	C:95.5%[S:94.8%,D:0.7%], F:3.0%,M:1.5%
Glyphotaelius pellucidus (DtoL)[41]	Integripalpia (tube case)	GCA_936435175.1	1,037,123,706	8,185,058/36,814,344	285/57	C:90.3%[S:89.5%,D:0.8%], F:6.8%,M:2.9%
Hesperophylax magnus[33]	Integripalpia (tube case)	GCA_026573805.1	1,215,205,050	11,205,906/n.a.	980/n.a.	C:91.4%[S:89.0%,D:2.4%], F:6.0%,M:2.6%
Limnephilus lunatus (DtoL)[42]	Integripalpia (tube case)	GCA_917563855.2	1,269,651,477	18,993,099/95,392,806	139/39	C:89.7%[S:88.9%,D:0.8%], F:7.4%,M:2.9%
Limnephilus marmoratus (DtoL)[43]	Integripalpia (tube case)	GCA_917880885.1	1,629,971,709	8,018,677/56,174,236	395/68	C:90.4%[S:89.3%,D:1.1%], F:6.7%,M:2.9%
Limnephilus rhombicus (DtoL)	Integripalpia (tube case)	GCA_929108145.2	1,578,808,083	10,796,652/54,234,467	272/62	C:89.8%[S:88.6%,D:1.2%], F:7.1%,M:3.1%
Leptonema lineaticorne[a]	Annulipalpia (retreat)	GCA_024500535.1	273,010,349	13,827,090/n.a.	65/n.a.	C:96.1%[S:95.3%,D:0.8%], F:2.3%,M:1.6%
Cheumatopsyche charites[44]	Annulipalpia (retreat)	GCA_024721215.1	223,232,897	2,851,765/13,966,006	207/68	C:96.4%[S:95.9%,D:0.5%], F:1.9%,M:1.7%
Arctopsyche grandis[37]	Annulipalpia (retreat)	GCA_029955255.1	485,663,687	6,470,670/n.a.	676/n.a.	C:97.3%[S:93.5%,D:3.8%], F:1.6%,M:1.1%

C, complete; S, single; D, duplicated; F, fragmented; M, Missing.
[a]This study. BUSCO: % of complete BUSCOs is given based on BUSCO 5.2.2[45] using the endopterygota_odb10 dataset.

631,263,655 bp (http://qb.cshl.edu/genomescope/genomescope2.0/analysis.php?code=c2kegs2IOSe5C6OQiSpO). Back-mapping approach revealed a genome size of 683.7 Mb (Figure S3). Our *de novo* assemblies of *L. lineaticorne* and *H. tibetana* rank among the highest quality assemblies for Trichoptera for gene completeness (i.e., more than 96% complete BUSCOs, assembly lengths congruent with estimated genome size, Table 1) and contiguity (e.g., contig N50: ~29 Mbp in *H. tibetana*; number of contigs: 65 in *L. lineaticorne*, Table 1). Re-mapping of the raw reads to the assembly revealed that 98.2% (*H. tibetana*) and 99.94% (*L. lineaticorne*) could be unambiguously placed with expected coverage distribution per position (Figures S1 and S3). BlobTools[39] detected no contaminations (Figures S2 and S4). We identified *h-fibroin* genes in the new *L. lineaticorne* and *H. tibetana* assemblies and in five previously published genome assemblies (Tables 2 and S1).

Characterization of the primary structure of h-fibroin

Using the newly identified as well as the four previously published full-length, high-quality *h-fibroin* sequences,[32,33,37] our final taxon sampling for comparing the primary structure of h-fibroins covers the main clades of Trichoptera with different silk usage: fixed retreat-making Annulipalpia (n = 3), cocoon-making basal Integripalpia (n = 2), and tube case-making Integripalpia (n = 6). The structure of the *h-fibroin* gene across the 11 species revealed a similar organization of introns and exons. The *h-fibroin* gene was characterized by a short exon (36–48 bp) followed by a single intron (61–1,656 bp) and a long second exon (12,709–29,502 bp) leading to a total length of 18,745–30,382 bp.

The h-fibroin protein consists of non-repetitive n- and c-termini that were highly conserved across the different clades of caddisflies as well their sister order, Lepidoptera (*B. mori*). The n-terminus contained 105–117 residues (without the signal peptide, Figure 1A). There were 42.5% identical sites (% of columns in the alignment where all sequences are identical) and 74.3% pairwise identity (% of pairwise residues

Table 2. Full-length *h-fibroins* of 11 caddisfly species derived from long-read sequencing ordered by silk usage

Species	GenBank Accession	Silk usage	Gene length	CDS	Exon 1	Intron 1	Exon 2	Protein size	Molecular weight (kD)
Atopsyche davidsoni[37]	OQ787677	cocoon	23,701	23,637	67	61	23,570	7,878	755.16
Himalospyche tibetana[a]	OQ983471	cocoon	20,892	19,233	36	1,656	19,197	6,411	628.28
Eubasilissa regina[32]	n.a., see[32]	case	25,256	25,161	40	92	25,121	8,386	815.86
Glyphotaelius pellucidus[a]	BK063450	case	24,412	23,592	39	817	23,553	7,864	801.63
Hesperophylax magnus[33]	OQ787679	case	26,666	25,872	42	791	25,830	8,624	875.02
Limnephilus lunatus[a]	BK063451	case	30,382	29,544	42	835	29,502	9,848	998.81
Limnephilus marmoratus[a]	BK063452	case	27,154	26,322	42	829	26,280	8,774	882.45
Limnephilus rhombicus[a]	BK063453	case	27,903	27,081	39	819	27,042	9,027	896.38
Leptonema lineaticorne[a]	OQ983470	retreat	20,538	20,451	42	84	20,409	6,817	675.22
Cheumatopsyche charites[a]	BK063449	retreat	18,745	18,669	42	73	18,627	6,223	633.908
Arctopsyche grandsis[37]	OQ787675	retreat	19,231	19,122	48	106	19,074	6,374	659.35

Gene length including introns and stop codon; CDS = protein coding DNA, gene length, CDS, exon, intron in bp; protein size in amino acids.
[a]This study.

that are identical in the alignment, including gap versus non-gap residues, but excluding gap versus gap residues) among the caddisfly species. Alignment of the n-terminus of Lepidoptera *B. mori* to the n-termini of Trichoptera leads to a decrease in pairwise identity (65.4%) and identical sites (11.7%). The c-terminus consisted of 40 residues with 32.5% (10% when including *B. mori*) identical sites, and 65.3% (57.4% with *B. mori*) pairwise identity. A conserved cysteine was detected at position 19 in the c-terminus alignment (Figure 1B).

The terminal domains flanked a central region, composed entirely of repeating sequence blocks, which have been represented in several ways in the literature to describe the primary structure of h-fibroin.[27,30,46] For example, Frandsen et al.,[31] represented each unique $(SX)_nE$ motif as the beginning of a repeat. Structurally, this would correspond to defining each repeat as beginning with a single $(SX)_nE$ ß-strand.[17] In presenting the new h-fibroin sequences reported here, we have defined the repeating modules as each having two parts: first, a region comprising a variable number (1–7) of $(SX)_nE$ motifs, each separated by short (8–24) stretches of intervening amino acids, and, second, a variable length (8–144 residues) G(glycine) - or G(glycine)-P(proline)-rich region. Schematically, in Figure 2, the two parts are represented by different symbols in each block. In the following paragraphs, we describe the repeating structural modules for each species analyzed in this study in detail. The full-length h-fibroin protein sequences are provided in Notes S1–S11. Schematic visualizations of each genus are presented in Figures S5–S13.

Simple repeating structural modules in basal Integripalpia (cocoon makers)

The structure of the repetitive central region in cocoon-building *A. davidsoni* was simple compared to the other sequences investigated in this study (Figure 2, cocoon). The repeating module was embedded between two transition regions (TR), each of which occurs once and directly flanks the termini. In these transition regions, the sequence transitions from one type of module to another. The TRs had similarities to the repeating module but their sequence was unique, e.g., they occurred only one time in the h-fibroin and, in most cases, form a transition between the conserved termini and the repeating structural modules (Figure 2: cocoon, tube case) or between repeat modules (Figure 2: retreat). The h-fibroin of *A. davidsoni* included a single repeating module consisting of a $(SX)_4E[15](SX)_3E$ region and a G-rich region of variable length (40–70 residues). The number in the square bracket refers to the short stretches of intervening amino acids. This module repeats 81 times across the sequence.

The repetitive region of *H. tibetana* was embedded between two transition regions (Figure S6). There were two repeating modules that are very similar. RM1 consists of a $(SX)_6E[16](SX)_4$**E**$[15](SX)_6E[58](SX)_5E$ or $(SX)_6E[16](SX)_4$**D**$[15](SX)_6E[58](SX)_5E$ (D: aspartic acid) motif and a glycine-rich motif of variable length (93–106 residues). RM2 was reduced to $(SX)_6E[16](SX)_4E[15](SX)_6E$. The glycine-rich motif was 56–90 residues long. RM1 occurred 26 and RM2 occurred 13 times.

