***Title:*** **Structural diversity of silk *heavy-chain fibroins* across major caddisfly (Trichoptera) clades**

***Author names and affiliations:***

Jacqueline Heckenhauer (jacqueline.heckenhauer@senckenberg.de)[1,2*], Russell J. Stewart (russell.stewart@utah.edu)[3], Blanca Ríos-Touma (briostouma@gmail.com)[4], Ashlyn Powell (ashlynpowell913@gmail.com)[5], Paul B. Frandsen (paul_frandsen@byu.edu)[1,5,6], Steffen U. Pauls (steffen.pauls@senckenberg.de)[1,2,7]

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

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

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

[6]Data Science Lab, Smithsonian Institution, Washington, DC, USA

[7]Institute for Insect Biotechnology, Justus-Liebig-University, Gießen, Germany

***Correspondence***

***Jacqueline Heckenhauer,*** jacqueline.heckenhauer@senckenberg.de; Phone: +49 (69) 7542-1811; LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG) & Department of Terrestrial Zoology, Senckenberg Research Institute and Natural History Museum Frankfurt, Senckenberganlage 25, 60325 Frankfurt, Germany

***Declarations of interest:*** none

## Highlights

- We report seven new complete sequences of the h-fibroin, the major silk protein, of caddisfly silk
- Primary structure of the h-fibroin shows conserved patterns
- Substantial variation occurs in repeat modules in clades with different silk usage
- Greater complexity in the h-fibroin in clades with more diverse silk usage
- This suggests evolutionary link between h-fibroin complexity and ecological diversity


## Abstract

Similar to their sister order Lepidoptera (moths and butterflies), the primarily aquatic larvae of Trichoptera (caddisflies) produce silk, which is used to build a wide variety of underwater architectures, such as cocoons, fixed retreats, and tube-cases. This diverse silk usage allows them to exploit a wide range of aquatic environments. Studies of the long and highly repetitive major silk gene, *heavy chain fibroin* (*h-fibroin*), are rare due to difficulties in sequence assembly. Only recently, new long-read sequencing techniques have been used to assemble high-quality, full-length sequences of this gene. Here, we report seven new gene and protein sequences of h-fibroin covering the breadth of the order Trichoptera, including the three major life history strategies characterized by different silk usage (fixed retreat-, cocoon- and tube case-making). When comparing the primary structure of the h-fibroin, we find conserved patterns, but also substantial diversity across groups. The signatures of conservation, such as conservation of N- and C-termini and characteristic repeating structural modules, suggest a common mechanism of silk formation and similar material properties across all three groups. The diversity in the number and order of the repeating modules, as well as the increasing structural complexity of h-fibroin in fixed retreat and tube-case building compared to cocoon-builders, may reflect differences in mechanical properties related to the diversity of silken architectures. This study presents a basis for future research into the role of fibroins in caddisfly silk function as a gauge of the potential of underwater silk in material applications.

**Keywords*:*** heavy chain fibroin, Insecta, arthropod genomics, Trichoptera, underwater silk, diversity, evolution, phylogeny

## 1. Introduction

In insects, labial glands, malpighian tubules, and a variety of dermal glands have evolved to produce silks (Sutherland et al., 2010). Silk production is widely distributed across arthropods and evolved multiple times in insects (Sutherland et al., 2010). Comparisons of silk produced across lineages of insects revealed common features which are thought to have evolved convergently and include the accumulation of silk proteins at high concentrations in specialized glands; similar composition of amino acids (high levels of glycine, alanine, or serine); and the folding of silk proteins into a single dominant secondary structure with high crystallinity, responsible for the mechanical properties (Sutherland et al., 2010).

Insects use silk for a variety of purposes and most research has focused on terrestrial Lepidoptera (moths and butterflies), especially on the commercially important silkworm *Bombyx mori.* The larvae of the latter build silken cocoons at the pupal stage. However, there are also aquatic insects that use silk, including the most specious aquatic insect order, Trichoptera (caddisflies), which exhibit the most diverse silk use. The properties of caddisfly silks are unique; they adhere underwater, have high tensile strength, extensibility, and toughness (Ashton et al., 2015, Ashton et al., 2019). These qualities make it of interest as potential biomimetic adhesives. Similar to their sister order Lepidoptera, several caddisfly species construct cocoons in the final larval instar. In larval stages, they produce a diverse array of underwater architectures which track their phylogeny. Trichoptera is divided into two suborders, Annulipalpia and Integripalpia (Thomas et al., 2020) which can be distinguished by differences in morphology, habitat, and use of silk. Larvae of Annulipalpia (fixed retreat makers) create stationary shelters of silk, often with mineral particles or plant material, which are fixed to the substrate (e.g., stones or aquatic plants). Larvae of most families of 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. 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 (Wiggins, 1978). Their diverse net- and case-making behavior allows caddisflies to exploit a range of ecological niches, which raises the question of how the properties of underwater silk evolved in Trichoptera. To answer this question, a comprehensive study of the basics, i.e. the primary molecular structure of the major silk protein, is necessary.

The silk fiber in Trichoptera consists of two filaments derived from a pair of labial glands. The fiber core is assembled from a large (200-500 kDa) heavy-chain fibroin (h-fibroin) and a smaller (~25 kDa) light chain protein. The h-fibroin is the major silk protein by size and mass. It consists of non-repetitive amino N- and C-terminal domains flanking a central region, composed of repeated structural modules (Yonemura et al., 2006; Ashton et al., 2013). These consist of repeating $(SX)_n$E motifs in which the S is often phosphorylated and where X is mostly amino acids with hydrophobic or aromatic side chains and sometimes Arginine, and n is 3-5. In $(SX)_n$E, E represents glutamic acid. The $(SX)_n$ E motifs are separated by Glycine(G)-rich regions of variable length (Stewart & Wang, 2010). The molecular structure of full-length *h-fibroin* sequences has not been studied extensively, due to challenges in the assembly of the *h-fibroin* caused by its repeated regions (Frandsen et al., 2019; Kawahara et al., 2022). Thus, in previous studies, the h-fibroin was derived from sequencing the ends of cDNAs resulting in partial sequences (Yonemura et al., 2006, 2009, Stewart & Wang, 2010). Two full-length h-fibroin sequences have been reported using a combination of long- and short-read sequencing techniques (Luo et al., 2018, Frandsen et al., 2019). Recently new genomic long-read sequencing technologies with low sequencing error rates (i.e. Pacific Biosciences HiFi) allow for full-length sequencing and assembly of *h-fibroin* gene sequences. However, only two full-length HiFi-derived *h-fibroin* sequences have been published to date, both for two case-making Trichoptera (Hotaling et al., 2022; Kawahara et al., 2022).

In this study, we used PacBio HiFi to sequence the full-length *h-fibroin* sequence of two caddisfly species, *Leptonema lineaticorne* OS Flint, 2008 (Annulipalpia, fixed retreat-maker) and *Micrasema longulum* R McLachlan, 1876 (Integripalpia, tube-case made of silk). In addition, we identified and

annotated complete protein-coding *h-fibroin* gene sequences from five high-quality genomes (four recently published by the Darwin Tree of Life Project and one from Ríos-Touma et al. (2022). This increases the number of high-quality, full-length h-fibroin sequences from two to nine. We used these full-length h-fibroin sequences to compare the primary structure of the major silk protein across three life history strategies of Trichoptera (fixed reteat-, cocoon- and tube case-making). Our study represents a first step in identifying structural features that reflect the evolution of Trichoptera and their clade-specific silk uses.

