School of Doctoral Studies in Biological Science

University of South Bohemia in České Budějovice

Faculty of Science

Comparative Analysis of Silk Proteins and Discovery of Novel Sericin Gene in Lepidopteran Moths

Ph.D. Thesis

MSc. Bulah Chia-hsiang Wu

Supervisor: Prof. RNDr. Michal Žurovec, CSc.

Faculty of Science, University of South Bohemia in České Budějovice

Institute of Entomology, Biology Centre of the Czech Academy of Sciences

České Budějovice 2023

This thesis should be cited as:

Wu, B.C. (2023). Comparative analysis of silk proteins and discovery of novel sericin gene in lepidopteran moths. Ph.D. Thesis, University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 94 pp.

Annotation

This thesis focuses on the silk components of the Mediterranean moth, *Ephestia kuehniella*, and the discovery of a novel silk gene, *P150/ser6*, in the silkworm, *Bombyx mori*. We analyzed and described the cocoon silk components in both species. In the first publication, we combined transcriptomic, genomic, and proteomic approaches to identify silk proteins in *E. kuehniella*. In the second publication, we described the discovery of gene *P150/sericin6* in *B. mori* based on microsynteny analysis.

Declaration

I hereby declare that I am the author of this dissertation and that I have used only those sources and literature detailed in the list of references.

České Budějovice, date Student's signature

....

This thesis originated from a partnership of the Faculty of Science, University of South Bohemia, and the Institute of Entomology, Biology Centre of CAS, supporting doctoral studies in the Integrative Biology study program.

Financial support

This research was supported by the European Community's Program Interreg Bayern-Tschechische Republik Ziel ETZ 2021-2022 no. 331, and Interreg Bayern-Tschechien Republik BYCZ01-039.

Acknowledgments

I would like to thank Prof. Michal Žurovec, my supervisor, for his immense patience and guidance throughout my doctoral study. My thanks also go to all my friends and colleagues who have been supportive from the very beginning.

List of papers and author's contribution

The thesis is based on the following papers:

I. **Wu, B. C.**, Sauman, I., Maaroufi, H. O., Zaloudikova, A., Zurovcova, M., Kludkiewicz, B., et al. (2022). Characterization of silk genes in *Ephestia kuehniella* and *Galleria mellonella* revealed duplication of sericin genes and highly divergent sequences encoding fibroin heavy chains. *Front. Mol. Biosci.* 9, 1–16. doi: 10.3389/fmolb.2022.1023381.

Bulah Chia-hsiang Wu isolated RNAs, performed qPCR, and conducted computer analyses. Ivo Sauman performed the electron microscopy and histology imaging. Barbara Kludkiewicz performed the northern blotting analysis. Miluse Hradilova prepared cDNA libraries and provided the sequencing data. Martina Zurovcova performed phylogeny analysis. Houda Ouns Maarouf performed qPCR and adjusted figures. Anna Zaloudikova performed silk solubility and hydroscopicity tests. Michal Zurovec supervised the entire project and wrote the manuscript with the input of all co-authors.

II. **Wu, B. C.**, Zabelina, V., Zurovcova, M., and Zurovec, M. (2023). Unravelling the complexity of silk sericins: P150/sericin 6 is a new silk gene in *Bombyx mori*. bioRxiv, 2023.09.22.558982. doi: 10.1101/2023.09.22.558982.

Bulah Chia-hsiang Wu isolated RNAs, performed qPCR, and conducted computer analyses. Valeriya Zabelina dissected silk glands. Martina Zurovcova performed phylogeny analysis. Michal Zurovec supervised the entire project and wrote the manuscript with the input of all co-authors.

Papers not included in this thesis:

I. Pivarciova, L., Vaneckova, H., Provaznik, J., **Wu, B. C.**, Pivarci, M., Peckova, O., et al. (2016). Unexpected Geographic Variability of the Free Running Period in the Linden Bug *Pyrrhocoris apterus*. *J. Biol. Rhythms* 31, 568–576. doi: 10.1177/0748730416671213.

II. Kotwica-Rolinska, J., Chodakova, L., Chvalova, D., Kristofova, L., Fenclova, I., Provaznik, J., **Wu, B. C.**, et al. (2019). CRISPR/Cas9 Genome Editing Introduction and Optimization in the Non-model Insect *Pyrrhocoris apterus*. *Front. Physiol.* 10, 1–15. doi: 10.3389/fphys.2019.00891.

III. Kotwica-Rolinska, J., Chodáková, L., Smýkal, V., Damulewicz, M., Provazník, J., **Wu, B. C.**, et al. (2022). Loss of Timeless Underlies an Evolutionary Transition within the Circadian Clock. *Mol. Biol. Evol.* 39, 1–10. doi: 10.1093/molbev/msab346.

IV. Maaroufi, H. O., Pauchova, L., Lin, Y.-H., **Wu, B. C.**, Rouhova, L., Kucerova, L., et al. (2022). Mutation in *Drosophila* concentrative nucleoside transporter 1 alters spermatid maturation and mating behavior. *Front. Cell Dev. Biol.* 10, 1–17. doi: 10.3389/fcell.2022.945572.

V. Kmet, P., Kucerova, L., Sehadova, H., **Wu, B. C.**, Wu, Y.-L., and Zurovec, M. (2023). Identification of silk components in the bombycoid moth *Andraca theae* (Endromidae) reveals three fibroin subunits resembling those of Bombycidae and Sphingidae. *J. Insect Physiol.* 147, 104523. doi: 10.1016/j.jinsphys.2023.104523.

VI. Smykal, V., Chodakova, L., Hejnikova, M., Briedikova, K., **Wu, B. C.**, Vaneckova, H., et al. (2023). Steroid receptor coactivator TAIMAN is a new modulator of insect circadian clock. *PLOS Genet.* 19, e1010924. doi: 10.1371/journal.pgen.1010924.

Co-author agreement

Michal Žurovec, the supervisor of this Ph.D. thesis and co-author of papers I and II, fully acknowledges the stated contribution of Bulah Chia-hsiang Wu to these manuscripts.

.....

Prof. RNDr. Michal Žurovec, CSc.

Content

List of abbreviations

ASG	Anterior silk gland
BAC	Bacterial artificial chromosome
BLAST	Basic Local Alignment Search Tool
Fhx/P25	Fibrohexamerin/P25
FibH	Fibroin heavy chain
FibL	Fibroin light chain
MSG	Middle silk gland
Muc	Mucin
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
PEBP	Phosphatidylethanolamine-binding protein
PSG	Posterior silk gland
Ser	Sericin
Serpin	Serine proteinase inhibitors
SG	Silk gland
SMRT	Single-molecule real-time
TGS	Third-generation sequencing
TIL	Trypsin inhibitor-like
WGS	Whole genome sequencing
wt	Wild-type

Introduction

Overview

Silk is composed of polymers secreted by ectodermal glands in organisms such as Insecta, Arachnida, and Myriapoda (Sehnal and Žurovec, 2004). The use of natural silk dates back approximately 5,000 years when the silkworm *Bombyx mori* was domesticated from the wild silkworm *B. mandarina* (Underhill, 1997). Biologically, organisms produce silk for various purposes. Ballooning, which involves aerial dispersal using silk, has been documented in the lepidopteran superfamily Tineoidea, Yponomeutoidea, Gelechioidea, Cossoidea, Sesioidea, Tortricoidea, Pyraloidea, Geometroidea, and Noctuoidea (Bell et al., 2005). In the case of the green lacewing *Nineta flava* (Neuroptera), females produce silk to create an egg stalk that aids in attaching eggs to the leaves of the host plant (Lucas et al., 1957). The pyralid moth *Bradyrrhoa gilveolella* is renowned for constructing feeding tubes to facilitate larval foraging (Caresche and Wapshere, 1975). Many moths construct silk cocoons during the late-instar larval stage to shield pupae from bacterial infections and potential predation during metamorphosis (Offord et al., 2016). Thanks to its impressive extensibility, tensile strength, sustainability, and biodegradability, silk has long been considered an exemplary material for high-performance artificial fibers and is widely utilized in the arms industry, biomedicine, composite materials manufacturing, and cosmetics industry (Holland et al., 2019; Eom et al., 2020; Das et al., 2021; Siengchin, 2023).

Silk components and structure

Silk of *B. mori* is synthesized in silk glands (SGs), a pair of tubular-shaped, highly specialized labial glands. SG is divided into three compartments with distinct boundaries: posterior silk gland (PSG), middle silk gland (MSG), and anterior silk gland (ASG), each of which constitutes approximately 350, 255, and 520 cells (Perdrix-Gillot, 1979). The fibrous core protein, fibroin, is synthesized in the PSG and subsequently moved into the MSG and coated with the glue protein sericin. After being transported to the ASG, the condensed, aqueous proteins undergo a phase transition and are converted to solid fibers and spun at the spinneret (Wang et al., 2017).

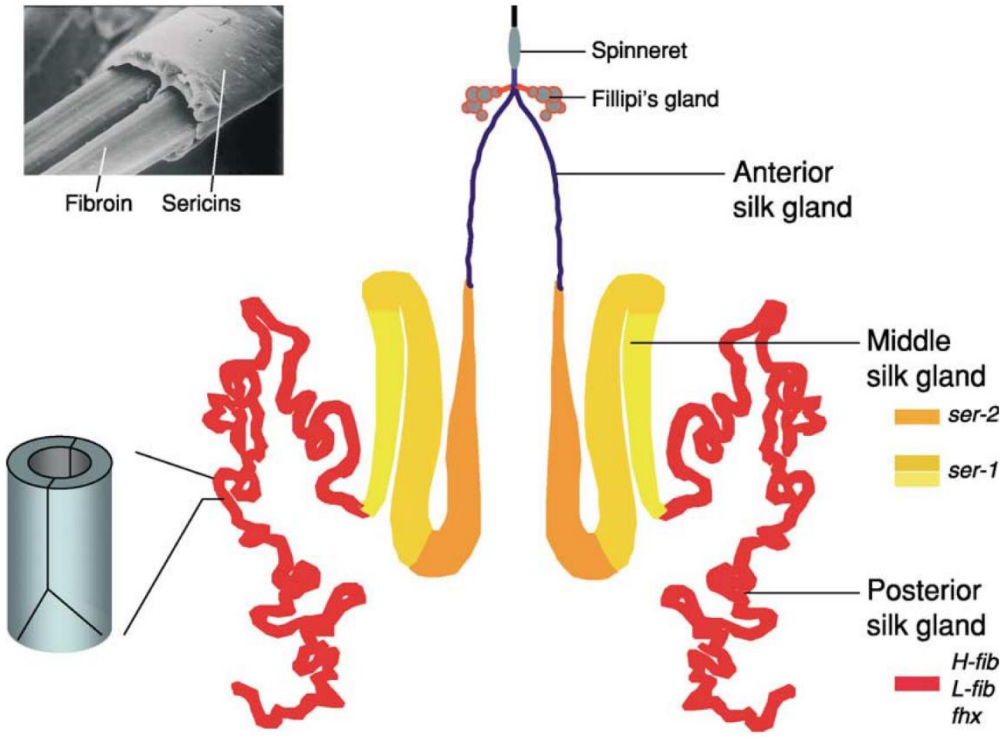


Figure 1. The silk glands of *B. mori*. The SG is divided into anterior silk gland (red), middle silk gland (dark orange, orange and yellow) and posterior silk gland (black). Major silk genes includes fibroin heavy chain (*H-fib*), fibroin light chain (*L-fib*), and *P25/fibroinhexamerin* (*fhx*), *sericin 1* (*ser-1*) and *sericin 2* (*ser-2*). Color blocks next to the gene names indicate compartments where genes primarily express (Julien et al., 2005).

The natural silk is composed of the core protein, fibroin, and the glue protein, sericin. Fibroin is a macromolecular complex, which typically consists of three proteins: fibroin heavy chain (FibH, 350 kDa), fibroin light chain (FibL, 26 kDa), and glycoprotein P25/ Fibrohexamerin (P25, approximately 30 kDa). By comparing the genomic and cDNA sequences of two naked pupa mutant strains, *Nd-s* and *Nd-s*^{D}, to the sequences of the wild-type strain J-139, it was shown that a disulfide bond plays a crucial role in associating the residue Cys-c20 of FibH and the residue Cys-172 of FibL (Mori et al., 1995). Mutants carrying homozygous FibL mutant alleles *Nd-s*^{D} fail to combine the FibH and FibL proteins, which severely decreases the amount of fibroin production to less than 0.3% (Takei et al., 1987). P25, on the other hand, is not associated with FibH via covalent bonds but via hydrophobic interactions; it contains Asn-linked oligosaccharide chains, which has been suggested could play a role in assisting the folding of FibH (Tanaka et al., 1999). An

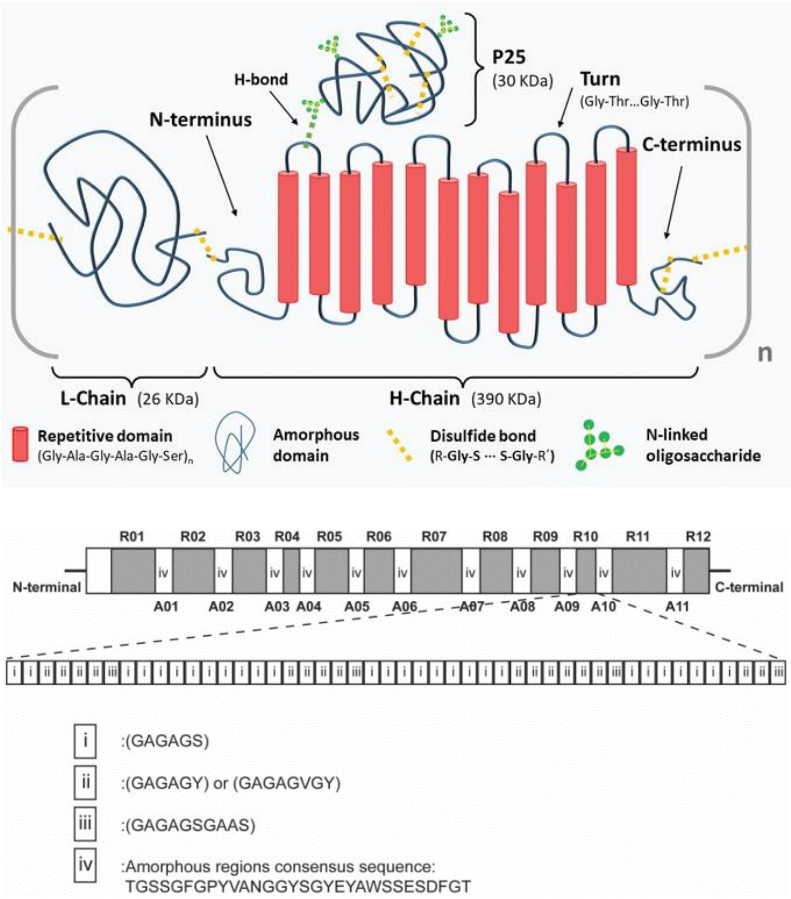


Figure 2. Schematic diagrams show (A) the components of an elementary unit of *B. mori* fibroin, including FibH, FibL and P25, and (B) the sequence structure of FibH. R1-R12 represent 12 repetitive blocks, interleaved with 11 amorphous regions A1-A11. Sequences of repeat types (i, ii and iii) and the consensus sequence of amorphous regions (iv) are indicated (Reizabal et al., 2023).

elementary unit of *B. mori* fibroin consists of FibH, FibL and P25 in a molar ratio of 6:6:1 (Inoue et al., 2000).

The complete sequence of *B. mori FibH* was resolved in 2000 by a hybrid strategy of shotgun sequencing and physical map-directed sequencing. The gene contains two exons, 67 bp and 15750 bp, intervened by a 971-bp intron (Zhou et al., 2000). The most abundant amino acids in the composition of FibH include Glycine (45.9%), Alanine (30.3%), Serine (12.1%), and Tyrosine (5.3%). Overall, the amino acid sequence can be delineated as N- and C-terminus non-repetitive regions and a central, core repetitive region. The repetitive region can be further subdivided into 12 repetitive blocks and 11 amorphous blocks. Three repetitive

motifs are observed throughout the full sequence: (i) Gly-Ala-Gly-Ala-Gly-Ser motif, (ii) Gly-Ala-Gly-Ala-(Gly-Val)-Gly-Tyr motif, and (iii) Gly-Ala-Gly-Ala-Gly-Ser-Gly-Ala-Ala-Ser motif. Due to amino acid composition, the repetitive blocks and the amorphous blocks are hydrophobic and hydrophilic, respectively (Koh et al., 2015; Reizabal et al., 2023).

Although it has been shown that P25 is a typical component of fibroin in various lepidopteran species, it was not detected in the Japanese oak silkmoth *Antheraea yamamai* (Saturniidae), the ghost moth *Hepialus californicus* (Hepialidae), and Trichoptera (Tanaka and Mizuno, 2001; Yonemura et al., 2009; Collin et al., 2010). In *B. mori*, it is suggested that the N-glycosylated P25 protein is involved in maintaining the structural stability of the fibroin complex (Inoue et al., 2000). Zabelina et al. (2021) generated *P25* mutants via transcription activator-like effector nuclease (TALEN)-mediated gene knockout in *B. mori*; subsequent characterization of these mutants corroborated the role of *P25* in stabilizing the fibroin proteins during luminal transport by, possibly, regulating the solubility of fibroin globules in the SG lumen.

Sericin is a glue protein that wraps around and holds silk fibers together. Because sericin contains a high ratio of polar amino acids (such as serine, aspartic acid, and lysine), it is water-soluble and the hydrophilic properties of the side chains are crucial for crosslinking and modification (Liu et al., 2022). Up to now, multiple sericins have been identified in *B. mori*. The sizes of the identified sericin proteins vary from approximately 123 kDa to 331 kDa. Alternative splicing to create protein isoforms has been reported in *Ser1*, *Ser2* and *Ser4*. The expression profile of these glue proteins has been well characterized. Spatially, the genes *Ser1*, *Ser4* and *Ser5* are expressed in the middle and posterior parts of the middle silk gland (MSG); *Ser2* is detected mainly in the anterior and the middle parts of MSG; *Ser3* is predominately expressed in the anterior part of MSG. Temporally, *Ser1* and *Ser2* are expressed throughout the larval developmental stage, mainly in the 5^{th}-instar larvae (*Ser1*) and in the 3^{rd} and 4^{th} instar larvae (*Ser2*), respectively; *Ser3* is specifically expressed in the 5^{th}-instar larvae; *Ser4* is expressed in the 1^{st}-4^{th}-instar larval stage; Ser5 is expressed only before the 5^{h} instar larval stage, a pattern similar to the expression of *Ser4*. The spatiotemporal expression pattern of sericin genes leads to the difference in the composition of non-cocoon silk and cocoon silk. According to the proposed model, the fibroin filaments of non-cocoon silk, which are secreted by young (the 1^{st}-4^{th} instar) larvae, are coated with, from the innermost to the outermost layer, Ser4, Ser5, and Ser2 glue proteins; the fibroin filaments

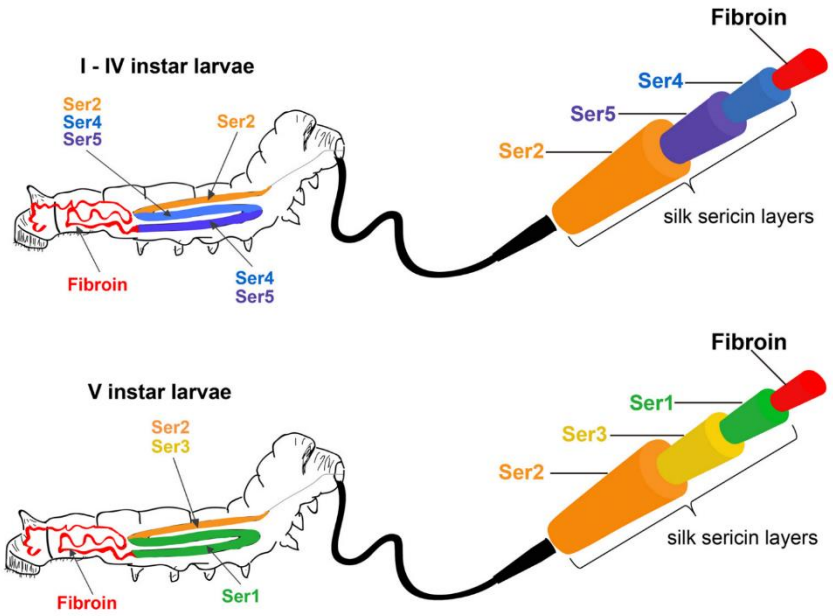


Figure 3. Proposed structure of multiple-layered non-cocoon silk secreted by the 1st-4th instar larvae (upper panel) and cocoon silk by the 5th instar larvae (lower panel). The core protein, fibroin, is produced in PSG; sericin 1-5 are secreted in different parts of MSG and ASG and covering fibroin (Guo et al., 2022).

of cocoon silk, which are secreted by the 5th-instar larvae, are coated with Ser1, Ser3 and Ser2 from the innermost to the outermost layer (Guo et al., 2022). In addition to silk structure, sericin is involved in the cellular immune system (Ye et al., 2021). Its photoluminescent property also makes it a candidate carrier for medical applications such as *in vivo* bio-imaging and tracking (Wang et al., 2014). Recent research successfully generated *piggyBac*-mediated transgenic silkworms to express *Ser3* ectopically in PSG; the resulting cocoon silk gained improved properties (tensile strength, moisture absorption and liberation, water solubility and stability), which demonstrates an efficient way to manipulate the silk structure in a living organism (Chen et al., 2022).

In addition to fibroin and sericin, several proteins have been identified in lepidopteran silk, especially the antimicrobial proteins. Seroin was first described in the greater wax moth *Galleria mellonella*. The name was coined because it was found in MSG and PSG, where sericin and fibroin were detected (Zurovec et al., 1998). Seroin 1 and 2 have shown antimicrobial activity and therefore might play a role in protecting pupae against pathogenic microbes (Singh et al., 2014). Seroin 3 was first identified by Dong et al. (2016), although its

low protein content implies insignificant antimicrobial activity. Kucerova et al. (2019) screened silk gland specific transcriptomes and publicly available datasets of multiple species to propose a comprehensive classification system of seroins in Lepidoptera. Other protease inhibitors were also identified in cocoon silk to inhibit the growth of fungi, including phosphatidylethanolamine-binding protein (PEBP) and serine proteinase inhibitors (serpins), trypsin inhibitor-like (TIL)-type protease inhibitors and Kunitz-type protease inhibitors (Guo et al., 2015; Li et al., 2015; Zhang et al., 2020b, 2020a).

Genome resources of *B. mori*

Nowadays, whole genome sequencing (WGS) data has become an inseparable part in many fields of biology. Since the 1990s, due to the improvement of automation technology in Sanger sequencing, it has become attainable to acquire complete genome sequencing data. Genome drafts of various model organisms, such as the budding yeast *Saccharomyces cerevisiae* (12.15 Mb), the nematode *Caenorhabditis elegans* (100.29 Mb), the fruit fly *Drosophila melanogaster* (142.73 Mb), the mouse-ear cress *Arabidopsis thaliana* (119.67 Mb), and the house mouse *Mus musculus* (2728.22 Mb), etc., were subsequently published (Goffeau et al., 1996; C. elegans Sequencing Consortium, 1998; Adams et al., 2000; Chinwalla et al., 2002).

The silkworm, *B. mori*, has been a lepidopteran model organism due to its great economic importance. Its first two genome drafts were released in 2004 by two independent research teams. Although derived from different strains (p50T and *Dazao*), both drafts were based on the whole genome shotgun sequencing technique, with 3- and 5.9-fold coverage respectively (Mita et al., 2004; Xia et al., 2004). An update of these two assemblies was released in 2008 in the wake of the collaboration between the two teams; the merged scaffold-level genome assembly, which was derived from the previous two drafts together with end sequences of fosmid and bacterial artificial chromosome (BAC) libraries, achieved coverage of 8.5 folds with an N50 size of about 3.7 Mb (The International Silkworm Genome Consortium, 2008).

One major disadvantage of next-generation sequencing (NGS) technologies is the short length of the reads, which cannot span the repetitive regions in genomes. In addition, relying on the polymerase chain reaction (PCR) to amplify the samples inescapably introduces bias against the high GC-rich regions (van Dijk et al., 2018). Different from next-generation

sequencing (NGS), third-generation sequencing (TGS) does not involve a PCR amplification step amid library preparation. On the contrary, it requires long fragments of high-molecular-weight DNA for library preparation. Long-read sequencing data generated on two most prominent platforms, single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio), and nanopore long-read sequencing from Oxford Nanopore Technologies (ONT), have been shown to overcome the challenges met in genome assembly, which are mainly caused by the complex, repetitive genomic regions (Goodwin et al., 2016). The latest *B. mori* reference genome (p50T strain) was released in 2019. This assembly, which was based on a hybrid assembly approach, combines the PacBio long reads and Illumina short reads. The assembly achieves a 140-fold sequencing depth with a scaffold N50 of 16.8 Mb (Kawamoto et al., 2019).

Given the advances in sequencing technologies, together with the decreasing sequencing cost, whole genome sequencing projects have moved from focusing on model organisms towards a more diverse, wider selection of individuals and populations. For example, genome phasing, a sequencing approach that considers both paternal and maternal haplotypes in a diploid genome, mitigates the reference bias and enables more accurate variant calling (Ballouz et al., 2019); generation of a pangenome to better represent the diversity within a species has been realized in many organisms (Ding et al., 2018; Badet and Croll, 2020; Bayer et al., 2020; Tettelin and Medini, 2020). The pan-genome of the silkworm, based on 545 high-quality, long-read reference assemblies, was published in 2022. This genomic resource uncovered additional 264 domestication-associated and 185 genes trait-improvement-associated genes; the genetic basis of economically important traits and adaptation-related traits, such as the fineness of silk and embryonic diapause, was also revealed (Tong et al., 2022). Moreover, the release of single-cell transcriptomic atlas of *B. mori* silk glands provides a new approach to deciphering the dynamics of silk synthesis and associating the genotype and phenotype (Ma et al., 2022).

Genome resources of *Ephestia kuehniella*

The Mediterranean flour moth *E. kuehniella* is a cosmopolitan storage pest, which causes economic damage to grains and grain products (CABI, 2023). Although genetic studies on this species date back to the 1950s (Caspari and Gottlieb, 1959), knowledge of *E. kuehniella*

genetic/genomic information was scarce. Since 2021, two research groups have released genomic data on *E. kuehniella*. Visser et al. (2021) used Oxford Nanopore technology to construct a long-read assembly of approximately 42× coverage depth, containing 165 contigs with an N50 size of 8.3 Mb. The assembly was annotated based on *ab initio* prediction and tissue-specific RNA-seq data, and identified 13882 genes with an N50 of 7.2 kb (Wu et al., 2022). Künstner et al. (2022) published a set of female/male genome and transcriptome assemblies based on short-reads data generated using Illumina paired-end and mate-pair sequencing technologies. These female/male assemblies have 90999/90445 contigs with N50 of 11860/12636 bp, respectively.

Synteny-assisted homology detection

Identifying homologous genes is a process of finding unknown sequences based on known sequences. This is the fundamental step to understand gene/protein functions and establish the evolutionary relationships among species. However, when the pairwise sequence identity is lower than ~30% and the signal falls from the safe zone into the twilight zone, the sequence similarity searching method could be problematic (Rost, 1999). Various tools, such as the Basic Local Alignment Search Tool (BLAST) and HMMER (Altschul et al., 1990; Johnson et al., 2010), have been developed to assist with the task. However, remote homology detection remains hindered by the low sequence identity (Kilinc et al., 2023). For instance, the anti-freeze glycoprotein (AFGP) homologs found in notothenioid fishes and codfishes consist of a simple glycotripeptide (Thr-Ala/Pro-Ala) repeat motif, and manual annotation was required to properly characterize members in this gene family (Chen et al., 1997a, 1997b; Baalsrud et al., 2018). Similar challenges were met in identifying the main components of silk, fibroin and sericin, which are known for their long and repetitive sequence composition. Due to the nature of the sequences of silk genes, the homology search of these genes requires additional manual investigations and adjustments (Rouhová et al., 2022; Heckenhauer et al., 2023). On the other hand, the subject sequences employed in homology search could also be the source of homology detection failure. It has been noted that in sequence databases there exists considerable erroneous information, including redundancy, inconsistency and misannotaion (Chen et al., 2017; Stoler and Nekrutenko, 2021). Such errors might be utilized in the automated annotation pipeline and result in error propagation from the source to the new records (Goudey et al., 2022). In addition, genetic

variation, such as single-nucleotide polymorphisms (SNPs) and structural variants (inversions, deletions, and duplications) in the reference genome may not exist in the individuals/populations where the query sequences originate, which results in the detection failure.

The term “synteny” initially refers to “gene loci on the same chromosome” (Passarge et al., 1999). This usage later expanded to mean a state such as “the conservation of co-localized genes in the same order within different genomes” (Vergara and Chen, 2010), “the conserved order of aligned genomic blocks between species” (Ensembl, 2023), and so on. Genes on a syntenic block shared between two genomes are not necessarily in perfect colinear order; micro rearrangements could interrupt the order and lead to dissimilar gap regions within the block (Pevzner and Tesler, 2003). Because conserved synteny can be preserved over the course of evolution, it has been employed to annotate genomes, infer homology, and reconstruct the ancestral karyotype (Zheng et al., 2005; Song et al., 2018; Damas et al., 2022; Simakov et al., 2022; Kirilenko et al., 2023). Synteny analysis also amends the traditional sequence alignment-based phylogenetic analysis; a recent study demonstrated that a microsynteny information-based approach is accurate in reflecting the phylogenetic relationships of flowering plants (Zhao et al., 2021).

It has been reported that, in *B. mori*, silk genes *Sericin 1-5* form a cluster on chromosome 11:2,534,704-4,923,394 (Dong et al., 2019; Guo et al., 2022). Also located within this region is another structural gene, *Mucin-12* (LOC101736082), whose homologs were confirmed as a silk component in several lepidopteran species (Kludkiewicz et al., 2019; Rouhova et al., 2021; Volenikova et al., 2022; Wu et al., 2022; Kmet et al., 2023). Syntenic analysis reveals that this specific gene cluster harbors on a syntenic block, which is seemingly well conserved across Lepidoptera (Figure 4). Application of such analysis in sericulture research provided a picture of the expansion of soluble silk components observed in *G. mellonella* (Kludkiewicz et al., 2019), and assisted in the discovery of a new silk gene in *B. mori* (Wu et al., 2023).

Figure 4. Syntenic analysis on 17 representative species to infer the origin of the sericin genes in Lepidoptera. To consider an overall scope, 17 genome assemblies, including 16 lepidopteran genomes of 16 superfamilies and 1 trichopteran genome, were selected to infer the syntenic blocks. The result shows that syntenic blocks were consistently preserved across Lepidoptera, even in Trichoptera. (*Continued on next page*)

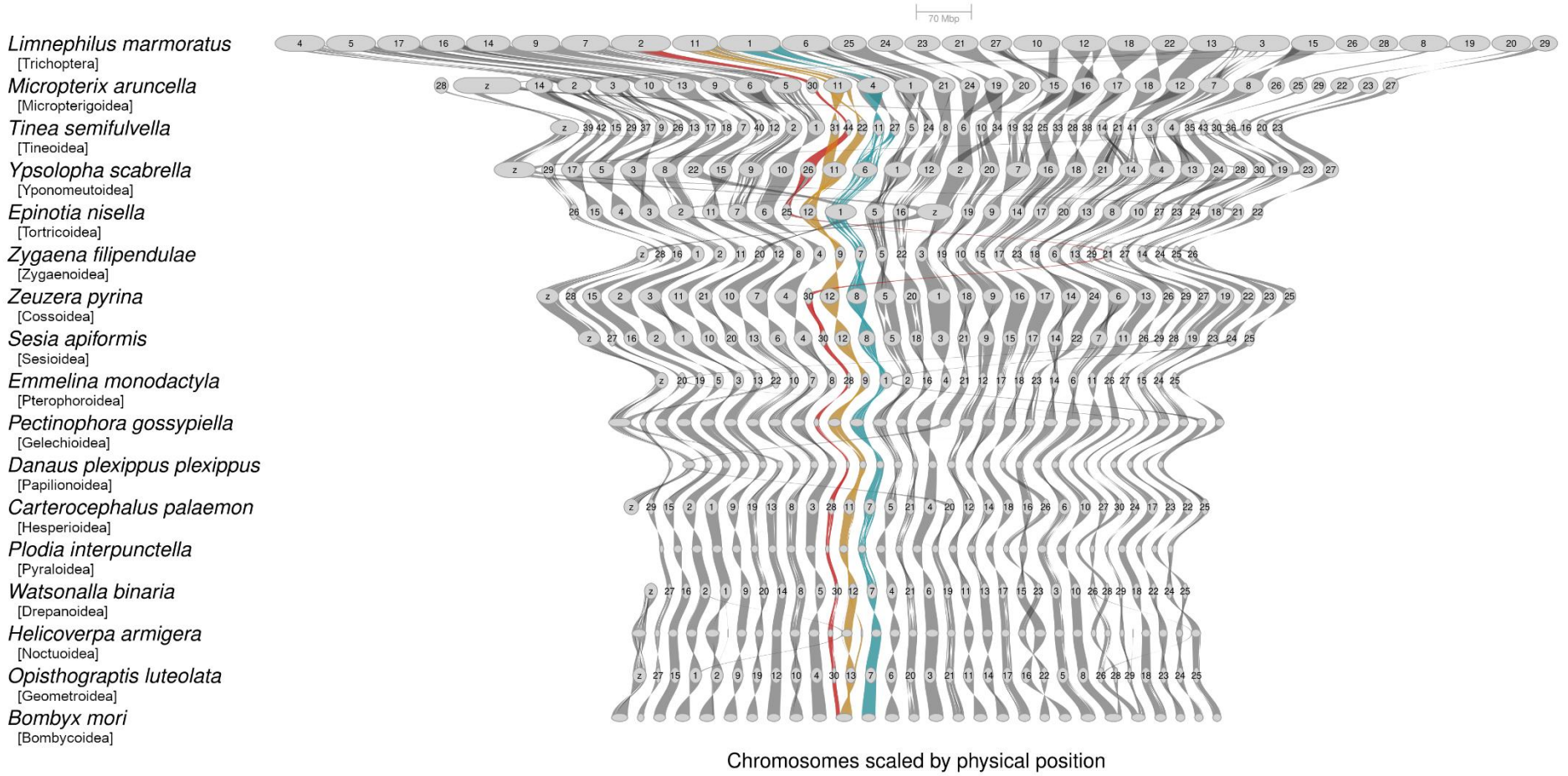


Figure 4. (*Continued from previous page*) A block fusion (red and gold ribbons) in *B. mori* leads to chromosome 11 (NC_051368.1), which hints a lineage specific fusion in Bombycoidea. *B. mori* silk genes *sericin 1-5* and *Mucin-12*, sits within the red syntenic block. The newly identified silk gene *P150/ser6* is located on chromosome 12, which belongs to the green syntenic block. Synteny analyses were implemented using Genespace (Lovell et al., 2022).