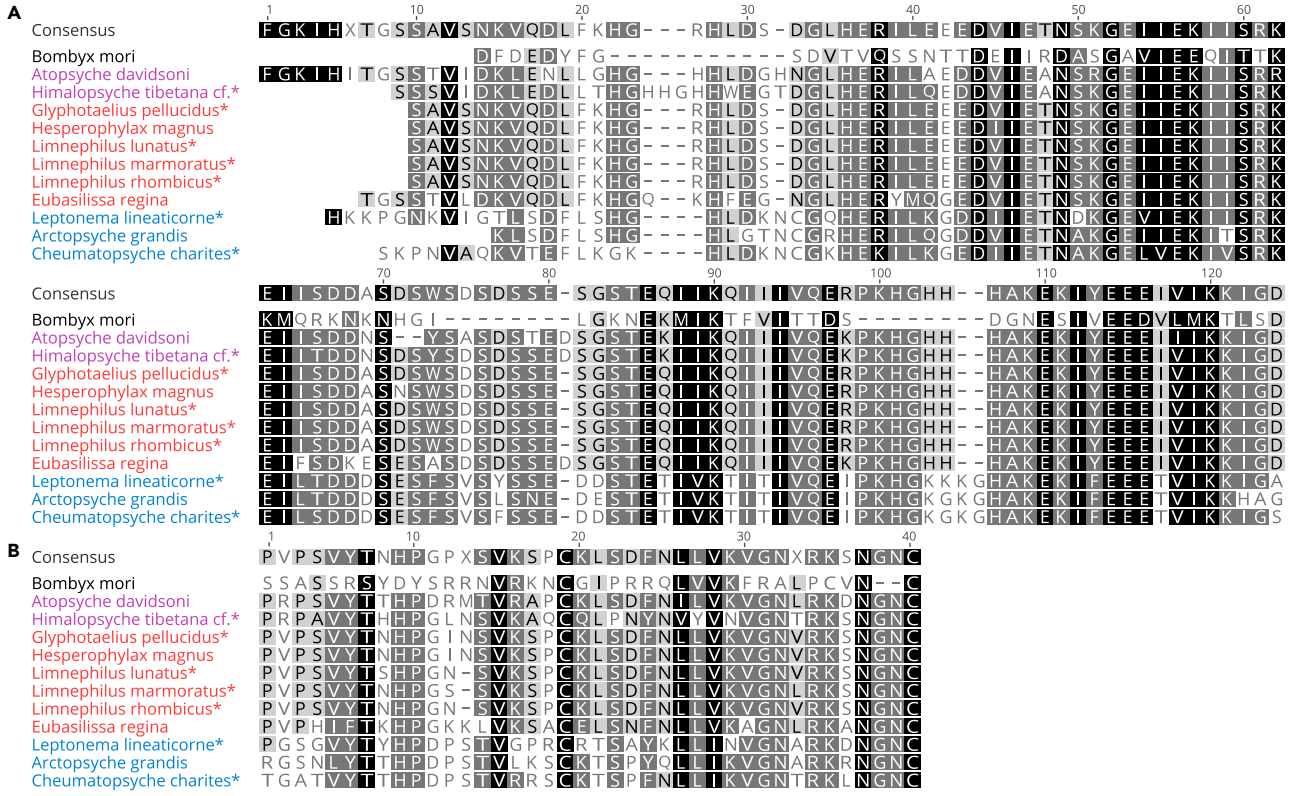


Figure 1. Highly conserved regions of the h-fibroin

(A) n-terminus without signal peptide.

(B) c-terminus. Different silk usage is color-coded: violet: cocoon-, red: tube case-, blue: retreat-making, black: *Bombyx mori* (Lepidoptera, outgroup).

Increasing structural complexity of repeating modules in Integripalpia (tube case makers) and Annulipalpia (fixed retreat makers)

The primary structure of the h-fibroin was variable within tube case-making basal Integripalpia. In all seven species sampled, the repetitive central region was flanked by two transition regions. However, the number and organization of repeat modules were diverse within this clade.

Specifically, in *Eubasilissa regina* which uses plant material (leaves) and silk to build a tube case, the h-fibroin had only one type of repeating module consisting of a $(SX)_4E[8](SX)_4E[17](SX)_4E[11](SX)_3E[20](SX)_3E$ motif and a glycine-rich region (14–144 residues). It was repeated 39 times (Figure S7).

In *Glyphotaelius pellucidula* whose silk usage is similar to *E. regina*, the h-fibroin consisted of two repeat modules (Figure S8). RM1 consisted of a $(SX)_4E[13](SX)_3E$ motif and glycine-rich region (38–81 residues), and RM2 consisted of a single $(SX)_5E$ motif and glycine-rich region (12–51 residues). Each RM was repeated 70.

The h-fibroin of *H. magnus* (Figure S9), which uses stones the build tube cases, consisted of two repeat modules. RM1 contains a $(SX)_5E[14](SX)_4E[11](SX)_4E$ motif and a glycine-rich region (41–81 residues) and occurred 69 times and was interrupted by 13 RM2, which consisted of a single $(SX)_5E$ motif and a glycine-rich region (45 residues).

Within the three *Limnephilus* species, the primary structure of the h-fibroin was comparable (Notes S6–S8, Figure S10). This genus uses a diverse array of plant materials (wood, moss, and leaves) for tube case-making. There were two repeat modules. In L. *lunatus*, RM1 consisted of a $(SX)_5E[14](SX)_4E[11](SX)_4E$ region and a glycine-rich region (14–82 residues). It occurred 96 times and was interrupted by ten RM2 which consisted of a single $(SX)_5E$ motif and a glycine-rich region (14 residues), similar to *G. pellucidula* (Figure 2). In *L. marmoratus*, RM1 occurred 68 and RM2 17 times. In *L. rhombicus* RM1 occurred 59 and RM2 51 times.

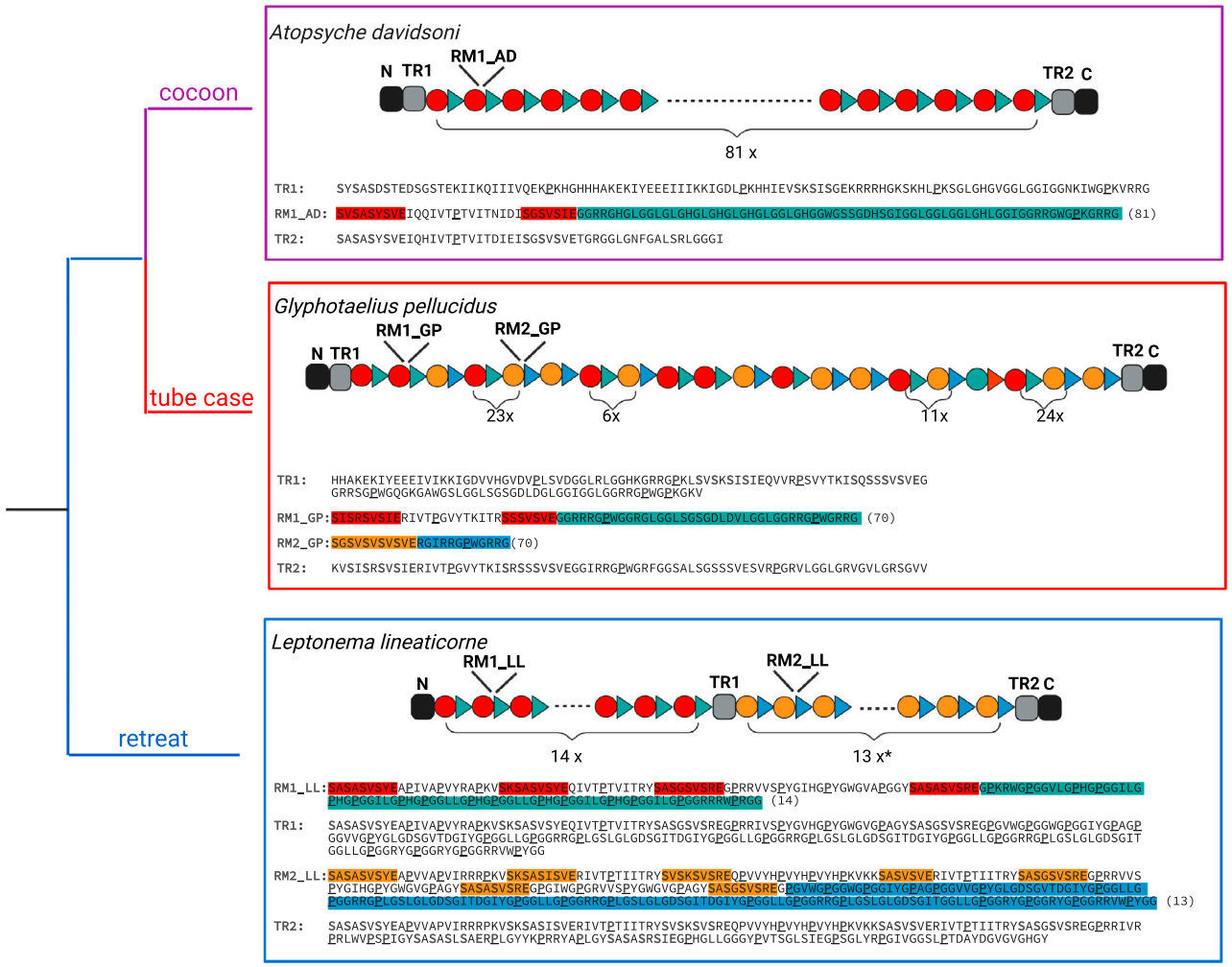


Figure 2. Schematic visualization of the primary structure of the h-fibroin gene of one representative species per clade