## 2. Material and methods

### *2.1 New genomic resources for two caddisfly species*

We generated PacBio HiFi reads for *M. longulum* (Integripalpia, family Brachycentridae) which builds a tube-case using only silk, and *L. lineaticorne* (Annulipalia, family Hydropsychidae) which uses silk to construct a fixed retreat and capture net. A single larva was collected for *M. longulum* in Germany, Hesse, Bieber (5009'39.7"N, 920'09.1"E), and a single adult individual was collected for *L. lineaticorne* in an Amazon Blackwater channel in Ecuador, Sucumbios, Sacha Lodge (Amazon Basin), Caño Anaconda (0º28'20.71"S; 76º27'59.08"W, elevation 237 m asl.). We extracted high molecular weight from single individuals 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 mode using 30-hour movie time. For *M. longulum*, genomic HiFi reads were generated by circular consensus sequencing (CCS) from the subreads.bam files provided by the sequencing facility using pbccs tool v6.6.0 in the pbbioconda package (https://github.com/PacificBiosciences/pbbioconda). For *L. lineaticorne* we generated HiFi reads from the raw data (reads with quality above Q20) using PacBio SMRTlink software (https://github.com/PacificBiosciences/pbbioconda). Genome size was estimated from sequencing reads using a *k-mer*-based statistical approach. After *k-mers* were counted with JELLYFISH v2.2.10 (Marçais and Kingsford, 2011) using jellyfish count -C -s 25556999998 -F 3 and a *k-mer* length of 21 (-m 21) with the ccs-reads, a histogram of *k-mer* frequencies was produced with jellyfish histo.

GenomeScope 2.0 (Ranallo-Benavidez et al., 2020) was run 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, Read length = 150, Max kmer coverage = 1000. We assembled the *L. lineaticorne* genome, with hifiasm (Cheng et al., 2021) with the default settings. Contaminations of adapters detected by NCBI were filtered out with samtools faidx (Li et al., 2009). We used flye v2.8.2 (Kolmogorov et al., 2019) with the –pacbio-hifi option to assemble the *M. longulum* genome. Assembly quality was evaluated based on continuity [QUAST v5.0.2 (Gurevich et al., 2013)] and completeness of Benchmarking Universal Single-Copy Orthologs (BUSCOs) with BUSCO 5.2.2 (Simão et al., 2015; Waterhouse et al., 2018) using the lineage dataset endopterygota_odb10 in genome mode.

*2.2 Heavy-chain fibroin gene identification and annotation*

Recently, the Wellcome Sanger Institute published four high-quality caddisfly genomes (*Glyphotaelius pellucidus*: GCA_936435175.1, GCA_936435185.1; *Limnephilus lunatus*: GCA_917563855.2, GCA_917563845.2; *Limnephilus marmoratus*: GCA_917880885.1, GCA_917880875.2; *Limnephilus rhombicus*: GCA_929108145.1, GCA_929113645.1) and one assembly was published from Ríos-Touma et al. (2022). We identified the *h-fibroin* genes from these assemblies by using tBLASTn to identify conserved N- and C-termini with query sequences from the termini of *Hesperophylax sp*. (Ashton et al., 2013), *Limnephilus decipiens* AB214509 (Yonemura et al., 2006) and *Rhyacophila obliterata* AB354689.1 and AB354588.1 (Yonemura et al., 2009) in Geneious Prime 2022.1.1 (https://www.geneious.com). We then extracted the sequences and 1,000 bps of flanking regions from the assembly using the sequence view “extract” in Geneious and annotated this region using Augustus v.3.3.2. (Stanke et al., 2008). Introns that did not affect reading frames were manually removed from the annotation. Protein coding nucleotide sequences were translated with Expasy Translate (https://web.expasy.org/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* (Frandsen et al., 2019).

Because the sequencing coverage and thus the assembly quality was low for *M. longulum*, we had to conduct several extra steps to identify the *h-fibroin.* With query sequences from the termini of *Hesperophylax sp*., we used tBLASTn to identify the conserved N- and C-termini and extracted the *h-fibroin* plus 500 bp flanking regions as described above. To assemble the error-free full-length sequence of the *h-fibroin*, we mapped all CCS reads back to the extracted region of the genome and used FreeBayes (Richter et al., 2020), a haplotype-based variant detector, to call variants and WhatsHap phase (Martin et al., 2016) to phase genomic variants. To reconstruct the primary haplotype from the phased VCF file, we used the bcftools consensus command. The resulting sequence was annotated with Augustus v.3.3.2 as described above.

Signal peptides and the location of their cleavage site of the h-fibroin protein sequences were predicted with the SignalP 6.0 server (https://services.healthtech.dtu.dk/service.php?SignalP, Teufel et al., 2022) using the following settings: organism = Eukarya, model mode = slow.

*2.3 Comparison of heavy-chain fibroins across and within caddisflies clades*

We used the previously published full-length h-fibroin sequences and the newly annotated *h-fibroin* sequences generated here to compare their primary structure. In order 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 (Edgar, 2004) plugin with a maximum of 1,000 iterations. We compared % pairwise identity and % of identical sites in Geneious. To visualize how the sequence similarity and divergence is reflected in the phylogeny of Trichoptera, we used the Geneious Tree Builder with the Jukes-Cantor genetic distance model and neighbor-Joining to construct phylogenetic trees based on the conserved terminal regions with *Bombyx mori* AF226688 (Zhou et al., 2000) as an outgroup. We used Expasy ProtParam (https://web.expasy.org/protparam/) to compute the molecular weight of each sequence. 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.

## 3. Results

### *3.1 New genomic resources*

Sequencing of *M. longulum* resulted in 751,710 ccs-sequences with a total of 9,746,879,982 bp. Due to low sequencing coverage, Genomscope2 did not converge. Based on a previous study (Heckenhauer et al., 2022), the genome size of *M. longulum* is ~590-650 Mbp. This means the sequencing coverage is 15x. The low sequencing coverage of *M. longulum* resulted in an assembly of poor contiguity (N50: 440,037 bp, 16,208 contigs) and was thus not pursued and not published at NCBI. Sequencing of *L. lineaticorne* resulted in 525,771 ccs-sequences with a total of 6,127,219,457 bp. Genomescope2 revealed a genome size of 249,938,531 bp (http://qb.cshl.edu/genomescope/genomescope2.0/analysis.php?code=kDbCJvi1CxiFht7B980V). This means the sequencing coverage was 24.5 x. To date, there are seven caddisfly HiFi genomes published at NCBI with an N50 of up to 32 Mb (*Eubasilissa regina*, (Kawahara et al., 2022). The *L. lineaticornis* assembly generated in this study is of similarly high quality. It has a total length of 273,010,394 bp, 65 contigs, a N50 of 13,827,090 and a L50 of 9. There were 96.1% complete (single: 95.3%, duplicated:0.8%), 2.3% fragmented, and 1.6% missing BUSCOS. The sequencing coverage was too low to assemble a highly contiguous genome of *M. longulum*. However, we were still able to manually assemble the *h-fibroin* allele using WhatsHap (Martin et al., 2016), rendering the species' sequences useful for the presented comparison of *h-fibroin* genes. Recent studies have shown that HiFi reads allow for assembly of the repetitive ~20-kb *h-fibroin* gene (Hotaling et al., 2022, Kawahara et al., 2022). We could identify the *h-fibroin* in the assemblies of seven additional caddisfly species. Our final taxon sampling thus covers the three main life history strategies of Trichoptera: fixed retreat-making Annulipalpia (N=1), cocoon-making basal Integripalpia (N=1), and tube-case-making Integripalpia (N=7). Full-length HiFi-sequencing derived *h-fibroin* sequences have not been published for fixed retreat- or cocoon-making species before and therefore we provide important resources in

this study that allowed us to characterize and compare the *h-fibroin* gene and h-fibroin protein sequences across the main clades and different ecologies, i.e. different silk usage, of caddisflies.

### *3.2 Gene structure of h-fibroin*

The comparison of the structure of the *h-fibroin* gene across the nine species revealed a similar organization of introns and exons. The *h-fibroin* gene is characterized by a short exon (39-42 bp) followed by a single intron (92-1036 bp) and a long second exon (12,709-29,508 bp) leading to a total length of 20,538-30,382 bp. *M. longulum* has an additional intron and exon (see Table 1).