The aim of my study

An integrated, multi-omic approach, which combines genomic, transcriptomic and proteomic tools, has been widely applied to systematically investigate the components of silk at different biological levels. In my first publication, I characterized the cocoon silk of the Mediterranean moth *E. kuehniella*. In parallel, I re-analyzed the cocoon silk of the greater wax moth *G. mellonella* and made intra-family comparisons. Genes *FibH* and encoded proteins of nine Pyraloidea species were presented and their conserved features were discussed.

Although the silkworm *B. mori* is one of the most studied species in sericulture, there remains a subset of unknown genes encoding its silk components. In my second publication, I conducted a microsynteny analysis and corrected an erroneous gene model in the current *B. mori* reference genome. This new gene model led to the discovery of a novel silk gene *P150/ser6*. The expression pattern of *P150/ser6* homologs is similar in *B. mori* and two pyraloid moths, *E. kuehniella* and *G. mellonella*. However, the difference in P150/ser6 protein intensities observed in silk cocoons implies other genetic factors (for example, cis-regulatory elements) may be involved in the utilization of P150/ser6 during silk production.

Chapter 1

Characterization of silk genes in *Ephestia kuehniella* and *Galleria mellonella* revealed duplication of sericin genes and highly divergent sequences encoding fibroin heavy chains

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 29 November 2022
DOI 10.3389/fmolb.2022.1023381

OPEN ACCESS

EDITED BY
Qiu-Ning Liu,
Yancheng Teachers University, China

REVIEWED BY
Najmeh Sahebzadeh,
Zabol University, Iran
Xianzhao Kan,
Anhui Normal University, China

*CORRESPONDENCE
Michal Zurovec,
zurovec@entu.cas.cz

SPECIALTY SECTION
This article was submitted to Molecular Evolution,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 19 August 2022
ACCEPTED 24 October 2022
PUBLISHED 29 November 2022

CITATION
Wu BC-h, Sauman I, Maaroufi HO, Zaloudikova A, Zurovcova M, Kludkiewicz B, Hradilova M and Zurovec M (2022), Characterization of silk genes in Ephestia kuehniella and Galleria mellonella revealed duplication of sericin genes and highly divergent sequences encoding fibroin heavy chains.
Front. Mol. Biosci. 9:1023381.
doi: 10.3389/fmolb.2022.1023381

Characterization of silk genes in *Ephestia kuehniella* and *Galleria mellonella* revealed duplication of sericin genes and highly divergent sequences encoding fibroin heavy chains

Bulah Chia-hsiang Wu[1,2], Ivo Sauman[1,2], Houda Ouns Maaroufi[1,2], Anna Zaloudikova[1], Martina Zurovcova[1], Barbara Kludkiewicz[1], Miluse Hradilova[3] and Michal Zurovec[1,2]*

[1]Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Ceske Budejovice, Czechia, [2]Faculty of Science, University of South Bohemia, Ceske Budejovice, Czechia, [3]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Praha, Czechia

Silk is a secretory product of numerous arthropods with remarkable mechanical properties. In this work, we present the complete sequences of the putative major silk proteins of *E. kuehniella* and compare them with those of *G. mellonella*, which belongs to the same moth family Pyralidae. To identify the silk genes of both species, we combined proteomic analysis of cocoon silk with a homology search in transcriptomes and genomic sequences to complement the information on both species. We analyzed structure of the candidate genes obtained, their expression specificity and their evolutionary relationships. We demonstrate that the silks of *E. kuehniella and G. mellonella* differ in their hydrophobicity and that the silk of *E. kuehniella* is highly hygroscopic. In our experiments, we show that the number of genes encoding sericins is higher in *G. mellonella* than in *E. kuehniella*. By analyzing the synteny of the chromosomal segment encoding sericin genes in both moth species, we found that the region encoding sericins is duplicated in *G. mellonella*. Finally, we present the complete primary structures of nine *fibH* genes and proteins from both families of the suborder Pyraloidea and discuss their specific and conserved features. This study provides a foundation for future research on the evolution of silk proteins and lays the groundwork for future detailed functional studies.

KEYWORDS

synteny, mucin, mediterranean flour moth, wax moth, pyralidae, crambidae

Introduction

The Pyraloidea are the third largest superfamily of the Ditrysian Lepidoptera order, containing about 16,000 species. They are found on all continents except Antarctica and consist of two families: Pyralidae and Crambidae. About 5,000 representatives of the Pyralidae family have been described, including a number of important pests. The Mediterranean flour moth, *E. kuehniella* Zeller (Lepidoptera: Pyralidae), is one of the most important pests of stored products. The larvae infest stocks of flour or cereal grains as a food source, but they do the most damage by producing silk that clogs machinery (Jacob and Cox, 1977). The larvae spend most of their lives in silk tubes that provide protection from parasitoids and reduce water loss (Fedic et al., 2003). *G. mellonella* (also from the family Pyralidae) is a pest of honey bees whose larvae live in hives protected by a maze of feeding tubes (Ellis and Hayes, 2009). Previous studies of *G. mellonella* helped to elucidate that the general protein composition of silk is conserved in moths (Zurovec et al., 1992; Zurovec et al., 1995; Zurovec et al., 1998). The study of sericin genes in *G. mellonella* also revealed that there is a high proportion of proteins surrounding the fibroin core, which is associated with an unusually high number of sericin genes (Kludkiewicz et al., 2019).

The silk of lepidoptera is produced in the transformed labial salivary gland of the larva, which is called the silk gland (SG). The posterior silk gland (PSG) produces filaments consisting of three proteins: fibroin heavy chain, fibroin light chain and fibrohexamerin (Shimura et al., 1982), while the middle silk gland (MSG) produces envelope proteins that mainly have an adhesive function and primarily consist of sericins (Gamo, 1982; Prudhomme et al., 1985). The anterior silk gland (ASG) is a tubular duct lined by a cuticle. The silk undergoes significant changes during evolution, both in the sequence of individual proteins and in the presence of individual protein components.

The fibroin heavy chain (FibH) is the best-studied silk component. It contains regions of regular protein secondary structures consisting of antiparallel beta-sheets and forms crystalline domains responsible for fiber strength (Deny, 1980). These tend to be primarily composed of the simple amino acids alanine, glycine, and serine, which enable the formation of beta structures (Craig, 1997). Despite the profound sequence differences between species, there are structural requirements needed for fiber strength, and a limited number of β-sheet configurations for suitable crystal domain motifs exist. This may have led to convergent evolution and the reappearance of motifs found in unrelated species (Lucas and Rudall, 1968). Previous experiments have shown that the silk fibers of *E. kuehniella* and *G. mellonella* have approximately the same tensile strength as those of *B. mori*, although both contain relatively short and scattered putative crystallites in the FibH (Fedic et al., 2003). Proteins produced by MSG are less well studied and seem to be subject to even greater changes than fibroins. These include large adhesive proteins, sericins, mucins and zonadhesin-like proteins, as well as seroins and protease inhibitors involved in protection against microorganisms. The low conservation of silk gene sequences makes it difficult to identify new proteins based on homology between more distant lepidopteran species; however, homology can be very useful for identifying genes in closely related species. Recent advances in proteomics and sequencing of lepidopteran genomes have provided a flood of information on new silk components and have made it possible to obtain complete sequences of large repetitive genes that were previously difficult to study (Davey et al., 2021).

In this study, we present the complete sequences of the putative major silk proteins from two members of the moth family Pyralidae, *E. kuehniella* (subfamily Phycitinae) and *G. mellonella* (subfamily Galleriinae). We identified silk genes in both moths based on proteomic analysis of cocoon silk and by searching for homologies in the transcriptomes and genomes of both species. We also compared genomic sequences of *E. kuehniella* and *G. mellonella* with genomic DNA from *Amyelois transitella* (subfamily Phycitinae), also from the Pyralidae family. We discovered a region containing clusters of sericin genes and identified blocks of synteny (colocalized gene clusters shared between genomes). The resulting microsynteny map allowed identification of duplication events in the sericin family. Finally, we present the complete primary structures of nine FibH proteins from both families of the suborder Pyraloidea and discuss their specific and conserved features.

Materials and methods

Insects and silk

Mediterranean flour moth (*E. kuehniella*) and waxmoth (*G. mellonella*) larvae were laboratory strains previously established from specimens found in České Budějovice, Czech Republic and kept in the Institute of Entomology, Biology Centre of the Czech Academy of Sciences. The *E. kuehniella* larvae were reared on a mixture of wheat flour and wheat bran (volume ratio of 4:1) supplemented with a small amount of dry yeast at 24°C without humidity control. The food was sterilized at 110°C for 2 h before adding the yeast (Marec and Traut, 1994). The *G. mellonella* larvae were reared on a semi-artificial diet at 30°C (Sehnal, 1966). The diet for *G. mellonella* consisted of wheat flower, corn and wheat meals in ratios 1:2:1 mixed with dry milk, dry yeast, beeswax, glycerol, and honey. *B. mori* cocoons were a gift from Dr. D. Zitnan, (Bratislava, Slovakia).

Histology and electron microscopy

Whole mount preparations of SGs from *E. kuehniella* were conducted as follows: SGs were dissected from water anesthetized last instar wandering stage larvae, transferred to a drop of phosphate-buffered saline (PBS) on a microscope slide, covered with a coverslip, and imaged under an Olympus BX63 microscope (Olympus Corporation, Tokyo, Japan) equipped with a CCD camera (Olympus DP74).

The histology of *E. kuehniella* larvae was carried out as follows: The cuticles of water anesthetized larvae were punctured with a fine needle under Bouin–Hollande fixing solution supplemented with mercuric chloride (Levine et al., 1995). After one hour of fixation, the larvae were cut into three parts and then fixed overnight at 4°C. Standard histological procedures were used for tissue dehydration, embedding in Paraplast, sectioning (7 µm), deparaffinization, and rehydration. Sections were treated with Lugol's iodine followed by 7.5% sodium thiosulfate solution to remove residual heavy metal ions, washed in distilled water, and stained with the HT15 Trichrome Staining Kit (Masson) (Sigma-Aldrich, Inc., St. Louis, MO, United States) according to the manufacturer's protocol. The stained sections were dehydrated and mounted using a DPX mounting medium (Fluka, Buchs, Switzerland). High-resolution images were acquired using a BX63 microscope, DP74 CMOS camera, and cellSens software (Olympus) by stitching multiple images together.

Semi-thin sections of cocoons were produced as follows: Pieces of freshly spun and degummed cocoons were prepared in PBS and fixed in 2.5% glutaraldehyde or at least 4 h at room temperature (RT) or overnight at 4°C. Specimens were then dehydrated and embedded in Epon resin as previously described (Kludkiewicz et al., 2019). Semi-thin sections were cut with a glass knife and placed onto a droplet of 10% acetone on a microscope slide. The dried sections were stained with toluidine blue and imaged under a light microscope.

The analysis of the ultrastructure of the silk was conducted as follows: Silk samples were cut from cocoons, glued to the surface of aluminum holders, sputter-coated with gold, and analyzed using a Jeol JSM-7401F scanning electron microscope (Jeol, Akishima, Japan).

Northern blotting and qPCR

Total RNA was extracted from dissected larvae and SG with TRIzol reagent (Invitrogen). RNA aliquots of 5 µg were collected for agarose electrophoresis, blotted onto a nylon membrane (Hybond N+, Sigma-Aldrich, St. Louis, United States), and hybridized under high stringency conditions as previously described (Zurovec et al., 2016). Probes for northern blotting were amplified using reverse transcription polymerase chain reaction and primers listed in Supplementary Table S1A, then labeled with a-^{32}P[dATP] using random priming with an Oligo labeling kit (Thermo Fisher Scientific, Prague). Autoradiographic detection was performed using the storage phosphor screen of a STORM 860 Phosphorimager (Molecular Dynamics, Chatsworth, United States).

qPCR was performed using HOT FIREPol EvaGreen qPCR Mix Plus (Solis BioDyne, Tartu, Estonia). Five individuals were used for each sample. All samples were collected in triplicate. The PCR reaction volume of 20 µL contained 5 µL of diluted cDNA and 250 nM primers. Amplification was carried out using a Rotor-Gene Q MDx 2plex HRM (Qiagen, Hilden, Germany) for 45 cycles (95°C for 15 s; annealing temperature adjusted to the primer pair for 30 s; 72°C for 20 s) following an initial denaturation/Pol activation step (95°C for 15 min). Each sample was analyzed in triplicate. Primers (Supplementary Table S1B) were designed using Geneious Prime software platform (Biomatters, Auckland, New Zealand; version 2021.2.2) to ensure that each amplicon was specific. The output was analyzed using the software Rotor Gene Q (version 2.3.5). Elongation factor 1 alpha (EF1a, NM_001044045.1) was used as a reference gene, and the relative expression of the target genes was calculated using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). Statistical analysis was performed using the Student's t-test in R (version 4.1.1); p-values < 0.05 were considered statistically significant. The detailed statistical analysis is shown in Supplementary Table S2.

Transcriptome preparation

RNA isolation, cDNA library synthesis, and RNA sequencing were performed as previously described (Rouhova et al., 2021). Briefly, last instar wandering larvae were dissected and tissues were separated. The RNA for transcriptome preparation was isolated using TRIzol reagent and further purified using a NucleoSpin RNA II kit (Macherey-Nagel, Duren, Germany). The mRNA was then isolated using Oligo(dT)25 Dynabeads (Ambion, Life Technologies). RNA integrity was checked, and concentration was measured using a Bioanalyser 2100 (Agilent, Waldbronn, Germany). The cDNA library was prepared using a NEXTflex Rapid RNA-Seq Kit (Bioo Scientific, Austin, TX, United States). Sequencing was performed using a MiSeq (Illumina, San Diego, United States), generating sequences in a 2 × 150 nt pair-end format. The BUSCO tool suite (version 3.0) (Simao et al., 2015) was used to assess the completeness of the assembly. A total of 1.6×10^7 reads were assembled *de novo* using Trinity software (version 2.9.1 + galaxy1) on the Galaxy platform (Afgan et al., 2018). The transcriptome was further improved with the genome annotation pipeline MAKER (version 2.28) by incorporating information on the *E. kuehniella* genome assembly (Visser et al., 2021), as well as protein datasets for *B. mori*, *G. mellonella*, and arthropoda

(Odb10, https://busco.ezlab.org/). Full-length transcripts were found using the genome annotation pipeline MAKER (version 2.28). The completeness of the resulting transcriptome assemblies was assessed using BUSCO (version 5.2.2, lineage dataset insecta_odb10). Transcripts were annotated using NCBI BLAST (version 2.12.0+), InterProScan (version 5.52–86.0), and Pepstats/Pepinfo from EMBOSS (version 6.5.7).

The transcriptome of *G. mellonella* was previously prepared (Kludkiewicz et al., 2019) using Roche GS-FLX 454 pyrosequencing according to the manufacturer's instructions. Three cDNA libraries were prepared: those from the SGs of the penultimate-instar larvae (PI), the post-feeding wandering last instar larvae (WS), and the apolyzing (initial phase of pupation) last instar larvae (ECD). These were then sequenced, and the data were concatenated.

Chromosomal localization and collinearity analyses

We used high-quality genome sequences of *E. kuehniella* (Visser et al., 2021) and *G. mellonella* (GenBank assembly accession GCA_003640425.2). Transcripts of *E. kuehniella* and *G. mellonella* silk genes were mapped to the genomes using minimap2 (version 2.24-r1122) to locate potential gene clusters (Li, 2018). Syntenic relationships were then built based on reciprocal best translated blast (tblastx, version 2.12.0+) hits among the transcriptome datasets of *E. kuehniella*, *G. mellonella*, *A. transiella* (GenBank assembly accession GCA_001186105.1), and *B. mori* (GCA_014905235.2) and visualized using R package ggplot2 (Wickham, 2009).

Silk degumming and hygroscopicity tests

To determine the percentage of soluble silk components, cocoon samples containing approximately 40 mg of dried *E. kuehniella* and *G. mellonella* silk material were cut into pieces, weighed, submerged in water, and boiled three times for 15 min. The samples were then centrifuged, and the soluble fraction was discarded. The undissolved silk remaining in the pellet was washed five times with water, vacuum dried, and weighed. The soluble fraction was measured by calculating the weight loss percentage before and after the degumming process as described previously (Kludkiewicz et al., 2019).

To measure the hygroscopicity of silk, cocoon samples approximately 40 mg each of *E. kuehniella*, *G. mellonella* and *B. mori* were vacuum dried, weighed, and then incubated in a jar with 75% humidity for 48 h. Moisture uptake was inferred from the percentage increase in the sample weight before and after the incubation. Six biological replicates were used for every sample. Statistical significance was tested by the Student's t-test in R (version 4.1.1).

The grand average of hydropathicity index (GRAVY) of the FibH was calculated using the ExPASy ProtParam server (https://web.expasy.org/protparam/) from the sum of the hydropathy values of all amino acids divided by the sequence length (Kyte and Doolittle, 1982).

Protein identification using mass spectrometry

Silk samples were dissolved in 8 M urea and further processed with SP3 as previously described (Hughes et al., 2014). After washing the samples, they were digested with trypsin, acidified with trifluoroacetic acid to a final concentration of 1%, and peptides were desalted with homemade C18 disk-packed tips (Empore, Oxford, United States) according to Rappsilber et al. (Rappsilber et al., 2007).

The samples were processed and analyzed using nanoscale liquid chromatography coupled to tandem mass spectrometry (nLC-MS/MS) as described elsewhere (Erban et al., 2020). The analysis and quantification of proteins were performed using MaxQuant label-free algorithms (MaxQuant, version 1.5.3.8) (Cox et al., 2014). The false discovery rate (FDR) was set to 1% for both proteins and peptides, and a minimum length of seven amino acids was set. The Andromeda search engine (Cox et al., 2011) was used to compare the MS/MS spectra with the transcriptome/genome-derived *E. kuehniella* protein database. Further data analysis was performed using Perseus 1.5.2.4 software (Tyanova et al., 2016).

Phylogenetic analysis

Coding sequences identified in the annotated genomes were used. Codon-based alignment was performed using MEGA7 software according to the MUSCLE method (Kumar et al., 2016). The phylogram was generated using the IQ-TREE server (Nguyen et al., 2015), which included both the selection of the best substitution model by ModelFinder (Kalyaanamoorthy et al., 2017) and tree inference using MLE (ultrafast bootstrap, 1,000 replicates).

Identification of *fibH, ser1* and *muc1* genes in other pyraloidea

G. mellonella FibH sequence was inferred from two long-read sequencing genome assemblies. The FibH sequences of *Acentria ephemerella, Acrobasis suavella, Chilo suppressalis,*

TABLE 1 Major silk proteins in *E. kuehniella* cocoons. GenBank – GenBank accession numbers; Intensity – MaxQuant peptide intensity; M. W. – molecular weight (kDa); pI – isoelectric point; H. I. – hydropathy index (GRAVY); 1st/2nd/3rd AA (%) – proportion of the three most frequent amino acids; Evid. – data to detect/infer proteins, where N, P, Q, and T represent northern blotting, proteome, qPCR, and transcriptome.

Gene/protein	GenBank	Intensity	Evid.	M. W.	pI	H. I.	1st AA (%)	2nd AA (%)	3rd AA (%)
Ek-Ser3	ON604819	8093000000	PQ	144.19	6.865	−0.823	S (52.6)	G (16.2)	T (8.3)
Ek-FibH	ON604816	5058700000	PQ	493.70	3.491	0.054	S (23.1)	G (22.0)	A (19.1)
Ek-FibL	ON604822	4001400000	NPQ	26.41	4.186	0.299	A (20.1)	G (11.0)	N (10.2)
Ek-Zon1	ON604824	3562800000	PQ	140.56	4.717	−0.542	C (15.1)	P (9.3)	G (8.6)
Ek-P25	ON604823	3071800000	NPQ	25.02	7.334	−0.076	L (8.7)	F (8.3)	N (8.3)
Ek-SerP150	ON604820	2846400000	PQ	173.26	4.212	−0.849	S (34.6)	Q (11.1)	G (9.2)
Ek-Muc1	ON604821	2764400000	NPQ	473.65	4.044	−1.378	Q (25.7)	S (23.5)	E (11.0)
Ek-Zon2	ON604825	1061200000	PQ	101.09	4.907	−0.235	C (8.9)	S (7.8)	N (7.3)
Ek-Ser1A	ON604817	231370000	NPQ	124.10	4.196	−0.545	T (19.7)	S (17.6)	A (13.1)
Ek-Lprcp66-1	ON604829	112370000	PQ	18.28	6.503	0.236	G (22.5)	A (16.6)	L (10.7)
Ek-Ser1B	ON604818	76048000	NPQ	149.79	4.068	0.032	A (17.3)	S (14.6)	P (13.5)
Ek-Sn1	ON604827	58061000	PQ	29.50	4.461	−0.603	P (13.9)	E (9.8)	V (9.8)
Ek-Lprcp66-2	ON604830	27834000	PQ	20.43	7.358	0.085	A (21.8)	G (18.9)	Y (11.7)
Ek-Zon3	ON604826	17266000	PQ	48.74	4.119	−0.431	P (11.0)	S (10.4)	C (9.7)
Ek-Sn4	ON604828	9657900	PQ	12.44	6.792	−0.323	A (11.3)	G (10.4)	N (9.3)
Ek-Sn3	OP185489	7321500	PQ	11.63	4.532	−0.679	E (10.4)	N (10.4)	S (10.4)
Ek-P47	OP185490	5865900	PQ	31.90	4.528	−0.647	E (12.0)	K (9.1)	N (8.3)
Ek-Pebp	OP185492	4880600	PQ	23.00	9.368	−0.111	V (11.3)	A (10.8)	P (8.5)
Ek-Fbn	OP185491	3261600	PQ	153.14	5.147	−0.797	P (10.8)	T (8.1)	E (8.0)
Ek-Pcp	OP185495	2008500	PQ	26.85	6.502	−0.574	A (28.0)	Q (14.2)	P (9.4)
Ek-Muc2	OP185487	764630	P	40.22	3.388	−0.866	T (36.9)	E (16.6)	S (9.1)
Ek-Csp1	OP185497	-	T	58.41	6.619	−1.298	S (10.9)	Q (9.9)	N (9.1)
Ek-Csp2	OP251358	-	T	31.29	7.353	−0.789	P (11.9)	S (8.6)	D (7.6)
Ek-Csp3	OP251359	-	T	44.91	4.582	−0.369	V (10.4)	P (10.1)	S (9.4)
Ek-Muc3	OP185493	-	T	85.32	8.361	−0.914	T (15.0)	A (11.4)	P (9.9)
Ek-P-12	OP185496	-	T	16.82	7.047	−0.625	G (39.5)	Q (9.6)	F (7.9)
Ek-P13	OP185503	-	T	10.68	9.988	−0.458	P (14.3)	V (10.2)	A (9.2)
Ek-Ser4	OP251360	-	NT	48.72	4.455	−0.613	S (40.6)	G (9.2)	A (8.4)
Ek-Sn2	OP185488	-	QT	24.71	6.226	−0.450	S (12.0)	N (10.2)	F (7.8)
Ek-Zon4	OP185494	-	T	22.19	7.112	−0.536	C (14.8)	P (10.8)	G (7.9)

Cnaphalocrocis exigua, Endotricha flammealis, Hypsopygia costalis and Plodia interpunctella were identified from high-quality genomes published by the Darwin Tree of Life Project (Blaxter and Project, 2022).

We identified *fibH* genes from these assemblies using TBLASTN and conserved N- and C-termini as query sequences. Fibroin sequences were predicted from the surrounding sequence using online Augustus (Stanke and Morgenstern, 2005). The software BioEdit (v 7.2) was used to visualize sequence alignments (Hall, 1999). Information on accession numbers of genome assemblies and sequences was shown in Table 3. We also used TBLASTN and conserved C-termini of Ser1 and Muc1 as query sequences.

Results

E. kuehniella silk and silk glands

The SG of *E. kuehniella* consists of a tube with large polyploid secretory cells. Like in other moth species, three regions can be distinguished morphologically: anterior, middle, and posterior (Figure 1A). The PSG is approximately 20% shorter than the MSG and is not folded. The SG extends about two-thirds of the length of the larval body. The ASG is narrow and gradually widens into the MSG. A large sigmoid loop forms at the junction between the MSG and the PSG. The diameter of the MSG remains more or less the same and decreases only slightly

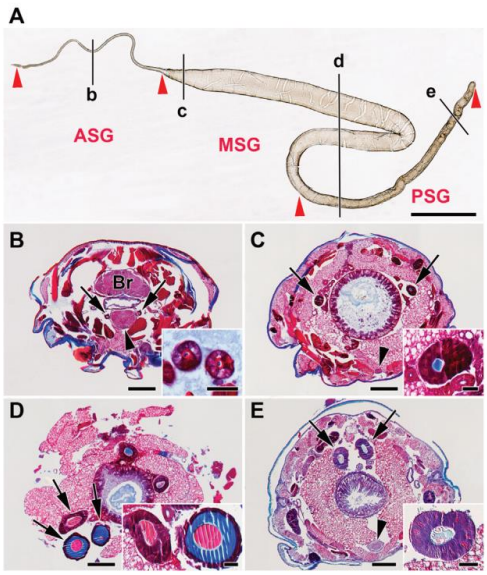


FIGURE 1
Morphology of the silk gland (SG) from *E. kuehniella* last instar larvae. **(A)** Whole mount preparation of one SG illustrates its overall morphology. Red arrowheads depict the boundaries of the SG compartments, where ASG = anterior SG, MSG = middle SG, and PSG = posterior SG. Black lines marked by the lowercase letters b–e refer to the whole-body sections B–E and show the approximate positions where the glands were cut in transverse Paraplast sections. **(B–E)** Transverse Paraplast sections through the body of the last larval instar stained with Masson trichrome stain (Sigma). The inset images show higher magnification of the SG sections marked by arrows. **(B)** ASG; brain (Br); the arrowhead depicts the suboesophageal ganglion (SOG). **(C)** Anterior portion of the MSG; arrowhead shows ventral nerve cord. **(D)** MSG in the region of the sigmoidal loop. **(E)** PSG; the arrowhead marks the ventral nerve cord ganglion. Red areas are acidic; blue areas are alkaline. Scale-bars: **(A)**, 1,000 µm; **(B–E)**, 200 µm; inset images, 50 µm.

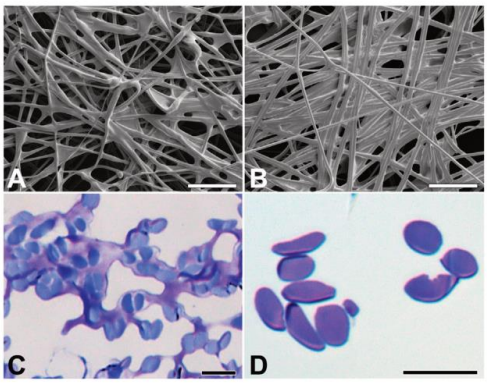


FIGURE 2
Silk of *E. kuehniella* cocoon. **(A,B)** Scanning electron micrographs of the outer and inner surfaces and inner surfaces of the cocoon, respectively. **(C,D)** Toluidine blue stained semi-thin sections of the silk fibers of the cocoon before and after degumming, respectively. Scale bars: A,B = 50 µm; C,D = 10 µm.

toward the PSG. The boundary between the rear part of the MSG and the PSG is less distinct than in *G. mellonella* or *B. mori*.

Silk fiber width varies among moth species, ranging from 12 µm in *B. mori* to 5 µm in *G. mellonella* and 0.5–1 µm in diameter in *Tineola bisselliella* (Kludkiewicz et al., 2019; Rouhova et al., 2021). To study the morphology of the silk cocoons and fibers, we characterized them using a scanning electron microscope (Figures 2A,B). The width of silk fiber of *E. kuehniella* is approximately 1 µm. The overall structure of the silk of *E. kuehniella* appears to be similar to that of *G. mellonella*.

Cells with polyploid nuclei are found along the entire length of the gland. The gland only produces fibroins in the PSG, and these are thought to be mixed with sericins and other silk components in the MSG and ASG. As can be seen on the paraffin sections stained with Masson trichrome stain, there are color differences in the liquid silk in the different glandular compartments. At least two types of secretion are seen on the glandular sections: a column of fibroin surrounded by a layer of sticky sericins (the color changes in the stained sections). While the fibroin in the PSG was stained blue, both the fibroin and sericin in the MSG and ASG were red (Figures 1B–E).

Silks differ in sericin content from 26% in *B. mori* to 48% in *G. mellonella* (Kludkiewicz et al., 2019). To compare the percentage of these proteins in the silks, we dissolved the coating proteins of *E. kuehniella*, *G. mellonella*, and *B. mori* by degumming the silk in water, which dissolved and removed most of the sericin layer. The removal of sericins was verified by microscopic examination. Interestingly, the degumming dissolved also part of the silk core of *E. kuehniella* (Figures 2C,D). Therefore, we concluded that the silk of *E. kuehniella* is more soluble than that of *G. mellonella* and, therefore, that this method is not suitable for measuring the exact proportion of sericins in *E. kuehniella* silk.

Because the properties of different silk species can be distinctive with respect to water, such as solubility or even hygroscopicity, we also tested water adsorption from the environment. As can be seen in Figure 3, *E. kuehniella* silk is highly hygroscopic compared to *G. mellonella* and *B. mori* silks, and it can retain twice as much water as *G. mellonella* silk (68% and 30%, respectively; Figure 3).

Transcriptome *de novo* assembly

The first *de novo* assembly of the silk gland-specific transcriptome of *E. kuehniella* revealed 43,923 contigs with an

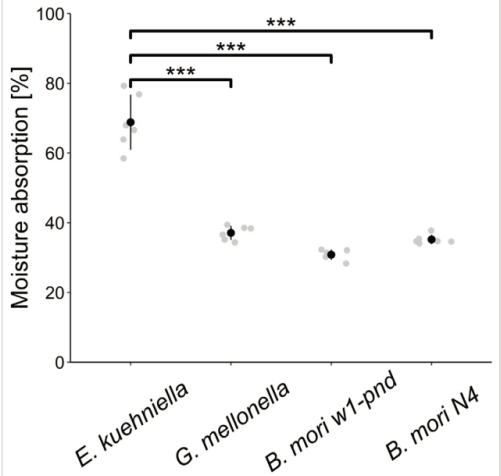


FIGURE 3
Hygroscopicity evaluation. The scatter plot shows the average amounts of water absorption in the dehydrated cocoon silk of *E. kuehniella*, *G. mellonella*, and *B. mori w1-pnd* and *N4* strains. Black dots represent the mean, and vertical bars represent the mean ± SD. *** indicates $p < 0.001$ (paired *t*-test comparing means, $n = 6$).

average length of 1,012.7 base pairs. The completeness of the non-redundant transcripts was assessed using the BUSCO tool suite. The results showed that the transcriptome was 81.6% complete. However, approximately 25% of the complete and duplicated BUSCOs indicated redundant isoforms, and approximately 19% of the incomplete BUSCOs (8% fragmented and 10.4% missing) indicated missing genes. To address this issue, we took advantage of the long-read genome assembly of *E. kuehniella* (Visser et al., 2021) and used a combined transcriptome assembly strategy to generate an improved transcriptome. The MAKER-annotated genome contained a total of 13,382 recovered protein-coding genes with an average gene length of 7,207.8 base pairs. The BUSCO statistics showed a completeness of 98.6%, while the levels of redundancy and incompleteness decreased to 0.5% and 1.4%, respectively. We concluded that this improved transcriptome contained high-quality data suitable for further analysis (see Supplementary Table S3).

Detection of *E. kuehniella* candidate silk proteins

Sequence annotation revealed that a substantial proportion of cDNAs encode ribosomal proteins or proteins involved in protein translation or transport. Potentially secreted proteins identified by the presence of a putative signal peptide accounted for approximately 10% of all annotated contigs. Because silk genes evolve rapidly, it is difficult to identify them in new moth species based on homology without information on genes from a closely related species. In this way, we were able to reliably identify the sequences of the FibH, the fibroin light chain, and P25/fibrohexamerin (P25) based on homology to known genes.

Because *E. kuehniella* silk was available to us, we chose proteomic analysis as the primary method for detecting the gene sequences that encode its components. The proteins of a silk cocoon were dissolved in urea and trypsinized, and the resulting peptides were analyzed using proteomic analysis. The MS/MS spectra of the peptides were aligned with the protein sequence database derived from the reference transcriptome. It was expected that most of the proteins detected in the silk would not be structural components because some of the housekeeping proteins are secreted from SG cells *via* apocrine-like secretion during silk spinning. We identified 140 proteins, 77 of which contained a predicted signal peptide sequence. BLAST-based annotations were performed using the NCBI nr database, and the annotations were manually verified. Based on the annotations, we excluded most proteins with close homologs in other moth species that were not associated with silk structure.

Expression specificity of candidate *E. kuehniella* proteins

Putative structural silk proteins are likely to be abundant. They carry signal peptides at their N-terminal sequences, and their transcripts are specific to the SG. We isolated total RNA from control larvae with ablated SGs, as well as from different parts of the SG. The SG specificity of the candidate transcripts encoding silk proteins was confirmed using northern blotting and qPCR analyses. The northern blotting analysis revealed that some candidate genes could produce more bands, suggesting alternative splicing (Figure 4A). Most of the transcripts showed distinct differences in expression in SG sub-regions. For example, *Ek-serP150* is highly expressed in the anterior part of the MSG, whereas the *Ek-Zon1* and *Ek-Ser1A* transcripts are predominantly expressed in the rear MSG (Figure 4B). Interestingly, maximal expression of the *Ek-P25* transcript was found in the rear MSG.