Black boxes: n- and c-terminus, for sequences, see Figure 1. Gray boxes: transition regions (TR). The repeating modules (RM) have two parts which are represented by different symbols: A region comprising a variable number (1–7) of $(SX)_nE$ motifs (red and orange circles correspond to red/orange residues in the sequence), each separated by short (8–24) stretches of intervening amino acids, and a variable length (8–144) G(glycine) - or G(glycine)-P(proline)-rich region (blue arrows correspond to blue residues in the sequence). Cocoon maker: *Atopsyche davidsoni* (AD): RM1_AD: Repeat module 1 occurs 81 times. Tube case maker: *Glyphothaelius pellicudulus* (GP): RM1_GP: Repeat module 1 and RM2_GP: Repeat module 2 occur 70 times each. Retreat maker: *Leptonema lineaticorne* (LL): RM1_LL: Repeat module 1 is repeated 14 times and RM2_LL: Repeat module 2 is repeated 13 times. The consensus sequence of each repeat module is given. The Glycine/Glycine-Proline-rich motifs vary in length. For full-length sequence see Notes S1–S11. Similar figures for each genus are given in Figures S5–S13. Phylogeny shown after.[14] Figure created with BioRender.com.

In Annulipalpia (fixed retreat makers), in *L. lineaticorne*, the sequence was divided into two parts each with a distinct repeat module (Figure 2). RM1 consisted of a $(SX)_4E[13](SX)_4E[11](SX)_4E[24](SX)_4E$ motif and glycine-rich region (68–92 residues) and is repeated 14 times. RM2 consisted of a $(SX)_4E[14](SX)_4E[11](SX)_4E[19](SX)_3E[11](SX)_4E[24](SX)_4E[23](SX)_4E$ motif and a glycine-rich region (75–84 residues) and was repeated 13 times. There were two transition regions. TR1 in *L. lineaticorne* separated the two repeat modules and TR2 was located between RM2_LL and the c-terminus.

In *Cheumatopsyche charites*, the repetitive region was surrounded by two transition regions (Figure S13). Similar to *L. lineaticorne*, the h-fibroin was divided into two parts. The first part of the gene consisted of six repeat modules (RM1: $SX_4E[20]$-SX_4E and a glycine-proline-rich motif (112–141 residues): occurred 1x, RM2: $(SX)_4E[15](SX)_4E[11](SX)_4E[25]$-$(SX)_4E[20](SX)_4E$ and a glycine-proline-rich motif (86–118 residues): occurred 2x, RM3: $(SX)_4E[15](SX)_4E[11](SX)_4E$ and a glycine-proline-rich motif (44 residues): occurred 1x, RM4: $(SX)_4E[15](SX)_4E[11](SX)_4E[25](SX)_4E[20](SX)_4E[20](SX)_4E$ and a glycine-proline-rich motif (86–118 residues): occurred

6x, RM5: $(SX)_4E[15](SX)_4E[13](SX)_3E[16](SX)_4E[13](SX)_3E[18](SX)_3E[11](SX)_4E$ and a glycine-proline-rich motif (89 residues,: occurred 2 times, RM6: $(SX)_4E[20](SX)_4E[20](SX)_4E$ and a G-P-rich motif (86 residues, occurred 1x). In the second part of the RM1 and RM3 are alternating 13 times.

In *Arctopsyche grandis* (Figure S12), the sequence included three internally repeating modules flanked by two transition regions. RM1 consisted of a $(SX)_3E[12](SX)_4E[12](SX)_4E$ and a glycine-proline-rich motif (40–132 residues) and was repeated 24 times. RM2 contained a $(SX)_3E$ and a glycine-proline-rich motif (22–63 residues). It occurred 31 times. RM3 comprised a $(SX)_4E[12](SX)_3E$ and glycine-proline-rich motif (15–16 residues) and was represented 20 times.

Amino acid composition *of the* h-fibroin

Higher percentage of proline in the h-fibroin of fixed-retreat makers

In general, the amino acid composition of the protein sequence was conserved across the taxa that we sampled. Glycine and serine were consistently the most abundant residues across all three clades. However, despite these consistent patterns in composition, we observed some differences among clades. In retreat-making caddisflies, h-fibroin was characterized by a high amount of proline which ranged from 9.9 to 12.3% (n = 3). In contrast, the proportion of proline was much lower in the h-fibroin of tube case makers, ranging from 4 to 5.6% (n = 6). In the h-fibroin of cocoon-making caddisflies, the content of proline was even lower, ranging from 2.1–2.7% (n = 2).

Differences in the amino acid composition of h-fibroins of aquatic Trichoptera and terrestrial Lepidoptera

When comparing the amino acid composition of h-fibroins in Trichoptera to those of various terrestrial Lepidoptera (pyraloid moth,[32] ermine moth,[47] bagworm,[48] silkworm,[11] butterfly[49]), we find some consistent differences (Table 3). Although h-fibroins of both orders had high proportions of glycine (Trichoptera: 21.2–35.6%, Lepidoptera: 18.3–45.9%) and serine (Trichoptera: 9.3–17.2%, Lepidoptera:6.9–18.5%), h-fibroins in Lepidoptera had much more alanine (Trichoptera:0.1–4.9%, Lepidoptera: 21.9–40.52%). In addition, the Lepidoptera sequences exhibited a smaller percentage of charged residues. Negatively charged amino acids (aspartic acid and glutamic acid) ranged from 4–7.7% in the Trichoptera h-fibroin but only 1.1–2.4% in Lepidoptera. Positively charged amino acids (arginine, lysine) summed up to 7.6–16.9% in Trichoptera h-fibroins but were much lower in Lepidoptera (0.5–1.1%). Moreover, the amount of hydrophobic residues valine, leucine, and isoleucine was higher in the h-fibroins of caddisflies compared to Lepidoptera. Specifically, the amount of valine was 8.43% in retreat, 6.25% in cocoon, and 9.75% in case makers, but only 3.01% in Lepidoptera. Leucine ranged from 5.2% in retreat makers to 6.93% in case makers culminating in 10.1% in cocoon makers, whereas in Lepidoptera it was as low as 2.23%. The amount of isoleucine rages from 4.47% (case makers), 5.2% (retreat makers) to 7.3% in cocoon makers. H-fibroins of Lepidoptera only contained 1.36% of isoleucine.

DISCUSSION

In this study, we report eight new full-length *h-fibroin* sequences for caddisflies across a diverse set of silk usage. Two of these were generated from new genomic resources, whereas five were mined from previously published genomes,[41–44] highlighting the relevance of genome sequencing projects for the wider scientific community. The new full-length *h-fibroins* represent a substantial increase in the number of genomic resources available for the study of caddisfly silk and allowed us to compare the major silk gene for eleven species across the primary clades of Trichoptera, in which species exhibit different silk usage.

The genetic structure of *h-fibroin* of the eleven Trichoptera species showed similar organization of introns and exons, in line with previously reported *h-fibroin* sequences of Lepidoptera.[32,47] In addition to resolving the genetic structure, we compared the primary protein sequence of the h-fibroins. Consistent with previous research,[27,31] we found that structural elements of h-fibroin are conserved across Trichoptera. For example, of the species sampled, the n- and c-termini exhibit a high pairwise identity (n-terminus: 74.3%, c-terminus: 65.3%). In Lepidoptera, the conservation of the termini is linked with function. The n-terminus dimerizes and the c-terminus has been reported to interact with the light chain fibroin: the terminal cysteine forms an intermolecular disulfide bond with the light chain fibroin in the silkworm *B. mori*.[4] Given similar patterns of conservation of the termini in Trichoptera, their function is likely conserved across both orders. The c-terminus of all Trichoptera was characterized by the presence of a cysteine at position 19 in

Table 3. Amino acid composition of full-length h-fibroins of Trichoptera and terrestrial sister order Lepidoptera