### *3.3 Primary structure of* h-fibroin

#### *3.3.1. Conserved termini*

H-fibroin consists of non-repetitive amino N- and C-termini which are highly conserved across the different clades. The N-terminus contains 109-115 residues (without signal peptide). There are 44.9% identical sites and 80% pairwise identity between the species . The C-terminus consists of 40 residues and exhibits 40% identical sites and 73.8% pairwise identity. The Neighbor-Joining tree based on the N-termini is congruent with existing phylogenetic trees in which the fixed retreat makers are basal to cocoon and tube-case makers (Thomas et al., 2020), whereas the Neighbor-Joining tree based on the C-terminus varies with the cocoon-making species being basal to fixed retreat and tube-case makers.

#### *3.3.2 Repeating structural modules*

The terminal domains flank 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 ( Yonemura et al., 2006, Wang et al., 2010; Ohkawa et al., 2013). For example, Frandsen et al. (2019) 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 (Ashton, 2013). In presenting the new h-fibroin sequences reported here, we have defined the repeating modules as each having two parts: an N-terminal region comprising a variable number (1-6) of $(SX)_nE$ motifs, each separated by short (8-24) stretches of intervening amino acids, and a variable length G-rich C-terminal region.

Schematically, the two parts are represented by different symbols in each block (Fig. 2-8). The following paragraphs describe the repeating structural modules for each species analyzed in this study in detail.

*3.3.2.1 Basal Integripalpia (cocoon makers)*

The structure of the repetitive central region in free-living, pupal cocoon-building *A. davidsoni* is simple compared to the other sequences investigated in this study (Figure 2). The repeating module is embedded between two transition regions (TR), each of which occurs once and directly flanks the termini. The TRs have similarities to the repeating module but their sequence is unique, e.g. they occur only one time in the h-fibroin and are in most cases transition between the conserved termini and the repeating structural modules (Figures 2-7) or between repeat modules (Figure 8). There is only one repeating module (RM1_AD) consisting of a $(SX)_4E[15](SX)_3E$ region and a G-rich region variable in length. The number in the square bracket refers to the short stretches of intervening amino acids. The repeating module RM1_AD occurs 81 times.

*3.3.2.2 Integripalpia (tube-case makers)*

The primary structure of the h-fibroin is variable within tube-case-making Integripalpia. In all seven species investigated the repetitive central region is flanked by two transition regions (TR). However, the number and organization of repeat modules are diverse within this clade.

Specifically, in *E. regina* which used plant material (leaves) and silk to build a tube case, the h-fibroin has 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 G-rich region (Figure 3). It is repeated 39 times.

In *G. pellucidula* whose silk use is similar to *E. regina*, the h-fibroin consists of two repeat modules (RM1_GP, RM2_GP, Figure 4). RM1_GP consists of a $(SX)_4E[13](SX)_3E$ motif and G-rich region, and RM2_GP consists of a single $(SX)_5E$ motif and G-rich region. Each RM is repeated 70.

Within the three *Limnephilus* species, the primary structure of the h-fibroin is comparable (Figure 5). This genus uses a diverse array of plant materials (wood, moss, leaves) for tube-case-making. There

are two repeat modules (RM1_L, RM2_L). In *L. lunatus*, RM1_L consists of a $(SX)_5E[14](SX)_4E[11](SX)_4E$ region and a G-rich region. It occurs 96 times and is interrupted by ten RM2_L which consist of a single $(SX)_5E$ motif and a G-rich region, similar to *G. pellucidula*. In *L. marmoratus*, RM1_L occurs 68 and RM2_L 17 times. In *L. rhombicus* RM1_L occurs 59 and RM2_L 51 times. The h-fibroin of *H. magnus*, which uses stones the build tube-cases, is constituted by two repeat modules (RM_HM, RM2_HM, Figure 6). RM1_HM contains a $(SX)_5E[14](SX)_4E[11](SX)_4E$ motif and a G-rich region and occurs 69 times and is interrupted by 13 RM2_HM which consist of a single $(SX)_5E$ motif and a G-rich region.

In the h-fibroin of *M. longulum* which constructs tube-cases purely made of silk, there are three repeat modules (RM1_ML, RM2_ML, RM3_ML, Figure 7). RM1_ML contains a $(SX)_4E[18](SX)_3E[11](SX)_4E[15](SX)_4E$ motif and a G-rich region and is repeated 24 times. RM2_ML consists of a $(SX(_4E[18](SX)_3E[11](SX)_4E[18](SX)_3E[11](SX)_4E[14](SX)_4E$ motif and a G-rich region which occurs 7 times. RM3_ML consists of a $(SX)_4E$ motif and G-rich region which occurs 36 times.

*3.3.2.3 Annulipalpia (fixed retreat makers)*

The structure of the repetitive central region of fixed retreat-making *L. lineaticorne* is more complex compared to free-living *A. davidsoni*. The sequence is divided into two structural units, i.e. repeating modules (Figure 8, RM1_LL, RM2_LL) which almost look like two different proteins fused together. RM1_LL consists of a $(SX)_4E[13](SX)_4E[11](SX)_4E[24](SX)_4E$ motif and G-rich region and is repeated 14 times. RM2_LL consists 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 G-rich region and is repeated 13 times. There are two transition regions. In these, the sequence transitions from one type of module to another. TR1 separates the two repeat modules and TR2 is located between RM2_LL and the C-terminus.

## 4. Discussion

In this study, we report new genomic resources for the order Trichoptera. This is important since aquatic insects have been underrepresented in genomics research (Hotaling et al., 2020, 2021a, 2021b). For the first time, we provide long high-fidelity sequences of a fixed retreat-making caddisfly species. In addition, we report seven new full-length h-fibroins from *de novo* assembled HiFi data. This data was either provided by the authors of this study or large genome sequencing initiatives (e.g. Darwin Tree of Life, European Reference Genome Atlas) showing the importance of these projects for the 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 across the main clades of Trichoptera, which are represented by species with different silk usage.

The genetic structure of the major silk gene, *h-fibroin*, of the nine Trichoptera species show similar organization of introns and exons, except for *M. longulum* which has a second intron splitting the second exon into two separate exons. Recently, the first HiFi-derived full-length *h-fibroin* of a representative of mainly terrestrial sister order Lepidoptera has been published (Kawahara et al., 2022). The genetic structure of this *h-fibroin* is similar (first exon: 42 bp, intron: 720 bp, second exon: 13,103 bp) which is interesting given the fact that the divergence of both orders occurred approx. 290 million years ago (Thomas et al., 2020). Further studies with more representatives of Lepidoptera and Trichoptera are needed to confirm this finding.

In addition to the genetic structure, we compared the primary protein structure of the h-fibroins. In particular, the h-fibroin protein has been shown to be both markedly conserved in its structure while also harboring substantial diversity across the various caddisfly clades.

We find, consistent with previous research (Yonemura et al., 2006; Frandsen et al., 2019), that structural elements of h-fibroin are conserved across all of the species that we sequenced. For example, of the species sampled the N- and C-termini exhibit a high pairwise identity (N-terminus: 73.8%, C-terminus: 80%), suggesting conservation of function of terminal regions. In addition, some conserved themes emerge in the central repetitive region, which consists of repeating two-part structural modules, each containing a characteristic region of $(SX)_n$E motifs interspersed with G-rich

regions of variable length. The serines of the $(SX)_nE$ motifs have been shown to be extensively phosphorylated (Stewart and Wang, 2010; Addison et al., 2013; Ohkawa et al., 2013). The phosphates bind multivalent metal ions, which stabilize the silk and are responsible for the strength of the silk (Ashton et al., 2013; Ashton and Stewart, 2015, Ashton et al., 2019). A structural model has been proposed, in which each (pSX)nE motif, where pS is phosphoserine, forms a ß-strand, which in turn associates into anti-parallel Ca2+-stabilized ß-sheets (Ashton et al., 2013). The ß-sheets stack through alternating hydrophobic and Ca2+-phosphate interfaces creating microcrystalline ß-domains. By this model, each repeating module would correspond to a structural module in which anti-parallel $[(pSX)_nE]_m$ß-sheets, where m refers to the number of $(SX)_nE$ motifs, are connected to a G-rich spacer region. The $(SX)_nE$/G-rich blocks may combine to form a higher-order ß-domain structure through intra- or intermolecular stacking of the $[(SX)_nE]_m$ ß-sheets. The crystalline ß-domains would be separated by flexible and extensible G-rich regions. These signatures of conservation in repeat modules despite 280 million years of divergence (Thomas et al., 2020) 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 Frandsen et al. (2019).