Comparison of candidate silk proteins between *E. kuehniella and G. mellonella*

To identify a complete set of candidate silk-encoding genes of *E. kuehniella*, we performed a parallel study on *G. mellonella*. Previous results on the silk of *G. mellonella* were supplemented

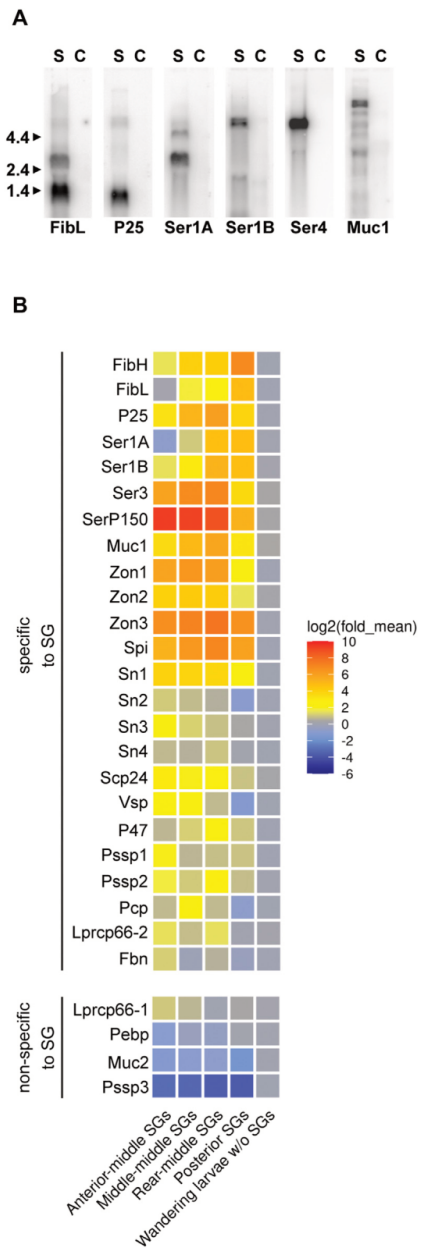


FIGURE 4
Expression of selected silk genes detected by **(A)** northern blotting and **(B)** qPCR. **(A)** Lanes: C – larva without SG, S – larval SG. Total RNA (5 μg) was separated on an agarose gel, blotted to a nylon membrane and probed with [^{32}P]-labelled cDNA
(*Continued*)

FIGURE 4 (Continued)
fragments from each of the indicated genes. The length (kb) of the size marker is indicated on the left side. **(B)** Relative expression of candidate tissue-specific genes in controls and parts of silk glands of last instar larvae examined *via* qPCR. mRNA expression levels were normalized to the internal reference gene elongation factor 1-α. Heatmap was plotted based on log2-transformed fold change between SG and control, indicated by the colored scale. Genes of significantly higher expression level in SG ($p < 0.05$) were classified as SG-specific genes (see Supplementary Table S2 for statistics). Gene names are shown in Table 1.

with a new cocoon protein proteomic analysis. The tryptic peptides were tested against the custom protein database derived from the NCBI dataset (GenBank assembly accession GCA_003640425.2) and the previously created transcriptome (Kludkiewicz et al., 2019). The resulting set of *G. mellonella* sequences was used to search for homologous sequences in the transcriptome of *E. kuehniella* and *vice versa*. Data for both species were then complemented based on homology. Thus, BLAST searches of the *G. mellonella* sequences of P-12 (LOC113521678), P13 (LOC113521978), mucin-5AC-like (LOC113516440), seroin 2 (LOC113518101) zonadhesin (LOC113516017), and candidate silk proteins (LOC113523011, LOC113515440, and LOC113511581) revealed new putative *E. kuehniella* homologs that were not detected *via* the proteomic approach (Table 1). Conversely, we discovered homologs of *E. kuehniella* proteins annotated as fibrillin (OP185491), pupal cuticle protein-like (OP185495), and phosphatidylethanolamine-binding protein (OP185492) in the *G. mellonella* transcriptome (not found in *G. mellonella* silk proteomics). In addition, the new proteomic analysis of *G. mellonella* cocoon silk revealed 18 silk protein candidates that were not previously detected in silk, including several sericins, mucin, zonadhesin, seroin, and cuticle proteins (Table 2).

Interestingly, we found no clear homologs of several silk proteins of *G. mellonella* in *E. kuehniella*, suggesting that these proteins may be putative species-specific genes. These include proteins such as P250 (LOC113513637), P17 (LOC113512752), MG5 (LOC113521079), P22 (LOC113513778), GMPiso00198 (LOC113512751), MG4 (LOC113509977), P-8 (LOC113513777), MG9 (LOC116413334), MG-2 (LOC113519334), P-7/P14 (LOC116413327), MG6 (LOC113509728), MG-1/MG-3 (LOC113517751), MG7 (LOC113522155), MG8 (LOC116413345), P-11 (LOC113511945), GMPiso00090/GMPiso00234 (LOC113513780) were previously characterized. Several other novel proteins were detected in this study in the silk of *G. mellonella* and automatically annotated as dentin sialophosphoprotein-like (LOC113509273), cuticle protein LPCP-23-like (LOC113509759), cell wall protein IFF6-like (LOC113511377), serine-rich adhesin for platelets-like (LOC113523571), protein PB18E9.04c-like (LOC113512274), and sericin-2-like (LOC113522365; see Table 2).

TABLE 2 Major silk proteins in *G. mellonella* cocoon. GenBank accession numbers; Intensity – MaxQuant peptide intensity; M. W. – molecular weight (kDa); pI – isoelectric point; H. I. – hydrophobicity index (GRAVY); 1st/2nd/3rd AA (%) – proportion of the three most frequent amino acids; Evid. – protein newly identified in this paper (A) or previously identified in (Kludkiewicz et al., 2019) (B).

Gene/protein	Genbank	Intensity	Evid.	M. W.	pI	H. I.	1st AA (%)	2nd AA (%)	3rd AA (%)
Gm-FibL	XM_026895755	22648000000	AB	27.01	3.994	0.352	A (19.1)	L (9.7)	N (9.4)
Gm-FibH	XM_026905081	20827000000	AB	487.83	3.723	0.553	G (31.3)	A (23.3)	S (17.6)
Gm-P250	XM_031911780	18581000000	AB	56.27	11.222	−0.564	S (34.0)	G (10.6)	P (9.9)
Gm-Muc4-L	MG770312	15626000000	AB	174.28	4.301	−0.883	S (19.7)	Q (11.9)	E (9.7)
Gm-P25	XM_026894481	11985000000	AB	24.84	5.301	−0.047	L (11.0)	N (10.6)	A (8.3)
Gm-Ser1B	XM_031911923	5701200000	AB	84.29	6.258	−0.805	S (35.2)	A (20.7)	N (14.0)
Gm-P17	XM_026896665	5607100000	AB	9.31	7.906	−0.666	G (29.9)	K (16.5)	D (14.4)
Gm-P150	XM_026908157	3074000000	AB	155.41	4.745	−1.247	S (22.4)	Q (17.3)	N (7.8)
Gm-Sn1	XM_031913182	1892500000	AB	30.26	4.715	−0.535	P (15.6)	N (9.1)	F (8.0)
Gm-MG5	MG770318	1173700000	AB	65.34	10.791	−1.257	S (45.1)	N (13.3)	R (7.9)
Gm-Ser1A	MG770315	941580000	AB	94.82	3.627	−0.521	S (22.3)	Q (14.4)	A (8.6)
Gm-P22	MG770325	629610000	AB	22.24	4.497	−0.370	T (18.0)	S (14.7)	P (10.6)
Gm-ZdA	XM_026895687	478590000	AB	104.52	4.442	−0.343	C (8.7)	N (7.8)	G (7.4)
Gm-P47	XM_026903570	468680000	AB	55.37	3.894	−1.033	N (16.1)	E (13.2)	S (9.0)
Gm-GMPiso00198	XM_026896664	383160000	AB	15.44	4.006	−0.804	E (18.4)	P (17.7)	C (9.9)
Gm-MG4	XM_026893397	104810000	AB	96.02	6.345	−1.040	S (40.2)	N (24.6)	G (6.0)
Gm-Sn3	XM_026903091	100400000	AB	11.56	8.498	−0.058	N (12.3)	V (12.3)	F (8.5)
Gm-Usp03	XM_031907470	85362000	A	122.76	3.827	−0.109	T (16.8)	S (11.4)	I (8.2)
Gm-P-8	MH464805	68619000	AB	8.41	8.864	−0.381	K (12.7)	P (11.4)	A (7.6)
Gm-GMPiso00148	XM_026905986	59040000	AB	38.03	4.258	−1.217	A (19.0)	E (15.2)	K (14.9)
Gm-Usp01	XM_031910267	39692000	A	78.93	7.319	−1.186	N (13.2)	P (9.6)	S (9.5)
Gm-Usp02	XM_031911792	29142000	A	78.80	4.260	−0.445	S (14.9)	P (13.3)	Q (10.7)
Gm-Ds-like	XM_031908961	27903000	A	106.79	4.419	−0.744	S (21.0)	N (8.9)	E (8.8)
Gm-MG9	XM_031911849	25465000	AB	53.65	4.522	−0.664	S (51.4)	A (18.0)	N (12.9)
Gm-Muc-6-like	XM_026903992	12493000	A	35.23	3.737	−0.343	P (12.0)	C (10.2)	S (8.6)
Gm-Eln-like	XM_026907481	11901000	A	27.18	10.646	0.539	A (30.1)	G (14.9)	L (8.3)
Gm-Lprcp66-like	XM_026899336	11741000	A	20.19	8.846	0.022	G (24.6)	A (12.3)	Y (12.3)
Gm-P13	XM_026907679	11723000	AB	10.22	9.408	−0.535	P (14.6)	G (11.5)	A (9.4)
Gm-Papln-like	XM_031914292	8296100	A	273.14	4.388	−0.645	E (9.4)	T (9.4)	G (9.0)
Gm-Iff6-like	XM_026895003	5793800	A	49.24	4.216	−0.933	S (29.2)	G (20.2)	N (18.6)
Gm-Cp21-like	XM_026905740	5698400	A	30.92	9.426	0.609	A (35.9)	V (14.4)	P (12.5)
Gm-P-12	MH464803	5559200	AB	11.77	8.505	−0.625	G (31.4)	S (17.4)	Q (8.3)
Gm-Sn4	XM_026903191	5067300	A	11.96	7.834	−0.243	S (11.6)	G (10.7)	V (10.7)
Gm-Imuc-C.1-like	XM_031908793	2629500	A	21.24	4.321	−0.323	T (17.7)	E (8.9)	V (8.3)
Gm-Sn2	XM_031913070	2473700	AB	19.70	9.317	−0.330	A (10.9)	S (10.9)	N (10.3)
Gm-Muc5AC-like	XM_026900881	2253000	A	96.92	7.745	−0.885	T (14.5)	A (9.8)	P (9.2)
Gm-Lpcp-23-like	XM_026893171	1278700	A	35.69	9.832	0.747	A (22.8)	S (11.9)	L (9.2)
Gm-Zon-like	XM_026900349	1113800	A	46.15	7.341	−0.393	C (14.6)	P (8.9)	K (6.9)
Gm-LcpA2B-like	XM_026905750	863070	A	15.56	6.512	0.302	A (15.5)	P (15.5)	V (13.5)
Gm-Cp3-like	XM_026892615	783140	A	23.16	6.801	0.525	A (35.4)	V (9.2)	G (7.5)
Gm-Ser2-like	XM_031911858	727170	A	24.05	3.881	−0.631	S (53.5)	A (9.8)	N (7.8)
Gm-MG-2	XM_031911783	727170	AB	30.75	3.896	−0.602	S (52.3)	A (10.4)	N (6.7)
Gm-Fbn	XM_026904452	-	A	150.36	5.483	−0.756	T (11.0)	P (8.7)	C (8.0)
Gm-Pcp-like	XM_026899249	-	A	25.50	6.292	−0.592	A (26.9)	Q (12.4)	P (11.2)
Gm-Pedp1-like	XM_026894507	-	A	14.19	9.149	−0.655	P (9.5)	G (8.7)	K (7.9)
Gm-MG-1/MG-3	XM_031911850	-	B	30.62	3.823	−0.479	S (42.1)	G (15.8)	N (8.8)

(Continued on following page)

TABLE 2 (*Continued*) Major silk proteins in *G. mellonella* cocoon. GenBank accession numbers; Intensity – MaxQuant peptide intensity; M. W. – molecular weight (kDa); pI – isoelectric point; H. I. – hydrophobicity index (GRAVY); 1st/2nd/3rd AA (%) – proportion of the three most frequent amino acids; Evid. – protein newly identified in this paper (A) or previously identified in (Kludkiewicz et al., 2019) (B).

Gene/protein	Genbank	Intensity	Evid.	M. W.	pI	H. I.	1st AA (%)	2nd AA (%)	3rd AA (%)
Gm-MG6	XM_026893136	-	B	91.96	10.571	−1.551	S (54.6)	N (16.0)	E (6.7)
Gm-MG7	XM_031911880	-	B	55.07	3.360	−0.862	S (38.3)	N (22.4)	G (11.1)
Gm-MG8	XM_031911957	-	B	28.34	3.517	0.082	S (32.0)	G (23.2)	A (17.4)
Gm-P-11	XM_026895673	-	B	10.95	4.986	−0.291	S (12.1)	L (11.1)	G (8.1)
Gm-P-7/P14	XM_031911797	-	B	14.34	3.830	−0.261	D (14.0)	I (12.4)	S (12.4)
Gm-GMPcon00005	XM_026895223	-	B	57.59	4.701	−0.243	A (10.8)	V (9.9)	S (9.5)
Gm-GMPiso00278	XM_026899644	-	B	32.09	6.502	−0.866	P (11.0)	D (7.8)	Q (7.4)
Gm-GMPiso00090/00234	XM_026897762	-	B	26.45	8.911	−0.474	L (10.4)	S (9.5)	A (7.9)

TABLE 3 Comparison of protein parameters of FibH including number of amino acid residues, molecular weight, percentage of three major amino acids, H.I. - hydropathy index (GRAVY) and isoelectric pont (pI). Genbank accession numbers: *G. mellonella* (XM_026905081), *E. kuehniella* (ON604816), *P. interpunctella* (JAJAFS010000023.1), *A. suavella* (OW971947.1), *E. flammealis* (LR990872.1), *H. costalis* (OW443343.1), *C. exigua* (CM032477.1), *Ch. Suppressalis* (OU963910.1), *A. ephemerella* (OW971889.1), *B. mori* (NM_001113262.1), *A. yamamai* (AB542805.1) and *S. cynthia* (AB971865).

Species	Family	Subfamily	Number a.a.	Mw	Ala (%)	Gly (%)	Ser (%)	H. I.	pI
Galleria mellonella	Pyralidae	Galleriinae	6020	487830	23.30	31.30	17.60	0.553	3.94
Ephestia kuehniella	Pyralidae	Phycitinae	5631	493710.7	19.10	21.90	23.10	0.054	3.81
Plodia interpunctella	Pyralidae	Phycitinae	4714	413334.4	26.30	18.30	18.50	0.084	4.04
Acrobasis suavella	Pyralidae	Phycitinae	6057	534495.5	15.50	23.9	24.20	-0.438	3.78
Endotricha flammealis	Pyralidae	Pyralinae	8027	685384.5	34.90	20.00	19.90	0.164	3.43
Hypsopygia costalis	Pyralidae	Pyralinae	5981	521977.6	22.20	24.9	25.70	-0.17	3.323
Cnaphalocrocis exigua	Crambidae	Pyraustinae	5418	485846.3	20.60	26.30	5.10	-0.452	4.31
Chilo suppressalis	Crambidae	Crambinae	4386	383147.9	43.10	16.00	13.00	0.011	5.33
Acentria ephemerella	Crambidae	Nymphulinae	6151	533043.56	21.9	27.8	10.3	-0.138	9.78
Bombyx mori	Bombycidae	Bombycinae	5263	391633.58	30.2	45.9	12.1	0.216	4.39
Antheraea yamamai	Saturniidae	Saturniinae	2856	234576.05	42.9	27.2	11.0	0.186	4.49
Samia cynthia	Saturniidae	Saturniinae	2880	227361.93	45.4	31.7	6.7	0.336	4.85

This comparative approach allowed us to overcome some limitations of the proteomic analysis and helped to identify additional silk genes. The final list of *E. kuehniella* and *G. mellonella* candidate silk proteins is shown in Tables 1, 2.

Some of the silk genes are arranged in clusters

Our results show that several silk genes including sericins, seroins, and zonadhesins, form clusters on different chromosomes. The genes for sericins were located in two clusters that are conserved between species to some extent. However, the microsynteny of one of these clusters appears to be impaired in *G. mellonella* compared to *E. kuehniella,* and *A. transitella*. Figure 5 shows the microsynteny of a genomic region containing mainly genes for sericins from three moths flanked by evolutionarily conserved genes (see Table 3). The lines connect the putative orthologs of the conserved genes. Comparison of sericin coding regions between species showed that a number of sericin genes are in tight cluster, which is present only in *G. mellonella* but not in the other two species. Our data support the hypothesis that the genomic region encoding the sericin gene cluster has been recently duplicated in *G. mellonella*.

Phylogenetic relationships among silk genes

The two major classes of putative adhesive proteins produced by the MSG are sericins and mucins. They are generally encoded by large genes with repetitive sequences that contain a high proportion of serine residues. We identified at least four putative sericin proteins (Ek-Ser1A, Ek-Ser1B, Ek-Ser3 and Ek-Ser4), three different mucins (Ek-Muc1, Ek-Muc2 and Ek-Muc3), and one protein (Ek-P150) in *E. kuehniella* that can be

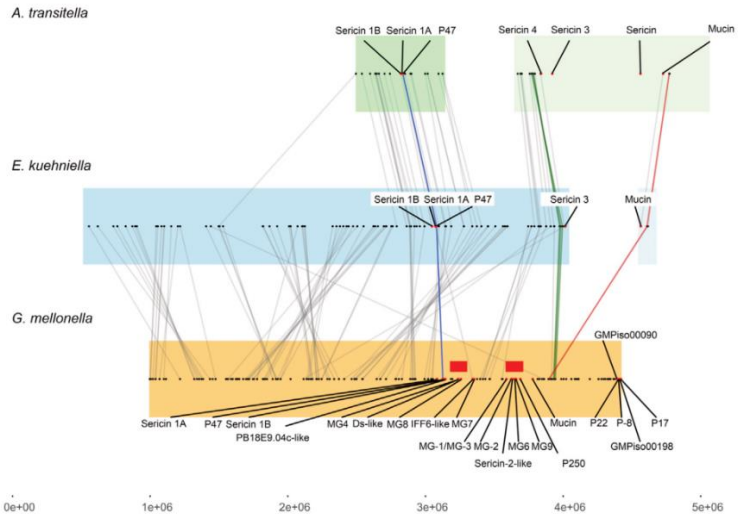


FIGURE 5
Collinearity analysis of sericin genes among *E. kuehniella*, *G. mellonella*, and *A. transitella*. Collinear blocks are highlighted with background colors. Silk genes are indicated with red dots. Red rectangles indicate the area of the *G. mellonella* genome containing recently duplicated silk genes. The color and gray lines connect the syntenic gene pairs. The proposed "landmark" genes (see Supplementary Table S5 for gene names) that are adjacent to the silk genes in three species are connected with blue, green and magenta lines.

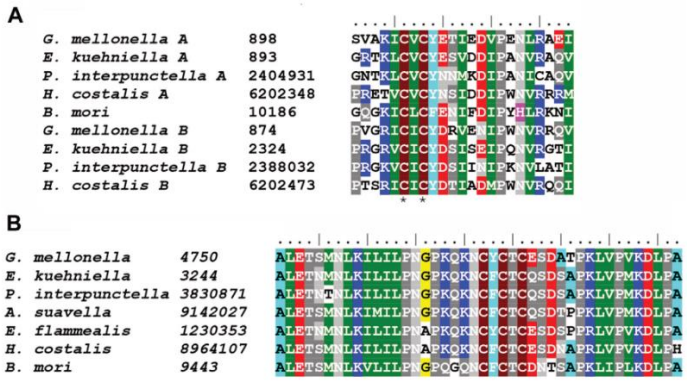


FIGURE 6
(A) Multiple sequence alignment of Ser1 C-terminal sequences containing conserved CXCX motif: *G. mellonella* Ser1A (MG770315), *E. kuehniella* Ser1A (ON604817) *P. interpunctella* Ser1A (JAJAFS010000011.1), *H. costalis* Ser1A (OW443355.1), *B. mori* Ser1 (XM_038013610.1), *G. mellonella* Ser1B (XM_031911923), *E. kuehniella* Ser1B (ON604818), *P. interpunctella* Ser1B (JAJAFS010000011.1), *H. costalis* Ser1B (OW443355.1). Conserved cysteine residues are marked with asterisks. **(B)** Multiple sequence alignment of Muc1 C-terminal sequences containing conserved CXCXCX motif: *G. mellonella* (MG770312.1), *E. kuehniella* (ON604821), *P. interpunctella* (JAJAFS010000011.1), *A. suavella* (OW971956.1), *E. flammealis* (LR990879.1), *H. costalis* (OW443355.1), *B. mori* (XM_021350004.2). Conserved cysteine residues are marked with asterisks.

classified as both a sericin and a mucin. Two sericin-1-like proteins, Ek-Ser1A and Ek-Ser1B, carry a CXCX motif near the C-terminus, whereas Ek-Muc1 has the three-Cys motif NCFCTC near the C-terminus (Figure 6), similar to the KCYCSC motif of Ek-P150. Such motifs seem to be conserved among species.

It has previously been suggested that sericin 1 (Ser1), mucin-1 (Muc1), and P150 loci might be related (Kludkiewicz et al., 2019). To explore possible phylogenetic relationships, we tested these sequences in one dataset. The phylogram (Supplementary Figure S1A) clearly shows, apart from mucin-1, a distinct group of Ser1 and P150. Within this group, the P150 loci formed well-supported subclusters, but the discrimination of Ser1 and P150 may not be complete because Gm-Ser1A is separated from the other Ser1-like proteins.

We found four seroin-like proteins localized in a single cluster in the genomes of both *E. kuehniella* and *G. mellonella*. All four seroin types contained putative signal peptides. Except for Ek-Sn2, all were found in the cocoons of both species *via* proteomic analysis (Ek-Sn2 was inferred from homology and its expression was validated by qPCR). Previously, only three seroin genes were identified in *G. mellonella* (Kludkiewicz et al., 2019). As shown in Supplementary Figure S1B, the newly discovered seroin 4 forms a separate branch.

Within the Pyraloidea superfamily, species have only one copy of the *fibrohexamerin (P25)* gene, and its genealogy follows the phylogeny of the Pyraloidea (Supplementary Figure S1C). At least four zonadhesin-like proteins have been detected in *E. kuehniella*. Zonadhesins apparently have EGF_2/TIL domains (these domains partially overlap; see Supplementary Table 4). Phylogenetic analysis revealed that these genes belong to three well-delineated clusters (Supplementary Figure S1D). A schematic diagram showing the evolutionary relationships of the Pyraloidea moths is also shown for comparison in Supplementary Figure 1E (adapted from Regier et al., 2012).

Comparison of FibH proteins from nine species of pyraloidea

To learn more about the specific and conserved features of FibH proteins, we identified seven additional *fibH* genes from pyralid moths in assemblies published by the Wellcome Sanger Institute. The species comprise two families (five members of Pyralidae and three Crambidae) and six subfamilies. The list of species and protein parameters is shown in Table 3. A schematic representation of the FibH sequences is shown in Supplementary Figure S2.

The length of fibroin proteins ranges from 4386 (*Ch. suppressalis*) to 8027 amino acids (*E. flammealis*). As expected, FibH molecules consist of nonrepetitive N- and C-termini that are well conserved, and large repetitive regions, that are highly species-specific (Figure 7). As shown in Supplementary Figure S2, the arrangements and lengths of the repetitive sequences vary considerably, but all sequences of the Pyralidae memebrs (*A. suavella*, *E. flammealis*, *H. costalis*, *G. mellonella* and *P. interpunctella*) have clusters of amino acid residues similar to *E. kuehniella* VIVIEENQSSAAAAASSSSS with 4 hydrophobic amino acids and 1-4 hydrophilic amino acids, and a crystalline domain with a block of alanine and serine residues. The hydrophobic motif also occurs in *A. ephemerella* which belongs to the family Crambidae and in the C-terminal part of *C. exigua* FibH. while in *Ch. suppressalis* this motif does not exist (Supplementary Figure S2).

There are major differences in the hydrophobicity of the FibH proteins, with *G. mellonella* having the most hydrophobic fibroin, whereas the silks of *A. suavella* and *C. exigua* are very hydrophilic (Table 3). The FibH of *E. kuehniella* is much less hydrophobic than the FibH of *G. mellonella* (see Supplementary Figures S2A,B), which is probably related to the high hygroscopicity and solubility of this silk (see above). The fibroin genes can also be divided into two categories according to the regularity of the arrangement of the repeated sequences, with the fibroins of *E. kuehniella* and *G. mellonella* showing a very regular arrangement of these sequences while *Ch. suppressalis* is the most irregular (Supplementary Figure S2I). Interestingly, *A. ephemerella* FibH contains 3.7% tyrosine residues and the isoelectric point (pI) is 9.62 which is reminiscent of the FibH from caddisfly *P. conspersa* (Rouhová et al., 2022). *A. ephemerella* is an aquatic insect whose larva pupates in an underwater cocoon filled with air.

Discussion

In this study, we identified 30 genes encoding major silk components of *E. kuehniella* cocoon and verified specificity of their expression. We also analyzed silk genes from *G. mellonella*, which belongs to the same moth family. By comparing the silk genes of the two species, we gained insight into the degree of divergence between the species and found that several orthologs of genes encoding sericins present in *G. mellonella* are absent in *E. kuehniella*. In addition, we annotated the *fibH* genes of several other members of the Pyraloidea and analyzed their sequences.

Specific features of silk from pyralid moths

The silks studied so far are characterized by the insolubility of the fibers and the solubility of the sericin coating. Consequently, the fibroins of *B. mori*, *A. yamamai*, and other saturniids are hydrophobic with an GRAVY index representing the hydrophobicity (Kyte and Doolittle, 1982) ranging from 0.186 to 0.336. Interestingly, the GRAVY indexes of fibroins in the Pyraloidea vary greatly from the extremely hydrophobic fibroin of *G. mellonella* (GRAVY = 0.553) to the hydrophilic fibroins of *A. suavella* or *C. exigua* with a negative GRAVY index of −0.440 and −0.452, respectively. The silk of *E. kuehniella*

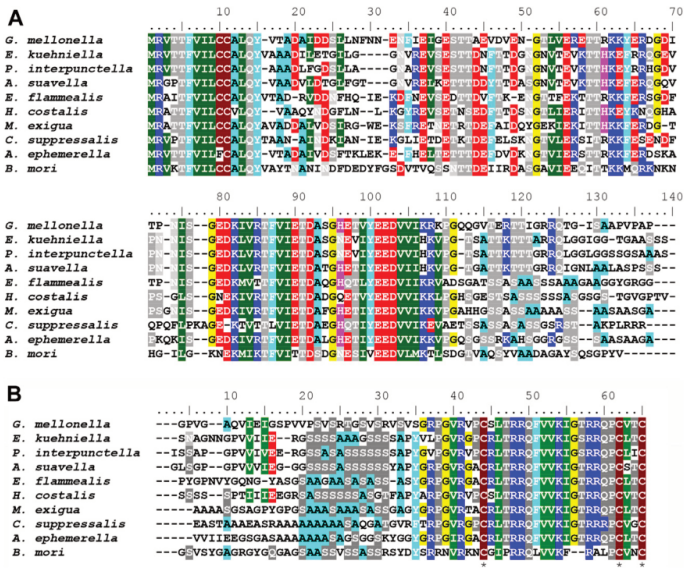


FIGURE 7
(A) Multiple sequence alignment of FibH N-terminal sequences **(B)** Multiple sequence alignment of FibH C-terminal sequences Conserved cysteine residues are marked with asterisks. Genbank accession numbers are listed in Table 3. Complete amino acid FibH sequences are shown in Supplementary Figure S2.

exhibits intermediate hydrophobicity (GRAVY = 0.054), and is readily soluble under the conditions used in degumming the silk of other species.

Comparison of available FibH sequences from Pyraloidea revealed remarkable differences in size, amino acid composition, structure of repeats etc. These molecules contain putative crystalline regions consisting of Ala and Ser residues typical of molecules of the X-ray class III (Lucas and Rudall, 1968), which are shorter than similar $S(A)_{13\text{-}15}S$ motifs in FibH proteins of *A. yamamai* or *S. ricini*. Crystalline sequences include the SSAAAAAASSSS motif in *E. kuehniella* and the SSAASAAAA motif in *G. mellonella* (Supplementary Figure S2). Previous experiments have shown that fibers with regularly ordered repeat sequences of fibroins from *G. mellonella* and *E. kuehniella* have much higher tensile strength than fibers from *P. interpunctella* with disordered repeat sequences (Fedic et al., 2003). Interestingly, *C. exigua*, contains a crystalline A_8S_2 sequence accompanied by $(PXX)_{8\text{–}21}$ motifs (Supplementary Figure S2). Such motifs have been shown to form so called polyproline II helices that can self-assemble and form compacted structures (Jin et al., 2009).

It has been consistently reported that the silks of some arthropod species can absorb considerable amounts of water and that they are quite hygroscopic; for example, the aggregate glue in spider webs absorbing atmospheric water and dissolving glycoproteins so that they spread and adhere upon contact with flying insects (Opell and Stellwagen, 2019). The silk of *B. mori* can absorb up to 30% of its weight in water (Hasan et al., 2019). Silk is considered as a highly hygroscopic material, and degummed silk is slightly less hygroscopic because the sericins absorb better than the fibre (Sonwalkar, 1993). Our results show that the hygroscopicity of *E. kuehniella* silk is extremely high, at least twice that of *B. mori* or *G. mellonella*. It is likely that both the sericins and fibroin core contribute to this property. The high hygroscopicity of *E. kuehniella* silk is possibly an adaptation to the dry environment in which Mediterranean flour moths and other members of subfamily Phycitinae live. The cocoon of *E. kuehniella* may help it absorb water from the air and protect the pupa from desiccation. It has previously been reported that *E. kuehniella* silk appears to increase moisture in stored agricultural products, increasing the likelihood of fungal outbreaks (https://www.internationalpheromones.com/product/meal-moths-ephestia-plodia-species/).

Genes encoding coating proteins

The adhesive proteins that form the envelope around the fibroin core can be formally divided into several classes, including

Ser1-like proteins, high-serine outer layer sericins, P150-like sericins, mucins, and zonadhesin-like proteins.

Ser1-like proteins are expressed in the rear part of MSG and are deposited on the fibroin core as the first sericin layer (Takasu et al., 2007; Kludkiewicz et al., 2019; Rouhova et al., 2021). *B. mori* contains a single Ser1-like gene consisting mainly of a repetitive sequence of 38 amino acid residues, of which 31% are serine residues. It is expressed in the middle and posterior regions of the MSG, and four to five Ser1 transcripts are generated by alternative splicing. The truncated Ser1 mutant of *B. mori* tends not to spin and often forms coarse cocoons (Takasu et al., 2017). Interestingly, Ser1-like proteins in other species, such as *E. kuehniella* and *G. mellonella*, are encoded by two genes (Ser1A and Ser1B), encoding proteins with 14–17% serine residues in *E. kuehniella* and 22–35% serine residues in *G. mellonella*. All Ser1 proteins contain CXCX motifs near their C-terminus, including those of *B. mori*, *A. yamamai*, and *T. bisselliella* (Kludkiewicz et al., 2019). Ser1-like proteins may represent a constant sericin component present in most moth silks and could have a specific function by directly covering the fibroin fiber as the innermost layer.

It has been reported that *B. mori* has only one sericin protein in the cocoon besides Ser1, namely sericin 3 (sericins 2 and 4 are not present in the cocoon silk) (Takasu et al., 2007; Dong et al., 2019). Ser3 is characterized by a high serine content, is localized in the outer silk layer, and possibly serves as a lubricant to reduce friction during secretion (Takasu et al., 2007). We discovered a putative sericin gene product from *E. kuehniella* (Ek-Ser4), which contains more than 40% serine residues, resembling Ser3 of *B. mori* or MG2 or MG6 of *G. mellonella*. In contrast, there are at least eight sericins with high-serine in *G. mellonella*. Homologs of these *G. mellonella* genes are most likely absent in *E. kuehniella*. Previous phylogenetic analyses support the idea that sericins resembling Ser3 in *B. mori* expand multiple times during evolution, as suggested by the species-specific branching of sericin proteins in the phylogenetic trees of *G. mellonella, A. yamamai* and *S. cynthia ricini* (Kludkiewicz et al., 2019). The high proportion of sericins in the silk of *G. mellonella* may contribute to the compactness of the cocoon in this species as required for its protection from bees. Our results suggest that sericins may serve as adhesives, lubricants or regulators of silk compactness and cross-linking of silk proteins.

Previous phylogenetic analyses have shown that the cocoon of *G. mellonella* contains a very abundant sericin-like protein called P150. It appeared to be phylogenetically quite distant from other sericin genes (Kludkiewicz et al., 2019). In this study, we discovered the putative homolog of this protein in *E. kuehniella* and named it Ek-P150. Phylogenetic analyses show that P150-like proteins form a distinct group that can be classified as both mucin and sericin because they contain serine rich repeats and a CXCXCX motif near their C-termini.

Mucins form a distinct group of silk proteins that contain serine-rich repetitive sequences. Unlike sericins they include repeats with threonine and proline residues (Syed et al., 2008). Such proline-threonine-serine motifs usually account for about one-third of the total length of the mucin protein (Perez-Vilar and Hill, 1999). We found at least three mucins in the silk of *E. kuehniella* and at least four in the silk of *G. mellonella*. At least one of them in each species could represent mucin-1-like proteins with a CXCXCX motif near the C-terminus. Our study is consistent with the idea of a common origin of the mucin and sericin families.

Impact of omics methods on silk research

Omics methods allow silk to be studied with great breadth and resolution by capturing large, if not complete, populations of genes or proteins that are related to their structure and elucidating the evolutionary and structural relationships among them. In addition, parallel study of related species allows comparison and completion of missing information. Comparative data can be also supplemented using sequences from high-quality genomes published by Wellcome Sanger Institute and other sources. However, due to the rapid diversification of these proteins, the similarities among silk proteins are not obvious, sequence conservation is rare and limited to species of the same family or subfamily. Identification of the structural components of silk still requires "traditional methods". The lack of similarity and correct annotation of genomic data is an important limitation. Here we were able to detect heavy chain fibroins in a number of pyralid moths based on sequence conservation in both ends of the fibroin molecules and compare their structures. Our results suggest that Ser1 and Muc1 can also be detected based on similarity. The use of BLAST methods for most other genes is still limited and needs more information.

Using omics data, we analyzed homologous regions on chromosomes and traced the synteny of putative sericin sequences in tight clusters. Detailed analysis revealed that the sericin region containing at least 12 genes in the *G. mellonella* genome is likely the result of a recent duplication. Synteny provides a robust framework for the identification of homologs to known genes, helps searching for new genes, and provides important information on annotation. For example, two of the genes localized in the *G. mellonella* genomic sericin cluster—annotated as dentin sialophosphoprotein-like (XM_031908961) and cell wall protein IFF6-like (XM_026895003)—are expressed in SGs, contain a signal peptide, a repetitive sequence, and a high proportion of serine residues (21% and 29%), and thus appear to belong to the sericin family of *G. mellonella*.

Overall, we identified the silk genes in *E. kuehniella* and *G. mellonella* based on proteomic analysis of cocoon silk and by homology searches in the transcriptomes and genomes of both species. We discovered a region containing clusters of sericin genes and identified blocks of synteny between both genomes (colocalized gene clusters shared between genomes). The resulting microsynteny map allowed identification of

duplication events in the sericin family. Finally, we present the complete primary structures of nine *fibH* genes and proteins from both families of the suborder Pyraloidea and discuss their specific and conserved features.

Data availability statement

The trancriptome assembly was deposited in the Dryad repository (https://doi.org/10.5281/zenodo.7273794). The experimental data that support the findings of this study is available within this article or its Supplementary Material. List of silk gene candidates their GenBank accession codes are listed in Tables 1, 2. The sequences of FibH from 9 pyralid species are available as Supplementary Data (Supplementary Figure S2).