Residue	Trichoptera - retreat (n = 3)	Trichoptera - cocoon (n = 2)	Trichoptera - case (n = 6)	Terrestrial Lepidoptera (n = 4)
Gly (G)	23.60 (1.71)	33.55 (2.05)	29.43 (1.73)	31.14 (9.73)
Arg (R)	9.43 (1.59)	7.5 (1.5)	13.12 (1.79)	0.44 (0.11)
Lys (K)	1.67 (0.48)	1.3 (0.3)	2.12 (0.52)	0.32 (0.21)
His (H)	2.03 (0.97)	5.25 (1.75)	1.12 (2.23)	0.07 (0.04)
Ala (A)	4.20 (0.50)	1.75 (0.65)	0.35 (0.30)	29.00 (6.34)
Val (V)	8.43 (0.37)	6.25 (1.05)	9.75 (0.74)	3.01 (1.01)
Leu (L)	5.20 (0.71)	10.1 (0.7)	6.93 (1.27)	2.23 (2.28)
Ile (I)	5.20 (0.22)	7.3 (0.8)	4.47 (0.84)	1.36 (0.99)
Pro (P)	10.77 (1.09)	2.4 (0.3)	4.78 (0.69)	2.46 (2.02)
Ser (S)	11.77 (0.50)	12.1 (2.8)	15.20 (1.27)	13.66 (3.97)
Tyr (Y)	5.10 (1.36)	2.45 (1.35)	2.95 (1.05)	4.21 (1.61)
Asp (D)	3.37 (1.39)	1.5 (0.3)	2.25 (0.48)	0.61 (0.19)
Glu (E)	2.60 (0.14)	2.5 (0.1)	2.97 (0.18)	1.33 (0.57)
Thr (T)	2.90 (0.37)	3 (0.2)	1.62 (0.30)	0.91 (0.38)
Trp (W)	1.87 (0.57)	1.35 (0.25)	2.52 (0.82)	0.06 (0.08)

The mean is given for each amino acid and group of Trichoptera/Lepidoptera. The standard derivation is given in parentheses. Amino acid composition for each species is given in Additional data Files: Data S1–S17.

the alignment (Figure 1). This suggests that disulfide crosslinking of fibroins occurs also in caddisflies and implies that covalent complex formation through the c-terminus is of similar importance for the structure, stability, and secretion of the fibroin complex as reported in *B. mori*.[28] In addition to the conservation of the termini, some conserved themes emerged in the central repetitive region, which consisted of repeating two-part structural modules, each containing a characteristic region of $(SX)_nE$ motifs interspersed with glycine-rich (in the cocoon- and case makers) or glycine-proline-rich (in retreat makers) regions of variable length. The serines of the $(SX)_nE$ motifs are extensively phosphorylated.[19,29,30] These phosphates then bind multivalent metal ions, which stabilize the silk and are responsible for the strength of the silk.[17,50,51] A structural model has been proposed, in which each (pSX)nE motif forms a β-strand, which in turn associates into anti-parallel Ca2+-stabilized ß-sheets.[17] The β-sheets stack through alternating hydrophobic and Ca2+-phosphate interfaces creating microcrystalline β-domains. By this model, each repeating module would correspond to a structure in which anti-parallel β-sheets are linked with a glycine-rich or glycine-proline-rich spacer region. The $(SX)_nE$/glycine-rich blocks may combine to form a higher-order β-domain structure through intra- or intermolecular stacking of the $[(SX)_nE]_m$ β-sheets. The crystalline β-domains would be separated by flexible and extensible glycine-/glycine-proline-rich regions. These signatures of conservation in repeat modules despite ~280 million years of divergence[14] and diverse silk usage, suggest a common mechanism for protein folding and silk formation across the underwater silks generated by all clades as suggested by.[31]

Despite the conservation in structure, we observed variation in the ordering and number of repeat modules. Our study builds on previous studies (i.e., [17,27,30,31,33,37,46]) by unveiling substantial diversity in the number and order of these repetitive structures. We observed a range of complexity in the repetitive structures across the phylogeny. In the simplest h-fibroin structure, sequenced from *A. davidsoni*, a single structural module was repeated 81 times (Figure 2). Slightly more complicated was the h-fibroin sequence of tube case-making *G. pellucidula*, which consisted of two repeat modules shuffled throughout the central region of the sequence (Figure 2). The sequences of the fixed retreat-makers were more variable, including *L. lineaticorne*, which was split into two parts, each of which consisted of a unique repeating module (Figure 2). The variations in repeating block patterns between the different clades may reflect adaptations for the diverse silk usage. For instance, cocoon-making *H. tibetana* and *A. davidsoni* are free-living as larvae and only produce pupal silk for building cocoons and pupal domes. The increase in h-fibroin sequence complexity scaled with silk usage diversity. The species in this study that construct both fixed retreats and capture nets exhibited the most variable h-fibroin sequences. The diversity in the number and order

of the repeat modules may hold clues to unraveling the unique applications of silk across clades (e.g., case-, cocoon-making versus fixed retreat building) and may be directly responsible for differences in mechanical properties.

The amino acid composition of the h-fibroins was largely conserved across samples, with some notable exceptions. The proportion of proline was clade-specific. Although proline was found in low proportions in the h-fibroin of Integripalpia (cocoon (2.1–2.7%) and tube case makers (4–5.6%)), higher proportions were found in Annulipalpia sequences (9.9–12.3%). Annulipalpian larvae are generally characterized by their fixed retreats which serve as shelters. In addition, some annulipalpian families, including all of the species investigated in this study, also construct silken capture nets, which are used for capturing food and would presumably require more extensible silk. Future work in caddisflies should focus on linking the physical properties of the silk with variations in the h-fibroin sequence. For example, orb-weaver spiders use silk to build prey-capture spirals with high fiber extensibility. This is necessary to catch insects in flight without breaking the web. The abundance of proline content in the major ampullate spidroin MaSp2 of orb-weaver spider silk was linked to enhanced extensibility of these fibers[52–57] because it increases the secondary structure disorder in the amorphous region.[52–54] Furthermore, Arakawa et al.[58] found that breaking strain was positively correlated with the presence of an amorphous, proline-rich region in the major ampullate spidroins MaSP1/2, which are often incorporated into dragline threads.[58]

Despite similarities in the gene structure and conservation of the terminal regions of *h-fibroin* in aquatic Trichoptera and terrestrial Lepidoptera, we found consistent differences in the amino acid composition of the protein sequences. In general, to prevent limitations by protein content in their diets, non-essential amino acids glycine, alanine, and serine are the dominant residues in insect silk genes.[2] However, in contrast with Lepidoptera h-fibroins, which have high alanine content, Trichoptera h-fibroin sequences were extremely low in alanine. This is likely because of differences in how the proteins are folded. In Lepidoptera, the β-sheet structures are mainly derived from repetitive polyglycine-alanine domains (family Bombicidae, e.g., *B. mori*), (non-)/polyalanine-domains (family Saturniidae: saturniid moths), or a combination of both domains (bagworm family Psychidae[48]) and, thus, these amino acids are important for the strength of the silk.[59,60] In contrast, as noted above, caddisfly h-fibroin β-sheets are primarily formed through the interaction of phosphorylated serine blocks with metal ions derived from their aquatic environment, an adaptation specific to aquatic dwelling species.[51] H-fibroins of both, Trichoptera (9.3–17.2%) and Lepidoptera (6.89–18.5%) contain a high % of serines. Therefore, Ashton et al.[22] suggested that one of the key molecular adaptations of a terrestrial ancestor silk to aquatic environments could have been kinase phosphorylating H-fibroin serines. We also detected a higher percentage of hydrophobic (Valine, Isoleucine, Leucine) and charged residues in caddisflies (Table 3), which has previously been hypothesized as an adaptation for aquatic silks.[27] Some previous research has been conducted on silk genes in aquatic insects.[61–63] For example, the aquatic black fly larva *Simulium vittatum* uses silk for larval adherence in rapidly flowing water and to construct pupal tents.[61] Characterization of its silk gland proteins did not show sequence similarity to fibroins but revealed multiple phosphorylated serine residues and a high amount of hydrophobic and charged amino acids in the central region of the proteins.[61] The authors argue that in an aquatic environment, hydrophobicity would lead to clumping and the greater proportion of charged amino acids might lead to proteins that are less hydrophobic compared to terrestrial silks.[61] Previous studies on water-associated spiders also revealed higher concentrations of hydrophobic amino acid motifs (Glycine-Valine) in silk gene sequences (spidroins, 20–38%) compared to terrestrial spiders (2–4%). The spidroins of the semi-aquatic spiders showed little sequence similarity to fibroins of caddisflies and authors did not find evidence for similar serine-rich motifs.[62] However, egg sacs of a semi-aquatic spider species contained calcium and phosphorus. These elements have not been detected in the egg sacs of non-aquatic spiders and might contribute to the water-repellant properties of the silk.[63]

The new genomic data provided in this study was used to investigate the primary structure of h-fibroins across caddisflies. Although we observed conserved patterns in the primary structure of the h-fibroin, the amino acid composition and the number and arrangement of repeating modules varied among species with different silk usages. To understand the role of this variation in generating the myriad silk phenotypes that we observed across Trichoptera the sequences of the h-fibroin need to be linked with experimental evidence, such as mechanical testing as shown for spider silk.[58] Such studies are essential to gauge the potential of caddisfly silk in material science. The sequences that we have compiled here represent an important step toward performing such analyses.