Despite the reported conservation, wide variation in the ordering and number of repeat modules are observed. Our study builds on previous studies (i.e. Yonemura et al., 2006; Wang et al., 2010; Ashton et al., 2013; Ohkawa et al., 2013; Frandsen et al., 2019) by unveiling substantial diversity in the number and order of these repetitive structures. In the h-fibroin structure from *A.davidsoni*, the single type of structural module was repeated 81 times (Fig. 2). However, elsewhere across the phylogeny, the repeat module structure complexity scales with silk usage diversity. For example, in fixed retreat-making *L. lineaticorne*, the sequence is divided into two repeating modules (Figure 8) which resemble two different proteins fused together. The sequence of tube-case-making *M. longulum* has three repeat motifs which are shuffled throughout the central region of the h-fibroin (Figure 7). The variations in repeating block patterns between the different clades may reflect adaptations for the diverse silk use. For example, *Atopsyche davidsoni* is free-living as a larva and only produces pupal silk for building cocoons. Current phylogenetic evidence suggests that the common ancestor of all Trichoptera was

free-living and created a pupal cocoon within a protective stone dome and the silk usage behavior diversified from this form (Frandsen et al., in preparation). Perhaps the patterns that we observe with increasing complexity in the h-fibroin diversified with caddisfly silk usage. The diversity in the number and order of the repeat modules may also hold clues to unraveling the unique applications of silk across clades (e.g., case-making vs. fixed retreat-building) and may be directly responsible for differences in mechanical properties.

The next step to a better understanding of the role of the variation across fibroins in caddisflies is to follow up sequencing studies, such as this one, with experimental evidence, such as mechanical testing and developmental research. Regardless, these new sequences provide the basis for future research into the role of fibroins in silk function and are essential to the eventual development of caddisfly silk in material science.

**Author contributions**

J. Heckenhauer: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation; Visualization, Writing - original draft; Writing - review & editing

R. J. Stewart: Conceptualization, Investigation, Methodology, Validation; Writing - original draft; Writing - review & editing

P. B. Frandsen: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Writing - original draft; Writing - review & editing

B. Ríos-Touma: Resources, Writing - review & editing

A. Powell: Software; Writing - review & editing

S. U. Pauls: Conceptualization, Funding acquisition, Project administration; Resources; Supervision; Writing - review & editing

**Funding**

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) that supported J.H. and S.U.P.

P. B. F. received internal funding from Brigham Young University, College of Life Science for sequencing the *Leptonema lineaticorne* genome.

**Acknowledgements**

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.19.02 ("Mountain Freshwater Diversity, from Taxonomy to Functional Genomics, and Approximation from Trichoptera"). We appreciate help from Sacha Lodge for field logistics.

**Availability of source code**

All custom-made scripts used in this study are available on GitHub.

Project name: h-fibroin-visual

Project home page: https://github.com/AshlynPowell/h-fibroin-visual

**Data Availability**

Raw sequence data, genome assembly (SUB10837191) and sample information (SAMN25408291) of *L. lineaticorne* has been submitted to NCBI under Bioproject PRJNA802365. Raw sequence data (SRR15840268) and sample information (SAMN20982427) of *M. longulum* are available from NCBI under Bioproject PRJNA558902. Full-length *H-fibroin* gene sequences including introns, *h-fibroin* protein coding sequences and h-fibroin protein sequences newly identified in this paper are available at Figshare Project: https://doi.org/10.6084/m9.figshare.20407101.v1

**References**

Addison, J.B., Ashton, N.N., Weber, W.S., Stewart, R.J., Holland, G.P., 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

Ashton, N.N., Roe, D.R., Weiss, R.B., Cheatham, T.E., 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

Ashton, N.N., 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

Ashton, N.N., Pan, H., Stewart, R.J., 2016. Connecting caddisworm silk structure and mechanical properties: combined infrared spectroscopy and mechanical analysis. Open Biol. 6, 160067. https://doi.org/10.1098/rsob.160067

Ashton, N.N. and R.J. Stewart, 2019. Aquatic caddisworm silk is solidified by environmental metal ions during the natural fiber-spinning process. FASEB J 33(1), 572-583.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., 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

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

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

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: Quality assessment tool for genome assemblies. Bioinforma. Oxf. Engl. 29. https://doi.org/10.1093/bioinformatics/btt086

Heckenhauer, J., Frandsen, P.B., Sproul, J.S., Li, Z., Paule, J., Larracuente, A.M., Maughan, P.J., Barker, M.S., Schneider, J.V., Stewart, R.J., Pauls, S.U., 2022. Genome size evolution in the

diverse insect order Trichoptera. GigaScience 11, giac011. https://doi.org/10.1093/gigascience/giac011

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

Hotaling, S., Kelley, J.L., Frandsen, P.B., 2021a. Toward a genome sequence for every animal: Where are we now? Proc. Natl. Acad. Sci. 118, e2109019118. https://doi.org/10.1073/pnas.2109019118

Hotaling, S., Sproul, J.S., Heckenhauer, J., Powell, A., Larracuente, A.M., Pauls, S.U., Kelley, J.L., Frandsen, P.B., 2021b. Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. Genome Biol. Evol. 13, evab138. https://doi.org/10.1093/gbe/evab138

Hotaling, S., Wilcox, E.R., Heckenhauer, J., Stewart, R.J., Frandsen, P.B., 2022. Realizing the promise of biodiversity genomics with highly accurate long reads. https://doi.org/10.1101/2022.07.10.499467

Kawahara, A.Y., Storer, C.G., Markee, A., Heckenhauer, J., Powell, A., Plotkin, D., Hotaling, S., Cleland, T.P., Dikow, R.B., Dikow, T., Kuranishi, R.B., Messcher, R., Pauls, S.U., Stewart, R.J., Tojo, K., Frandsen, P.B., Storer, C.G., Markee, A., Heckenhauer, J., Powell, A., Plotkin, D., Hotaling, S., Cleland, T.P., Dikow, R.B., Dikow, T., Kuranishi, R.B., Messcher, R., Pauls, S.U., Stewart, R.J., Tojo, K., Frandsen, P.B., 2022. Long-read HiFi sequencing correctly assembles repetitive heavy fibroin silk genes in new moth and caddisfly genomes. Gigabyte 2022, 1–14. https://doi.org/10.46471/gigabyte.64

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540–546. https://doi.org/10.1038/s41587-019-0072-8

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., 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

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

Marçais, G., 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

Martin, M., Patterson, M., Garg, S., Fischer, S.O., Pisanti, N., Klau, G.W., Schöenhuth, A., Marschall, T., 2016. WhatsHap: fast and accurate read-based phasing. https://doi.org/10.1101/085050

Ohkawa, K., Miura, Y., Nomura, T., Arai, R., Abe, K., Tsukada, M., 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

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

Richter, F., Morton, S.U., Qi, H., Kitaygorodsky, A., Wang, J., Homsy, J., DePalma, S., Patel, N., Gelb, B.D., Seidman, J.G., Seidman, C.E., Shen, Y., 2020. Whole Genome De Novo Variant Identification with FreeBayes and Neural Network Approaches. https://doi.org/10.1101/2020.03.24.994160

Ríos-Touma, B., Holzenthal, R.W., Rázuri-Gonzales, E., Heckenhauer, J., Pauls, S.U., Storer, C.G., 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

Simão, F., Waterhouse, R., Ioannidis, P., Zdobnov, E., 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. 31. https://doi.org/10.1093/bioinformatics/btv351

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stewart, R.J., 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

Sutherland, T.D., Young, J.H., Weisman, S., Hayashi, C.Y., 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

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., 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

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

Wang, Y., Wang, H., Zhao, T., 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

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., 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

Wiggins, G.B., Mackay, R.J. 1978. Some relationships between systematics and trophic ecology in nearctic aquatic insects, with special reference to trichoptera. Ecology 59(6):1211–20.