Author contributions

BC-hW isolated RNAs, performed the computer analyses, and wrote most of the manuscript. IS performed the electron microscopy and histology imaging. BK performed the northern blotting analysis. MH prepared cDNA libraries and provided the sequencing data. MaZ performed phylogeny analysis. HM performed qPCR and adjusted figures. AZ performer silk solubility and hydroscopicity tests. MiZ supervised the entire project.

Funding

This research was supported by European Community's Program Interreg Bayern-Tschechische Republik Ziel ETZ 2021–2022 no. 331. This publication is supported by the project "BIOCEV – Biotechnology and Biomedicine Centre of the Academy of Sciences and Charles University" (CZ.1.05/1.1.00/02.0109), from the European Regional Development Fund. We also acknowledge the core facility Laboratory of Electron Microscopy, Biology Centre CAS supported by the MEYS CR (LM2018129 Czech-BioImaging) and ERDF (No. CZ.02.1.01/0.0/0.0/16_013/0001775). Computational resources were supplied by "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) funded by the Ministry of Education, Youth and Sports of the Czech Republic, and by the ELIXIR-CZ project (LM2018131), part of the international ELIXIR infrastructure.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1023381/full#supplementary-material

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46 (W1), W537–W544. doi:10.1093/nar/gky379

Blaxter, M. L., and Project, D. T. L. (2022). Sequence locally, think globally: The Darwin tree of life project. *Proc. Natl. Acad. Sci. U. S. A.* 119 (4), e2115642118. doi:10.1073/pnas.2115642118

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 13 (9), 2513–2526. doi:10.1074/mcp.M113.031591

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 10 (4), 1794–1805. doi:10.1021/pr101065j

Craig, C. L. (1997). Evolution of arthropod silks. *Annu. Rev. Entomol.* 42, 231–267. doi:10.1146/annurev.ento.42.1.231

Davey, P. A., Power, A. M., Santos, R., Bertemes, P., Ladurner, P., Palmowski, P., et al. (2021). Omics-based molecular analyses of adhesion by aquatic invertebrates. *Biol. Rev. Camb. Philos. Soc.* 96 (3), 1051–1075. doi:10.1111/brv.12691

Deny, M. W. (1980). Silks--their properties and functions. *Symp. Soc. Exp. Biol.* 34, 247–272.

Dong, Z. M., Guo, K. Y., Zhang, X. L., Zhang, T., Zhang, Y., Ma, S. Y., et al. (2019). Identification of *Bombyx mori* sericin 4 protein as a new biological adhesive. *Int. J. Biol. Macromol.* 132, 1121–1130. doi:10.1016/j.ijbiomac.2019.03.166

Ellis, A. M., and Hayes, G. W. (2009). Assessing the efficacy of a product containing Bacillus thuringiensis applied to honey bee (hymenoptera: Apidae) foundation as a control for Galleria mellonella (Lepidoptera: Pyralidae). *J. Entomol. Sci.* 44 (2), 158–163. doi:10.18474/0749-8004-44.2.158

Erban, T., Klimov, P., Talacko, P., Harant, K., and Hubert, J. (2020). Proteogenomics of the house dust mite, Dermatophagoides farinae: Allergen repertoire, accurate allergen identification, isoforms, and sex-biased proteome differences. *J. Proteomics* 210, 103535. doi:10.1016/j.jprot.2019.103535

Fedic, R., Zurovec, M., and Sehnal, F. (2003). Correlation between fibroin amino acid sequence and physical silk properties. *J. Biol. Chem.* 278 (37), 35255–35264. doi:10.1074/jbc.M305304200

Gamo, T. (1982). Genetic variants of the *Bombyx mori* silkworn encoding sericin proteins of different lengths. *Biochem. Genet.* 20 (1-2), 165–177. doi:10.1007/BF00484944

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41, 95–98. doi:10.14601/Phytopathol_Mediterr-14998u1.29

Hasan, K., Kayumov, J., Zhu, G., Khatun, M., Nur, A., and Dings, X. (2019). An experimental investigation to examine the wicking properties of silk fabrics. *J. Text. Sci. Technol.* 5 (4), 108–124. doi:10.4236/jtst.2019.54010

Hughes, C. S., Foehr, S., Garfield, D. A., Furlong, E. E., Steinmetz, L. M., and Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* 10 (10), 757. doi:10.15252/msb.20145625

Jacob, T. A., and Cox, P. D. (1977). The influence of temperature and humidity on the life-cycle of Ephestia kuehniella zeller (Lepidoptera: Pyralidae). *J. Stored Prod. Res.* 13 (3), 107–118. doi:10.1016/0022-474x(77)90009-1

Jin, T. Q., Ito, Y., Luan, X. H., Dangaria, S., Walker, C., Allen, M., et al. (2009). Elongated polyproline motifs facilitate enamel evolution through matrix subunit compaction. *PLoS Biol.* 7 (12), e1000262. doi:10.1371/journal.pbio.1000262

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14 (6), 587–589. doi:10.1038/Nmeth.4285

Kludkiewicz, B., Kucerova, L., Konikova, T., Strnad, H., Hradilova, M., Zaloudikova, A., et al. (2019). The expansion of genes encoding soluble silk components in the greater wax moth, Galleria mellonella. *Insect biochem. Mol. Biol.* 106, 28–38. doi:10.1016/j.ibmb.2018.11.003

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157 (1), 105–132. doi:10.1016/0022-2836(82)90515-0

Levine, J. D., Sauman, I., Imbalzano, M., Reppert, S. M., and Jackson, F. R. (1995). Period protein from the giant silkmoth *Antheraea pernyi* functions as a circadian clock element in *Drosophila melanogaster*. *Neuron* 15 (1), 147–157. doi:10.1016/0896-6273(95)90072-1

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18), 3094–3100. doi:10.1093/bioinformatics/bty191

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25 (4), 402–408. doi:10.1006/meth.2001.1262

Lucas, F., and Rudall, K. M. (1968). "Extracellular fibrous proteins: The silks," in *Comprehensive biochemistry*. Editors M. Florkin and E. H. Stotz (Amsterdam: Elsevier), 475–558.

Marec, F., and Traut, W. (1994). Sex-chromosome pairing and sex-chromatin bodies in W-Z translocation strains of ephestia-kuehniella (Lepidoptera). *Genome* 37 (3), 426–435. doi:10.1139/g94-060

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32 (1), 268–274. doi:10.1093/molbev/msu300

Opell, B. D., and Stellwagen, S. D. (2019). Properties of orb weaving spider glycoprotein glue change during Argiope trifasciata web construction. *Sci. Rep.* 9 (1), 20279. doi:10.1038/s41598-019-56707-1

Perez-Vilar, J., and Hill, R. L. (1999). The structure and assembly of secreted mucins. *J. Biol. Chem.* 274 (45), 31751–31754. doi:10.1074/jbc.274.45.31751

Prudhomme, J., C., Couble, P., Garel, J. P., and Daillie, J. (1985). "Silk synthesis," in *Comprehensive insect physiology, biochemistry and pharmacology*. Editors G. A. Kerkut and L. I. Gilbert (New York: Pergamon), 571–594.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* 2 (8), 1896–1906. doi:10.1038/nprot.2007.261

Regier, J. C., Mitter, C., Solis, M. A., Hayden, J. E., Landry, B., Nuss, M., et al. (2012). A molecular phylogeny for the pyraloid moths (*Lepidoptera*: *Pyraloidea*) and its implications for higher-level classification. *Syst. Entomol.* 37 (4), 635–656. doi:10.1111/j.1365-3113.2012.00641.x

Rouhova, L., Kludkiewicz, B., Sehadova, H., Sery, M., Kucerova, L., Konik, P., et al. (2021). Silk of the common clothes moth, Tineola bisselliella, a cosmopolitan pest belonging to the basal ditrysian moth line. *Insect biochem. Mol. Biol.* 130, 103527. doi:10.1016/j.ibmb.2021.103527

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

Sehnal, F. (1966). Kritisches studium der bionomie und biometrik der in verschiedenen lebensbedingungen gezuchteten wachsmotte Galleria mellonella L (lepidopera). *Z. Fur Wiss. Zool.* 174 (1-2), 53.

Shimura, K., Kikuchi, A., Katagata, Y., and Ohotomo, K. (1982). The occurrence of small component proteins in the cocoon fibroin of *Bombyx mori*. *J. Seric. Sci. Jap.* 51 (1), 20–26. doi:10.11416/kontyushigen1930.51.20

Simao, 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. *Bioinformatics* 31 (19), 3210–3212. doi:10.1093/bioinformatics/btv351

Sonwalkar, T. N. (1993). *Hand book of silk technology*. New Delhi: Wiley Eastern.

Stanke, M., and Morgenstern, B. (2005). Augustus: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi:10.1093/nar/gki458

Syed, Z. A., Hard, T., Uv, A., and van Dijk-Hard, I. F. (2008). A potential role for Drosophila mucins in development and physiology. *Plos One* 3 (8), e3041. doi:10.1371/journal.pone.0003041

Takasu, Y., Iizuka, T., Zhang, Q., and Sezutsu, H. (2017). Modified cocoon sericin proteins produced by truncated Bombyx Ser1 gene. *J. Silk Sci. Technol. Jpn.* 25, 35–47. doi:10.11417/silk.25.35

Takasu, Y., Yamada, H., Tamura, T., Sezutsu, H., Mita, K., and Tsubouchi, K. (2007). Identification and characterization of a novel sericin gene expressed in the anterior middle silk gland of the silkworm *Bombyx mori*. *Insect biochem. Mol. Biol.* 37 (11), 1234–1240. doi:10.1016/j.ibmb.2007.07.009

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., et al. (2016). The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nat. Methods* 13 (9), 731–740. doi:10.1038/nmeth.3901

Visser, S., Volenikova, A., Nguyen, P., Verhulst, E. C., and Marec, F. (2021). A conserved role of the duplicated Masculinizer gene in sex determination of the Mediterranean flour moth, Ephestia kuehniella. *PLoS Genet.* 17 (8), e1009420. doi:10.1371/journal.pgen.1009420

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Berlin, Germany: Springer Science & Business Media, 1–212. doi:10.1007/978-0-387-98141-3

Zurovec, M., Kodrik, D., Yang, C., Sehnal, F., and Scheller, K. (1998). The P25 component of Galleria silk. *Mol. Gen. Genet.* 257 (3), 264–270. doi:10.1007/s004380050647

Zurovec, M., Sehnal, F., Scheller, K., and Kumaran, A. K. (1992). Silk gland specific cdnas from galleria-mellonella L. *Insect Biochem. Mol. Biol.* 22 (1), 55–67. doi:10.1016/0965-1748(92)90100-S

Zurovec, M., Vaskova, M., Kodrik, D., Sehnal, F., and Kumaran, A. K. (1995). Light-chain fibroin of Galleria mellonella L. *Mol. Gen. Genet.* 247 (1), 1–6. doi:10.1007/BF00425815

Zurovec, M., Yonemura, N., Kludkiewicz, B., Sehnal, F., Kodrik, D., Vieira, L. C., et al. (2016). Sericin composition in the silk of Antheraea yamamai. *Biomacromolecules* 17 (5), 1776–1787. doi:10.1021/acs.biomac.6b00189

Supplementary Material

Supplementary Figure 1. Phylogenetic analysis of selected silk proteins in Pyraloidea species and evolutionary relationships among Pyraloidea.

Supplementary Figure 2. Supplementary Figure 2. FibH protein sequences of nine pyraloid moths predicted from the genomic regions using online Augustus software (Stanke and Morgenstern, 2005).

Supplementary Table 1. Summary of primer sequences used in *E. kuehniella* for (A) northern blot probes and (B) real-time qPCR.

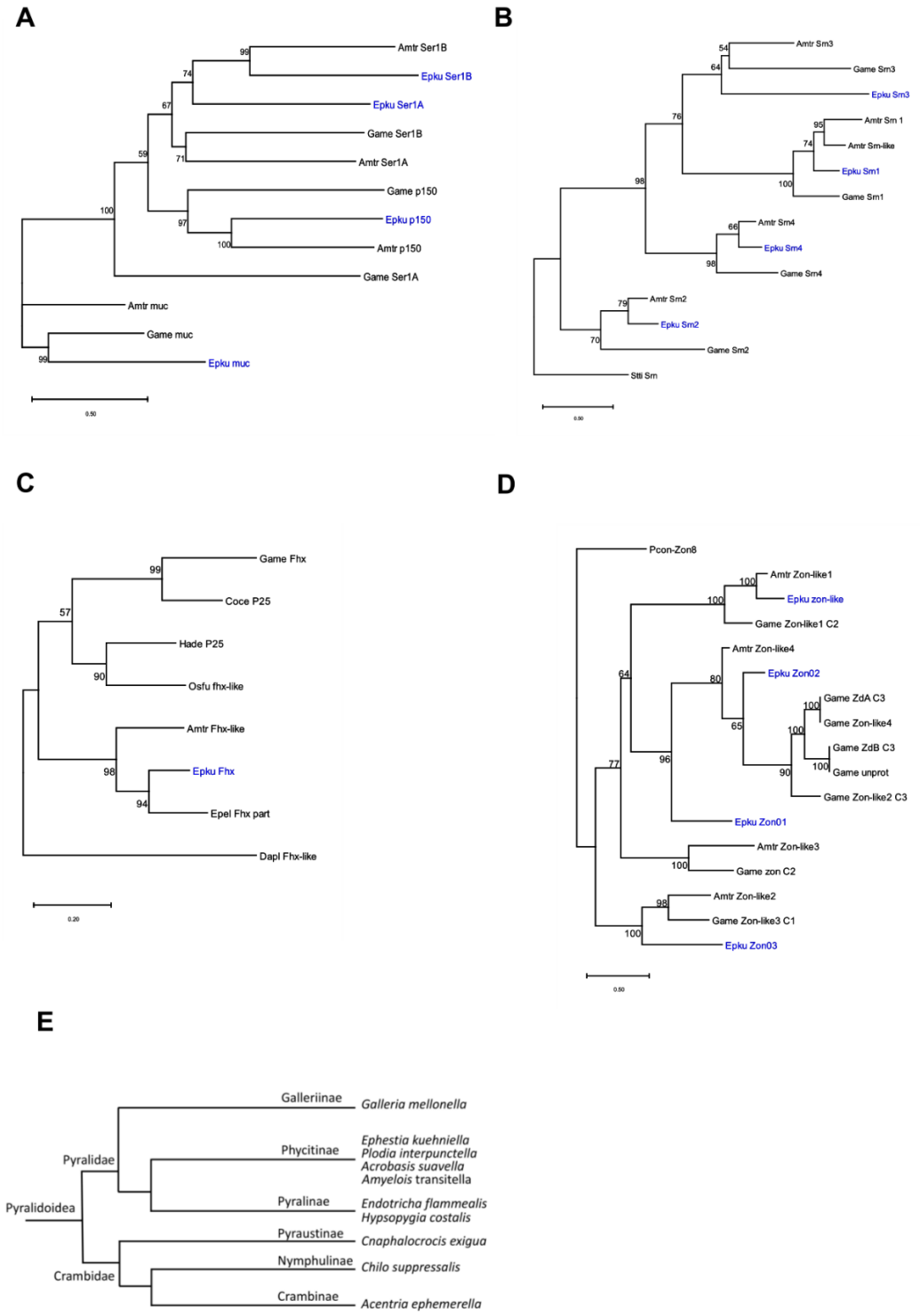
Supplementary Table 2. Statistics of expression levels of selected silk genes detected by qPCR in *E. kuehniella*. AMS, anterior-middle SG; MMS, middle-middle SG; RMS, rear-middle SG; PS, posterior SG; PS, anterior-middle SG; WLWS, wandering larva without SG. Statistical differences (T-test; $P < 0.05$) are indicated by asterisk (*).

Supplementary Table 3. (A) BUSCO assessment of initial and improved transcriptome of *E. kuehniella.* (B) Genome assembly statistics for *E. kuehniella*.

Supplementary Table 4. Summary of domains identified in zonadhesin protein sequences of *E. kuehniella*, *G. mellonella*, *A. transitella*, and *Danaus plexippus plexippus*. The search was performed using the web tool MOTIF Search (https://www.genome.jp/tools/motif/) against motif library PROSITE Pattern, PROSITE Profile and Pfam.

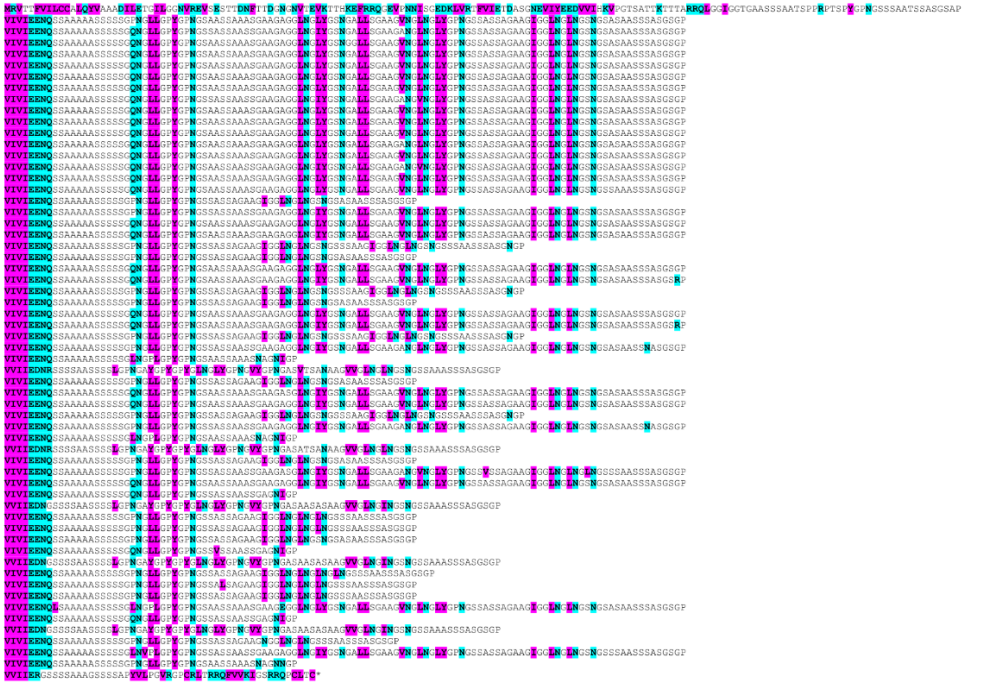
Supplementary Table 5. Summary of proposed landmark genes linked with silk genes in *E. kuehniella*, *G. mellonella* and *A. transitella*. These genes are evolutionarily conserved and their positions in the genomes of the three species are shown in Figure 5. The corresponding genes in Figure 5 are connected by coloured lines (blue, green and magenta).

Supplementary Figure 1. Phylogenetic analysis of selected silk proteins in Pyraloidea species and evolutionary relationships among Pyraloidea. (**A-D**) Phylograms of selected silk proteins. The evolutionary history was inferred by the Maximum Likelihood method with 1000× bootstrap. The percentage of trees in which the associated taxa clustered together is shown below the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Abbreviations of species are as follows: Amtr, *A. transitella*; Epku, *E. kuehniella*; Game, *G. mellonella*; Stti, *Stenopsyche tienmushanensis*; Epel, *Ephestia elutella*; Coce, *Corcyra cephalonica*; Hade, *Haritalodes derogata*; Osfu, *Ostrinia furnicalis*; Dapl, *Danaus plexippus*; Pcon, *Plectrocnemia conspersa*. (A) Sericins / mucins / P150. Based on TIM+F+I transition model. Amtr – Ser1A (LOC106134399), Ser1B (LOC106134400 transcript X1), Muc (LOC106133704), P150 (LOC106132366); Epku – Ser1A (ON604817), Ser1B (ON604818), Muc (ON604821), P150 (ON604820); Game – Ser1A (LOC113519119, transcript X1), Ser1B (LOC113512273), Muc (MG770312), P150 (LOC113522468). (B) Seroins. Based on HKY+F+G4 model. Amtr – Sro1 (LOC106133523), Sro2 (LOC106133521), Sro3 (LOC106133479), Sro4 (LOC106133522), Sro-like (LOC106133524); Epku – Sro1 (ON604827), Sro2 (OP185488), Sro3 (OP185489), Sro4 (ON604828); Game – Sro1 (LOC113518338), Sro2 (LOC113518101), Sro3 (LOC113518258), Sro4 (LOC113518326); Stti – Sro (Steno.02678-RA). (C) P25. Based on TN+F+I+G4 model. Epku – Fhx (ON604823); Epel – Fhx part (00014683-RA); Amtr – Fhx-like (LOC106131755); Game – Fhx (LOC113510933); Coce – P25 (GQ901976); Hade – P25 (KY792994); Osfu – Fhx-like (LOC114362712); Dapl – Fhx-like (LOC116776386). (D) Zonadhesins. Based on GTR+F+I+G4 model. Amtr – zon-like1 (LOC106138817), zon-like2 (LOC106136372), zon-like3 (LOC106130156), zon-like4 (LOC106131409); Epku – Zon01 (ON604824), Zon02 (ON604825), Zon03 (ON604826), zon-like (OP185494); Game – zon-like1 C2 (LOC113516017), zon-like2 C3 (LOC113511957), zon-like3 C1 (LOC113519003), zon-like4 (LOC113511955), ZdB C3 (MG770321), ZdA C3 (MG770320), uncharacterized protein (LOC113509084), zon_C2 (LOC113511802); Dapl – zon-like C3 (LOC116765598), zon C2 (LOC116771299); Pcon – Zon8 (OL405648). (**E**) The evolutionary relationships of the Pyraloidea (adapted from Regier et al., 2012, DOI: 10.1111/j.1365-3113.2012.00641.x.).

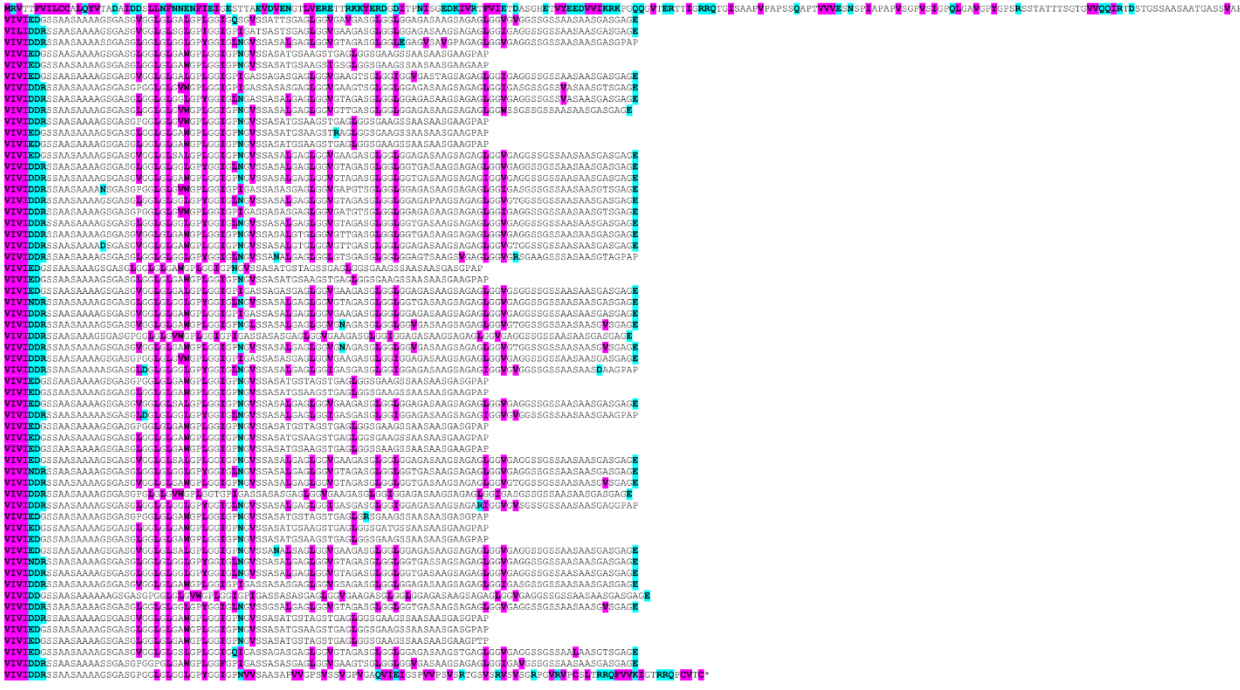

A
Amtr Ser1B
Epku Ser1B
Epku Ser1A
Game Ser1B
Amtr Ser1A
Game p150
Epku p150
Amtr p150
Game Ser1A
Amtr muc
Game muc
Epku muc
B
Amtr Sm3
Game Sm3
Epku Sm3
Amtr Sm 1
Amtr Sm-like
Epku Sm1
Game Sm1
Amtr Sm4
Epku Sm4
Game Sm4
Amtr Sm2
Epku Sm2
Game Sm2
Stli Sm
C
Game Fhx
Coce P25
Hade P25
Osfu fhx-like
Amtr Fhx-like
Epku Fhx
Epel Fhx part
Dapl Fhx-like
D
Pcon-Zon8
Amtr Zon-like1
Epku zon-like
Game Zon-like1 C2
Amtr Zon-like4
Epku Zon02
Game ZdA C3
Game Zon-like4
Game ZdB C3
Game unprot
Game Zon-like2 C3
Epku Zon01
Amtr Zon-like3
Game zon C2
Amtr Zon-like2
Game Zon-like3 C1
Epku Zon03
E
Galleriinae
Galleria mellonella
Phycitinae
Ephestia kuehniella
Plodia interpunctella
Acrobasis suavella
Amyelois transitella
Pyralidae
Pyralinae
Endotricha flammealis
Hypsopygia costalis
Pyralidoidea
Pyraustinae
Cnaphalocrocis exigua
Nymphulinae
Chilo suppressalis
Crambidae
Crambinae
Acentria ephemerella

Supplementary Figure 2. FibH protein sequences of nine pyraloid moths predicted from the genomic regions using online Augustus software (Stanke and Morgenstern, 2005). Amino acids are highlighted with cyan (hydrophilic: RKDENQ), magenta (hydrophobic: YVMCLFIW), or no shadings (netural: SGHTAP). Genbank accession codes: *E. kuehniella* (ON604816), *G. mellonella* (NHTH01000008.1/NHTH01000133.1), *E. flammealis* (LR990872.1), *H. costalis* (OW443343.2), *P. interpunctella* (JAJAFS010000023.1), *A. suavella* (OW971947.1), *A. ephemerella* (OW971889.1), *C. exigua*, (CM032477.1), *Ch. suppressalis* (OU963910.1).

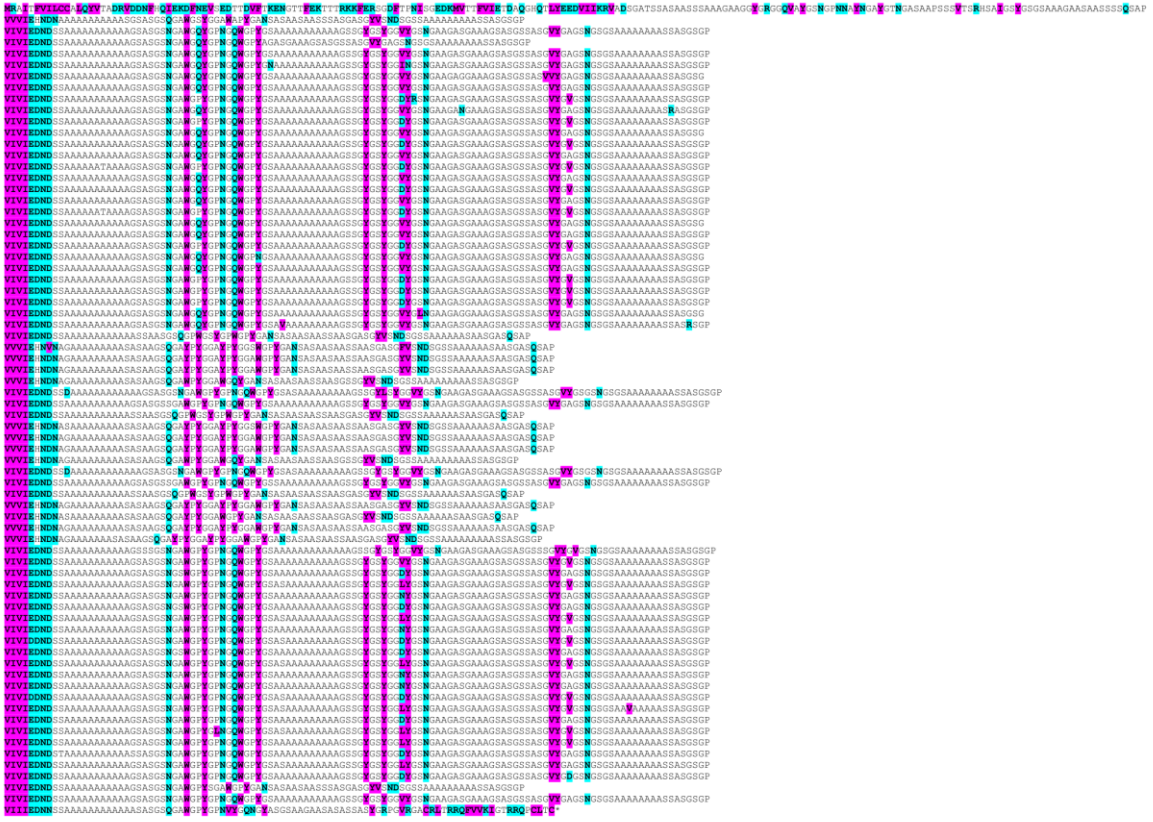
(A) ***Ephestia kuehniella* FibH**



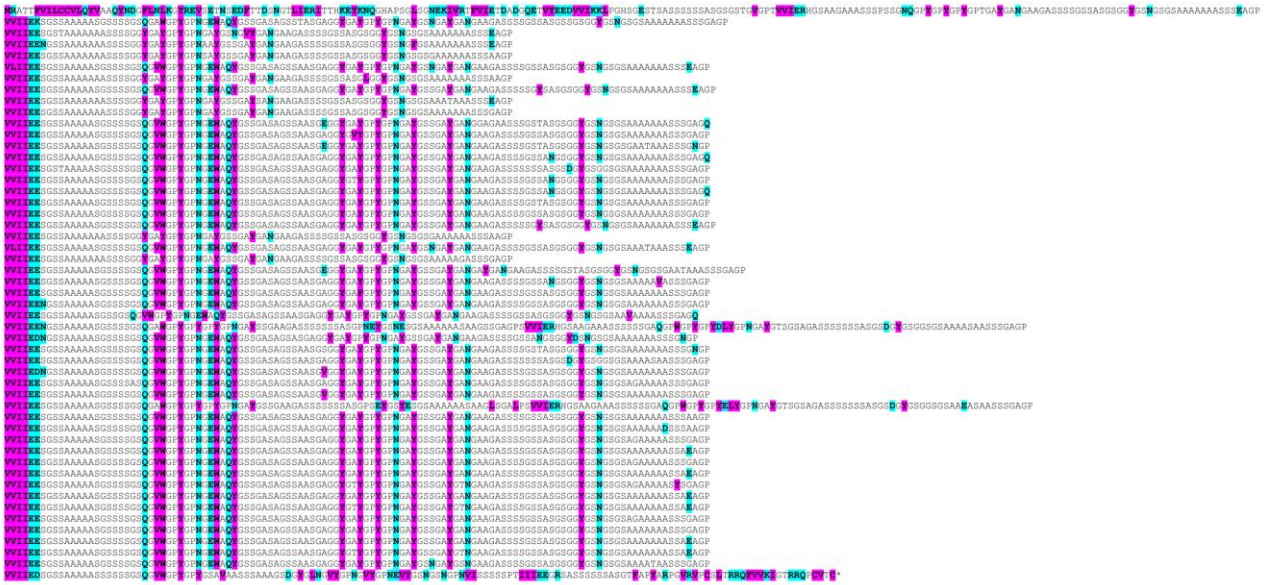
(B) ***Galleria mellonella*** **FibH**



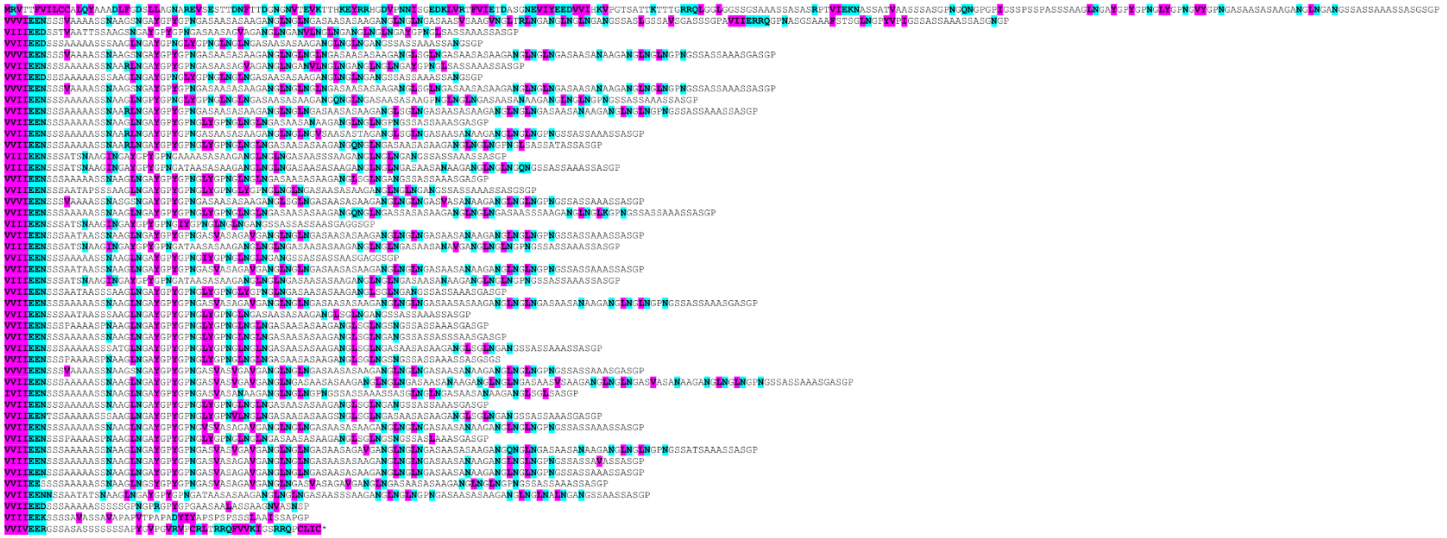
(C) *Endotricha flammealis* FibH



(D) ***Hypsopygia costalis*** **FibH**

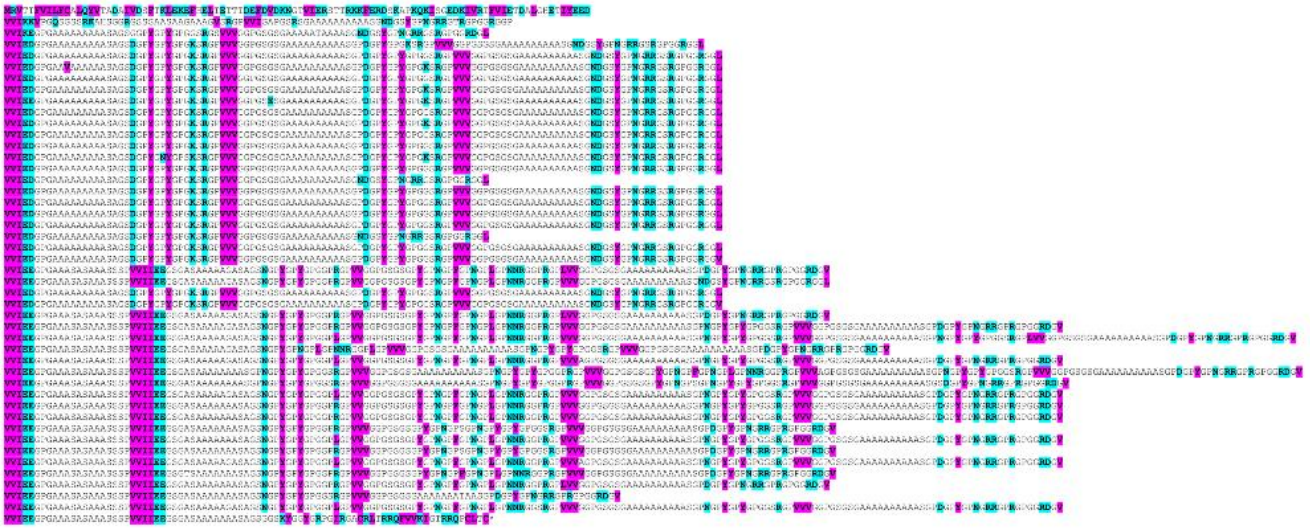


(E) ***Plodia interpunctella*** **FibH**

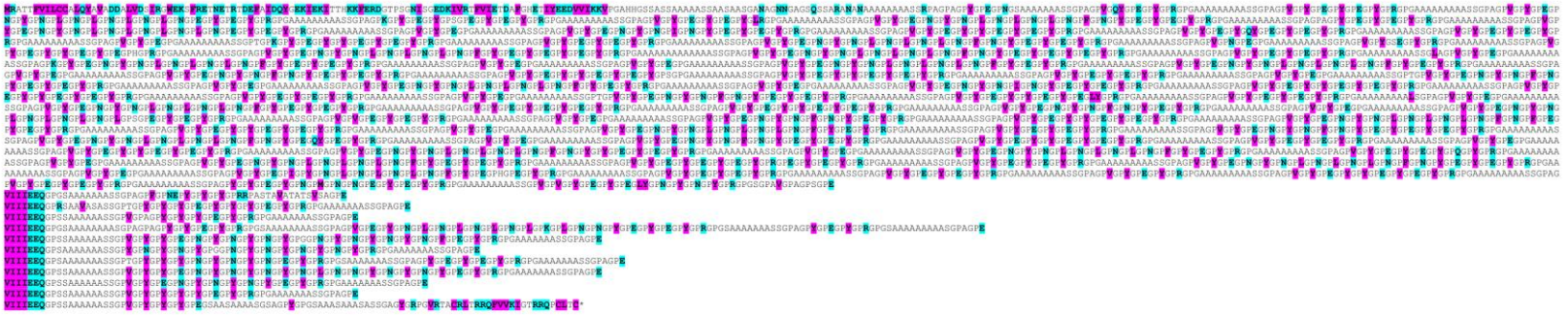


(F) ***Acrobasis suavella*** **FibH**

(G) *Acentria ephemerella* FibH



(H) ***Cnaphalocrocis (Marasmia) exigua*** **FibH**



(I) ***Chilo suppressalis*** **FibH**

Supplementary Table 1. Summary of primer sequences used in *E. kuehniella* for (A) northern blot probes and (B) real-time qPCR.