Limitations of the study

This study provides characterizations of the primary structure of h-fibroin from a diverse set of caddisflies with different silk usage strategies (fixed retreat: n = 3, tube case: n = 6, cocoon builders: n = 2). We acknowledge that future studies including additional species of the three main clades of caddisflies (especially fixed retreat and cocoon builders) are important to underpin our findings. In addition, the study lacks comparisons with terrestrial Trichoptera (e.g., North American *Philocasca demita* or *European Enoicyla pusilla*) whose silk genes have not been sequenced to date. We, instead, compared the amino acid composition of the h-fibroin of Trichoptera with those of terrestrial species of sister order Lepidoptera (moths and butterflies). Despite these limitations, we hope that our study forms a foundation for future studies to screen and correlate amino acid motifs and other sequence features with quantifiable silk properties, such as mechanical measurements.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample acquisition
- METHOD DETAILS
  - DNA extraction and whole-genome sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Raw data processing, genome-size estimation, and whole-genome assemblies
  - Assembly quality control
  - Identification and annotation of heavy-chain fibroins
  - Comparison of heavy-chain fibroins across caddisflies clades

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107253.

ACKNOWLEDGMENTS

This work is a result of the LOEWE-Centre for Translational Biodiversity Genomics funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK). J.H. acknowledges funding from Deutsche Forschungsgemeinschaft - Project number 502865717. P.B.F. received internal funding from Brigham Young University, College of Life Science for sequencing the Leptonema lineaticorne genome. P.B.F. and T.D. received funding for fieldwork from the USAID and US National Academies of Sciences PEER program (BH-035). We thank Brigham Young University and LOEWE-Centre for Translational Biodiversity Genomics (TBG) for providing the computational resources needed to complete this study. The Leptonema lineaticorne specimen was collected in Ecuador under the "Genetic Resources Access Contract" No. MAAE-DBI-CM-2021-0161, and the support of project AMB.BRT.23.02 ("Mountain Freshwater Diversity, from Taxonomy to Functional Genomics, and Approximation from Trichoptera-II"). We appreciate help from Sacha Lodge for field logistics. Graphical abstract, figures, and supplementary figures were created with BioRender.com. Photo of capture net of retreat-making caddisfly in graphical abstract was provided by Sven Tränker.

AUTHOR CONTRIBUTIONS

J.H.: Conceptualization, Data curation, Formal analysis, Investigation, Validation; Visualization, Writing - original draft; Writing - review and editing.

R.J.S.: Conceptualization, Investigation, Validation; Writing - original draft; Writing - review and editing.

P.B.F.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Resources: computational and DNA sequencing, Writing - original draft; Writing - review and editing.

B.R-T.: Resources: material for DNA extraction, Writing - review.

T.D.: Resources: material for DNA extraction, Writing - review.

A.P.: Software; Writing - review.

S.U.P.: Conceptualization, Funding acquisition, Project administration; Resources: computational; Writing - review and editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: February 6, 2023
Revised: April 5, 2023
Accepted: June 27, 2023
Published: July 4, 2023

REFERENCES

1. Sehnal, F., and Sutherland, T. (2008). Silks produced by insect labial glands. Prion *2*, 145–153. https://doi.org/10.4161/pri.2.4.7489.

2. Sutherland, T.D., Young, J.H., Weisman, S., Hayashi, C.Y., and Merritt, D.J. (2010). Insect silk: one name, many materials. Annu. Rev. Entomol. *55*, 171–188. https://doi.org/10.1146/annurev-ento-112408-085401.

3. Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., Zhao, P., Zha, X., Cheng, T., Chai, C., et al. (2004). A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science *306*, 1937–1940. https://doi.org/10.1126/science.1102210.

4. Tanaka, K., Kajiyama, N., Ishikura, K., Waga, S., Kikuchi, A., Ohtomo, K., Takagi, T., and Mizuno, S. (1999). Determination of the site of disulfide linkage between heavy and light chains of silk fibroin produced by Bombyx mori. Biochim. Biophys. Acta *1432*, 92–103. https://doi.org/10.1016/s0167-4838(99)00088-6.

5. Tsujimoto, Y., and Suzuki, Y. (1979). The DNA sequence of Bombyx mori fibroin gene including the 5' flanking, mRNA coding, entire intervening and fibroin protein coding regions. Cell *18*, 591–600. https://doi.org/10.1016/0092-8674(79)90075-8.

6. Wang, S.-P., Guo, T.-Q., Guo, X.-Y., Huang, J.-T., and Lu, C.-D. (2006). Structural analysis of fibroin heavy chain signal peptide of silkworm Bombyx mori. Acta Biochim. Biophys. Sin. *38*, 507–513. https://doi.org/10.1111/j.1745-7270.2006.00189.x.

7. Inoue, S., Tanaka, K., Arisaka, F., Kimura, S., Ohtomo, K., and Mizuno, S. (2000). Silk fibroin of Bombyx mori is secreted, assembling a high molecular mass elementary unit consisting of H-chain, L-chain, and P25, with a 6:6:1 molar ratio. J. Biol. Chem. *275*, 40517–40528. https://doi.org/10.1074/jbc.M006897200.

8. Long, D., Lu, W., Zhang, Y., Guo, Q., Xiang, Z., and Zhao, A. (2015). New insight into the mechanism underlying fibroin secretion in silkworm, Bombyx mori. FEBS J. *282*, 89–101. https://doi.org/10.1111/febs.13105.

9. Mita, K., Ichimura, S., and James, T.C. (1994). Highly repetitive structure and its organization of the silk fibroin gene. J. Mol. Evol. *38*, 583–592. https://doi.org/10.1007/BF00175878.

10. Tong, X., Han, M.-J., Lu, K., Tai, S., Liang, S., Liu, Y., Hu, H., Shen, J., Long, A., Zhan, C., et al. (2022). High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. Nat. Commun. *13*, 5619. https://doi.org/10.1038/s41467-022-33366-x.

11. Zhou, C.Z., Confalonieri, F., Medina, N., Zivanovic, Y., Esnault, C., Yang, T., Jacquet, M., Janin, J., Duguet, M., Perasso, R., and Li, Z.G. (2000). Fine organization of Bombyx mori fibroin heavy chain gene. Nucleic Acids Res. *28*, 2413–2419. https://doi.org/10.1093/nar/28.12.2413.

12. Mackay, R.J., and Wiggins, G.B. (1979). Ecological Diversity in Trichoptera. Annu. Rev. Entomol. *24*, 185–208. https://doi.org/10.1146/annurev.en.24.010179.001153.

13. Morse, J.C., Frandsen, P.B., Graf, W., and Thomas, J.A. (2019). Diversity and Ecosystem Services of Trichoptera. Insects *10*, 125. https://doi.org/10.3390/insects10050125.

14. Thomas, J.A., Frandsen, P.B., Prendini, E., Zhou, X., and Holzenthal, R.W. (2020). A multigene phylogeny and timeline for Trichoptera (Insecta). Syst. Entomol. *45*, 670–686. https://doi.org/10.1111/syen.12422.

15. Wiggins, G.B. (2004). Caddisflies: The Underwater Architects (University of Toronto Press).

16. Ashton, N.N., and Stewart, R.J. (2015). Self-recovering caddisfly silk: energy dissipating, Ca(2+)-dependent, double dynamic network fibers. Soft Matter *11*, 1667–1676. https://doi.org/10.1039/c4sm02435d.

17. Ashton, N.N., Roe, D.R., Weiss, R.B., Cheatham, T.E., and Stewart, R.J. (2013). Self-Tensioning Aquatic Caddisfly Silk: Ca2+-Dependent Structure, Strength, and Load Cycle Hysteresis. Biomacromolecules *14*, 3668–3681. https://doi.org/10.1021/bm401036z.

18. Brown, S.A., Ruxton, G.D., and Humphries, S. (2004). Physical properties of Hydropsyche siltalai (Trichoptera) net silk. J. North Am. Benthol. Soc. *23*, 771–779. https://doi.org/10.1899/0887-3593(2004)023<0771:PPOHST>2.0.CO;2.

19. Stewart, R.J., and Wang, C.S. (2010). Adaptation of Caddisfly Larval Silks to Aquatic Habitats by Phosphorylation of H-Fibroin Serines. Biomacromolecules *11*, 969–974. https://doi.org/10.1021/bm901426d.

20. Lane, D.D., Kaur, S., Weerasakare, G.M., and Stewart, R.J. (2015). Toughened hydrogels inspired by aquatic caddisworm silk. Soft Matter *11*, 6981–6990. https://doi.org/10.1039/C5SM01297J.

21. Tszydel, M., Zabłotni, A., Wojciechowska, D., Michalak, M., Krucińska, I., Szustakiewicz, K., Maj, M., Jaruszewska, A., and Strzelecki, J. (2015). Research on possible medical use of silk produced by caddisfly larvae of Hydropsyche angustipennis (Trichoptera, Insecta). J. Mech. Behav. Biomed. Mater. *45*, 142–153. https://doi.org/10.1016/j.jmbbm.2015.02.003.