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

Yonemura, N., Mita, K., Tamura, T., 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

Zhou, C.Z., Confalonieri, F., Medina, N., Zivanovic, Y., Esnault, C., Yang, T., Jacquet, M., Janin, J., Duguet, M., Perasso, R., 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

**Figure legends**

Figure 1: Conserved regions of the *h-fibroin*. A: N-terminus without signal peptide. B: C-terminus. Phylogenetic trees were reconstructed using Geneious Tree Builder with the Genetic distance model Jukes-Cantor and Tree Builder Method Neighbor-Joining.*from this study, *B. mori* (Lepidoptera) was used as an outgroup.

Figure 2: Schematic visualization of the *h-fibroin*e of *Atopsyche davidsoni* (7870 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_AD: Repeat module consisting of $(SX)_4E[15](SX)_3E$ motif (red circles) and G-rich region (grey-blue triangle), is repeated 81 times. The G-rich region varies in length. The consensus sequence of RM1_AD is given.

Figure 3: Schematic visualization of the primary allele of the *h-fibroin* gene of *Eubasilissa regina* (8386 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_ER: Repeat module 1 consists of a $(SX)_4E[8](SX)_4E[17](SX)_4E[11](SX)_3E[20](SX)_3E$ (red circles) motif and a G-rich region (grey-blue triangle) and occurs 39. The G-rich region varies in length. The consensus sequence is given for each repeat module.

Figure 4: Schematic visualization of the primary allele of the *h-fibroin* gene of *Glyphotaelius pellucidus* (7864 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_GP: Repeat module 1 consists of a $(SX)_4E[13](SX)_3E$ (red circles) motif and a G-rich region (grey-blue triangle) and occurs 70 times. RM2_GP: Repeat module 2 consists of a $(SX)_5E$ (orange circles) motif and a G-rich region (blue triangle) and occurs 70 times. The G-rich region varies in length. The consensus sequence is given for each repeat module.

Figure 5: Schematic visualization of the h-fibroin gene of *Limnephilus marmoratus* (8775 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_L, consisting of $(SX)_5E[14](SX)_4E[11](SX)_4E$ (red circles) motif and a G-rich region (grey-blue triangle), occurs 96 times. RM2_L: Repeat module 2 consists of a $(SX)_5E$ motif (orange circles) and a G-rich region (blue triangles) and occurs 10 times. The first $(SX)_5E$ is missing in three RM1 module (white asterisk). The G-rich region varies in length for RM1 and RM2. The consensus sequence is given for each repeat module. The sequence is very similar to the other two *Limnephilus* species.

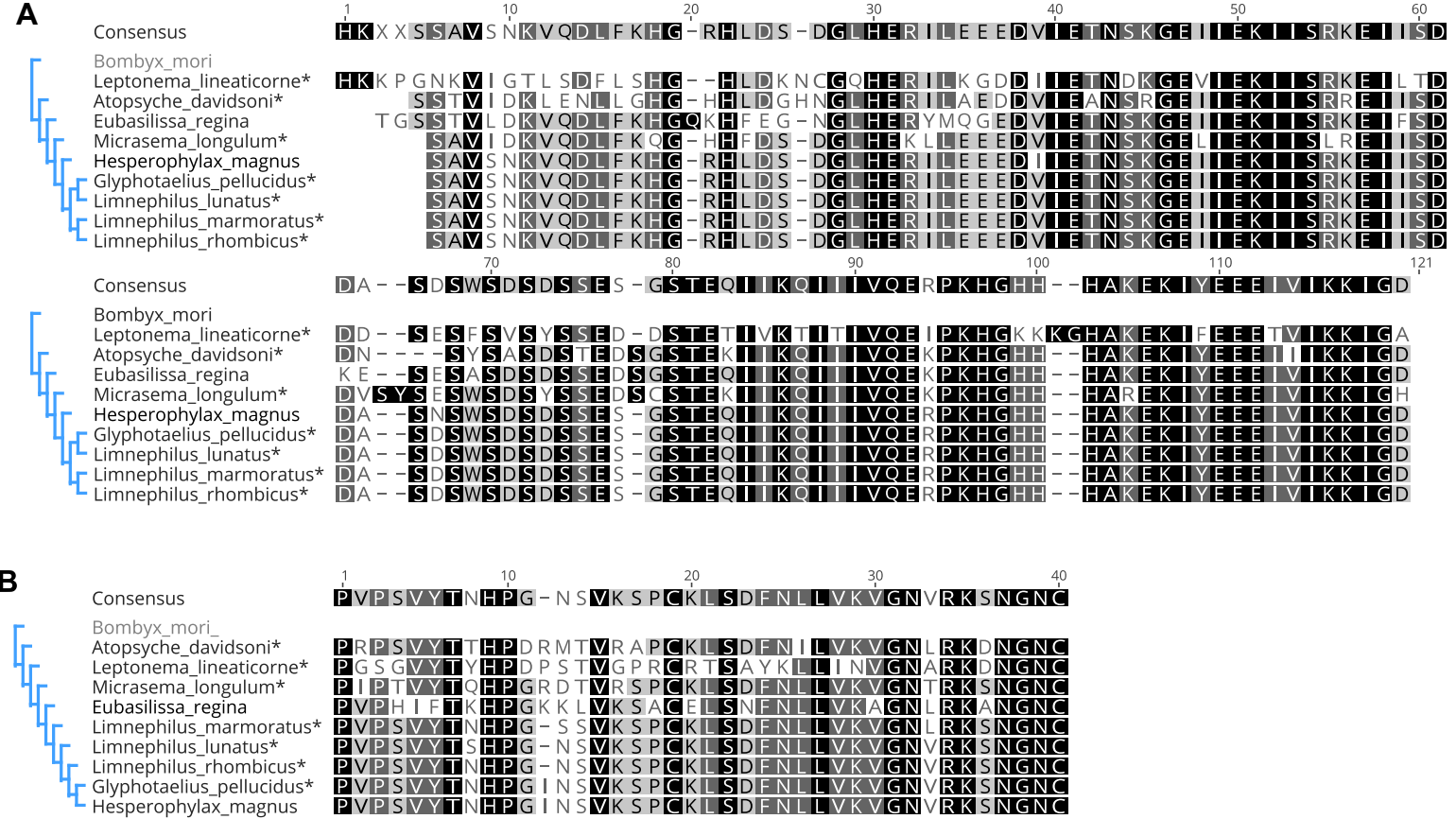
Figure 6: Schematic visualization of the *h-fibroin* gene of *Hesperophylax magnus* (8624 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_HM: Repeat module 1, consists of a $(SX)_5E[14](SX)_4E[11](SX)_4E$ (red circles) motif and a G-rich region (grey-blue triangle) and occurs 69 times. RM1_HM: Repeat module 2 consists of a single $(SX)_5E$ motif (orange circles) and a G-rich region (blue triangles) and occurs 13 times. The first $(SX)_5E$ is missing in three RM1_HM modules (asterisk). The G-rich region varies in length for RM1_HM and RM1_HM. The consensus sequence is given for each repeat module.