(A) Northerns

Gene	Name	Forward	Reverse
FibL	Fibroin light chain	TGCTGCCTATCGTTTTGGTATTAT	CGGGCAGCGTTGGGGTCGTT
P25	P25	TGCTACGTTTCTGAGTTCTTCTTC	TGGGACCATAATCTTCACAAT
Ser1A	Sericin 1A	TCCCAGGATGCGAAAATAAATACT	GGCCGCCTGGTTGAATGAT
Ser1B	Sericin 1B	CGACGCCAACGGAAACAC	GCTCTGCGCATTACTTACG
Dana-1	Mucin 1	AGCGCTCAAGCATCTCAATC	GCGCTGCAATGCCAACTTCA
Dana-2	Mucin 1	ATCCGAGGAATCTACGACACTT	
Ser4-1	Ser4	CCCTTGCTTTGGTACTCCTCACG	CCACTGCTACTGCCGCTACTTTGT
Ser4-2	Ser4	CGGCTCCGGCTCAGAAACA	GAGGAGCCATCAGCCGAGACT

(B) qPCR

Gene	Name	Forward	Reverse
EF-1α	Elong. factor 1-alpha	TCAAACGGTTACACGCCTGT	GGACTTGGGGTTGTCCTCAG
Fbn	Fibrillin	CACGCTCGTCTCCAACATCT	CACTTCCACCATCCGCATCT
FibH	Fibroin heavy chain	CAACGGACCAGTTGTCATCA	TGTCTACGGGTAAGACGGCA
FibL	Fibroin light chain	AGCCTTGAACAACCGTAGCA	GCGACTGCTCTCAGGAAGTT
Lprcp66-1	rigid cut. protein 66-like 1	CGTTTCAGTCCACCCTGTCA	GCGACAGTAGTAGCGTACCC
Lprcp66-2	rigid cut. protein 66-like 2	AAGTCAAGTCGTCGGCCAAT	CGCCATCCTTCCTCACAACT
Muc1	Mucin 1	ATCATGCACGCCTTGGAAAC	CATTGGCACGAGTTTAGGCG
Muc2	Mucin 2	ACCGAACCATCTACCTCCGA	TTCGGTGGTAGTTTCGGTGG
P25	P25	ACCTTCGCTGACATCACCAC	GTGGAAGTTCTCCCGTTCGT
P47	P47	ATGATGCTGCAGTGGGTTCA	TCTTCACCCACGCCTGTTAG
Pebp	PEBP	CCCGGGAAACTCAACTTCGA	CTGCGAACTTTGCGATGGAG
Pssp1	salivary secr. peptide 1	AATGCCCTGGTTCTGGGTAC	TGTCCTCCCAAAGCCATGAC
Pssp2	Salivary secr. peptide 2	CTTCTCTGACCCTGGACAGC	CGTGATGATGGCTGTGGAGT
Pssp3	Salivary secr. peptide 3	ACCAAGACTTCGTGTACCGC	TAGCGAGGGTACAGCGTACT
Pcp	Pupal cut. protein-like	CTCATGGCAAGGCAATGCTG	CTCGGGGGTTTCCAGGATG
Ser1A	Sericin 1A	AAATCTTGGTCCAGGCACCG	ACTAACGGAAACAGCAATGCT
Ser1B	Sericin 1B	CTCAGTTGCCCCAGGATCTG	TGAAGAAGCAGAGCTGGTGG
Ser3	Sericin 3	ACCACCAAACACAGCCATGA	CTTGATCACCTCTCCGGTCG
SerP150	Sericin P150	CCCAAAGCAGCACCAGTCTT	ATCGAGCTGGACTGTGCATT
Sn1	Seroin 1	GTCGTCAAACGTGAACGGTG	TCCTTGGCATTCTTCGGGTC
Sn2	Seroin 2	TAGTCCTCCCAATCCGAGCA	AGGAGGAGGAGTAGGCTGTG
Sn3	Seroin 3	CCGCTTTCGTTTGTGCTTCA	TTCTCAGTTGATGGACCGGT
Sn4	Seroin 4	TGTGAAGAGCCTCAAGCCAG	CCATGGTCTTCCCGTTCACA
Scp24	Sialomucin core prot. 24	CGCTAAGACACATAGAGACAGACA	TGCGTGTCGTTTCATTGAGG
Spi	Silk proteinase inh.	GGCCGCCATTGTTTCATCAA	TTCCAGAAGCATATCCCGGC
Vsp	Venom serine protease	TACACGCCTGGGAAAGACAC	GTCCATGCCGCAGTAGTCTC
Zon1	Zonadhesin 1	AACGGCTGCGACTGTATTGA	TCCGTTAACGCACGTCGAAT
Zon2	Zonadhesin 2	TACGAAACCAAGACCCTGCC	GTTGACTTCGCAGGTTGGTG
Zon3	Zonadhesin 3	TCGGCAAATTCTTCAAATTCTTCCA	TCCAACAGTCACATTATTGCTGA

Supplementary Table 2. Statistics of expression levels of selected silk genes detected by qPCR in *E. kuehniella*. AMS, anterior-middle SG; MMS, middle-middle SG; RMS, rear-middle SG; PS, posterior SG; PS, anterior-middle SG; WLWS, wandering larva without SG. Statistical differences (T-test; $P < 0.05$) are indicated by asterisk (*).

		Fold-change					*P*-value			
Gene	**Name**	**AMS**	**MMS**	**PS**	**RMS**	**WLWS**	**AMS**	**MMS**	**RMS**	**PS**
Fbn	Fibrillin	2,203	0,903	0,768	1,501	1,001	*0.001	0,701	0,155	0,069
FibH	Fibroin heavy chain	2,87	16,685	119,12	17,668	1,023	0,63	*0.011	*0.000	*0.000
FibL	Fibroin light chain	1,125	3,546	29,657	5,481	1,003	0,968	*0.041	*0.004	*0.000
Lprcp66-1	Larval / pupal rigid cuticle protein 66-like 1	2,043	1,611	1,322	1,024	1,132	0,056	0,093	0,826	0,931
Lprcp66-2	Larval / pupal rigid cuticle protein 66-like 2	2,771	1,78	1,038	2,709	1,14	*0.041	*0.048	*0.018	0,242
Muc1	Mucin 1	11,514	31,387	8,811	55,267	1,342	*0.031	*0.010	*0.013	*0.028
Muc2	Mucin 2	0,429	0,513	0,3	0,571	1,036	0,079	0,166	0,25	*0.026
P25	P25	9,418	35,525	14,334	67,263	1,031	*0.001	*0.000	*0.001	*0.003
P47	P47	1,63	2,24	2	4,615	1,009	*0.016	0,168	*0.006	*0.010
Pebp	Phosphatidylethanolamine-bin. Prot	0,499	0,78	1,086	0,678	1,077	0,263	0,586	0,39	0,697
Pssp1	Probable salivary secreted peptide 1	3,879	1,587	1,914	1,823	1,006	*0.000	0,345	0,159	*0.005
Pssp2	Probable salivary secreted peptide 2	3,432	2,106	1,777	6,804	1,054	*0.019	0,082	0,09	0,132
Pssp3	Probable salivary secreted peptide 3	0,114	0,102	0,076	0,079	1,016	*0.003	*0.002	*0.000	*0.001
Pcp	Pupal cuticle protein-like	1,684	4,278	0,573	1,686	1,019	0,568	*0.017	0,241	0,353
Ser1A	Sericin 1A	0,569	2,064	28,324	26,681	1,077	0,229	0,059	*0.007	0,058
Ser1B	Sericin 1B	2,764	7,875	27,865	36,649	1,129	0,1	0,094	*0.008	*0.011
Ser3	Sericin 3	59,288	110,316	12,281	127,197	1,309	*0.025	*0.013	*0.023	0,061
SerP150	Sericin P150	676,983	611,45	39,29	448,051	1,247	*0.005	*0.004	*0.005	*0.012
Sn1	Seroin 1	15,837	14,222	5,167	13,37	1,015	*0.000	*0.001	*0.000	*0.001
Sn2	Seroin 2	2,09	1,746	0,526	1,473	1,009	*0.006	0,051	*0.022	*0.036
Sn3	Seroin 3	4,021	2,294	1,334	1,848	1,038	*0.004	0,09	*0.020	0,341
Sn4	Seroin 4	1,678	1,506	1,083	1,949	1,024	*0.033	0,07	*0.011	0,349
Scp24	Sialomucin core protein 24	7,77	3,925	1,911	5,235	1,231	*0.019	*0.013	*0.005	0,102
Spi	Silk proteinase inhibitor	37,426	83,184	63,73	127,507	1,002	*0.000	*0.000	*0.001	*0.002
Vsp	Venom serine protease-like	4,874	4,942	0,434	1,692	1,01	*0.002	*0.000	0,191	0,469
Zon1	Zonadhesin 1	58,363	85,918	4,94	68,127	1,006	*0.000	*0.000	*0.000	*0.006
Zon2	Zonadhesin 2	16,218	19,306	2,811	18,863	1,173	*0.015	*0.008	*0.008	0,154
Zon3	Zonadhesin 3	106,037	142,901	96,997	205,674	1,028	*0.000	*0.000	*0.000	*0.001

Supplementary Table 3. (A) BUSCO assessment of initial and improved transcriptome of *E. kuehniella.* (B) Genome assembly statistics for *E. kuehniella*.

(A) *E. kuehniella* transcriptome	**Initial**	**Improved**
Element	Number (%)	Number (%)
Total BUSCO groups searched	1367 (100)	1367 (100)
Complete BUSCOs	1115 (81.6)	1348 (98.6)
Complete and single-copy BUSCOs	771 (56.4)	1341 (98.1)
Complete and duplcated BUSCOs	344 (25.2)	7 (0.5)
Fragmented BUSCOs	110 (8.0)	5 (0.4)
Missing BUSCOs	142 (10.4)	14 (1.0)

(B) *E. kuehniella* genome assembly	**Statistics**
Assembly size (Mb)	351,8
Number of contigs	165
Largest contig (Mb)	15,1
GC content (%)	36,1
N50 contig length (Mb)	8,3
Number of protein coding genes	13382
Mean gene length (bp)	7207,8
Repetitive elements (%)	37,8

Supplementary Table 4. Summary of domains identified in zonadhesin protein sequences of *E. kuehniella*, *G. mellonella* and *A. transitella* from family Pyralidae, and *Danaus plexippus* from family Nymphalidae. The search was performed using the web tool MOTIF Search (https://www.genome.jp/tools/motif/) against motif library PROSITE Pattern, PROSITE Profile and Pfam.

Lepidopteran species	GenBank	GenPept	Name	Prosite Pattern	Prosite Profile
Danaus plexippus	XM_032655128.1	XP_032511019	zonadhesin-like C3	EGF_2 x4; SERPIN	CYS_RICH; PROKAR_LIPOPROT.
	XM_032663119.1	XP_032519010	zonadhesin C2	ZINC_FINGER_C2H2_1	
Amyelois transitella	XM_013340097.1	XP_013195551	zonadhesin-like1	EGF_2; ZINC_FINGER_C2H2_1	CYS_RICH
	XM_013336902.1	XP_013192356	zonadhesin-like2	EGF_2 x6; ASX_HYDROXYL x2	CYS_RICH x3; SER_RICH
	XM_013328929.1	XP_013184383	zonadhesin-like3	EGF_2 x3	ANTISTASIN; CYS_RICH x2
	XM_013330525.1	XP_013185979	zonadhesin-like4	EGF_2 x25; serpin	CYS_RICH x2
Galleria mellonella	XM_026900349.2	XP_026756150	zonadhesin-like1 C2	EGF_2 x2; ZINC_FINGER_C2H2_1	CYS_RICH
	XM_031913577.1	XP_031769437	zonadhesin-like2 C3	EGF_2 x6	CYS_RICH x2; SER_RICH
	XM_031914733.1	XP_031770593	zonadhesin-like3 C1	EGF_2 x3	CYS_RICH
	XM_026895687.2	XP_026751488	zonadhesin-like4	EGF_2 x6; SERPIN	CYS_RICH; SER_RICH
	MG770321.1	AXY94923	zonadhesin-like B (ZdB) C3	EGF_2 x5; SERPIN	CYS_RICH
	MG770320.1	AXY94922	zonadhesin-like A (ZdA) C3	EGF_2 x6; SERPIN	CYS_RICH x2; SER_RICH
	NW 022276972.1	XP_026748171.2	uncharacterized protein	EGF_2 x5	
	XM_031913565	XP_031769425	zonadhesin_C2	EGF_2 x6	
Ephestia kuehniella	ON604824		zonadhesin 01	EGF_2 x13	CYS_RICH; SER_RICH
	ON604825		zonadhesin 02	EGF_2 x7; SERPIN	CYS_RICH; SER_RICH
	ON604826		zonadhesin 03	EGF_2 x3	SER_RICH
	OP185494		zonadhesin-like	EGF_2	CYS_RICH

Supplementary Table 5. Summary of proposed landmark genes adjacent to silk genes in *E. kuehniella*, *G. mellonella* and *A. transitella*. These genes are evolutionarily conserved and their positions in the genomes of the three species are shown in Figure 5. The corresponding genes shown in Figure 5 are connected by colored lines (blue, green and magenta).

Color code	Species	Contig / Scaffold	Gene	Name
Blue	*Ephestia kuehniella*	contig_172	OP185490	P47 (silk gene)
	Galleria mellonella	NW_022271951.1	LOC113518611 | XP_026759371.2	P47 (silk gene)
	Amyelois transitella	NW_013535448.1	LOC106134372 | XP_013189852.1	P47 (silk gene)
Green	*Ephestia kuehniella*	contig_172	EKM1R5v1_00011257	Similar to protein-lysine N-methyltransferase mettl10
		contig_172	EKM1R5v1_00011258	Similar to histidine triad nucleotide-binding protein 3-like
		contig_172	EKM1R5v1_00011259	Similar to motile sperm domain-containing protein 2-like
	Galleria mellonella	NW_022271951.1	LOC113513742 | XP_031767745.1	EEF1A lysine methyltransferase 2
		NW_022271951.1	LOC113513812 | XP_026753595.1	histidine triad nucleotide-binding protein 3-like
		NW_022271951.1	LOC113513811 | XP_026753594.1	motile sperm domain-containing protein 2-like
	Amyelois transitella	NW_013535442.1	LOC106133710 | XP_013188979.1	protein-lysine N-methyltransferase mettl10
		NW_013535442.1	LOC106133711 | XP_013188980.1	histidine triad nucleotide-binding protein 3-like
		NW_013535442.1	LOC106133738 | XP_013189019.1	motile sperm domain-containing protein 2-like
Magenta	*Ephestia kuehniella*	contig_493	EKM1R5v1_00013302	Similar to uncharacterized protein LOC106133714
	Galleria mellonella	NW_022271951.1	LOC113513985 | XP_026753770.1	uncharacterized protein LOC113513985
	Amyelois transitella	NW_013535442.1	LOC106133714 | XP_013188983.1	uncharacterized protein LOC106133714

Chapter 2

Unravelling the complexity of silk sericins: P150/sericin 6 is a new silk gene in *Bombyx mori*

ABSTRACT

Sericins are a small family of highly divergent proteins that serve as adhesives and coatings for silk fibers and are produced in the middle part of the silk gland. So far, five genes encoding sericin proteins have been found in *Bombyx mori*. Sericins 1 and 3 are responsible for silk adhesion in the cocoon, while sericins 2, 4, and 5 are present in non-cocoon spun silk of younger larvae (including the early last instar). We found a new gene, which we named *P150*/*sericin 6*, which appears to be an ortholog of the sericin-like protein previously found in *Galleria mellonella*. The *B. mori* sequence of the *P150/sericin 6* ORF was previously incorrectly predicted and assigned to two smaller, uncharacterized genes. We present a new *P150/sericin 6* gene model and show that it encodes a large protein of 467 kDa. It is characterized by repeats with a high proportion of threonine residues and a short conserved region with a cysteine knot motif (CXCXCX) at the C-terminus. Expression analysis has shown that *B. mori P150/ser6* has a low transcriptional level in contrast to its *G. mellonella* homolog. We also discuss the synteny of homologous genes on corresponding chromosomes between moth species and possible phylogenetic relationships between *P150/ser6* and cysteine knot mucins. Our results improve our understanding of the evolutionary relationships between adhesion proteins in different lepidopteran species.

Keywords: Synteny, sericin, mucin, gene duplication, Bombyx mori, wax moth, Pyralidae

1. Introduction

Silk is a secretory product of several arthropod groups, including insects with the best known being produced by specialized larval salivary glands (the silk glands, SG) of the silkworm. Silk fibers are composed mainly of two types of proteins: fibroins and sericins. Fibroins are the central structural proteins that give silk strength and durability. In most moth species, fibroin is a complex of three protein subunits, fibroin heavy chain (Fib-H), fibroin light chain (Fib-L), and fibrohexamerin/P25 (Fhx/P25)[1-3]. They are produced in the posterior part of SG [4]. Sericins, on the other hand, are coating proteins produced in several layers in the middle part of the SG [5]. Sericins are a small highly divergent family of adhesives that help to glue the fibers together so that silk can form intricate structures such as cocoons or feeding tubes. The use of sensitive proteomic methods and the sequencing of transcriptomes and genomes of individual species indicate that the family of sericin proteins is larger than previously thought. The large divergence among sericin proteins suggests that they have evolved to perform slightly different functions depending on the specific needs of the silk-producing organism.

Two major sericin genes of *B. mori* have been found to produce cocoon sericins: Sericin 1 (Ser1) and Sericin 3 (Ser3) [6, 7]. Silkworm mutants carrying a truncated ser1 gene failed to spin or produced coarse cocoons, suggesting that it is involved in reducing friction during spinning [8]. The product of another sericin gene, sericin 2 (Ser2), and the products of two recently identified genes sericin 4 (Ser4) and sericin 5 (Ser5), have been described in non-cocoon silk and are spun by younger larvae (including early last instar larvae) [9-12]. The presence of Ser2, Ser4, and Ser5 in non-cocoon silks suggests that they may have a specific function during these developmental stages. In this study, we describe another adhesive protein from *B. mori* silk that is homologous to a previously identified sericin-like protein, P150, found in the larval cocoons of *G. mellonella* and *Ephestia kuehniella* (superfamily Pyraloidea) [13, 14].

We show here that there is at least one additional sericin-like protein in the silk of *B. mori*, which we named *P150/ser6* and which is similar to the previously discovered sericin protein P150 of *G. mellonella* [13]. Our results suggest that *B. mori* sequences similar to P150 were misannotated in both GenBank and Silk Base and assigned to two different genes with unknown functions. We present a new model of the *P150/ser6* gene and show that it is a large gene with 4 exons expressed in the middle part of the SG. The P150/ser6 protein product is

found in the pre-cocoon silk, and to a lesser extent in the inner cocoon layer formed at the end of the last instar. The newly discovered protein adds to the existing sericin family of the silkworm. Our results improve our understanding of the functions of silk proteins and the evolutionary relationships among adhesion proteins in different lepidopteran species.

2. Materials and methods

2.1. Silkworm strains and datasets used

A non-diapausing *B. mori* strain, *w1-pnd* (*white egg 1, non-pigmented and non-diapausing egg*), was used in the experiments as a wild type (*wt*) strain. Larvae were reared on mulberry leaves at 25°C. A list of RNA-seq datasets from the NCBI Sequence Read Archive (SRA) is provided in Table S1. (Supplementary Table S1).

2.2. RNA extraction and RT-PCR

To verify the structure of *P150/ser6*, total RNA was extracted from the SGs from SGs of 3-5-day-old fifth-instar larvae using Trizol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. The first cDNA strand was synthesized using 0.5 μg of total RNA as templates. The cDNA product was used to verify the last exon junction of *B. mori P150/ser6* and its expression level in different tissues. Primers were designed using the Geneious Prime software platform (Biomatters, Auckland, New Zealand; version 2021.2.2) and are listed in Supplementary Table S2.

qPCR was performed using HOT FIREPol EvaGreen qPCR Mix Plus (Solis BioDyne, Tartu, Estonia). The PCR reaction volume of 20 μl contained 5 μl diluted cDNA and 250 nM primer. Amplification was performed using a Rotor-Gene Q MDx 2plex HRM (Qiagen, Hilden, Germany) for 45 cycles (95°C for 15 s; annealing temperature matched to the primer pair for 30 s; 72°C for 20 s) after an initial denaturation step (95°C for 15 min). Each sample was analyzed in triplicate. Results were analyzed using Rotor-Gene Q software (version 2.3.5). Elongation factor 1 alpha (EF1a, NM _001044045.1) was used as a reference gene, and the relative expression of target genes was calculated using the $2^{-\Delta\Delta CT}$ method [15]. Statistical analysis was performed using Student's t-test in R (version 4.1.1); p values < 0.05

were considered statistically significant. Detailed statistical analysis is provided in Supplementary Table S3.

2.3. Proteomic analysis and data mining

Protein analysis of *B. mori* cocoons and database searches were performed at the Proteomics Core Facility (BIOCEV, Vestec, Czech Republic) as previously described [16]. Approximately 10 mg of the silk cocoon was boiled in 8 M urea, and samples were further processed using solid-phase enhanced sample preparation technology (SP3 beads) [17]. Samples were then digested with trypsin, and the resulting peptides obtained were subjected to liquid chromatography - MS. Four *wt* cocoons were analyzed in parallel. In addition, raw proteomic data deposited in public databases [18] on the composition of individual silk layers in the *B. mori* cocoon were reanalyzed. The obtained MS/MS spectra were matched against the UniProt database (hwww.uniprot.org), which was enriched for the newly discovered P150/ser6. Quantification was performed using label-free algorithms, and data were analyzed using MaxQuant and Perseus v.1.5.2.4 [19, 20].

2.4. Chromosomal localization and collinearity analysis

The genome assemblies and annotated information of *B. mori* (GCF_014905235.1), *E. kuehniella* [14, 21], and *G. mellonella* (GCF_026898425.1) were processed and submitted to the GENESPACE software [22] for syntenic analysis. Plots showing the microsyntenic relationships were then generated based on the best mutual hits between the three species and visualized using the R package ggplot2 [23].

2.5. Phylogenetic analysis

Codon-based alignment of the P150/ser6 and cysteine knot mucin 3’ ends was performed using MEGA7 software following the MUSCLE method [24]. The phylogram was generated using the IQ-TREE server [25, 26], which included both the selection of the best substitution model by ModelFinder [27] and tree inference using MLE (ultrafast bootstrap, 1,000 replicates).

3. Results

3.1. Identification of *P150/ser6* gene in *B. mori*

To identify the homolog of the major sericin gene *P150* described previously in *G. mellonella* and *E. kuehniella*, we performed a BLAST search against the *B. mori* genome. We found two adjacent homologous regions in the genomic sequence, which were predicted to be parts of two *B. mori* genes. Most of the homologous sequence belonged to LOC101737213; the remaining C-terminus was predicted to be a separate gene LOC119629229. The *B. mori* sequence shared 47.5% identity per 70 C-terminal amino acid residues with *G. mellonella* P150. We hypothesized that the predicted gene models were incorrect and that the sequences of both *B. mori* genes were part of one large *P150/ser6* gene. To test that the two putative *B. mori* genes represent a single gene and produce a single large mRNA, we designed RT-PCR primers that fuse the last two exons of the predicted LOC101737213 with the second exon LOC119629229 and amplified a fragment that confirmed the integrity of the putative large exon (Fig. 1A). We also designed primers that bridge both exons of LOC119629229. As shown in Figure 1B, the amplified cDNA fragments supported the hypothesis of a single large *P150/ser6* gene.

The new gene model of *P150/ser6* is shown in Figure 1A. The entire gene spans approximately 20 kb and consists of four exons and three introns (Fig. 1A). The first two exons of *P150/ser6* encode a signal peptide and part of the short N-terminal nonrepetitive sequence. The third exon is very large (containing 94% of the ORF) and contains two central repetitive regions flanked by unique sequences. The last exon contains a short ORF and a stop codon. The entire gene contains an ORF encoding 4552 amino acids including a signal peptide (19 amino acid residues).

3.2. Putative P150/ser6 protein

The deduced protein product of the P150/ser6 gene is a large protein of 467 kDa consisting of 4552 amino acid residues. It contains a signal peptide of 19 amino acid residues in length, followed by a central portion consisting of a 616 amino acid non-repetitive region, followed by 45 copies of a 30-amino acid repeat 1, a 34-amino acid non-repetitive linker, a

repeat 2 consisting of 73 copies of 35 amino acids, and a nonrepetitive C-terminus of 388 amino acids (the complete sequence is shown in Text. S1). As shown in shown in Fig. 2, P150/ser6 consists of two types of highly conserved, threonine-rich repeat blocks (Supplementary Table S4). The P150/ser6 protein is relatively highly hydrophilic (hydropathy index = -0.672 compared to -1.118 of Ser1). The *B. mori* P150/ser6 contains more than 27% threonine, 14% serine, and 12% alanine residues. The C-terminus (encoded by the last exon) contains a short, conserved cysteine knot motif (Supplementary Table S5). Compared to the P150 proteins of *G. mellonella* and *E. kuehniella*, the *B. mori* P150/ser6 is almost three times larger, less hydrophilic, and contains fewer serine residues. There is very little conservation between the P150/ser6 proteins except for the C-terminal amino acids (Supplementary Table S6).

3.3. *P150/ser6* mRNA is specifically expressed in MSG

To determine whether the *B. mori* homolog of *P150/ser6* is specifically expressed in silk glands, we isolated mRNAs from different parts of the silk glands and control tissues (intestine and ovary) of day 3-5 last instar larvae, prepared cDNA, and performed qPCR. As shown in Figure 3, *P150/ser6* mRNA is highly specific for MSG and ASG, whereas it is absent in PSG and control tissues.

In addition, we reanalyzed the publicly available RNA-seq data for silk gene expression from previous experiments [18, 28]. We used a new annotation of *P150/ser6* and quantified transcript abundance using Kallisto software, and estimated fold changes using DESeq2 (see material and methods). As shown in Figure S1, the results support our data above and indicate that *P150/ser6* mRNA is highly specific to MSG and its maximal expression occurs in the middle part of MSG, similar to that of *Ser2* and *Mucin-12* (*Muc-12*). In addition, the maximal expression of *Ser1* and *Ser3* is found in the middle part of MSG, with a low expression level also in PSG (Fig. S1). The presence of sericin mRNAs in the posterior SG sample may be caused by different separation sites between the posterior and middle SG during tissue dissection.

3.4. Quantitative proteomic analysis of silk samples

To test whether the P150/ser6 protein is present in *B. mori* cocoons, we performed MS proteomic analyses of silk from *wt* cocoons. The results were examined using the Andromeda search engine integrated into the MaxQuant software. The relative abundance of proteins was determined by label-free quantification. We identified 118 proteins using the false discovery rate (FDR) of 1% for protein identification. The intensity of each protein in the biological samples was in strong agreement. A summary of the identified proteins from a triplicate analysis of each cocoon is shown in Table 1.

The results of the proteomic analysis show that the expression of P150/ser6 protein is quite low, similar to those of Ser2, and Muc-12, which are also present at low intensities at the instrumental detection limit (IDL). In contrast, the data (Fig. 4A) showed that Ser1 and Ser3 are the most abundant components of the cocoon silk with concentrations at least five orders of magnitude higher than P150/ser6 (Fig. 4A).

To determine the protein abundance in the different cocoon layers and whether P150/ser6, Muc-12, and Ser2 are coordinately expressed, we also re-analyzed the existing proteomics data in the public repository [18] using our new annotation of P150/ser6. The abundance of P150/ser6, Muc-12 and Ser2 in cocoons is shown in Figure 4B. All three proteins are found at very low levels, with P150/ser6 and Ser2 being the most abundant in the innermost cocoon layer (layer 1), whereas Muc-12 is found in the outermost layer. Taken together, our data show that the abundance of P150/ser6 and Muc-12 in cocoons is low and differs in localization and timing of secretion, with the outermost layer being secreted first, whereas the innermost layer is produced at the end of spinning.

3.5. Synteny in regions coding for *P150/ser6* genes in Lepidoptera

A previous study on the pyralid moths *G. mellonella* and *E. kuehniella* showed that the known sericin genes, except for *P150/ser6*, lie within the cluster of orthologous genes in the corresponding chromosomal regions [14]. Such microsynteny can be also observed between *B. mori* and *G. mellonella* or *E. kuehniella* (Fig 5). The results also show a number of local rearrangements and duplications in this region including the expansion of several sericin genes in *G. mellonella* compared to related moths [14].

In contrast, the *P150/ser6* gene is located on a different chromosome in a more conserved region of a separate cluster, between the genes encoding metalloprotease 1 and croquemont 1.

As shown in Figure 5, the region on chromosome 12 of *B. mori* has a well-conserved synteny, except for an inversion that places the *P150/ser6* region in the opposite direction to the surrounding genes.

3.6. *P150/ser6* may be related to *Muc-12*

To investigate the evolution of *P150/ser6*, we searched for homologous sequences in insect genomes using BLAST and the C-terminal conserved sequence as bait. We found no obvious orthologs in non-lepidopteran insects, suggesting that the *P150/ser6* gene appears to be specific to Lepidoptera. We also found no *P150/ser6* ortholog in members of the superfamily Papilionoidea.

P150/ser6 proteins are highly divergent, and homologous proteins from different Lepidopteran families are difficult to align, except for the conserved C-terminus (the consensus sequence is shown in Fig. 3). The most prominent conserved motif is the cysteine knot (CXCXCX) sequence, which is located 12–29 amino acid residues away from the C-terminus.

Interestingly, there are other silk gland-specific proteins, which also contain cysteine knot sequences. One of these is Muc-12, which is reminiscent of P150/ser6 because of its size and the repetitive structure of its molecule. To learn more about the relationship between cysteine knot mucins and P150/ser6 sequences, we constructed a dendrogram of P150/ser6 and Muc-12 C-termini from representatives of several lepidopteran families (Fig. 6). The resulting phylogenetic tree is robust and distinguishes both clades—P150/ser6 and Muc-12—with high support except for the sequences from the most primitive species (Fig. 6). The sequence alignment and consensus are shown in Figure 6B.

4. Discussion

We discovered a new sericin-like gene *P150/ser6* in the genome of *B. mori*, based on homology with a similar gene in *G. mellonella*. We also found that the region on chromosome 12 where this gene is located was not correctly annotated. Previous gene models of *P150/ser6* contained a homologous sequence that was split into two putative *B. mori* genes. Here, we

present the correct *P150/ser6* gene model and show that its ORF encodes a large protein with a repetitive structure that is specifically expressed in SG.

Structure prediction of large genes is still a problem [29]. The best results are obtained by comparing genomic and cDNA sequences or by using proteomic data [30]. Previous identification of *P150/ser6* sequences in *G. mellonella* and *E. kuehniella* was successfully performed using this approach [13, 14]. In contrast, finding the *B. mori P150/ser6* homolog has been more difficult due to its relatively low expression. Here we show that *P150/ser6* and possibly some other silk genes can be identified on the basis of the homology of a sufficiently conserved, relatively short motif. The identification of *P150/ser6* was supported by microsynteny between the corresponding genomic regions in *B. mori* and *G. mellonella* (Fig. 5).