22. Ashton, N.N., Taggart, D.S., and Stewart, R.J. (2012). Silk tape nanostructure and silk gland anatomy of trichoptera. Biopolymers *97*, 432–445. https://doi.org/10.1002/bip.21720.

23. Yonemura, N., Mita, K., Tamura, T., and Sehnal, F. (2009). Conservation of Silk Genes in Trichoptera and Lepidoptera. J. Mol. Evol. *68*, 641–653. https://doi.org/10.1007/s00239-009-9234-5.

24. Engster, M.S. (1976). Studies on silk secretion in the Trichoptera (F. Limnephilidae). II. Structure and amino acid composition of the silk. Cell Tissue Res. *169*, 77–92. https://doi.org/10.1007/BF00219309.

25. Wang, C.S., Ashton, N.N., Weiss, R.B., and Stewart, R.J. (2014). Peroxinectin catalyzed dityrosine crosslinking in the adhesive underwater silk of a casemaker caddisfly larvae, Hysperophylax occidentalis. Insect Biochem. Mol. Biol. *54*, 69–79. https://doi.org/10.1016/j.ibmb.2014.08.009.

26. Rouhová, L., Sehadová, H., Pauchová, L., Hradilová, M., Žurovcová, M., Šerý, M., Rindoš, M., and Žurovec, M. (2022). Using the multi-omics approach to reveal the silk composition in Plectrocnemia conspersa. Front. Mol. Biosci. *9*, 945239. https://doi.org/10.3389/fmolb.2022.945239.

27. Yonemura, N., Sehnal, F., Mita, K., and Tamura, T. (2006). Protein Composition of Silk Filaments Spun under Water by Caddisfly Larvae. Biomacromolecules *7*, 3370–3378. https://doi.org/10.1021/bm060663u.

28. Stewart, R.J., Frandsen, P.B., Pauls, S.U., and Heckenhauer, J. (2022). Conservation of Three-Dimensional Structure of Lepidoptera and Trichoptera L-Fibroins for 290 Million Years. Molecules *27*, 5945. https://doi.org/10.3390/molecules27185945.

29. Addison, J.B., Ashton, N.N., Weber, W.S., Stewart, R.J., Holland, G.P., and Yarger, J.L. (2013). β-Sheet Nanocrystalline Domains Formed from Phosphorylated Serine-Rich Motifs in Caddisfly Larval Silk: A Solid State NMR and XRD Study. Biomacromolecules *14*, 1140–1148. https://doi.org/10.1021/bm400019d.

30. Ohkawa, K., Miura, Y., Nomura, T., Arai, R., Abe, K., Tsukada, M., and Hirabayashi, K. (2013). Long-range periodic sequence of the cement/silk protein of Stenopsyche marmorata: purification and biochemical characterisation. Biofouling *29*, 357–367. https://doi.org/10.1080/08927014.2013.774376.

31. Frandsen, P.B., Bursell, M.G., Taylor, A.M., Wilson, S.B., Steeneck, A., and Stewart, R.J. (2019). Exploring the underwater silken architectures of caddisworms: comparative silkomics across two caddisfly suborders. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20190206. https://doi.org/10.1098/rstb.2019.0206.

32. Kawahara, A.Y., Storer, C.G., Markee, A., Heckenhauer, J., Powell, A., Plotkin, D., Hotaling, S., Cleland, T.P., Dikow, R.B., Dikow, T., et al. (2022). Long-read HiFi sequencing correctly assembles repetitive heavy fibroin silk genes in new moth and caddisfly genomes. Gigabyte *2022*. gigabyte64–14. https://doi.org/10.46471/gigabyte.64.

33. Hotaling, S., Wilcox, E.R., Heckenhauer, J., Stewart, R.J., and Frandsen, P.B. (2023). Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. BMC Genom. *24*, 117. https://doi.org/10.1186/s12864-023-09193-9.

34. Wang, Y., Sanai, K., Wen, H., Zhao, T., and Nakagaki, M. (2010). Characterization of unique heavy chain fibroin filaments spun underwater by the caddisfly Stenopsyche marmorata (Trichoptera; Stenopsychidae). Mol. Biol. Rep. *37*, 2885–2892. https://doi.org/10.1007/s11033-009-9847-1.

35. Luo, S., Tang, M., Frandsen, P.B., Stewart, R.J., and Zhou, X. (2018). The genome of an underwater architect, the caddisfly Stenopsyche tienmushanensis Hwang (Insecta: Trichoptera). GigaScience *7*, giy143. https://doi.org/10.1093/gigascience/giy143.

36. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

37. Frandsen, P.B., Hotaling, S., Powell, A., Heckenhauer, J., Kawahara, A.Y., Baker, R.H., Hayashi, C.Y., Ríos-Touma, B., Holzenthal, R., Pauls, S.U., and Stewart, R.J. (2023). Allelic resolution of insect and spider silk genes reveals hidden genetic diversity. Proc. Natl. Acad. Sci. USA *120*, e2221528120. https://doi.org/10.1073/pnas.2221528120.

38. Hotaling, S., Kelley, J.L., and Frandsen, P.B. (2020). Aquatic Insects Are Dramatically Underrepresented in Genomic Research. Insects *11*, 601. https://doi.org/10.3390/insects11090601.

39. Laetsch, D.R., and Blaxter, M.L. (2017). BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Res. *6*, 1287. https://doi.org/10.12688/f1000research.12232.1.

40. Ríos-Touma, B., Holzenthal, R.W., Rázuri-Gonzales, E., Heckenhauer, J., Pauls, S.U., Storer, C.G., and Frandsen, P.B. (2022). De Novo Genome Assembly and Annotation of an Andean Caddisfly, Atopsyche davidsoni Sykora, 1991, a Model for Genome Research of High-Elevation Adaptations. Genome Biol. Evol. *14*, evab286. https://doi.org/10.1093/gbe/evab286.

41. McSwan, E., Broad, G.R., Wallace, I., and Price, B.W. (2023). The genome sequence of the Mottled Sedge, *Glyphotaelius pellucidus* (Retzius, 1783) [version 1; peer review: awaiting peer review]. Wellcome Open Res. *8*, 102. https://doi.org/10.12688/wellcomeopenres.19076.18.

42. Austin, M., Clifford, C., Rutt, G., and Price, B.W. (2023). The genome sequence of a caddisfly, *Limnephilus lunatus* (Curtis, 1834) [version 1; peer review: awaiting peer review]. Wellcome Open Res. *8*, 25. https://doi.org/10.12688/wellcomeopenres.18752.18.

43. Clifford, C., Friend, K., Skipp, S., and Wallace, I. (2023). The genome sequence of the cinnamon sedge caddisfly, *Limnephilus marmoratus* (Curtis, 1834) [version 1; peer review: 1 approved]. Wellcome Open Res. *8*, 64. https://doi.org/10.12688/wellcomeopenres.18753.18.

44. Ge, X., Jin, J., Peng, L., Zang, H., Wang, B., and Sun, C. (2022). The First Chromosome-level Genome Assembly of Cheumatopsyche charites Malicky and Chantaramongkol, 1997 (Trichoptera: Hydropsychidae) Reveals How It Responds to Pollution. Genome Biol. Evol. *14*, evac136. https://doi.org/10.1093/gbe/evac136.

45. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

46. Wang, Y., Wang, H., Zhao, T., and Nakagaki, M. (2010). Characterization of a Cysteine-Rich Protein Specifically Expressed in the Silk Gland of Caddisfly Stenopsyche marmorata (Trichoptera; Stenopsychidae). Biosci. Biotechnol. Biochem. *74*, 108–112. https://doi.org/10.1271/bbb.90606.

47. Volenikova, A., Nguyen, P., Davey, P., Sehadova, H., Kludkiewicz, B., Koutecky, P., Walters, J.R., Roessingh, P., Provaznikova, I., Sery, M., et al. (2022). Genome sequence and silkomics of the spindle ermine moth, Yponomeuta cagnagella, representing the early diverging lineage of the ditrysian Lepidoptera. Commun. Biol. *5*, 1281–1313. https://doi.org/10.1038/s42003-022-04240-9.

48. Kono, N., Nakamura, H., Ohtoshi, R., Tomita, M., Numata, K., and Arakawa, K. (2019). The bagworm genome reveals a unique fibroin gene that provides high tensile strength. Commun. Biol. *2*, 148–149. https://doi.org/10.1038/s42003-019-0412-8.

49. Lohse, K., García-Berro, A., and Talavera, G. (2021). The genome sequence of the red admiral, *Vanessa atalanta* (Linnaeus, 1758) [version 1; peer review: 1 approved, 1 approved with reservations]. Wellcome Open Res. *6*, 356. https://doi.org/10.12688/wellcomeopenres.17524.16.

50. Ashton, N.N., and Stewart, R.J. (2019). Aquatic caddisworm silk is solidified by environmental metal ions during the natural fiber-spinning process. FASEB J. *33*, 572–583. https://doi.org/10.1096/fj.201801029r.