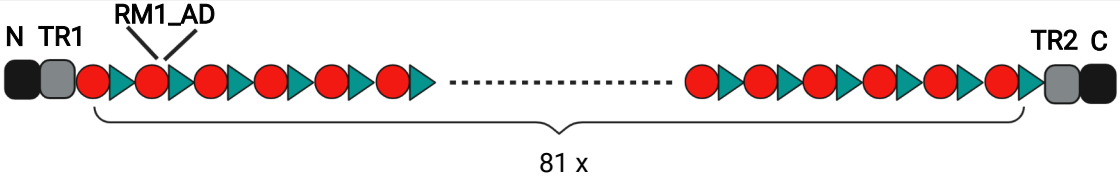
Figure 7: Schematic visualization of the *h-fibroin* gene of *Micrasema longulum* (7209 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_ML: Repeat module 1, consists of a $(SX)_4E[18](SX)_3E[11](SX)_4E[14](SX)_4E$ motif (red circles) and a G-rich region (grey-blue triangle) and is repeated 24 times. The first $(SX)_4E$ is missing in two RM1_ML module (white asterisk). RM2_ML: Repeat module 2 consists of a

$(SX)_4E[18](SX)_3E[11](SX)_4E[18](SX)_3E[11](SX)_4E[14](SX)_4E$ motif (orange circle) and a G-rich region (blue triangle). It occurs 7 times. RM3_ML: Repeat module 3, consisting of a $(SX)_4E$ motif (purple circle) and G-rich region (green triangle). It occurs 36 times. The G-rich region varies in length for RM1_ML, RM2_ML, and RM3_ML.

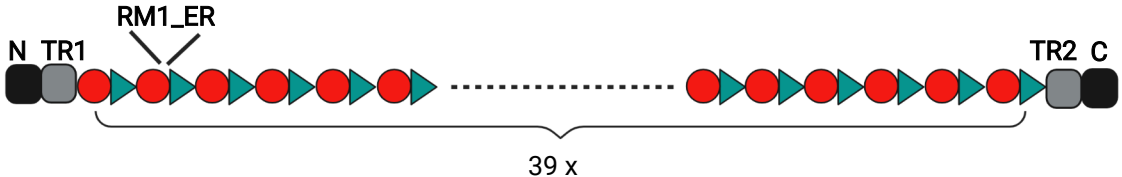
Figure 8: Schematic visualization of the *h-fibroin* gene of *Leptonema lineaticornis* (6818 residues). Black boxes: N- and C-terminus, signal peptide is underlined; Grey boxes: transition regions (TR); RM1_LL: Repeat module 1, consisting of a $(SX)_4E[13](SX)_4E[11](SX)_4E[24](SX)_4E$ motif (red circles) and a G-rich region(grey-blue triangle), is repeated 14 times. RM2_LL: Repeat module 2, consisting of a $(SX)_4E[14](SX)_4E[11](SX)_4E[19](SX)_3E[11](SX)_4E[24](SX)_4E[23](SX)_4E$ motif (orange circles) and a G- rich region (dark blue triangle), is repeated 13 times.The G-rich regions vary in length. The consensus sequence of each RM is given; *: In two of the RM2_LL, the $(SX)_4E$ motifs are reduced.

**Table 1:** Full-length h-fibroins of nine caddisfly species derived from PacBio HiFi sequencing ordered by suborder and silk use. Gene length including introns and stop codon; CDS= protein coding DNA, gene length, CDS, exon, intron in bp; protein size in amino acids

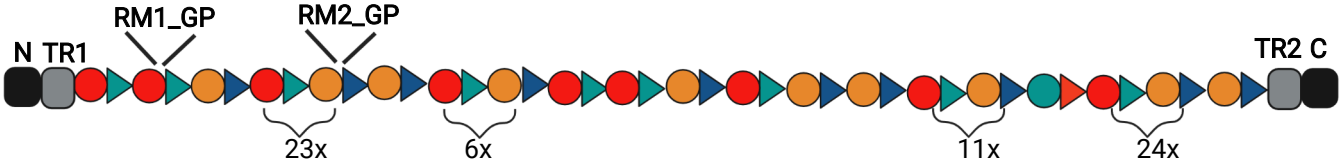

A
Consensus
Bombyx_mori
Leptonema_lineaticorne*
Atopsyche_davidsoni*
Eubasilissa_regina
Micrasema_longulum*
Hesperophylax_magnus
Glyphotaelius_pellucidus*
Limnephilus_lunatus*
Limnephilus_marmoratus*
Limnephilus_rhombicus*
B
Consensus
Bombyx_mori_
Atopsyche_davidsoni*
Leptonema_lineaticorne*
Micrasema_longulum*
Eubasilissa_regina
Limnephilus_marmoratus*
Limnephilus_lunatus*
Limnephilus_rhombicus*
Glyphotaelius_pellucidus*
Hesperophylax_magnus




```
N:       MRASTLLILFCALQIHITGSSTVIDKLENLLGHGHHLDGHNGLHERILAEDDVIEANSRGEIIEKIISRREIISDDN
TR1:     SYSASDSTEDSGSTEKIIKQIIIVQEKPKHGHHHAKEKIYEEEIIIKKIGDLPKHHIEVSKSISGEKRRRHGKSKHLPKSGLGHGVGGLGGIGGNKIWGPKVRRG
RM1_AD:  SVSASYSVEIQQIVTPTVITNIDISGSVSIEGGRRGHGLGGLGLGHGLGHGLGHGLGGLGHGGWGSSGDHSGIGGLGGLGGLGHLGGIGGRRGWGPKGRRG (81)
TR2:     SASASYSVEIQHIVTPTVITDIEISGSVSVETGRGGLGNFGALSRLGGGI
C:       PRPSVYTTHPDRMTVRAPCKLSDFNILVKVGNLRKDNGNC
```




```
N:       MRASFLLIVLCYWQIHATGSSTVLDKVQDLFKHGQKHFEGNGLHERYMQGEDVIETNSKGEIIEKIISRKEIFSDKESES
         ASDSDSSEDSGSTEQIIKQIIIVQEKPKHGHHHAKEKIYEEEIVIKKIGDV

TR1:     KHGTKSSVEKKCSGKKRKWGGRGSKTSVSVSVSVSVERVVGPTVLTRVRKGPRISVSRSVSIERVVQPGVRSKVSVSVSVEG
         HHGRRGRGRGPRGHGGRSVSVSVSVSVEGHHGPRGHGKRPWGHGKGPLGSLSGLGGSGDLGGLGLSGLGGHKGHGKGPWGPG
         KGRRGGSVSISVSVEGRRRHPRGSVSISVSVERGVGPTVLTSIRKGPRISVSQSVSIERIVRPGAWSKVSVSVSVEGHRG
         PRRHGGHSVSVSVSVEGHRGPRRHGYGPLGYGKGPLGSLSGLGGSGDLGGLGLGGLGGHGRHGKGPWGLGKG

RM1_ER:RGGSISISVSVEGRRRHPRGSVSISVSVERAVGPTVLTSIRKGPRISVSRSVSIERIVRPGAWSKVSVSVSVEGRHGPRR
         HGYGPRRYGGSSVSVSVSVEGHHGPRGHGYGPLGHGKGPLGSLGGLGGSGDLGGLGGLGGLGGLGGLGGHGKGPHGHGKG
         PHGGRSGSGDYDGLGGLGGLGGLGGLGGLGGHGGHGKGPLGLGRGR (39)

TR2:     PRISVSRSVSIERVVRPGVWSKVSLSVSVEGHHGGLGGILSHGKRPHRGVSRSVSIESGRRWGGRRGPSIAISRSVSVER
         VIRPGAWSLTSGSLSVSVEGIRQPSRRSEGVDISVSSSLESIRPGRVLGGLGRVGALGRSGV

C:       PVPHIFTKHPGKKLVKSACELSNFNLLVKAGNLRKANGNC
```