Previous results showed that the homolog of P150/ser6 is one of the most abundant silk proteins in *G. mellonella* [13]. In contrast, our results show that it is present only at trace levels in *B. mori* silk. The P150/ser6 protein product of *B. mori* is found mainly in the inner cocoon layer formed at the end of the last instar and also in the non-cocoon silk from earlier instar larvae, as suggested by the data of Dong et al. [28]. Consistently, Ser2, Ser4 and Ser5 are present in the non-cocoon silk [9-11]. Non-cocoon silk has previously been associated with the initial stage of silk spinning and is responsible for holding the molting larva and fixing the cocoon to a suitable substrate [11]. Alternatively, the sericin sequences of *B. mori* sericin sequences could have high serine content (Supplementary Table S5), as only simple sericins are digestible by cocoonase [31]. Sericins containing less serine would be restricted to non-cocoon silk. Mutants in cocoonase result in adults being trapped in the cocoon.

Why are there multiple sericins in moth silks? Unlike spiders, moths have only one fibroin gene, which is considered to be the main structural component of silk. However, silk is modified differently in different moth species, and sericins appear to be the most variable components, important for building three-dimensional silk structures and contributing to silk strength and toughness [32]. For example, *G. mellonella* builds very dense feeding tubes and cocoons that are needed for larval protection in hives. Sericins and other soluble silk components make up about 48% of the cocoon mass of *G. mellonella*, whereas *B. mori* cocoons contain only about 26% soluble cocoon proteins. Some moths, including those of *Samia ricini,* contain as little as 16 % soluble proteins [11, 33]. The number of sericin genes

can also vary widely, for example, *G. mellonella* contains twice as many sericin genes as *E. kuehniella* [14].

The cysteine knot motif (CXCXCX) encoded by the C-termini of some SG-specific genes is similar to the motif described in some mammalian growth factors, including the VEGF family [34]. The function of cysteine knot motifs has been suggested for protein structural integrity. In addition to P150/ser6 and Muc-12, there are at least two other *B. mori* SG-specific proteins that carry this motif. One of these, a putative 53 kDa non-repetitive protein, is also expressed in silk glands and has homologs in other moths and even caddisflies. The other protein, designated egalitarian protein homolog (LOC101741849), is more conserved and appears to be well separated from other cysteine knot proteins [35].

The general similarity of P150/ser6 and Muc-12 suggests that they may have a common origin. P150/ser6 and Muc-12 are both large, highly divergent proteins. They have repetitive sequences encoding ORFs with simple amino acids and both are likely to serve as silk adhesives [13]. Our phylogram (Fig. 6) distinguishes the two clades P150/ser6 and Muc-12 with good support, except for the sequences in the most primitive species, where it is difficult to place them confidently in the phylogram. The evolutionary relationship between P150/ser6 and mucins with a cysteine node motif is not clear. The question of whether P150/ser6 and cysteine knot mucins share a common ancestor and have diverged extensively or whether they are the product of convergent evolution remains an important question for future research.

Author contributions

MiZ: supervision writing; BChW: investigation, data analysis; writing; MaZ: phylogenetic analysis; VZ: *B. mori* rearing and staging.

Data and materials availability

Data will be made available on request.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the European Community's Program Interreg Bayern – Tschechien BYCZ01-039. This publication is also supported by the project "BIOCEV – Biotechnology and Biomedicine Centre of the Academy of Sciences and Charles University" (CZ.1.05/1.1.00/02.0109), from the European Regional Development Fund.

References

[1] K. Tanaka, K. Mori, S. Mizuno, Immunological Identification of the Major Disulfide-Linked Light Component of Silk Fibroin, J Biochem-Tokyo 114(1) (1993) 1-4.

[2] M. Zurovec, D. Kodrik, C. Yang, F. Sehnal, K. Scheller, The P25 component of *Galleria* silk, Mol Gen Genet 257(3) (1998) 264-70.

[3] M. Zurovec, M. Vaskova, D. Kodrik, F. Sehnal, A.K. Kumaran, Light-chain fibroin of *Galleria mellonella* L, Mol Gen Genet 247(1) (1995) 1-6.

[4] F. Lucas, K.M. Rudall, Extracellular fibrous proteins: The silks., in: M. Florkin, E.H. Stotz (Eds.), Comprehensive Biochemistry, Elsevier, Amsterdam, 1968, pp. 475-558.

[5] A. Shibukawa, Studies on the silk substance within the silk gland in the silkworm, *Bombyx mori* L., Bull. Sericult. Exp. Sta. 15 (1959) 401.

[6] H. Okamoto, E. Ishikawa, Y. Suzuki, Structural analysis of sericin genes. Homologies with fibroin gene in the 5' flanking nucleotide sequences, J Biol Chem 257(24) (1982) 15192-9.

[7] Y. Takasu, H. Yamada, T. Tamura, H. Sezutsu, K. Mita, K. Tsubouchi, Identification and characterization of a novel sericin gene expressed in the anterior middle silk gland of the silkworm *Bombyx mori*, Insect Biochem Mol Biol 37(11) (2007) 1234-40.

[8] Y. Takasu, T. Iizuka, Q. Zhang, H. Sezutsu, Modified cocoon sericin proteins produced by truncated *Bombyx* Ser1 gene, The Journal of Silk Science and Technology of Japan 25 (2017) 35-47.

[9] Z.M. Dong, K.Y. Guo, X.L. Zhang, T. Zhang, Y. Zhang, S.Y. Ma, H.P. Chang, M.Y. Tang, L.N. An, Q.Y. Xia, P. Zhao, Identification of *Bombyx mori* sericin 4 protein as a new biological adhesive, Int J Biol Macromol 132 (2019) 1121-1130.

[10] K. Guo, X. Zhang, D. Zhao, L. Qin, W. Jiang, W. Hu, X. Liu, Q. Xia, Z. Dong, P. Zhao, Identification and characterization of sericin5 reveals non-cocoon silk sericin components with high beta-sheet content and adhesive strength, Acta Biomater 150 (2022) 96-110.

[11] B. Kludkiewicz, Y. Takasu, R. Fedic, T. Tamura, F. Sehnal, M. Zurovec, Structure and expression of the silk adhesive protein Ser2 in *Bombyx mori*, Insect Biochem Molec 39(12) (2009) 938-946.

[12] J.J. Michaille, A. Garel, J.C. Prudhomme, Cloning and Characterization of the Highly Polymorphic Ser2 Gene of *Bombyx mori*, Gene 86(2) (1990) 177-184.

[13] B. Kludkiewicz, L. Kucerova, T. Konikova, H. Strnad, M. Hradilova, A. Zaloudikova, H. Sehadova, P. Konik, F. Sehnal, M. Zurovec, The expansion of genes encoding soluble silk components in the greater wax moth, *Galleria mellonella*, Insect Biochem Mol Biol 106 (2019) 28-38.

[14] B.C. Wu, I. Sauman, H.O. Maaroufi, A. Zaloudikova, M. Zurovcova, B. Kludkiewicz, M. Hradilova, M. Zurovec, Characterization of silk genes in *Ephestia kuehniella* and *Galleria mellonella* revealed duplication of sericin genes and highly divergent sequences encoding fibroin heavy chains, Front Mol Biosci 9 (2022) 1023381.

[15] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method, Methods 25(4) (2001) 402-8.

[16] V. Zabelina, Y. Takasu, H. Sehadova, N. Yonemura, K. Nakajima, H. Sezutsu, M. Sery, M. Zurovec, F. Sehnal, T. Tamura, Mutation in *Bombyx mori* fibrohexamerin (P25) gene causes reorganization of rough endoplasmic reticulum in posterior silk gland cells and alters morphology of fibroin secretory globules in the silk gland lumen, Insect Biochem Mol Biol 135 (2021) 103607.

[17] C.S. Hughes, S. Moggridge, T. Muller, P.H. Sorensen, G.B. Morin, J. Krijgsveld, Single-pot, solid-phase-enhanced sample preparation for proteomics experiments, Nat Protoc 14(1) (2019) 68-85.

[18] Y. Zhang, P. Zhao, Z. Dong, D. Wang, P. Guo, X. Guo, Q. Song, W. Zhang, Q. Xia, Comparative proteome analysis of multi-layer cocoon of the silkworm, *Bombyx mori*, PLoS One 10(4) (2015) e0123403.

[19] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ, Mol Cell Proteomics 13(9) (2014) 2513-26.

[20] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M.Y. Hein, T. Geiger, M. Mann, J. Cox, The Perseus computational platform for comprehensive analysis of (prote)omics data, Nat Methods 13(9) (2016) 731-40.

[21] S. Visser, A. Volenikova, P. Nguyen, E.C. Verhulst, F. Marec, A conserved role of the duplicated Masculinizer gene in sex determination of the Mediterranean flour moth, *Ephestia kuehniella*, PLoS Genet 17(8) (2021) e1009420.

[22] J.T. Lovell, A. Sreedasyam, M.E. Schranz, M. Wilson, J.W. Carlson, A. Harkess, D. Emms, D.M. Goodstein, J. Schmutz, GENESPACE tracks regions of interest and gene copy number variation across multiple genomes, Elife 11 (2022).

[23] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Use R (2009) 1-212.

[24] S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, Mol Biol Evol 33(7) (2016) 1870-4.

[25] D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation, Mol Biol Evol 35(2) (2018) 518-522.

[26] L.T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, Mol Biol Evol 32(1) (2015) 268-74.

[27] S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, L.S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates, Nature Methods 14(6) (2017) 587-+.

[28] Z. Dong, P. Zhao, C. Wang, Y. Zhang, J. Chen, X. Wang, Y. Lin, Q. Xia, Comparative proteomics reveal diverse functions and dynamic changes of *Bombyx mori* silk proteins spun from different development stages, J Proteome Res 12(11) (2013) 5213-22.

[29] J. Wang, S. Li, Y. Zhang, H. Zheng, Z. Xu, J. Ye, J. Yu, G.K. Wong, Vertebrate gene predictions and the problem of large genes, Nat Rev Genet 4(9) (2003) 741-9.

[30] M. Levin, F. Butter, Proteotranscriptomics - A facilitator in omics research, Comput Struct Biotec 20 (2022) 3667-3675.

[31] T.T. Gai, X.L. Tong, M.J. Han, C.L. Li, C.Y. Fang, Y.L. Zou, H. Hu, H. Xiang, Z.H. Xiang, C. Lu, F.Y. Dai, Cocoonase is indispensable for Lepidoptera insects breaking the sealed cocoon, Plos Genetics 16(9) (2020).

[32] Y.R. Li, Y.K. Wei, G.Z. Zhang, Y.S. Zhang, Sericin from Fibroin-Deficient Silkworms Served as a Promising Resource for Biomedicine, Polymers-Basel 15(13) (2023).

[33] S. Prasong, S. Yaowalak, S. Wilaiwan, Characteristics of silk fiber with and without sericin component: A comparison between *Bombyx mori* and *Philosamia ricini* silks, Pak. J. Biol. Sci. 12 (2009) 872−876.

[34] Y.A. Muller, C. Heiring, R. Misselwitz, K. Welfle, H. Welfle, The cystine knot promotes folding and not thermodynamic stability in vascular endothelial growth factor, J Biol Chem 277(45) (2002) 43410-6.

[35] S.M. Hong, S.K. Nho, N.S. Kim, J.S. Lee, S.W. Kang, Gene expression profiling in the silkworm, *Bombyx mori*, during early embryonic development, Zoolog Sci 23(6) (2006) 517-28.

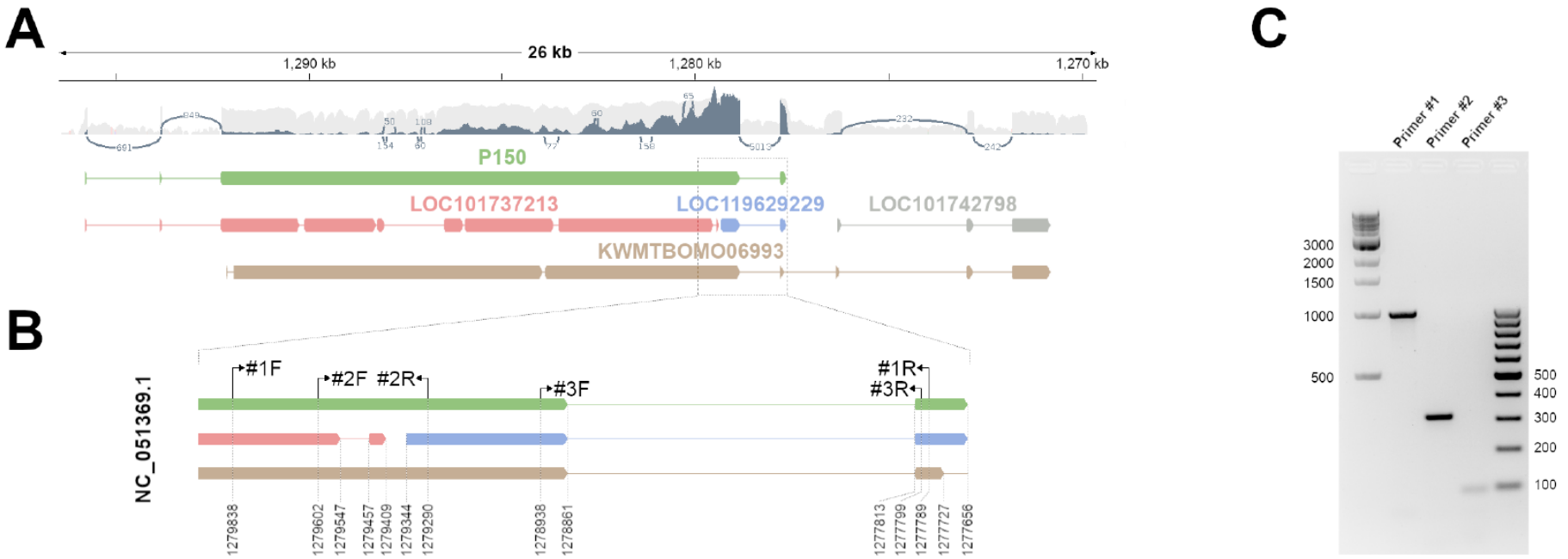


Figure. 1. Different models of a *B. mori* genomic region containing sequences similar to the P150 protein of *G. mellonella*. (A) The genomic region is part of chromosome 12 (NC_051369.1: 1270000-1280000, strand flipped). The exon-intron structure of the P150/ser6 gene, consisting of four exons and three introns, as predicted by our analysis, is shown in green. Below - comparison with current models in the NCBI database the following gene models are shown: LOC101737213 (pink); LOC119629229 (blue); LOC101742798 (gray); and the model of the gene KWMTBOMO06993 in Silkbase (brown). The size bar in kilobase pairs is shown at the top. (B) The lower inset shows an approximately 2.2-kb region with the chromosomal coordinates and positions of the primer pairs used in this study. (C) Verification of the 3' region of the newly designed P150/ser6 gene model by RT-PCR and agarose gel electrophoresis. The expected product sizes of primer pairs #1F-1R, #2F-2R, and #3F-3R were 1003 bp, 313 bp, and 93 bp, respectively. The electrophoretogram contains a 1 kb ladder (lane 1) and 100 bp ladder (lane 5).

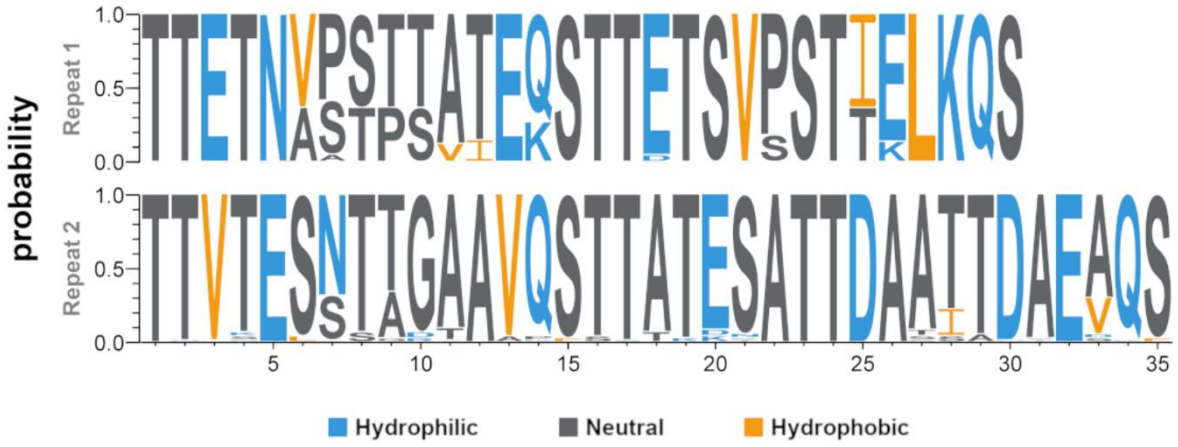


Figure. 2. Amino acid sequence logos. The sequence logos show the conservation of amino acids in two types of repeats found in the *B. mori* P150/ser6 protein. Hydrophobicity of an amino acid is indicated by color: hydrophilic (blue; RKDENQ); neutral (dark gray; SGHTAP); hydrophobic (orange; YVMCLFIW).

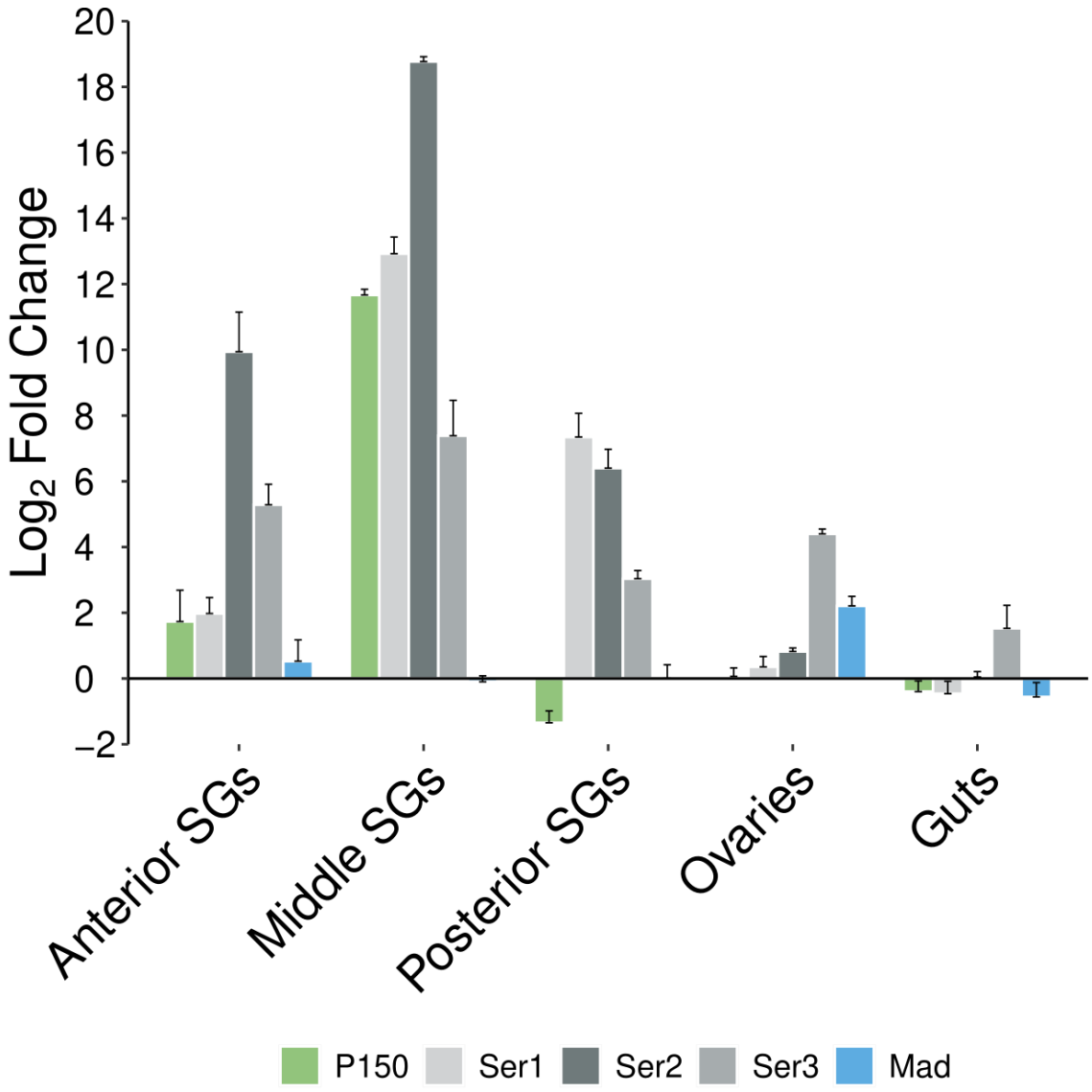


Figure. 3. Quantitative PCR (qPCR) analysis of gene expression of silk gland-specific genes (ser1, ser2, ser3, and P150/ser6) and control (non-silk gland-specific gene MAD) in different tissues. Statistical differences were evaluated using Student's t-test (see Supplementary data); error bars are SD.

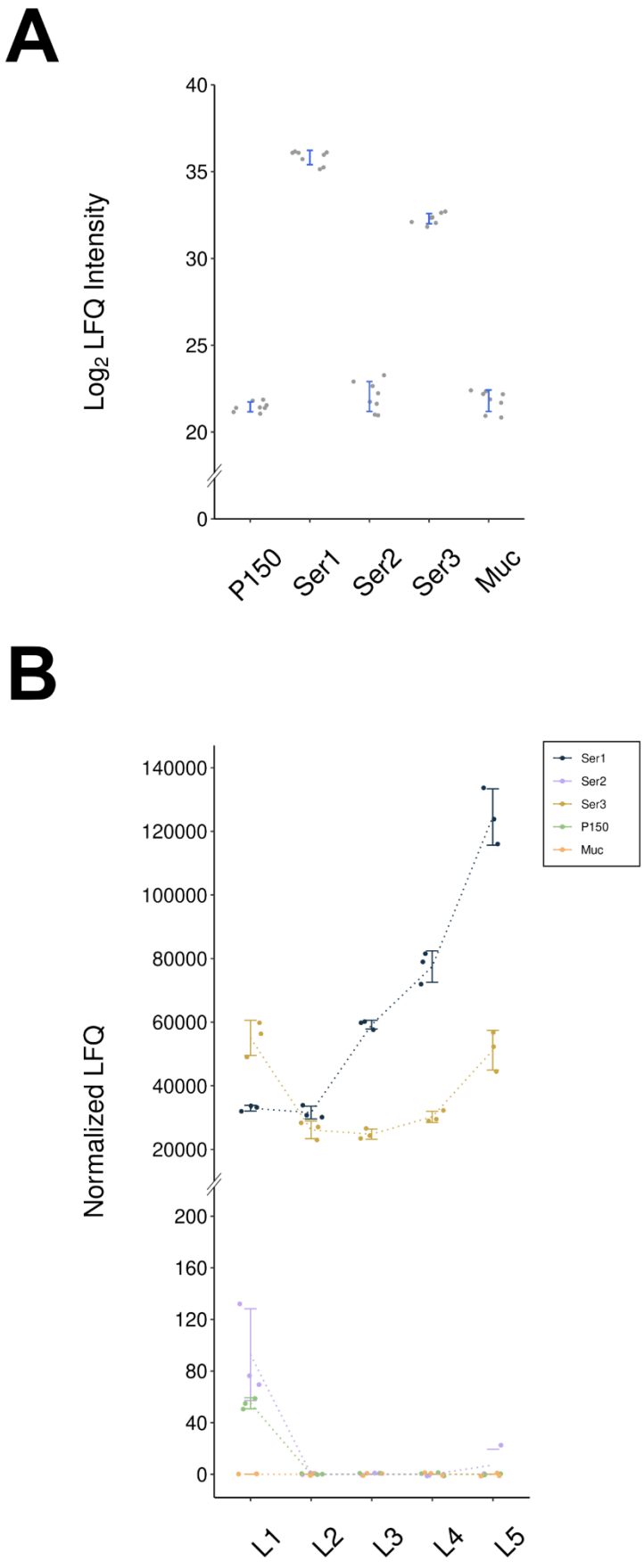


Figure. 4. Proteomic analysis of *B. mori* silk proteins (A) from wt cocoons as previously described [16]; and (B) individual cocoon layer data from a public repository [18]. Label-free quantification (LFQ) of silk proteins from cocoons was calculated using MaxQuant. LFQ intensities were log2-transformed. Relative protein contents in cocoon silk were analyzed using MaxQuant/Andromeda (eight experiments). Error bars indicate the standard deviation. Proteomic analysis confirmed that Ser1 and Ser3 were the most abundant silk components. The other proteins, including P150/ser6, Ser2, and Muc-12, were present at low levels at the instrumental detection limit (IDL).

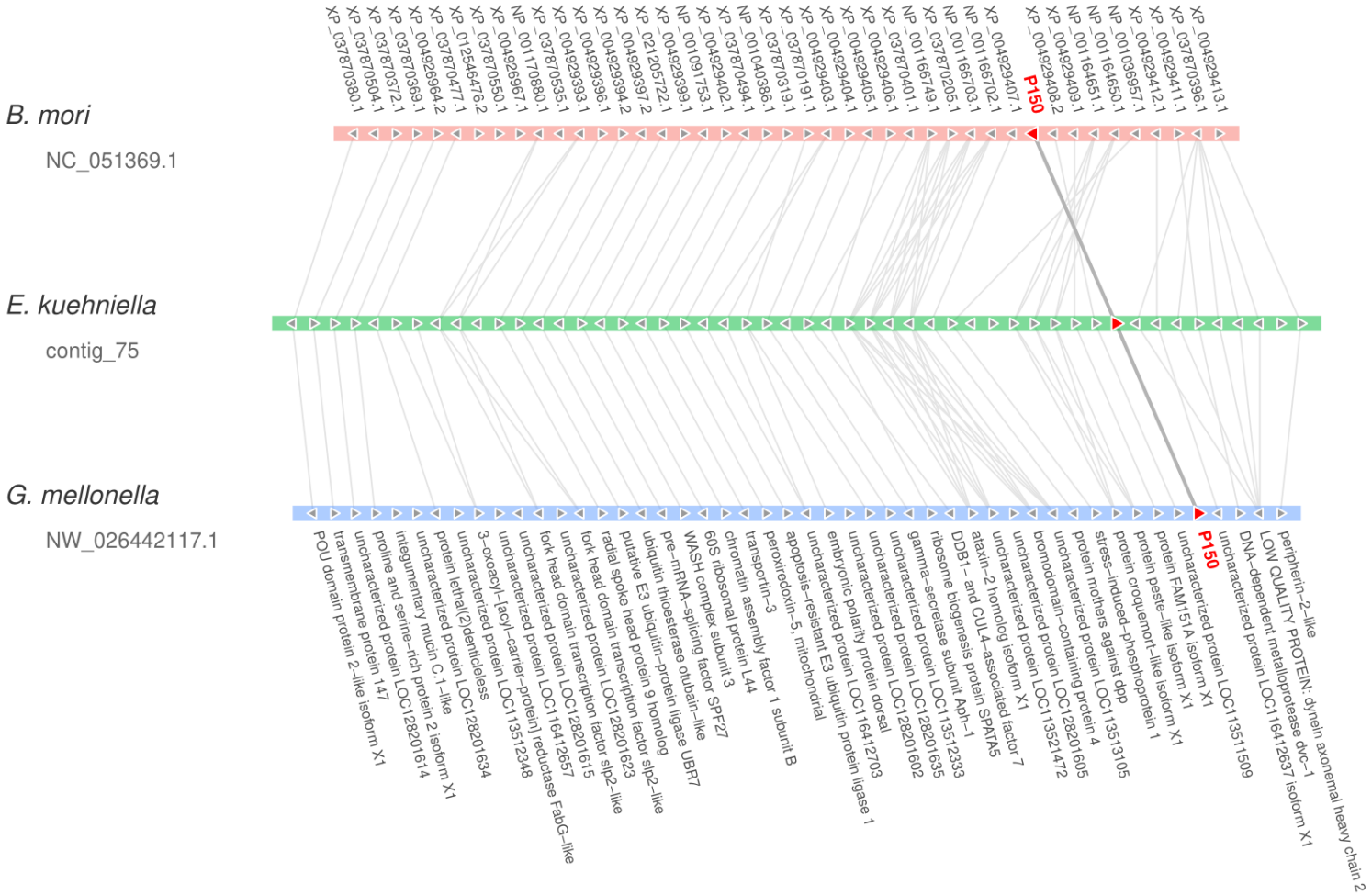


Figure. 5. Microsynteny maps of P150/ser6 and their flanking genes. Horizontal color blocks indicate chromosomal segments in each species. Homologous genes and gene orientation are represented by left- and right-pointing triangles and connected by lines. Detected P150/ser6 homologs are shown in red.

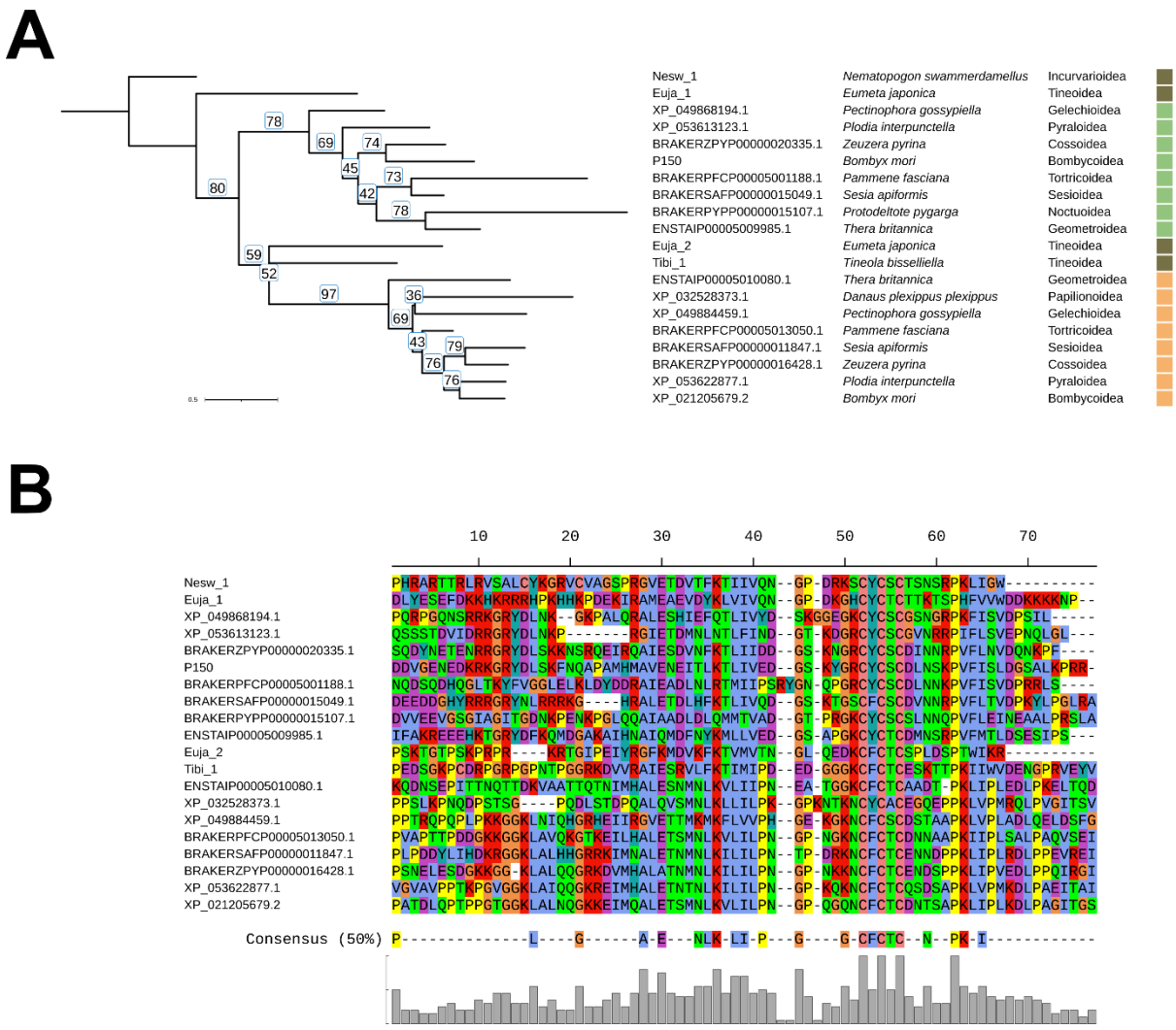


Figure. 6. Relationship between P150/ser6 and cysteine knot mucin sequences. (A) Maximum likelihood phylogenetic tree based on the alignment of the C-terminal amino acid sequences of P150/ser6 and Muc-12 homologs in selected lepidopteran species. The Nesw_1 transcript from *N. swammerdamellus*, (Incurvarioidea), the most primitive in this group, was selected for tree rooting. See Supplementary Text S2 and Supplementary Table S7 for sequences. (B) Alignment of the C-terminal regions of P150/ser6 proteins from representative lepidopteran species. Sequences include the characteristic CXCXCX region, which is well-conserved between species.

Table 1. List of proteins detected in *B. mori* cocoon silk by MaxQuant. Intensity is the sum of intensity values of all replicates.