51. Addison, J.B., Weber, W.S., Mou, Q., Ashton, N.N., Stewart, R.J., Holland, G.P., and Yarger, J.L. (2014). Reversible Assembly of β-Sheet Nanocrystals within Caddisfly Silk. Biomacromolecules *15*, 1269–1275. https://doi.org/10.1021/bm401822p.

52. Rauscher, S., Baud, S., Miao, M., Keeley, F.W., and Pomès, R. (2006). Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. Struct. Lond. Engl. *14*, 1667–1676. https://doi.org/10.1016/j.str.2006.09.008.

53. Creager, M.S., Jenkins, J.E., Thagard-Yeaman, L.A., Brooks, A.E., Jones, J.A., Lewis, R.V., Holland, G.P., and Yarger, J.L. (2010). Solid-State NMR Comparison of Various Spiders' Dragline Silk Fiber. Biomacromolecules *11*, 2039–2043. https://doi.org/10.1021/bm100399x.

54. Liu, Y., Sponner, A., Porter, D., and Vollrath, F. (2008). Proline and processing of spider silks. Biomacromolecules *9*, 116–121. https://doi.org/10.1021/bm700877g.

55. Savage, K.N., and Gosline, J.M. (2008). The effect of proline on the network structure of major ampullate silks as inferred from their mechanical and optical properties. J. Exp. Biol. *211*, 1937–1947. https://doi.org/10.1242/jeb.014217.

56. Savage, K.N., and Gosline, J.M. (2008). The role of proline in the elastic mechanism of hydrated spider silks. J. Exp. Biol. *211*, 1948–1957. https://doi.org/10.1242/jeb.014225.

57. Marhabaie, M., Leeper, T.C., and Blackledge, T.A. (2014). Protein composition correlates with the mechanical properties of spider (Argiope trifasciata) dragline silk. Biomacromolecules *15*, 20–29. https://doi.org/10.1021/bm401110b.

58. Arakawa, K., Kono, N., Malay, A.D., Tateishi, A., Ifuku, N., Masunaga, H., Sato, R., Tsuchiya, K., Ohtoshi, R., Pedrazzoli, D., et al. (2022). 1000 spider silkomes: Linking sequences to silk physical properties. Sci. Adv. *8*, eabo6043. https://doi.org/10.1126/sciadv.abo6043.

59. Malay, A.D., Sato, R., Yazawa, K., Watanabe, H., Ifuku, N., Masunaga, H., Hikima, T., Guan, J., Mandal, B.B., Damrongsakkul, S., and Numata, K. (2016). Relationships between physical properties and sequence in silkworm silks. Sci. Rep. *6*, 27573. https://doi.org/10.1038/srep27573.

60. Guo, C., Zhang, J., Jordan, J.S., Wang, X., Henning, R.W., and Yarger, J.L. (2018). Structural Comparison of Various Silkworm Silks: An Insight into the Structure-Property Relationship. Biomacromolecules *19*, 906–917. https://doi.org/10.1021/acs.biomac.7b01687.

61. Papanicolaou, A., Woo, A., Brei, B., Ma, D., Masedunskas, A., Gray, E., Xiao, G.G., Cho, S., and Brockhouse, C. (2013). Novel aquatic silk genes from Simulium (Psilozia) vittatum (Zett) Diptera: Simuliidae. Insect Biochem. Mol. Biol. *43*, 1181–1188. https://doi.org/10.1016/j.ibmb.2013.09.008.

62. Correa-Garhwal, S.M., Clarke, T.H., Janssen, M., Crevecoeur, L., McQuillan, B.N., Simpson, A.H., Vink, C.J., and Hayashi, C.Y. (2019). Spidroins and Silk Fibers of Aquatic Spiders. Sci. Rep. *9*, 13656. https://doi.org/10.1038/s41598-019-49587-y.

63. Correa-Garhwal, S.M., Chaw, R.C., Dugger, T., Clarke, T.H., Chea, K.H., Kisailus, D., and Hayashi, C.Y. (2019). Semi-aquatic spider silks: transcripts, proteins, and silk fibres of the fishing spider, *Dolomedes triton* (Pisauridae). Insect Mol. Biol. *28*, 35–51. https://doi.org/10.1111/imb.12527.

64. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. *35*, 543–548. https://doi.org/10.1093/molbev/msx319.

65. Schell, T., Feldmeyer, B., Schmidt, H., Greshake, B., Tills, O., Truebano, M., Rundle, S.D., Paule, J., Ebersberger, I., and Pfenninger, M. (2017). An Annotated Draft Genome for Radix auricularia (Gastropoda, Mollusca). Genome Biol. Evol. *9*, 0–592. https://doi.org/10.1093/gbe/evx032.

66. Pfenninger, M., Schönnenbeck, P., and Schell, T. (2022). ModEst: Accurate estimation of genome size from next generation sequencing data. Mol. Ecol. Resour. *22*, 1454–1464. https://doi.org/10.1111/1755-0998.13570.

67. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432. https://doi.org/10.1038/s41467-020-14998-3.

68. Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294. https://doi.org/10.1093/bioinformatics/btv566.

69. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

70. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

71. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

72. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

73. Hoff, K.J., and Stanke, M. (2019). Predicting Genes in Single Genomes with AUGUSTUS. Curr. Protoc. Bioinforma. *65*, e57. https://doi.org/10.1002/cpbi.57.

74. Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gíslason, M.H., Pihl, S.I., Tsirigos, K.D., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat. Biotechnol. *40*, 1023–1025. https://doi.org/10.1038/s41587-021-01156-3.

75. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

76. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In The Proteomics Protocols Handbook, J.M. Walker, ed. (Humana Press), pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571.

77. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011.

78. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinforma. Oxf. Engl. *29*, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

79. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Leptonema lineaticorne tissue of whole body	This study	Sample ID EC20210919-2, NCBI Biosample ID: SAMN25408291
Himalopsyche tibetana tissue of thorax	This study	Sample ID BH20221015-01, NCBI Biosample ID: SAMN31697150
Critical commercial assays		
Qiagen Genomic-tip extraction kit	Quiagen	NA
SMRTbell Express Prep kit v2.0	Pacific Biosciences, Menlo Park, CA	NA
Deposited data		
Figshare deposited data for main text and supplemental analyses	This study	Figshare repository: https://figshare.com/s/03f88091eda258465d2b
GitHub Project: h-fibroin-visual including all custom-made scripts used in this study	This study	https://github.com/AshlynPowell/h-fibroin-visual
Leptonema lineaticorne sequence reads	This study	NCBI: Short Read Archive: SRR20711493
*Leptonema lineaticorne*genome assembly	This study	GenBank Accession: GCA_024500535.1
*Leptonema lineaticorne*h-fibroin	This study	GenBank Accession:OQ983470
*Himalopsyche tibetana*sequence reads	This study	NCBI: Short Read Archive: SRR22537910
*Himalopsyche tibetana*genome assembly	This study	GenBank Accession: JAPJYX000000000
*Himalopsyche tibetana*h-fibroin	This study	GenBank Accession:OQ983471
Atopsyche davidsoni genome assembly	Ríos-Touma et al.[36]	GenBank Accession: GCA_022113835.1
Atopsyche davidsoni h-fibroin	Ríos-Touma et al.[36]	GenBank Accession: OQ787677
*Eubasilissa regina*genome assembly	Kawahara et al.[32]	GenBank Accession: GCA_022840565.1
*Glyphotaelius pellucidus*genome assembly	McSwan et al.[48]	GenBank Accession: GCA_936435175.1
Glyphotaelius pellucidus h-fibroin	This study	GenBank Accession:BK063450
*Hesperophylax magnus*genome assembly	Hotaling et al.[42]	GenBank Accession: GCA_026573805.1
Hesperophylax magnus h-fibroin	Hotaling et al.[42]	GenBank Accession:OQ787679
*Limnephilus lunatus*genome assembly	Austin et al.[46]	GenBank Accession: GCA_917563855.2
*Limnephilus lunatus*h-fibroin	This study	GenBank Accession:BK063451
*Limnephilus marmoratus*genome assembly	Clifford et al.[47]	GenBank Accession: GCA_917880885.1
Limnephilus marmoratus h-fibroin	This study	GenBank Accession:BK063452
*Limnephilus rhombicus*genome assembly	WELLCOME SANGER INSTITUTE[11]	GenBank Accession: GCA_929108145.2
*Limnephilus rhombicus*h-fibroin	This study	GenBank Accession:BK063453
*Cheumatopsyche charites*genome assembly	Ge et al.[49]	GenBank Accession: GCA_024500535.1
Cheumatopsyche charites h-fibroin	This study	GenBank Accession:BK063449
*Arctopsyche grandis*genome assembly	Frandsen et al.[39]	GenBank Accession: GCA_029955255.1
Arctopsyche grandis h-fibroin	Frandsen et al.[39]	GenBank Accession: OQ787675
Software and algorithms		
pbccs tool v6.6.0	NA	https://github.com/PacificBiosciences/pbbioconda
JELLYFISH v2.2.10	Marçais and Kingsford[64]	https://github.com/gmarcais/Jellyfish
GenomeScope 2.0	Ranallo-Benavidez[65]	http://qb.cshl.edu/genomescope/genomescope2.0/
Hifiasm v.0.13-r307	Cheng et al.[66]	https://github.com/chhylp123/hifiasm
BUSCO v5.2.2	Simão et al.[67]; Waterhouse et al.[68]	https://gitlab.com/ezlab/busco/-/releases/5.2.2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
backmap.pl v0.5	Schell et al.[69]; Pfenninger et al.[70]	https://github.com/schellt/backmap
minimap2 v2.24	Li et al.[71]	https://github.com/lh3/minimap2
qualimap 2.2.1	Okonechnikov et al.[72]	http://qualimap.conesalab.org/
MultiQC v1.10	Ewels et al.[23]	
bedtools v2.30.0	Quinlan et al.[73]	https://multiqc.info/
R v4.0.3	R Core Team 2021	https://www.r-project.org/contributors.html
BlobTools v1.1.1	Laetsch et al.[43]	https://github.com/DRL/blobtools
blastn 2.10.0+	Camacho et al.[74]	https://www.ncbi.nlm.nih.gov/books/NBK279690/
Geneious Prime 2022.1.1	NA	https://www.geneious.com/prime/
Augustus v.3.3.3	Hoff and Stanke[73]	https://bioinf.uni-greifswald.de/augustus/
SignalP 6.0	Teufel et al.[74]	https://services.healthtech.dtu.dk/service.php?SignalP
Muscle 3.8.425	Edgar et al.[75]	https://kbase.us/applist/apps/kb_muscle/MUSCLE_nuc/release
Expasy ProtParam	Gasteiger et al.[76]	https://web.expasy.org/protparam/
samtools v1.13	Li et al.[71]	https://github.com/samtools/