**N:** MRAILLLIVFCSWQVHRTGSSAVSNKVQDLFKHGRHLDSDGLHERILEEEDVIETNSKGEIIEKIISRKEIISDDAS
DSWSDSDSSESGSTEQIIKQIIIVQERPKHGH

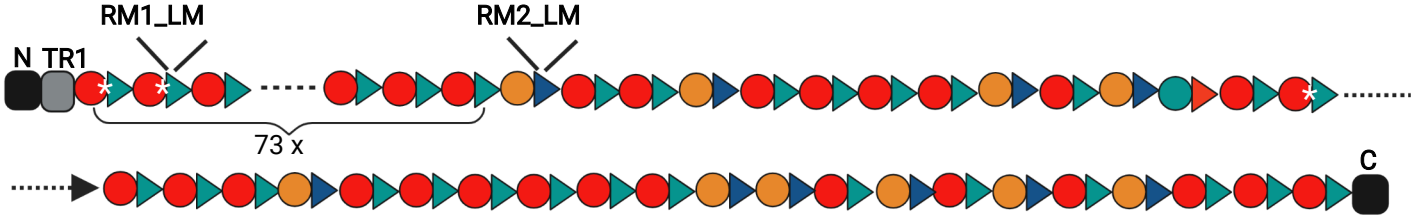
**TR1:** HHAKEKIYEEEIVIKKIGDVVHGVDVPLSVDGGLRLGGHKGRRGPKLSVSKSISIEQVVRPSVYTKISQSSSVSVEG
GRRSGPWGQGKGAWGSLGGLSGSGDLDGLGGIGGLGGRRGPWGPKGKV

**RM1_GP:** SISRSVSIERIVTPGVYTKITRSSSVSVEGGRRRGPWGGRGLGGLSGSGDLDVLGGLGGRRGPWGRRG (70)

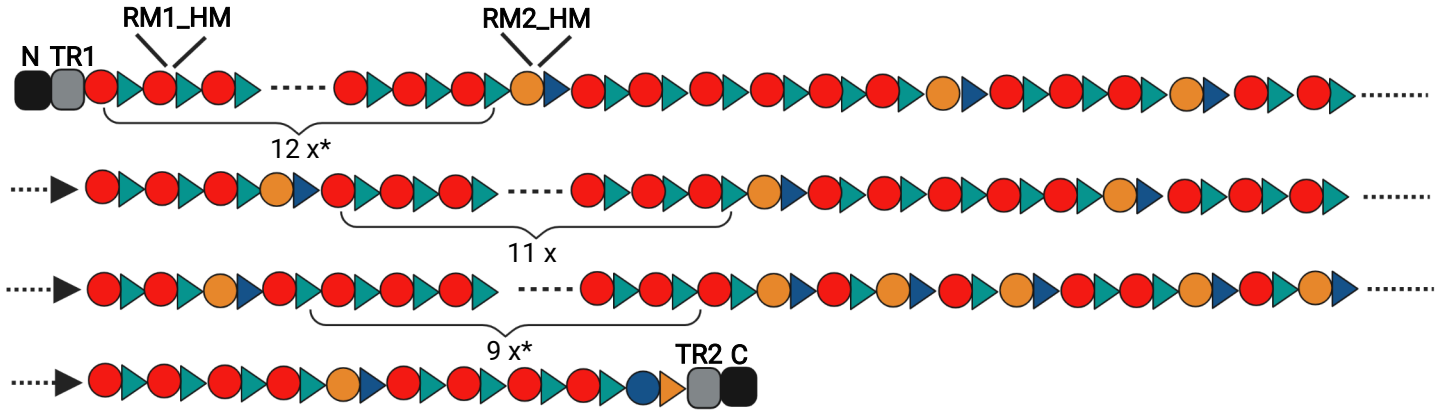
**RM2_GP:** SGSVSVSVSVERGIRRGPWGRRG(70)

**TR2:** KVSISRSVSIERIVTPGVYTKISRSSSVSVEGGIRRGPWGRFGGSALSGSSSVESVRPGRVLGGLGRVGVLGRSGVV

**C:** PVPSVYTNHPGINSVKSPCKLSDFNLLVKVGNVRKSNGNC




```
N:       MRAILLLIVFCSWQVHRTGSSAVSNKVQDLFKHGRHLDSDGLHERILEEEDVIETNSKGEIIEKIISRKEIISDDASDSWSDSDSSESGSTEQIIKQIIIVQERP
         KHGHHHAKEKIYEEEIVIKKIGDV

TR1:     IHGVDVPLSIDGGLRRGGHSGRRGPKLSISKSISIEQVVRPSVYTKISRSSSVSVEGGRQRGPWGKGKGAWGSLGGLSGSGDLDSLGGLGGLGGLRGPWGRRGKV

RM1_LM:  SGSVSVSVSVEGGRRRGPWGRRGKVSISRSVSIERIVTPGVYTKISRSSSVSVEGGRRRGPWGYGRGLGGLSGSGDLDGLGGVGGLGGLGGRRGPWGRGYGS (96)

RM2_LM:  SGSASVSVSVEGGRRRGPWGRGYGS (10)

C:       SRSSSIESVRPGRILAGLGRVGALGRSGVVPVPSVYTSHPGNSVKSPCKLSDFNLLVKVGNVRKSNGNC
```




**N:** MRAILLLIVFCSWQVHRTGSSAVSNKVQDLFKHGRHLDSDGLHERILEEEDIIETNSKGEIIEKIISRKEIISDDASNSWSDSDSSESGSTEQIIKQIIIVQ
ERPKHGHHHAKEKIYEEEIVIKKIGDV

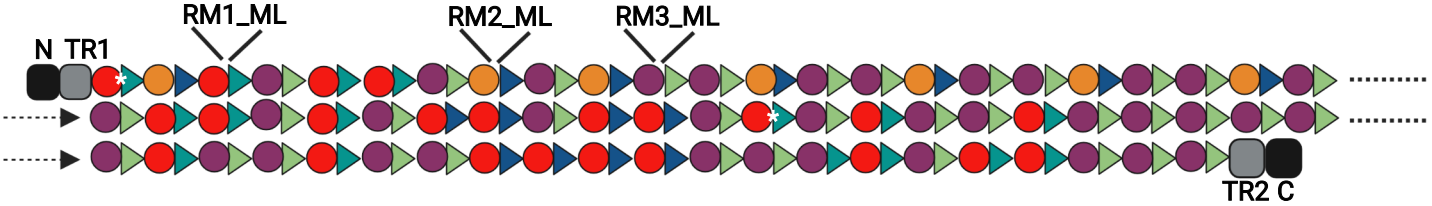
**TR1:** VHGVDVPITTDGGIRRIGHGGRRGPNISISRSISIEQVVRPRVYTKISRSSSVSVEGVRRHGPWGHGKGAWGPLGDLSSSGDLDGLGGLGSLGGRRGPWGRRG
KVSISRSVSIERIVTPGVYTKISRSSSVSVEGGRRPWGYGRGIGGLSGSGDLDGLGGLGGLGGYGGLGGRRGPWGRRGKV

**RM1_HM:** SVSVSLSVSVERGIRRGPWGRRGKVSISRSVSIERIVTPGVYTKISRSSSVSVEGGRRPWGYGRGLGGLSGSGDYDGLGGYGLGGYGGLGGYGGLGGRRGPWG
RRGXXGLSGSGDLDGLDGYGGLGGYGGLGGRRGPWGR (69)

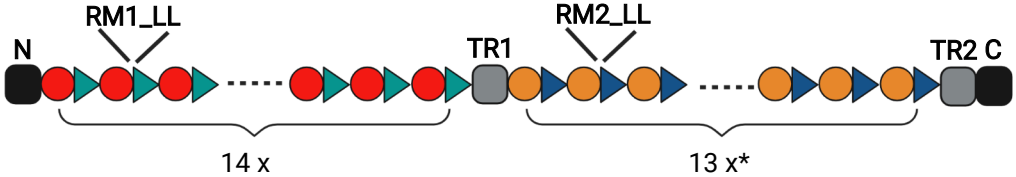
**RM2_HM:** DVSVSLSVSVEGGRRPWGYGRGLGGLSGSGDLDGLGGLGGLGGYGGLGGRRGPWGR (13)