Intensity	Accession	Gene	Residue	M. w. (kDa)	GRAVY	pI	Top3 AA	Signal peptide
380194000000	XP_037869538.1	sericin 1 isoform X1	3385	331.903	-1.118	4.773	Ser (38.405%); Gly (12.644%); Thr (9.365%)	MRFVLCCTLIALAALSVKAFG
41472000000	XP_004926112.1	uncharacterized protein LOC101740726	239	25.909	-0.016	8.824	Leu (15.900%); Ala (10.460%); Gly (8.368%)	MKLLVVLAVCAVAAA
27399600000	XP_004922636.1	fibrohexamerin-like	241	27.201	-0.058	4.555	Leu (9.959%); Asn (7.884%); Ser (7.884%)	MKLMLCLCLLSFFGLTVA
22181230000	XP_037874919.1	LOW QUALITY PROTEIN: zonadhesin	3973	432.78	-0.563	4.644	Cys (14.372%); Gly (8.759%); Asn (8.558%)	MLKTFLFVACIANVLLVDSTEA
17189800000	NP_001108116.1	sericin 3 precursor	1271	123.298	-1.465	5.708	Ser (43.588%); Gly (11.959%); Gln (7.396%)	MNCKVALFLIVAIVAVQA
11367900000	NP_001106733.1	fibroin heavy chain precursor	5263	391.593	0.216	4.128	Gly (45.886%); Ala (30.268%); Ser (12.065%)	MRVKTFVILCCALQYVAYTNA
7049000000	NP_001037488.1	fibroin light chain precursor	262	27.754	0.024	5.188	Ala (13.740%); Ser (9.542%); Gly (8.397%)	MKPIFLVLLVATSAYA
6734500000	XP_004921814.1	uncharacterized protein LOC101741789	474	50.925	-0.448	5.343	Lys (9.705%); Ser (9.494%); Val (8.228%)	MKVAAALFTILQIVNA
4998900000	NP_001037045.1	seroin 1 precursor	108	11.882	-0.388	4.059	Ser (12.037%); Asp (10.185%); Glu (9.259%)	MAFTKFLFVITLITIASA
3343110000	NP_001078833.1	beta-N-acetylglucosaminidase 1 precursor	611	69.655	-0.287	6.124	Leu (9.493%); Val (7.529%); Arg (6.710%)	MALHRLLALMTTMHLAGVLC
2228430000	XP_004921708.1	autotransporter adhesin BpaC-like	261	24.119	-0.133	3.879	Ala (29.502%); Ser (19.923%); Thr (17.241%)	MSLSRVAVFASILVCYYVSIVA
1978960000	XP_004926642.2	apolipophorins isoform X1	3317	369.106	-0.315	7.973	Leu (9.828%); Lys (9.014%); Ser (7.266%)	MGTISFSLSVILLISVLWEPALA
1776180000	NP_001106745.1	carboxypeptidase inhibitor precursor	113	12.561	-0.296	4.244	Cys (13.274%); Glu (10.619%); Gly (8.850%)	MKFLLFCAMCILVYG
1700100000	XP_004925684.1	rh5-interacting protein isoform X1	394	43.157	-0.34	4.366	Cys (13.959%); Pro (10.660%); Asn (8.376%)	MFILFCFISTVAA
1198871000	XP_012546338.1	glycine-rich cell wall structural protein 1.0-like	208	18.431	-0.801	3.207	Gly (36.538%); Asp (18.750%); Ala (12.500%)	MKLKWLLLLVVVAVTYVEA
1102090000	XP_037874647.1	zonadhesin isoform X1	683	75.835	-0.606	4.663	Cys (14.641%); Glu (9.956%); Pro (7.174%)	MGRFCFIVLTAAWLVVAKADA
1100470000	XP_037867528.1	zonadhesin isoform X1	2721	296.36	-0.549	5.952	Cys (14.627%); Pro (13.120%); Asn (8.012%)	MLVLFVGLLMLISSSTA
1081719000	XP_004928913.2	vanin-like protein 1	515	57.684	-0.124	5.014	Val (8.932%); Leu (7.573%); Ser (7.379%)	MKLIFYGLLIFSLTKATS
1060700000	XP_004923912.1	alcohol dehydrogenase 2	319	35.344	-0.202	6.788	Val (9.404%); Ile (8.464%); Asn (8.464%)	MFVITYFNILFVALFFVNVTST
1034207000	XP_021204310.2	uncharacterized protein LOC101744556 isoform X1	1979	222.161	-0.727	4.753	Asn (9.803%); Pro (8.489%); Ser (8.439%)	MEYKVNLIFLVVFSSKIVVSIP
961800000	XP_037871220.1	fibrohexamerin-like	240	27.327	-0.179	4.591	Leu (9.583%); Asn (8.333%); Glu (7.500%)	MRLVWYSCLALFGLITA
880960000	XP_037873722.1	glucose dehydrogenase [FAD, quinone]	577	64.896	-0.155	7.01	Leu (10.052%); Val (8.146%); Lys (7.452%)	-
749953000	XP_037869877.1	uncharacterized protein LOC101740721 isoform X1	2270	251.441	-0.988	6.716	Lys (9.824%); Thr (9.604%); Ser (9.471%)	MDWTKYLVLFILMVVVSS
717188000	XP_037874060.1	myrosinase 1	567	65.135	-0.404	6.173	Leu (8.466%); Ser (8.113%); Ala (7.760%)	MFGLLSMWISWDLFLFERPASV
700756000	NP_001139413.1	fibrohexamerin precursor	220	25.167	-0.102	7.162	Leu (10.000%); Ala (7.273%); Phe (6.818%)	MLARCLAVAAVAVLASA
677915000	NP_001108339.1	integument esterase 1 precursor	561	63.237	-0.425	6.835	Leu (8.021%); Gly (7.843%); Pro (7.843%)	MLLKLIFICAIVYYADA
673315000	XP_012551240.1	uncharacterized protein LOC105842476	234	26.439	-0.007	4.168	Leu (9.402%); Asn (8.547%); Val (7.692%)	MEFELCLCILSVLGLIAA
474186000	XP_004924748.1	proton-coupled folate transporter	508	57.334	0.229	8.584	Leu (10.827%); Ile (9.646%); Ser (8.661%)	-
456522000	XP_012546430.1	lysosomal acid glucosylceramidase isoform X1	536	61.151	-0.243	7.051	Leu (9.142%); Lys (7.836%); Ile (7.090%)	-
403273500	XP_004928916.1	glucose dehydrogenase [FAD, quinone]	657	73.389	-0.294	8.726	Leu (8.676%); Val (7.610%); Gly (7.154%)	MHTRRKIIFMWFLLFVNANCCDA
352649000	XP_037868215.1	beta-glucuronidase isoform X1	688	79.06	-0.304	7.015	Val (7.849%); Leu (6.977%); Thr (6.831%)	-
342382000	XP_004927001.2	lipase 1	399	46.491	-0.293	5.885	Leu (9.023%); Glu (8.020%); Ile (7.519%)	MVKIGLRTFFMILSIGLHLASA
342254000	XP_004922749.1	protein D2	209	23.055	0.077	6.528	Val (11.005%); Leu (10.048%); Ala (8.612%)	MGALIRCAIVLLTVATIVDFRVLTRA
319436000	XP_037872749.1	dehydrogenase mpl7 isoform X1	585	65.476	-0.206	9.092	Lys (9.573%); Leu (9.060%); Val (8.547%)	-
314065000	NP_001037046.2	seroin 2 precursor	112	12.273	-0.287	9.883	Ser (12.500%); Lys (10.714%); Pro (10.714%)	MAFTKFLFMLSLITIASA
298483000	XP_037869036.1	beta-galactosidase isoform X1	2942	328.542	-0.211	6.857	Leu (9.007%); Val (7.988%); Gly (7.104%)	MTLLIATFALLCFLGQTYS
283373000	XP_012545566.1	yellow-d isoform X1	475	53.869	-0.208	6.511	Val (8.211%); Leu (7.158%); Ser (7.158%)	-
252201950	NP_001036850.1	cathepsin B precursor	337	37.556	-0.441	6.382	Gly (9.792%); Ser (7.122%); Asp (6.825%)	MFISRAAYVTLVCVLAAA
245267000	XP_004924612.2	esterase FE4	550	61.673	-0.231	5.06	Leu (8.000%); Gly (7.636%); Ile (7.273%)	MLIKTFFVIASVVYVLG
181921000	XP_037876844.1	latent-transforming growth factor beta-binding protein 4	1359	150.321	-0.652	5.589	Thr (10.302%); Ser (8.389%); Cys (7.873%)	MKFIVVIFTFLFSTVNA
165783000	XP_004929234.1	venom dipeptidyl peptidase 4	740	83.919	-0.199	5.527	Leu (8.243%); Val (7.973%); Ile (7.162%)	MAMEQAVFLLMGLICHTQSSA
149592000	NP_001093080.1	ecdysteroid-regulated 16 kDa protein precursor	145	15.824	0.179	6.336	Leu (11.724%); Val (8.276%); Ala (7.586%)	MLFFITAAVLLASAEA
148759000	XP_037874096.1	myrosinase 1 isoform X1	495	57.471	-0.473	5.019	Leu (8.081%); Glu (7.879%); Ser (7.071%)	MNLSWQVAFSALMACAWG
145757000	XP_037868236.1	juvenile hormone esterase-like	562	63.516	-0.189	9.306	Leu (9.253%); Asn (8.541%); Ile (7.295%)	MSNTFYLITVLILVLNTINA
143543000	NP_001268822.1	neutral alpha-glucosidase AB-like precursor	925	104.635	-0.371	5.706	Leu (8.432%); Val (8.324%); Ala (8.000%)	MKVWALVLVAAFVIIGISA
137424000	XP_004926880.2	lysosomal acid glucosylceramidase isoform X1	530	59.677	-0.19	4.426	Leu (8.113%); Asp (7.358%); Ile (7.170%)	MDINKCIISASLWAAVYLLFGSADA
134477000	XP_012547740.1	heat shock 70 kD protein cognate isoform X1	960	106.18	-0.4	5.486	Lys (8.542%); Val (8.229%); Ala (8.125%)	-
121943000	XP_037868883.1	arylsulfatase B	537	60.465	-0.397	6.796	Leu (8.752%); Gly (8.194%); Lys (7.635%)	MFVFRVLLISVLMLVAES
116001000	XP_012544831.3	uncharacterized protein LOC101741510	1999	228.101	-0.53	5.916	Val (8.054%); Leu (8.004%); Lys (7.404%)	MSIKSAYLLIFQIIVIYNCSC
114229100	NP_001037047.1	silk proteinase inhibitor precursor	65	6.962	0.554	8.134	Cys (13.846%); Gly (10.769%); Thr (10.769%)	MKTSIVLIFLLVACCTLGAES
113920600	XP_021206228.1	poly(U)-specific endoribonuclease-D isoform X1	536	60.005	-0.79	9.886	Ser (10.634%); Thr (9.888%); Asn (8.769%)	-
102791000	NP_001106744.1	antennal binding protein precursor	140	15.49	-0.11	7.206	Lys (11.429%); Ala (10.000%); Leu (7.857%)	MMYLSFVVLICLAFAVFNCGA
97291000	XP_004927018.1	C3 and PZP-like alpha-2-macroglobulin domain-containing protein 8	1448	155.934	-0.031	5.938	Ala (11.119%); Leu (10.152%); Ser (8.840%)	MIEKIKTLLLFILSVPAVTQC
96514000	XP_037874655.1	inducible metalloproteinase inhibitor protein-like isoform X1	200	22.342	-0.704	5.769	Cys (10.500%); Glu (9.500%); Ala (7.000%)	-
93201000	XP_021203913.1	uncharacterized protein LOC101747119 isoform X1	722	80.01	-0.235	6.67	Val (9.695%); Leu (7.895%); Ser (7.895%)	MNSNEIIYFVILTLCSSVAG
77323000	XP_012545193.1	uncharacterized protein LOC101735738 isoform X1	671	75.143	-0.309	6.349	Leu (10.581%); Gly (8.197%); Val (7.899%)	MSVRFVRTILVLLAIVHNSRA
75014000	NP_001040174.1	alpha-esterase 13 precursor	540	61.416	-0.241	5.012	Leu (9.815%); Phe (7.593%); Lys (7.593%)	MLFAIIICVQVLSVFG
74413000	XP_037876352.1	4-hydroxy-tetrahydrodipicolinate synthase-like	328	35.694	0.094	4.9	Leu (11.585%); Ala (8.537%); Ile (7.927%)	MYSISGLTFLCVLIINVFDTKC
73489010	XP_037867171.1	15-hydroxyprostaglandin dehydrogenase [NAD(+)]-like	289	31.605	-0.025	9.45	Lys (10.727%); Ile (10.035%); Ala (7.958%)	MTLSLLKTFFVSCFIINIVSA
71931700	NP_001037075.1	calreticulin precursor	398	45.802	-0.992	4.245	Asp (13.819%); Lys (12.060%); Glu (11.307%)	MKAVVLVVVSLLALSSINC
68856900	XP_021206073.2	uncharacterized protein LOC105842244	1131	127.974	-0.974	9.852	Thr (10.698%); Ser (9.991%); Pro (9.107%)	MCGARLLLAATALQALILAVSC
67160000	XP_012547984.1	uncharacterized protein LOC101744060	934	101.918	-0.275	4.433	Leu (9.850%); Ser (8.565%); Ala (8.458%)	MNANKKAIVLLTLINILLVRS
66627000	XP_037874369.1	fibrillin-1 isoform X1	3047	325.331	-0.468	4.496	Gly (12.504%); Cys (12.143%); Asp (7.844%)	MGDGAKMSYWRLLVAAALLASVAG
60053000	XP_012550868.1	uncharacterized protein LOC778506 isoform X1	292	30.66	-0.599	4.054	Ala (16.781%); Glu (12.329%); Lys (11.644%)	-
59817000	XP_037868167.1	uncharacterized protein LOC101736658 isoform X1	456	51.569	-0.228	6.318	Leu (9.649%); Ser (8.333%); Val (8.114%)	MGSIETLVLLQLVYISSC
59417900	XP_004929039.1	phospholipase A2	177	20.411	-0.24	4.652	Glu (7.910%); Asp (7.345%); Gly (7.345%)	MFRHLFCLIVLIAVHTNKA
58891600	XP_037867041.1	alpha-crystallin B chain	241	26.543	-0.263	4.361	Val (11.618%); Ala (9.129%); Thr (9.129%)	MFSPRLFAVFIAFAGFATVTA
55656900	XP_021207905.1	uncharacterized protein LOC101743953 isoform X1	472	52.872	-0.063	7.143	Leu (11.441%); Ile (8.263%); Ala (8.051%)	MAVSLYVVLVLATTALG
55294500	XP_037868227.1	mucin-5AC	1768	198.075	-0.847	6.971	Thr (14.989%); Ser (8.484%); Glu (8.258%)	MGIFDYKFIGKLFKMESTIWLTLLLAVIATTPVSS
50392000	NP_001037090.1	serine protease inhibitor 4 precursor	410	46.3	-0.095	7.322	Leu (9.024%); Ser (7.805%); Ile (7.561%)	MCLLKFLVLAIVPLSFA
44188900	NP_001037171.1	protein disulfide isomerase precursor	494	55.589	-0.278	4.325	Glu (11.134%); Ala (8.704%); Lys (8.502%)	MRVLIFTAIALLGLALG
41498000	NP_001119727.1	actin, cytoplasmic A4	376	41.822	-0.193	5.156	Ala (7.979%); Glu (7.447%); Gly (7.447%)	-
40515000	XP_037874307.1	basement membrane-specific heparan sulfate proteoglycan core protein isoform X1	4228	463.254	-0.538	4.432	Gly (8.609%); Ser (8.538%); Asp (7.427%)	MRAGGLLAAALLLLSTFTIQVLKA
39652000	XP_004926666.2	alaserpin	410	45.953	-0.194	5.018	Leu (10.244%); Ile (9.024%); Glu (8.293%)	-
38903380	XP_012551293.2	heme peroxidase 2	1344	149.897	-0.304	6.792	Leu (10.045%); Ala (8.929%); Pro (7.068%)	MGAFKASIRSLLLFTLALT
37924900	NP_001155191.2	silk gland derived serine protease 1 precursor	392	42.098	0.134	7.393	Gly (9.694%); Ser (9.439%); Val (8.163%)	MCLQLVLVVLALNGVLS
37469400	NP_001040479.1	peptidylprolyl isomerase B precursor	205	22.397	-0.267	8.842	Gly (12.195%); Lys (10.244%); Thr (8.780%)	MGTLTMALGILLFIASAKS
37262200	NP_001037430.1	yellow-b precursor	457	50.803	-0.156	5.119	Leu (9.409%); Ala (8.972%); Asn (7.440%)	MRYTLTSLSMMIAIVLLAAVAAAAA
36225300	NP_001037073.1	glucosidase precursor	491	55.596	-0.375	4.575	Asp (7.943%); Leu (7.943%); Gly (7.536%)	MAWLTTLSILAVCHTGLA
34377200	NP_001268827.1	low molecular mass 30 kDa lipoprotein 21G1-like precursor	256	29.858	-0.714	4.942	Asp (10.938%); Lys (8.984%); Ser (8.594%)	MKLKILTTLSVIIATFFNMSVEA
30436100	XP_037870111.1	probable polyamine oxidase 5 isoform X1	485	53.978	-0.216	6.632	Leu (9.485%); Gly (7.835%); Thr (7.835%)	MIVRVAILVLCVAVQGRG
29882500	NP_001036944.1	juvenile hormone binding protein an-0895 precursor	243	27.294	-0.047	4.558	Leu (13.992%); Val (9.053%); Asn (8.642%)	MWTGLFLVLGLYESVVG
27055300	XP_021205588.1	probable cyclin-dependent serine/threonine-protein kinase DDB_G0292550 isoform X2	544	59.257	-1.054	6.833	Ser (18.199%); Asn (14.154%); Asp (8.272%)	MNIARSIIFSVYVILLVGLSPTSG
25743100	XP_004926879.1	apolipoprotein D	269	29.851	-0.199	6.508	Val (8.922%); Ser (8.550%); Leu (8.178%)	MWKLTVLGVFLTVTYVYS
25471600	XP_004926990.1	ATP-dependent (S)-NAD(P)H-hydrate dehydratase	327	35.913	-0.099	7.438	Ser (10.092%); Ile (9.786%); Gly (8.257%)	MTPSFNIILVILQVILFTFQITNG
25363500	XP_037876745.1	calsyntenin-1 isoform X1	953	105.096	-0.324	5.94	Ala (9.654%); Val (8.919%); Leu (8.185%)	MIVRFLSVLCVGVFLTAVYG
23648900	XP_004923234.1	putative alpha-L-fucosidase	477	55.387	-0.509	7.75	Leu (7.757%); Ser (7.128%); Gly (6.918%)	MKSFFLVFLFGFVNG
20809000	XP_004933848.1	15-hydroxyprostaglandin dehydrogenase [NAD(+)]	283	31.156	-0.108	6.252	Ile (8.834%); Ala (8.481%); Lys (7.774%)	MAVELTFLFCLVILSVRA
19455700	NP_001091826.1	ubiquitin and ribosomal protein S27a	155	17.847	-0.797	10.483	Lys (16.129%); Leu (7.742%); Gly (7.097%)	-
19071700	NP_001166287.1	sericin 2 isoform 1 precursor	1758	198.657	-2.152	9.158	Lys (17.179%); Ser (15.131%); Asp (11.832%)	MKIPYVLLFLVGVAVVNA
17735800	XP_004928346.2	uncharacterized protein LOC101735388	111	12.342	-0.008	5.119	Ile (9.910%); Asp (9.009%); Cys (8.108%)	MFSRLVVLGFVICVANA
17105900	XP_037867498.1	chaoptin isoform X1	1367	155.077	-0.135	6.335	Leu (15.435%); Ser (8.925%); Asn (7.901%)	MSLMAVVKFGYTLVAVTLLLLMIWASLSRA
16636200	P150	P150	4552	467.241	-0.672	3.815	Thr (27.307%); Ser (14.455%); Ala (12.017%)	MKVLCAIVLYIALMQPALC
16537500	XP_004921556.2	mucin-5AC isoform X1	856	92.849	-0.557	4.199	Glu (10.748%); Ser (10.748%); Thr (10.397%)	MIRSKGDMASPRFSTILLAAFHLFTVCWG
16068900	XP_004923220.1	endoplasmic reticulum resident protein 44	410	47.029	-0.391	6.032	Leu (10.000%); Lys (8.537%); Glu (7.561%)	MLKMNWKCPNLFNFKLSSCLFILLCHSFYNPTDS
14948100	XP_037872311.1	nardilysin	1146	132.148	-0.49	6.333	Leu (9.686%); Glu (8.551%); Ser (8.115%)	-
14014160	XP_004932517.1	extracellular serine/threonine protein CG31145 isoform X1	579	66.067	-0.285	7.713	Leu (10.363%); Arg (8.117%); Ala (7.772%)	-
12870000	NP_001243928.1	heat shock protein 68	628	69.367	-0.436	5.674	Ala (9.713%); Asp (7.643%); Leu (7.643%)	-
12693000	XP_037869639.1	alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase isoform X1	522	60.708	-0.538	8.814	Leu (8.238%); Asn (7.088%); Val (6.513%)	-
12590400	XP_021205679.2	mucin-12 isoform X1	3173	368.381	-1.498	4.137	Thr (18.374%); Glu (16.451%); Gln (15.065%)	MTAKTCCLLWVIVLASILATVVT
11544300	XP_004930613.1	peroxidase isoform X1	799	90.376	-0.4	6.855	Leu (9.387%); Ser (8.385%); Thr (6.383%)	-
10500500	XP_004930766.1	lachesin isoform X1	370	41.237	-0.2	7.444	Ile (8.378%); Ala (8.108%); Val (7.297%)	MDRNIYVYVFYTLLLYFTCNVSA
10166180	XP_012544866.1	zinc finger CCCH domain-containing protein 13	1004	117.836	-1.437	8.884	Arg (13.247%); Pro (11.554%); Thr (9.661%)	MKRSRHRWLGFTIFVCLLAILASELPNTAA
9468900	XP_004925158.1	mannosyl-oligosaccharide glucosidase	807	91.5	-0.417	9.002	Leu (10.905%); Gly (8.302%); Ser (6.939%)	-
9318700	XP_037871889.1	chitinase-like protein EN03 isoform X1	455	50.569	-0.292	7.002	Leu (11.209%); Ala (8.132%); Gly (7.692%)	MKLFIALVGLLALAKA
8995100	NP_001037584.1	ferritin precursor	229	26.033	-0.359	7.28	Leu (11.790%); Ala (7.860%); Lys (7.860%)	MKVYALIVACLALGVLA
8964600	XP_037869963.1	low-density lipoprotein receptor-related protein 2 isoform X1	4625	512.561	-0.477	5.013	Asp (9.319%); Gly (7.892%); Cys (6.962%)	MRSAREAFCLLLLAAAAAA
7899300	XP_004925541.1	uncharacterized protein LOC100500745	110	12.313	0.187	9.144	Ile (10.909%); Leu (10.909%); Lys (8.182%)	MKFAPVLIFIFLSLVCVIQC
7344340	NP_001037486.1	low molecular 30 kDa lipoprotein PBMHP-6 precursor	256	29.734	-0.444	6.523	Lys (9.375%); Leu (7.812%); Glu (7.422%)	MRLTLFAFVLAVCALASNA
6479200	NP_001153666.1	histone H2A-like protein 2	124	13.377	-0.318	11.361	Ala (12.097%); Leu (12.097%); Gly (11.290%)	-
6435200	NP_001243674.1	thrombospondin type-1 domain-containing protein 4-like precursor	574	62.763	-0.464	8.876	Arg (9.059%); Gly (8.885%); Ser (8.885%)	MAGADLYLLFVIFVTMVVG
5897400	XP_004929845.2	peroxidasin isoform X1	1379	157.181	-0.35	5.959	Leu (10.007%); Asp (6.744%); Glu (6.526%)	MRLCNFFKFLLIFNLFSLILS
4868770	XP_037871221.1	fibrohexamerin-like	235	26.815	-0.281	4.483	Leu (10.638%); Ser (7.660%); Asp (7.234%)	MKYKLCLFLSSLLNLIAA
4099550	NP_001037351.1	cathepsin D precursor	384	41.623	0.154	6.732	Leu (9.896%); Ala (9.375%); Gly (9.115%)	MGKISLFFLALIASSVMA
3173900	XP_012549547.1	malate dehydrogenase isoform X1	645	70.89	-0.186	6.641	Leu (9.457%); Ala (8.992%); Gly (8.992%)	-
3103050	XP_012551451.2	uncharacterized protein LOC105842505	1097	122.723	-0.805	6.678	Pro (12.124%); Thr (8.204%); Ser (7.840%)	MVRGATLALTLALAWVCVVRT
2356030	NP_001037057.1	BCP inhibitor precursor	105	12.249	-0.735	6.519	Glu (9.524%); Leu (9.524%); Ala (8.571%)	MNFVSVALLIATVVMASSA
1717200	XP_004925279.1	uncharacterized protein LOC101741292	235	27.212	-0.601	8.743	Arg (10.213%); Leu (9.787%); Ser (8.085%)	MGARGCISCLLLLTCGQIVRS

Supplementary Text S1. The full sequence of *B. mori* P150 protein. The sequence is divided into 5 regions: N-terminus (orange), Repeat 1 (green), linker (magenta), Repeat 2 (grey), and C-terminus (blue). Sequences of the N-terminal signal peptide and C-terminal conserved motif (CXCXCX) are underlined. Triangles in the schematic figure below indicate the variants of Repeat 1 (R1) and Repeat 2 (R2), which are designated as R1A, R1B, R2A and R2B.

[N-Terminus]-

15[R1]-1[R1B]-4[R1]-1[R1B]-8[R1]-1[R1B]-5[R1]-1[R1B]-8[R1]-1[R1A]-

[Linker]-

11[R2]-1[R2A]-2[R2]-1[R2A]-15[R2]-1[R2B]-5[R2]-1[R2B]-2[R2]-1[R2B]-5[R2]-1[R2B]-1[R2]-1[R2B]-1[R2]-1[R2B]-2[R2]-1[R2B]-4[R2]-1[R2B]-15[R2]-

[C-Terminus]

>BmP150

MKVLCAIVLYIALMQPALCDPPFAKKSNEHHTLDHLLNKGPENGNRYRAPSSFFENSAINKHFQNYFVNQQGRVQ
PVQPTRSARHLNAKPFRAAENKHNLHANARIPVRHPSIQPKDTQSETNHAKNSNLPTRVSLNLDDVTTESHINEN
TVPKQPTEGNGQIDSVTNTIDPIIKKPNEVNGTDKKPEIQWTTNTHEQTTSAVGKGETSSNNQLKLNKDTVLKGH
YIVRPATTNALDKTQNVEPLTTDSAYQIPLLPPSTQPADSLLSKESNQEVLEKIIEEVIKDNKYENSDDTNPVKF
IYKEELVPNAAIENQENVTTTTDLKGFLKNIKEIKHSEEKIENFKENKTQEWKSKKNMTIKKAESITQHDVSKKT
NEYTIEKRNETREESDFESNKQLSKYTEESQFELNTRISLTENSDEFAALNEYLEEVKKIENQNKAYEKVEQKSL
QAEKFDDLEFWRQEAEKEEKKREKEIREMQSKLNGGRHTEMTEKELVSKAEELVENDEDEFWNQEAIYLDNNYEN
SKTLNKTNSSISSTTNPKQPITGANIAKTTATEQSTTEAKVLSTTELKQSTTATSVPSTVASEQSTSEISVPSTT
GTEQPTTETSVPSTTATEQSRTDIKLPSTTSTVQSTTETNVPSTSVTEKSTTETSVSSTTKLKQSTTETNASTPT
ATEQSTTETSVPSTIELKQSTTETNVPSTTATEQSTTETSVPSTIELKQSTTETNVPSTTATEQSTTETSVPSTI
ELKQSTTETNVPSTSATEKSTTETSVPSTIELKQSTTETNVPSTSAIEKSTTDTSVPSTTELKQSTTETNASTPT
ATEQSTTETSVPSTIELKQSTTETNVPSTTATEQSTTETSVPSTIELKQSTTETNVPSTSVTEKSTTETSVSSTT
KLKQSTTETNASTPTATEQSTTETSVPSTIELKQSTTETNVPSTTATEQSTTETSVPSTIELKQSTTETNVPSTS
AIEKSTTETSVPSTTELKQSTTETNASTPTATEQSTTETSVPSTIELKQSTTETNVPSTSVTEKSTTETSVSSTT
ELKQSTTETNVPSTTATEQSTTETSVPSTIELKQSTTETSVPSTIELKQSTTETNVPSTSATEQSTTETSVPSTI
ELKQSTTETNVPSTSVTEKSTTETSVSSTTELKQSTTETNASTPTATEQSTTETSVPSTIELKQSTTETNVPSTT
ATEQSTTETSVPSTIELKQSTTETSVPSTIELKQSTTETNVPSTSATEQSTTETSVPSTIELKQSTTETNVPSTS
VTEKSTTETSVSSTTELKQSTTETNASTPTATEQSTTETSVPSTTKLKQSTTETNASTPTATEQSTTETSVPSTI
ELKQSTTETNVASTTATEQSTTETSVPSTIELKQSTTETNVPSTSAIEKSTTDTSVPSTTELKQSTTETNASTPT
ATEQSTTETSVSSTTELKQSTTETNASTPTATEQSTTETSVPSTIELKQSTTETSVPSTIELKQSTTETNVPSTS
ATEQSTTETSVSSTTKLKQSTTETNASTPTATEQSTTETSVPSTIELKQSTTETNVASTTATEQSTTETSVPSTI
ELKQSTTETNVPSTTATEQSTTETSVPSTTELKQSTTETNASTPTATEQSTTETSVPSTIELKQSTTETSVPSTI
ELKQSTTETNVPSTSATEQSTTETSVPSTIELKQSTTETNVPSTSAIEKSTTETSVPSTTELKQSTTETNASTPT
ATEQSTTETSVPSTTKLKQSTTETNASTPTATEQSTTETSVSSTTKLKQSTTETNASTPTATEQSTTETSVPSTI

ELKQSTTETNVPSTTATEQSTTETSVPSTIELKQSTTETNVPSTSAIEKSTTETSVPSTTELKQSTTETNASTPT
ATEQSTTETSVPSTIELKQSTTETNVPSTTAIEQSTNETSVPSTTDDNVQPVTKEDVTEPTAAYIKVQSTTVTES
NTTGAAVQSTTATENTTTDAEAQSTTVTESNTTGAAVQSTTATESGTTDAEVRSTTVNESNTTGAAVQSTTATES
ATTDAEAQSTTVTESNTTGTAVRSTTATESAITDAEVQSTTVTESNTTGAAVQSSTATESATTDAEAQSTTVTES
NTTGTAVQSTTATESAITDAEVQSTTVTESNTTGAAVQSTTATESATTDAEVQSTTVTESNTTGAAVQSTTATES
ATTDAEAQSTTVTESNTTDAAVQSTTATESASTDAEVQSTTVTESNTTGTAVQSTTATESAITDAEVQSTTVTES
NTTGAAVQSTTATESATTDAESQSTTVTESSTAGAAVQSTTATDAEDQSTTITESNTTGAAVQSTTATESATTDA
EAQSTTVTESNTTGAAVQSTTATESATTDAESQSTTVTESSTAGAAVQSTTATDAEDQSTTVTESNTTGAAVQST
TATESATTDAEAQSTTVTESNTTGAAVQSTTATESAITDAEVQSTTVKESNSAGAAVQSSTATESATTDAEAQST
TVTESNTTGTAVQSTTATESAITDAEVQSTTVTESNTTGAAVQSTTATESATTDAEVQSTTVTESNTTGAAVQFT
TATESATTDAEAQSTTVTESNTTDAAVQSTTATESASTDAEVQSTTVTESNTTGAAVQSTTATESATTDAEAQST
TVTESNTTDAAVQSTTATESASTDAEVQSTTVTESNTTGAAVQSTTATESATTDAEAQSTTVTESNTTGAAVQST
TATESATTDAEAQSTTVTESNTTGAAVQSTTATESAITDAEVQSTTVKESNSAGAAVQSSTATESATTDAEAQST
TVTESNTTGTAVQSTTATESAITDAEVQSTTVTESNTTGAAVQSTTTTESATTDAEAQSTTVTESNTTGAAVQST
TATESATTDAATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTESNTTGAAVQSTTATESATTDA
EAQSTTVTESNTTGTAVQSTTATESAITDAEVQSTTVTESNTTGAAVQSTTATESATTDAEVQSTTVTESNTTGA
AVQSTTATESATTDAEAQSTTVTESNTTGAAVQSTTATESATTDAATTDAEAQSTTVTESSTAGAAVQSTTATES
ATTDAEAQSTTVTESNTTGAAVQSTTATESAITDAEAQSTTVTESNTTGAAVQSTTATESATTDAATTDAEAQST
TVTESSTAGAAVQSTTATESATTDSEAQSTTVTESNTTGAAVQSTTATESAITDAEAQSTTVTELSTAGAAVQST
TATENATTDAEAQSTTVTESNTTGAAVQSTTTTESATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQST
NVTESNTTGAAVQSTTATESATTDAATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTESNTTGA
AVQSTTATESATTDAATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTESNTTGAAVQSTTATES
ATTDAATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTESNTTGAAVQSTTATESAITDAEAQST
TVTESNTTGAAVQSTTATESATTDAATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTESNTTGA
AVQSTTATETAITDAEAQSTTVTESNTTGAAVQSTTAKKSTTADAEAQFTTVSESNTGRAAVQSTTATESATTDA
EAQSTTVTESNTTGAAVQSTTATESATTDAATTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTES
NTTGAAVQSTTAKKSTTADAEAQFTTVSESNTGRAAVQFTTATESATTDAEAQSTTVTESYTTGAAVQSTTATES
ATTDAEAQSTTVTESNTAGAAVQSTTTTESATTDAEAQSTTVTESSTAGAAAQSTTATDSATTDAEAQSTTVTES
SSAGAAVQSTTATESATTNAEAQSTTVTESSSAGAAVQSTTTTESATTDAEAQSTTVTESSTAGAAAQSTTATDS
ATADAEVQSTTVTELSTTGAAVPSTTATESSTTDAEAQSTTVTESSTAGAAVQSTTATESATTDAEAQSTTVTES
SSAGAAVQSTTTTESATTDAEAQSTTVTESSTAGAAAQSTTATDSATADAEVQSTTVTELSTTGAAVPSTTATKS
STTDAEAQSTTVTESSTAGVAVQSTSATERAITDTEVQSATVTESNTAAKVHDQPTRTVFTSPCGNHPTSVATDN
TSEKESKQTNTPTEHTVIVAHNQLHKTQTKTTASAHPILETVSGMPSIHLTSKNDPPTTDEPIEIWEQSTKQFAP
TTFISESLVLKNKQTQFKVIPLEIPTEATQKPVQSTIEPIHPKTKPEQSTIKTVQSTDQSTQATMGSFPSIVKPI
QSRESNVFNVKAISSIEPLKSTTEPVHSITQSTVVSKQSSNEPLQVTIELLQSSIEPTQSTVQTQQSSAQPGQFA
TKPVNRIETTTENPESKKSNEHRPFIQVTKSTKSPIGSTLKGYYREINPTQGFGDDVGENEDKRKGRYDLSKFNQ
APAMHMAVENEITLKTLIVEDGSKYGRCYCSCDLNSKPVFISLDGSALKPRR

>BmP150_N_Terminus

MKVLCAIVLYIALMQPALCDPPFAKKSNEHHTLDHLLNKGPENGNRYRAPSSFFENSAINKHFQNYFVNQQGRVQ
PVQPTRSARHLNAKPFRAAENKHNLHANARIPVRHPSIQPKDTQSETNHAKNSNLPTRVSLNLDDVTTESHINEN
TVPKQPTEGNGQIDSVTNTIDPIIKKPNEVNGTDKKPEIQWTTNTHEQTTSAVGKGETSSNNQLKLNKDTVLKGH
YIVRPATTNALDKTQNVEPLTTDSAYQIPLLPPSTQPADSLLSKESNQEVLEKIIEEVIKDNKYENSDDTNPVKF
IYKEELVPNAAIENQENVTTTTDLKGFLKNIKEIKHSEEKIENFKENKTQEWKSKKNMTIKKAESITQHDVSKKT
NEYTIEKRNETREESDFESNKQLSKYTEESQFELNTRISLTENSDEFAALNEYLEEVKKIENQNKAYEKVEQKSL
QAEKFDDLEFWRQEAEKEEKKREKEIREMQSKLNGGRHTEMTEKELVSKAEELVENDEDEFWNQEAIYLDNNYEN
SKTLNKTNSSISSTTNPKQPITGANIAKTTATEQSTTEAKVLSTTELKQSTTATSVPSTVASEQSTSEISVPSTT
GTEQPTTETSVPSTTATEQSRTDIKLPSTTSTVQS