RESOURCE AVAILABILITY

Lead contact

Further information can be requested via the lead contact, Jacqueline Heckenhauer (jacqueline.heckenhauer@senckenberg.de).

Materials availability

This study did not generate any new reagents.

Data and code availability

Genomic data (sequence reads and assemblies) have been deposited to GenBank and are publicly available. For accession numbers, see key resources table. The data supporting the results of this article (all data associated with quality control of the assemblies, full-length *h-fibroin* gene sequences including introns, *h-fibroin* protein-coding nucleotide sequences and h-fibroin protein sequences newly identified in this paper) are available on Figshare https://doi.org/10.6084/m9.figshare.20407101. This paper utilizes existing, publicly available data. For accession numbers of these datasets see key resources table.

All original code has been deposited at GitHub and is publicly available at: https://github.com/AshlynPowell/h-fibroin-visual

Any other additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample acquisition

We collected a single adult individual of *L. lineaticorne* in an Amazon Blackwater channel in Ecuador, Sucumbios, Sacha Lodge (Amazon Basin), Caño Anaconda (0°28′20.71″S; 76°27′59.08″W, elevation 237 m asl.) and a single larva of *H. tibetana* in Gasa, Bhutan (28°03.9477′N, 89°39.0019′E, elevation 3824 m asl.).

METHOD DETAILS

DNA extraction and whole-genome sequencing

We extracted high molecular weight from single individuals of *L. lineaticorne and H. tibetana* using the Qiagen Genomic-tip extraction kit and prepared DNA sequencing libraries following the instructions of the SMRTbell Express Prep. For each individual, one SMRT cell sequencing run was performed on the Sequel System II in CCS (circular consensus sequencing) mode using 30-h movie time.

QUANTIFICATION AND STATISTICAL ANALYSIS

Raw data processing, genome-size estimation, and whole-genome assemblies

We generated HiFi reads from the raw data (reads with quality above Q20) using PacBio SMRTlink software (https://github.com/PacificBiosciences/pbbioconda). We estimated genome size using sequencing reads and a *k-mer*-based statistical approach. After counting *k-mers* with JELLYFISH v2.2.10[77] using jellyfish count -C -s 25556999998 -F 3 and a *k-mer* length of 21 (-m 21) with the ccs-reads, we produced a histogram of *k-mer* frequencies with jellyfish histo. We ran GenomeScope 2.0[67] with the exported *k-mer* count histogram within the online web tool (http://qb.cshl.edu/genomescope/genomescope2.0/) using the following parameters: *k-mer* length = 21, Ploidy = 2, Max kmer coverage = 10000. We assembled the *L. lineaticorne* and *H. tibetana* genomes, with hifiasm v0.13-r307[36] with the default settings.

Assembly quality control

We evaluated the assembly quality based on continuity (QUAST v5.0.2[78]) and completeness of Benchmarking Universal Single-Copy Orthologs (BUSCOs) with BUSCO v5.2.2[45,64] using the lineage dataset endopterygota_odb10 in genome mode. In addition, we calculated the back-mapping rate of the HiFi reads to the assemblies using backmap.pl v0.5[65,66] with the parameter -hifi. Other parameters were kept as default. This wrapper script automatically maps the reads to the assembly with minimap2 v2.24[79] and executes qualimap v 2.2.1,[68] MultiQC v1.10,[69] bedtools v 2.30.0,[70] and RScript v 4.0.3 (R Core Team 2021) to create the mapping quality report and a coverage histogram. In addition, it plots the coverage distribution and estimates of genome size from mapped nucleotides divided by the mode of the coverage distribution (>0). The final genome assemblies were screened for potential contaminations with taxon-annotated GC-coverage (TAGC) plots using BlobTools v1.1.1.[39] For this purpose, the bam file resulting from the back-mapping analysis was converted to a BlobTools readable.cov file with "blobtools map2cov". Taxonomic assignment for BlobTools was done with blastn 2.10.0+[71] using -task megablast and -e-value 1e-25. The blobDB was created and plotted from the cov file and blast hits. Contaminations of adapters detected by NCBI in the *L. lineaticorne* genome assembly were filtered out with samtools v1.13 faidx.[72]

Identification and annotation of heavy-chain fibroins

Recently, the Wellcome Sanger Institute published four high-quality caddisfly genomes (*Glyphotaelius pellucidus*: GCA_936435175.1, *Limnephilus lunatus*: GCA_917563855.2, *Limnephilus marmoratus*: GCA_917880885.1, *Limnephilus rhombicus*: GCA_929108145.1) in the course of the Darwin Tree of Life Project. An additional high-quality genome of a retreat-making caddisfly was published by Ge et al.[44] We identified the *h-fibroin* genes in these assemblies by using tBLASTn to search the assemblies with the conserved n- and c-termini with query sequences from previously published species *Hesperophylax* sp.,[17] *Limnephilus decipiens* AB214509[27] and *Rhyacophila obliterata* AB354689.1 and AB354588.1[23] in Geneious Prime 2022.1.1 (https://www.geneious.com) with default settings. After verifying that both BLAST hits (hit with n- and hit with c-terminus) were isolated to the same contig in the genome assembly, we extracted the sequences and 1,000 bp of flanking regions from the assembly using the sequence view "extract" in Geneious and annotated this region using Augustus v.3.3.3.[73] Introns that did not affect reading frames were manually removed from the annotation and h-fibroins were manually curated (see supplementary table). Protein coding nucleotide sequences were translated with the Geneious Tool "Translate" using the standard genetic code. We used the same approach to extract and annotate the *h-fibroin* in *L. lineaticorne* but using the termini of *Parapsyche elsis*.[31] For *H. tibetana* we used termini of *Rhyacophila obliterata*. For details see Table S1. We predicted signal peptides and the location of their cleavage site of the h-fibroin protein sequences with the SignalP 6.0 server (https://services.healthtech.dtu.dk/service.php?SignalP,[74]) using the following settings: organism = Eukarya, model mode = slow.

Comparison of heavy-chain fibroins across caddisflies clades

We used the previously published full-length h-fibroin sequences[32,33,37] and the newly identified h-fibroin sequences generated here to compare their primary structure. To compare conserved regions of the h-fibroin proteins, we aligned the n- and c-terminus each (without the signal peptide) in Geneious using the Muscle 3.8.425[75] plugin with a maximum of 1,000 iterations. We compared % pairwise identity and % of identical sites in Geneious. For each species, we used custom-made scripts to split the silk gene into repeat modules (https://github.com/AshlynPowell/silk-gene-visualization/tree/main). For schematic visualization of the primary structure, we generated a consensus sequence of all representative sequences of each repeat module by aligning these in Geneious using the Muscle 3.8.425 plugin with a maximum of

1,000 iterations. We used Expasy ProtParam[76](https://web.expasy.org/protparam/) to compute the molecular weight and the amino acid composition of each sequence. For comparison, we also calculated the amino acid composition of the four available Lepidoptera full-length h-fibroins [(silkworm *Bombyx mori*[11]; Indianmeal moth Plodia interpunctella,[32]; painted lady butterfly Vanessa carduii[37] and spindle ermine moth Yponomeuta cagnagella[47]] using ProtParam. The amino acid composition of the bagworm *Eumeta veriegata* was obtained from Kono et al. 2019.[48]