**TR2:** SVSVSLSVSVERGIRRGPWGRRGKVSISRSVSIERIVTPGVYSKISRSSSVSVEGGIRRGPWGSFGGYAVSASSSSEYVRPGRVLAGLGRVGALGRSGAV

**C:** PVPSVYTNHPGINSVKSPCKLSDFNLLVKVGNVRKSNGNC




```
N:       MRAILLLIVFCSWQVHTTGSSAVIDKVQDLFKQGHHFDSDGLHEKLLEEEDVIETNSKGELIEKII
         SLREIISDDVSYSESWSDSYSSEDSCSTEKIIKQIIIVQEKPKHGHHHAREKIYEEEIVIKKIGHA

TR1:     NHGVNAISGHGLSGSVLGTKASYSASIESGRRGGYLVGGRRAKASISKSISIERVVRPSSYTKISKSSSVSVESSRGPWGYGPYGRRDVEASSSVSIE
         RVVLPSSYTKISKSSSVSVERSRGPWGYGSYGRRDVKV

RM1_ML:LYGSRGGYSASSSVSVEGGRGLWGYGPYGRRDVEVSSSVSVENVVLPGSYTKISKSSSVSVESRRGPWGYGSRGGYSASSSVSIEGGRGLYGHGLLGR
         GPYGSLGGVYGSGDLDIVDGIDGIGGYGGIGGIGGYGGIGGLGHG (24)

RM2_ML:LYGSRGGYSASSSVSVEGGRGLWGYGPYGRRDVEVSSSVSIERVVRPGSYTKISKSSSVSVESRRGPWGYGPYGRRDVEVSSSVSVENVVLPGSYTKI
         SKSSSVSVESRRGPWGYGSRGGYSASSSVSIEGGRGLYGHGLLGRGPYGSLGGVYGSGDLDIVDGIDGIGGYGGIGGIGGYGGIGGLGHG (7)

RM3_ML:LYGSRGGYSASSSVSVEGGRGLYGHGDLDIVDGIDGIGYGGIGGIGGYGGIGGLGHGV (36)

TR2:     LYGSRGGYSASSSVSIEGGRGLWGYGPYGRRDVEVSSSVSIERVVRPGSYTKISKSSSVSVESERGPWGYGPYGRRDVEVSSSVSIERVVRPGSYTKI
         SKSSSVSVERERGPWGYGPYGRRDVEVSSSVSIERVVRPGSYTKISKSSSVSVESERGPWGYGPYGRRGSDIEISSSVSIESGRGPWGYGRGGYEYSS
         SVSLGGRGLLGYGLGLESSISVENSYGGRILGGLGRVGALGRSVGV

C:       PIPTVYTQHPGRDTVRSPCKLSDFNLLVKVGNTRKSNGNC
```




```
N:       MRAAILLILFCSLQIHITGAHKKPGNKVIGTLSDFLSHGHLDKNCGQHERILKGDDIIETNDKGEVIEKIISRKEILTDDDSESFSVSYSSEDDSTETIVKTITI
         VQEIPKHGKKKGHAKEKIFEEETVIKKIGAASAAPTIPIHIGPHKVS

RM1_LL:  SASASVSYEAPIVAPVYRAPKVSKSASVSYEQIVTPTVITRYSASGSVSREGPRRVVSPYGIHGPYGWGVAPGGYSASASVSREGPKRWGPGGVLGPHGPGGILG
         PHGPGGILGPHGPGGLLGPHGPGGLLGPHGPGGILGPHGPGGILGPGGRRRWPRGG (14)

TR1:     SASASVSYEAPIVAPVYRAPKVSKSASVSYEQIVTPTVITRYSASGSVSREGPRRIVSPYGVHGPYGWGVGPAGYSASGSVSREGPGVWGPGGWGPGGIYGPAGP
         GGVVGPYGLGDSGVTDGIYGPGGLLGPGGRRGPLGSLGLGDSGITDGIYGPGGLLGPGGRRGPLGSLGLGDSGITDGIYGPGGLLGPGGRRGPLGSLGLGDSGIT
         GGLLGPGGRYGPGGRYGPGGRRVWPYGG

RM2_LL:  SASASVSYEAPVVAPVIRRRPKVSKSASISVERIVTPTIITRYSVSKSVSREQPVVYHPVYHPVYHPKVKKSASVSVERIVTPTIITRYSASGSVSREGPRRVVS
         PYGIHGPYGWGVGPAGYSASASVSREGPGIWGPGRVVSPYGWGVGPAGYSASGSVSREGPGVWGPGGWGPGGIYGPAGPGGVVGPYGLGDSGVTDGIYGPGGLLG
         PGGRRGPLGSLGLGDSGITDGIYGPGGLLGPGGRRGPLGSLGLGDSGITDGIYGPGGLLGPGGRRGPLGSLGLGDSGITGGLLGPGGRYGPGGRYGPGGRRVWPYGG (13)

TR2:     SASASVSYEAPVVAPVIRRRPKVSKSASISVERIVTPTIITRYSVSKSVSREQPVVYHPVYHPVYHPKVKKSASVSVERIVTPTIITRYSASGSVSREGPRRIVR
         PRLWVPSPIGYSASASLSAERPLGYYKPRRYAPLGYSASASRSIEGPHGLLGGGYPVTSGLSIEGPSGLYRPGIVGGSLPTDAYDGVGVGHGY

C:       PGSGVYTYHPDPSTVGPRCRTSAYKLLINVGNARKDNGNC
```

**Table 1:** Full-length h-fibroins of nine caddisfly species derived from PacBio HiFi sequencing ordered by suborder and silk use. Gene length including introns and stop codon; CDS= protein coding DNA, gene length, CDS, exon, intron in bp; protein size in amino acids

| Species | Assembly | Reference | H-fibroin | Clade | Silk usage | Gene length | CDS | Exon 1 | Intron 1 | Exon 2 | Intron 2 | Exon 3 | Protein size | Molecular weight (kD) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Atopsyche davidsoni*** | GCA_022113835.1 | Ríos-Touma et al, 2022 | This study | Basal Integripalpia | cocoon | 23698 | 23610 | 42 | 88 | 23568 | | | 7870 | 754.24 |
| ***Eubasilissa regina*** | GCA_022840565.1 | Kawahara et al., 2022 | Kawahara et al., 2022 | Integripalpia | tube case | 25253 | 25161 | 40 | 92 | 25121 | | | 8387 | 815.86 |
| ***Glyphotaelius pellucidus*** | GCA_936435175.1 | Wellcome Sanger Institute | This study | Integripalpia | tube case | 24409 | 23592 | 42 | 817 | 23550 | | | 7864 | 801.63 |
| ***Hesperophylax magnus*** | JAIUSX000000000 | Hotaling et al, 2022 | Hotaling et al, 2022 | Integripalpia | tube case | 26666 | 25875 | 42 | 791 | 25833 | | | 8625 | 875.02 |
| ***Leptonema lineaticorne*** | SUB10837191 | This study | This study | Annulipalpia | fixed-retreat | 20538 | 20454 | 42 | 84 | 20412 | | | 6818 | 675.22 |
| ***Limnephilus lunatus*** | GCA_917563855.2 | Wellcome Sanger Institute | This study | Integripalpia | tube case | 30382 | 29547 | 39 | 835 | 29508 | | | 9849 | 998.81 |
| ***Limnephilus marmoratus*** | GCA_917880885.1 | Wellcome Sanger Institute | This study | Integripalpia | tube case | 27154 | 26325 | 39 | 829 | 26286 | | | 8775 | 882.45 |
| ***Limnephilus rhombicus*** | GCA_929108145.1 | Wellcome Sanger Institute | This study | Integripalpia | tube case | 27903 | 27084 | 39 | 819 | 27045 | | | 9028 | 896.38 |
| ***Micrasema longulum*** | n.a. | This study | This study | Integripalpia | tube case | 22770 | 21627 | 42 | 1036 | 12709 | 106 | 8876 | 7209 | 709.43 |