>BmP150_Repeat1

TTETNVPSTSVTEKSTTETSVSSTTKLKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETNVPSTSATEKSTTETSVPSTIELKQS
TTETNVPSTSAIEKSTTDTSVPSTTELKQS

TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETNVPSTSVTEKSTTETSVSSTTKLKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETNVPSTSAIEKSTTETSVPSTTELKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTSVTEKSTTETSVSSTTELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETSVPSTIELKQS
TTETNVPSTSATEQSTTETSVPSTIELKQS
TTETNVPSTSVTEKSTTETSVSSTTELKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETSVPSTIELKQS
TTETNVPSTSATEQSTTETSVPSTIELKQS
TTETNVPSTSVTEKSTTETSVSSTTELKQS
TTETNASTPTATEQSTTETSVPSTTKLKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVASTTATEQSTTETSVPSTIELKQS
TTETNVPSTSAIEKSTTDTSVPSTTELKQS
TTETNASTPTATEQSTTETSVSSTTELKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETSVPSTIELKQS
TTETNVPSTSATEQSTTETSVSSTTKLKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVASTTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTTELKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETSVPSTIELKQS
TTETNVPSTSATEQSTTETSVPSTIELKQS
TTETNVPSTSAIEKSTTETSVPSTTELKQS
TTETNASTPTATEQSTTETSVPSTTKLKQS
TTETNASTPTATEQSTTETSVSSTTKLKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTTATEQSTTETSVPSTIELKQS
TTETNVPSTSAIEKSTTETSVPSTTELKQS
TTETNASTPTATEQSTTETSVPSTIELKQS
TTETNVPSTTAIEQS

>BmP150_Linker

TNETSVPSTTDDNVQPVTKEDVTEPTAAYIKVQS

>BmP150_Repeat2

TTVTESNTTGAAVQSTTATENTTTDAEAQS
TTVTESNTTGAAVQSTTATESGTTDAEVRS
TTVNESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGTAVRSTTATESAITDAEVQS
TTVTESNTTGAAVQSSTATESATTDAEAQS
TTVTESNTTGTAVQSTTATESAITDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTDAAVQSTTATESASTDAEVQS
TTVTESNTTGTAVQSTTATESAITDAEVQS
TTVTESNTTGAAVQSTTATESATTDAESQS
TTVTESSTAGAAVQSTTATDAEDQS

TTITESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAESQS
TTVTESSTAGAAVQSTTATDAEDQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESAITDAEVQS
TTVKESNSAGAAVQSSTATESATTDAEAQS
TTVTESNTTGTAVQSTTATESAITDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEVQS
TTVTESNTTGAAVQFTTATESATTDAEAQS
TTVTESNTTDAAVQSTTATESASTDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTDAAVQSTTATESASTDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESAITDAEVQS
TTVKESNSAGAAVQSSTATESATTDAEAQS
TTVTESNTTGTAVQSTTATESAITDAEVQS
TTVTESNTTGAAVQSTTTTESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGTAVQSTTATESAITDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEVQS
TTVTESNTTGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESAITDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDSEAQS
TTVTESNTTGAAVQSTTATESAITDAEAQS
TTVTELSTAGAAVQSTTATENATTDAEAQS
TTVTESNTTGAAVQSTTTTESATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TNVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESAITDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATETAITDAEAQS
TTVTESNTTGAAVQSTTAKKSTTADAEAQF
TTVSESNTGRAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTATESATTDAATTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESNTTGAAVQSTTAKKSTTADAEAQF
TTVSESNTGRAAVQFTTATESATTDAEAQS
TTVTESYTTGAAVQSTTATESATTDAEAQS
TTVTESNTAGAAVQSTTTTESATTDAEAQS
TTVTESSTAGAAAQSTTATDSATTDAEAQS
TTVTESSSAGAAVQSTTATESATTNAEAQS
TTVTESSSAGAAVQSTTTTESATTDAEAQS
TTVTESSTAGAAAQSTTATDSATADAEVQS
TTVTELSTTGAAVPSTTATESSTTDAEAQS
TTVTESSTAGAAVQSTTATESATTDAEAQS
TTVTESSSAGAAVQSTTTTESATTDAEAQS
TTVTESSTAGAAAQSTTATDSATADAEVQS
TTVTELSTTGAAVPSTTATKSSTTDAEAQS

TTVTESSTAGVAVQSTSATERAITDTEVQS

>BmP150_C_Terminus

ATVTESNTAAKVHDQPTRTVFTSPCGNHPTSVATDNTSEKESKQTNTPTEHTVIVAHNQLHKTQTKTTASAHPIL
ETVSGMPSIHLTSKNDPPTTDEPIEIWEQSTKQFAPTTFISESLVLKNKQTQFKVIPLEIPTEATQKPVQSTIEP
IHPKTKPEQSTIKTVQSTDQSTQATMGSFPSIVKPIQSRESNVFNVKAISSIEPLKSTTEPVHSITQSTVVSKQS
SNEPLQVTIELLQSSIEPTQSTVQTQQSSAQPGQFATKPVNRIETTTENPESKKSNEHRPFIQVTKSTKSPIGST
LKGYYREINPTQGFGDDVGENEDKRKGRYDLSKFNQAPAMHMAVENEITLKTLIVEDGSKYGRCYCSCDLNSKPV
FISLDGSALKPRR

Supplementary Text S2. Sequences of P150 and Mucin-12 homologs used for the alignment and dendrogram construction in **Figure 6**.

>P150

GTAACAAAATCAACAAAGAGTCCGATCGGGAGCACCCTAAAAGGATATTACCGAGAAATCAACCCTACCCAAGGC
TTCGGTGATGATGTCGGGGAAAATGAAGATAAACGCAAAGGAAGATATGATTTATCTAAATTTAATCAAGCTCCG
GCAATGCATATGGCCGTTGAAAACGAAATAACATTGAAGACTCTGATCGTCGAAGACGGATCCAAGTATGGACGC
TGCTACTGTTCGTGTGACCTGAATAGCAAGCCAGTCTTCATTAGCTTGGATGGAAGTGCTTTGAAACCTAGACGA
TGA

>XP_021205679.2

CATGCAGATGGATCTAGCACTGAACCTAATAGGACAGTGCCTTGCAATAAGCCGCCTGCAGGACCTGGGCAGAAG
GAGCCGGCGACTGACTTGCAGCCTACCCCGCCTGGTACTGGAGGCAAATTGGCCTTGAATCAAGGCAAGAAAGAA
ATAATGCAAGCTTTGGAAACCAGCATGAATCTGAAGGTTTTAATTCTGCCAAATGGTCCTCAAGGGCAGAACTGT
TTCTGTACTTGTGACAACACGAGCGCACCTAAACTTATTCCTTTAAAAGATTTGCCGGCTGGTATCACTGGCAGT
TGA

>BRAKERZPYP00000020335.1

GAAAGCAGTGAAACTATTAAGACCGCACAAGTTAATCAGTTCCAACCACCTAAAAAACCAGGACATAATGATATT
GAGAGTCATCAGGCATCACAAGATTATAACGAAACAGAAAATAGAAGAGGACGATATGATCTTTCAAAGAAAAAT
TCTAGACAAGAAATCCGACAGGCTATTGAAAGTGATGTTAACTTCAAGACACTTATCATCGATGATGGTTCCAAG
AATGGACGATGTTACTGTTCGTGTGATATCAACAACAGACCAGTGTTTCTTAATGTGGATCAAAATAAACCTTTT
TAG

>BRAKERZPYP00000016428.1

GCGACCCCGCCGCTACAACCAGAAAGACCAGGGCAGTCTGCTTCAGGGCCACAACCAGTCGCTCCAAGCAACGAA
CTCGAGAGTGACGGCAAAAAAGGAGGAAAATTAGCATTGCAACAAGGCCGCAAAGATGTAATGCATGCATTGGCA
ACCAACATGAACCTGAAAATTCTCATACTGCCGAACGGACCAAATAAGAAAAACTGCTTCTGCACATGCGAGAAT
GACAGCCCTCCCAAACTGATACCAGTGGAGGACTTACCACCACAAATCAGAGGAATTAGAGCAAAACCCTTAGCC
TAA

>XP_049868194.1

GTAGACAAACCAGGTCAAGACACGCCGGGACAAGATCAACCAGTCCAAGATAAGCCAGGAAATGATAAACCTGGA
CGGCAACCAGATCAATGCAGACCACAAAGACCAGGCCAGAATTCTCGCAGAAAGGGTCGCTATGATTTAAACAAG
GGCAAACCAGCTCTTCAAAGGGCCCTGGAGAGCCACATTGAGTTTCAGACCCTGATAGTTTACGACAGTAAAGGA
GGAGAAGGAAAATGCTACTGCTCTTGCGGCAGTAACGGCAGACCTAAATTTATTAGTGTTGATCCTAGCATCCTA
TAG

>XP_049884459.1

ACAGAAAATACGGAAAGTGAAACGCAATGGAGCAAGGACAGCAAGACTCCAAGACCTGGTCGTGGAACTCCGCCA
ACGCGACAACCGCAGCCACTGCCAAAGAAGGGAGGGAAACTCAACATCCAACATGGGAGACATGAAATCATTCGC
GGTGTTGAAACCACAATGAAGATGAAGTTCCTGGTGGTTCCTCATGGTGAGAAGGGCAAGAACTGCTTCTGTTCG
TGCGACAGCACTGCTGCTCCCAAACTGGTACCGCTTGCAGACCTCCAAGAGCTAGACTCCTTTGGGAAAGAAAAC
TAA

>ENSTAIP00005009985.1

CCTGCTGAAAGTGAAGCCACATTGAACCGCAGCAATAAAGATAAGAAACAAAATGTTTACACTTCGGATAATCGT
GATACTAAATCTATATTTGCAAAGAGAGAAGAGGAACACAAAACAGGGCGTTATGATTTCAAACAAATGGATGGA
GCCAAAGCGATCCACAACGCGATACAGATGGATTTCAACTATAAGATGTTGTTGGTGGAAGACGGCTCTGCCCCA
GGGAAGTGCTACTGCACCTGTGATATGAACAGCCGACCTGTCTTTATGACGTTAGATTCTGAGTCCATACCTTCC
TAG

>ENSTAIP00005010080.1

ACAACACAAACTAGCTGGACCAATGAAGTAAGCAATAATCATGATAGCTCAGAAGAATCCTCATTCGAACAAACT
ACTGATAAACAGGATAATTCAGAACCAATCACAACGAATCAAACTACAGATAAGGTAGCAGCAACCACACAAACA
AACATAATGCACGCGCTTGAGTCCAATATGAATCTGAAGGTGCTTATAATTCCCAACGAGGCCACCGGTGGGAAA
TGCTTCTGCACTTGCGCCGCAGACACGCCCAAGCTGATTCCTCTCGAGGATCTGCCAAAAGAACTCACGCAGGAT
TGA

>BRAKERPYPP00000015107.1

GTTGGCGATGCTGTTGGAGGAATCGCAGCTGGCGTTGGCGGGATCGTTGGAGGAGTTGAAGCTGGTGTTGGTGAT
GTCGTTGAAGAAGTTGGGTCTGGTATTGCCGGAATCACCGGAGATAACAAACCAGAAAACAAACCAGGTCTTCAA
CAAGCCATCGCTGCTGATTTGGACCTTCAAATGATGACGGTGGCAGATGGGACTCCCAGAGGGAAGTGCTACTGC
TCTTGCAGTCTGAACAATCAGCCGGTATTCTTGGAGATAAATGAAGCAGCTCTGCCACGTTCCCTAGCGCTTACA
TAG

>XP_032528373.1

TTAGACCCAAAAGACAAACCTGAAGCCTCTGGACCAAAAGACGAAGTACAATCTACCGAGTCTCAGTCTACACCT
GCGCCACCCAGCCTAAAACCTAACCAGGACCCAAGTACTTCTGGACCCCAGGATCTTTCAACAGACCCGCAAGCT
CTCCAAGTGTCGATGAATCTGAAACTCCTCATCCTACCCAAGGGACCAAAAAACACAAAGAATTGCTACTGCGCC
TGCGAGGGCCAGGAACCGCCCAAACTGGTCCCGATGCGACAACTTCCGGTCGGGATTACTAGTGTAGATAAAAAT
TAA

>XP_053613123.1

GCCCAAAGCAACACCTACCTATATGGACAACCAGAACAACCTGGTTACCCAGGACCAGGGCAATCAGGCTCAAGC
TCAGGTTCTTCCTCACAAAGCCAAAGCAATGCTCAGTCCAGCTCAACAGACGTCATCGACAGACGAGGCCGTTAC
GATCTGAACAAACCTAGAGGTATTGAAACTGACATGAACCTGAATACACTATTTATCAACGACGGGACGAAGGAT
GGAAGGTGTTATTGCTCATGTGGAGTCAACCGCCGCCCTATCTTCTTGAGCGTCGAGCCCAACCAACTTGGACTT
TAA

>XP_053622877.1

AGTGAATCAGAATCTCGTGCAGAATCTGAATCGAGAAGAGATTCTGAATTGGAGTCCGAAAAGGAAGTAGGCGTT
GCTGTGCCGCCGACCAAACCTGGAGTGGGAGGCAAGCTGGCTATTCAACAAGGAAAACGCGAAATCATGCACGCC
TTGGAAACTAACACGAACTTGAAAATCCTGATCTTGCCAAACGGACCCAAGCAAAAGAACTGCTTCTGCACCTGT
CAAAGTGACAGCGCGCCAAAACTTGTACCAATGAAAGACTTGCCGGCCGAAATCACCGCCATCGAGTCAAAATCG
TAA

>BRAKERSAFP00000015049.1

CCTGGCGAACCAGAGCAACCAGGTGAACCAGGTGAAACAGCAGAGCCAGGCCAGCCGGGTGAACCAGAAGAATAT
GATGATGATGAAGAGGACGACGGACACTACAGAAGAAGGGGACGTTACAACTTGAGACGTAGGAAAGGACACCGT
GCCCTCGAAACTGACCTCCATTTCAAGACTTTGATCGTGCAAGACGGATCAAAGACCGGAAGTTGCTTCTGTTCC
TGTGACGTCAACAACCGTCCAGTATTCCTCACCGTAGACCCCAAGTACTTGCCCGGACTTCGCGCCCCACTGAAC
TAA

>BRAKERSAFP00000011847.1

TCTGGTCTCCCTAGACCAGAACACCCAGGTTCAGGACCACAGCCATCTGTTCCGAACAGACCTTTGCCAGATGAC
TATTTAATACATGACAAGAGAGGCGGCAAGTTAGCTTTGCATCATGGTCGACGAAAAATAATGAATGCACTGGAA
ACCAACATGAATCTCAAAATATTGATTCTGCCCAACACTCCCGATCGCAAGAACTGCTTCTGTACCTGTGAAAAC
AACGACCCGCCAAAACTCATACCGCTTCGCGATTTGCCGCCGGAAGTACGAGAAATCGGCGCGAAGGAATCAGAC
TAG

>BRAKERPFCP00005001188.1

GACAGCCAGGCCAGCCAGAACAACCAAATCAGCCAGAACAGCCAGATCAGCCAGGCCAGCCAGGACAAGGAGGCC
ATTCAGGGCAATCAGGACAGCCAGGACCACCAGGGACTGACAAAATATTTTGTAGGAGGTTTGGAGTTAAAACTT
GACTACGATGATCGTGCGATTGAGGCCGACCTCAACCTGCGGACCATGATCATCCCGAGCCGCTACGGCAACCAG
CCGGGCCGCTGCTACTGCTCCTGCGACCTCAACAACAAGCCCGTCTTCATCAGCGTCGACCCCAGGCGGCTCTCG
TGA

>BRAKERPFCP00005013050.1

GAAGAGGAGGAAGAGGTTACAGAAGGGTTCGCGTCAGGCAAGAAACCAGCGGCTCCTGGTACGCCAGTTGCACCG
ACGACGCCAGACGATGGCAAGAAGGGAGGAAAGCTAGCGGTCCAGAAAGGAACTAAAGAAATACTGCACGCCCTG
GAAACCAGCATGAACCTCAAGGTCCTCATCCTGCCCAACGGCCCTAACGGCAAGAACTGCTTCTGCACTTGCGAC
AACAACGCTGCGCCTAAGATCATCCCTCTGTCTGCTCTGCCAGCACAGGTCTCTGAGATTGCCGCAGAAAAAAAC
TAA

>Tibi_1

CGCCCAGGAGATGCAGGACGACCGGGAAGTCCAGGTCAACCAGAAGATTCGGGCAAGCCTTGTGATCGCCCAGGA
AGACCAGGCCCCAATACACCTGGGGGTCGTAAAGATGTCGTACGTGCAATTGAATCGCGAGTCCTATTTAAGACA
ATTATGATTCCCGACGAAGACGGCGGCGGCAAATGTTTCTGTACTTGTGAAAGTAAAACTACACCGAAAATTATT
TGGGTCGACGAAAACGGTCCCCGAGTTGAATACGTCCATAATAAAAATGGTGACCAAAACGAGAAAGGAATTAAA
TAA

>Euja_1

TCAGAATTCGATAAAATCCATAAATCAAAAATAAAAAAGGAAACTGATATAGAGGAAGATGATGAGACAAGCGAA
TATGAAGACGATCTATATGAAAGCGAATTTGATAAAAAACACAAGCGGCGACGTCATCCCAAGCATCACAAGCCA
GATGAAAAAATTAGAGCTATGGAGGCCGAAGTAGATTACAAGTTGGTGATCGTCCAGAACGGGCCGGATAAAGGC
CACTGCTACTGCACGTGCACGACCAAAACCAGTCCGCACTTCGTCGTCTGGGACGACAAGAAAAAGAAGAATCCT
TAA

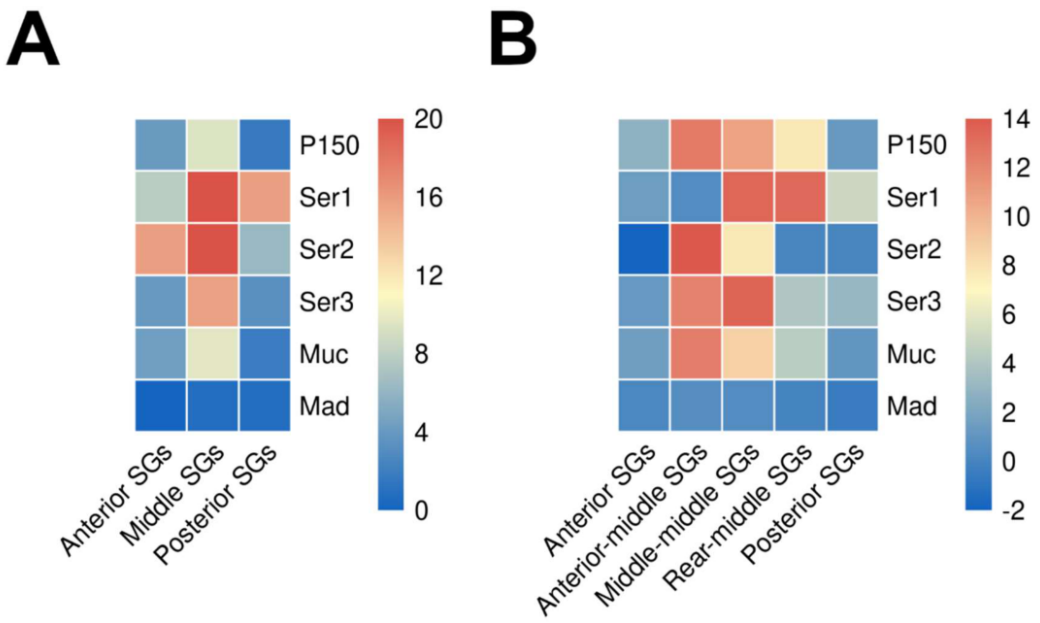
>Euja_2

GGCACAACGAAACAACCAGGTGGTGGGGTTAAACCTGGGGATTCTGGAAAACCGCAAGTATCGGGCAAGACAGAA
CAACCGAAAAAACCGAATGATGGCAAAATAGGTGCTTCTACCGGACCTAGTAAAACAGGCACACCAAGCAAACCT
CGTCCCCGCAAGCGGACTGGTATTCCGGAAATCTATCGCGGTTTCAAAATGGATGTCAAATTCAAGACGGTGATG
GTGACGAACGGATTGCAGGAGGACAAGTGCTTCTGCACATGCTCGCCGCTCGATTCGCCGACGTGGATTAAGCGT
TGA

>Nesw_1

CCGCACCGCGCCCGCACCACGCGACTGCGAGTGAGTGCACTGTGTTACAAGGGGCGCGTGTGTGTTGCAGGGTCG
CCGCGCGGCGTCGAGACGGACGTCACCTTCAAGACGATCATAGTGCAGAACGGACCCGACCGCAAGTCCTGCTAC
TGCTCGTGCACCAGCAACTCGCGGCCCAAACTCATCGGCTGG

Supplementary Figure S1. Expression levels of *B. mori* P150 gene along with major silk genes (Sericin 1-3 and Mucin-12) and the control gene (mothers against dpp; Mad) in different parts of silk glands. Heat maps represent the log2 fold-change values. As shown in both figures, P150 is highly expressed in middle silk glands. Sources of the raw RNA-seq data were listed in **Supplementary Table S1**.



Supplementary Table S1. Sources of the RNA-seq datasets utilized to verify the expression of *B. mori* P150 and other genes.

(A) datasets used in **Supplementary Figure S1A**

Run	BioProject	BioSample	Age	Instrument	LibraryLayout	LibrarySource	Organism	Tissue
SRR10035619	PRJNA559726	SAMN12608940	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior-Silk-Gland
SRR10035620	PRJNA559726	SAMN12608939	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior-Silk-Gland
SRR10035621	PRJNA559726	SAMN12608938	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior-Silk-Gland
SRR10035669	PRJNA559726	SAMN12609122	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Middle-Silk-Gland
SRR10035670	PRJNA559726	SAMN12609121	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Middle-Silk-Gland
SRR10035671	PRJNA559726	SAMN12609120	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Middle-Silk-Gland
SRR10035726	PRJNA559726	SAMN12609071	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Midgut
SRR10035727	PRJNA559726	SAMN12609070	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Midgut
SRR10035728	PRJNA559726	SAMN12609069	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Midgut
SRR10035756	PRJNA559726	SAMN12609044	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior-Silk-Gland
SRR10035757	PRJNA559726	SAMN12609043	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior-Silk-Gland
SRR10035758	PRJNA559726	SAMN12609042	5th instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior-Silk-Gland

(B) datasets used in **Supplementary Figure S1B**

Run	BioProject	BioSample	Age	Instrument	LibraryLayout	LibrarySource	Organism	Tissue
DRR186474	PRJDB8614	SAMD00180408	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior silk gland
DRR186475	PRJDB8614	SAMD00180409	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior silk gland
DRR186476	PRJDB8614	SAMD00180410	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior silk gland
DRR186477	PRJDB8614	SAMD00180411	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior part of the middle silk gland
DRR186478	PRJDB8614	SAMD00180412	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior part of the middle silk gland
DRR186479	PRJDB8614	SAMD00180413	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Anterior part of the middle silk gland
DRR186480	PRJDB8614	SAMD00180414	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Middle part of the middle silk gland
DRR186481	PRJDB8614	SAMD00180415	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Middle part of the middle silk gland
DRR186482	PRJDB8614	SAMD00180416	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Middle part of the middle silk gland
DRR186483	PRJDB8614	SAMD00180417	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior part of the middle silk gland
DRR186484	PRJDB8614	SAMD00180418	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior part of the middle silk gland
DRR186485	PRJDB8614	SAMD00180419	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior part of the middle silk gland
DRR186486	PRJDB8614	SAMD00180420	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior silk gland
DRR186487	PRJDB8614	SAMD00180421	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior silk gland
DRR186488	PRJDB8614	SAMD00180422	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Posterior silk gland
DRR186492	PRJDB8614	SAMD00180426	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Midgut
DRR186493	PRJDB8614	SAMD00180427	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Midgut
DRR186494	PRJDB8614	SAMD00180428	fifth instar day 3	Illumina NovaSeq 6000	PAIRED	TRANSCRIPTOMIC	Bombyx mori	Midgut

Supplementary Table S2. Sequences of primers used in (A) exon junction sequencing and (B) qPCR analysis.

(A) Sequencing primers	
P150 #1F	GAATCGAGCACTGCAGGAGT
P150 #1R	AGTAGGCTGATCGTGGACCT
P150 #2F	CCACGAAACAGTTTGCTCCA
P150 #2R	TAGGCTCGATTGTGGACTGC
P150 #3F	CCCAAGGCTTCGGTGATGAT
P150 #3R	TATGCATTGCCGGAGCTTGA

(B) qPCR primers	
P150 F	CCCAAGGCTTCGGTGATGAT
P150 R	TATGCATTGCCGGAGCTTGA
Ser1 F	CACAACCGATAAGACGAG
Ser1 R	GACGAAGTGGAGGAAGC
Ser2 F	CATCGGCTGACTACCA
Ser2 R	AGAGTTGCTGCCCTTAC
Ser3 F	TGTCTCGTCGGTGGAA
Ser3 R	TTGTTGTATGACTGGCTCT
Mad F	ACACAAGGCGTCACATAGGG
Mad R	TGGTGATTGCAGTTACGGCT
EF1a F	CAAGTCTGGAGATGCAGCCA
EF1a R	GGGGTGGGAATTCCTGGAAG

Supplementary Table S3. Statistical analysis of quantitative PCR analysis of gene expression of genes in different tissues. Statistical differences were evaluated by Student´s t-test. ASG: anterior silk glands; MSG: middle silk glands; PSG: posterior silk glands; OVA: ovaries. Backgound colors indicate statistical significance: orange (p < 0.001); yellow (p < 0.01); blue (p < 0.05); blank (p >=0.05).

	ASG					MSG					PSG					OVA					GUT				
ASG																									
MSG	6.89E-03	9.38E-05	5.64E-02	3.66E-01	8.29E-01																				
PSG	1.16E-01	3.66E-02	8.92E-01	8.28E-02	8.81E-01	1.14E-04	1.98E-02	1.07E-03	9.82E-02	8.89E-01															
OVA	3.89E-01	4.14E-02	1.84E-01	4.36E-01	1.22E-01	2.57E-06	3.05E-04	1.37E-07	2.23E-01	1.99E-03	2.64E-02	2.54E-02	7.48E-03	6.18E-03	9.12E-03										
GUT	2.28E-01	1.06E-02	1.50E-01	1.59E-02	4.47E-01	4.10E-05	1.20E-04	3.08E-08	2.81E-02	1.96E-01	1.13E-01	1.80E-02	4.87E-03	1.10E-01	2.42E-01	1.40E-01	7.94E-02	4.12E-03	4.05E-02	2.71E-03					
	P150	**Ser1**	**Ser2**	**Ser3**	**Muc**	**P150**	**Ser1**	**Ser2**	**Ser3**	**Muc**	**P150**	**Ser1**	**Ser2**	**Ser3**	**Muc**	**P150**	**Ser1**	**Ser2**	**Ser3**	**Muc**	**P150**	**Ser1**	**Ser2**	**Ser3**	**Muc**

Supplementary Table S4. Properties of 2 types of repeat sequences in *B. mori* P150 protein. GRAVY: grand average of hydropathy; M. w.: molecular weight; pI: isoelectric point.

	GRAVY	Residue	M. w. (Da)	pI	Top1 AA	Top2 AA	Top3 AA	Top4 AA	Top5 AA
Repeat 1	-0.8512	1275	133254.12	3.8738	Thr: 417 (32.706%)	Ser: 237 (18.588%)	Glu: 162 (12.706%)	Pro: 75 (5.882%)	Val: 75 (5.882%)
Repeat 2	-0.4541	2220	215927.31	2.6857	Thr: 696 (31.351%)	Ala: 440 (19.820%)	Ser: 320 (14.414%)	Glu: 211 (9.505%)	Val: 161 (7.252%)

Supplementary Table S5. Properties of *B. mori* sericins and mucin-12 proteins. GRAVY: grand average of hydropathy; M. w.: molecular weight; pI: isoelectric point. The conserved cystein residues at the C-terminal region were labeled in red.

Gene	Accession	GRAVY	Residue	M. w. (Da)	pI	Top1 AA	Top2 AA	Top3 AA	Top4 AA	Top5 AA	C-terminus
Sericin 1	XP_037869538.1	-1.1181	3385	331903.34	4.7726	Ser: 1300 (38.405%)	Gly: 428 (12.644%)	Thr: 317 (9.365%)	Asn: 270 (7.976%)	Asp: 221 (6.529%)	LLHKPGQGKI**C**L**C**FENIFDIPYHLRKNIGV
Sericin 2	NP_001166287.1	-2.1516	1758	198657.39	9.1575	Lys: 302 (17.179%)	Ser: 266 (15.131%)	Asp: 208 (11.832%)	Glu: 195 (11.092%)	Asn: 123 (6.997%)	SSSSSSSSSSSSSRSSSSSTYTGSHDDSSEE
Sericin 3	NP_001108116.1	-1.4651	1271	123298.08	5.7076	Ser: 554 (43.588%)	Gly: 152 (11.959%)	Gln: 94 (7.396%)	Asn: 85 (6.688%)	Lys: 79 (6.216%)	QSKASSSFSASSASESSSLSSDVNFEEKTD
Sericin 4	BGIBMGA011896	-0.9876	2270	251440.57	6.7156	Lys: 223 (9.824%)	Thr: 218 (9.604%)	Ser: 215 (9.471%)	Glu: 205 (9.031%)	Gly: 169 (7.445%)	IHGGENTEAGGHSPGLVGGLFKYLLGGKTQ
Sericin 5	KWMTBOMO06271	-2.0155	1976	218868.3	9.6116	Lys: 399 (20.192%)	Asp: 254 (12.854%)	Glu: 211 (10.678%)	Ser: 203 (10.273%)	Gly: 202 (10.223%)	SESASSSRSSYHSSSMQRNQKTSFDDDQNE
Sericin 6 / P150	Sericin 6 / P150	-0.6718	4552	467241.21	3.8151	Thr: 1243 (27.307%)	Ser: 658 (14.455%)	Ala: 547 (12.017%)	Glu: 479 (10.523%)	Val: 297 (6.525%)	SKYGR**C**Y**C**S**C**DLNSKPVFISLDGSALKPRR
Mucin-12	XP_021205679.2	-1.4977	3173	368381.36	4.137	Thr: 583 (18.374%)	Glu: 522 (16.451%)	Gln: 478 (15.065%)	Pro: 218 (6.870%)	Leu: 202 (6.366%)	QGQN**C**F**C**T**C**DNTSAPKLIPLKDLPAGITGS

Supplementary Table S6. Properties of P150 proteins in *B. mori*, *E. kuehniella* and *G. mellonella*. GRAVY: grand average of hydropathy; M. w.: molecular weight; pI: isoelectric point.

Accession	Species	Family	Superfamily	GRAVY	Residue	M. w. (Da)	pI	Top1 AA	Top2 AA	Top3 AA	Top4 AA	Top5 AA
P150	*Bombyx mori*	Bombycidae	Bombycoidea	-0.6718	4552	467241.21	3.8151	Thr: 1243 (27.307%)	Ser: 658 (14.455%)	Ala: 547 (12.017%)	Glu: 479 (10.523%)	Val: 297 (6.525%)
WDD44665.1	*Ephestia kuehniella*	Pyralidae	Pyraloidea	-0.8492	1741	173261.02	4.212	Ser: 603 (34.635%)	Gln: 193 (11.086%)	Gly: 160 (9.190%)	Pro: 154 (8.845%)	Ile: 94 (5.399%)
XP_026763958.2	*Galleria mellonella*	Pyralidae	Pyraloidea	-1.2812	1465	160622.85	4.7732	Ser: 334 (22.799%)	Gln: 251 (17.133%)	Asn: 116 (7.918%)	Glu: 101 (6.894%)	Thr: 101 (6.894%)

Supplementary Table S7. Coding sequences of P150/ser6 and Mucin-12 homologs for alignment and tree construction.

Class	ID	Species	Superfamily
-	Nesw_1	*Nematopogon swammerdamellus*	Incurvarioidea
-	Euja_1	*Eumeta japonica*	Tineoidea
P150	XP_049868194.1	*Pectinophora gossypiella*	Gelechioidea
P150	XP_053613123.1	*Plodia interpunctella*	Pyraloidea
P150	BRAKERZPYP00000020335.1	*Zeuzera pyrina*	Cossoidea
P150	P150	*Bombyx mori*	Bombycoidea
P150	BRAKERPFCP00005001188.1	*Pammene fasciana*	Tortricoidea
P150	BRAKERSAFP00000015049.1	*Sesia apiformis*	Sesioidea
P150	BRAKERPYPP00000015107.1	*Protodeltote pygarga*	Noctuoidea
P150	ENSTAIP00005009985.1	*Thera britannica*	Geometroidea
-	Euja_2	*Eumeta japonica*	Tineoidea
-	Tibi_1	*Tineola bisselliella*	Tineoidea
Mucin	ENSTAIP00005010080.1	*Thera britannica*	Geometroidea
Mucin	XP_032528373.1	*Danaus plexippus plexippus*	Papilionoidea
Mucin	XP_049884459.1	*Pectinophora gossypiella*	Gelechioidea
Mucin	BRAKERPFCP00005013050.1	*Pammene fasciana*	Tortricoidea
Mucin	BRAKERSAFP00000011847.1	*Sesia apiformis*	Sesioidea
Mucin	BRAKERZPYP00000016428.1	*Zeuzera pyrina*	Cossoidea
Mucin	XP_053622877.1	*Plodia interpunctella*	Pyraloidea
Mucin	XP_021205679.2	*Bombyx mori*	Bombycoidea

Conclusions

This Ph.D. thesis expands the knowledge of silk components in lepidopteran species and provides a perspective on the origin of silk genes. In Chapter 1, we have conducted a thorough study focused on the cocoon silk components of pyralid moths, specifically *E. kuehniella* (subfamily Phycitinae) and *G. mellonella* (subfamily Galleriinae). This study involved a thorough analysis utilizing proteomic, transcriptomic, and genomic data. Our investigation unveiled intriguing findings about the cocoon silk properties of *E. kuehniella*. We discovered that this particular silk is remarkably hygroscopic, exhibiting a capacity to absorb moisture that surpasses that of *G. mellonella* silk by 38%. In addition, we have presented complete sequences for nine FibH proteins derived from the suborder Pyraloidea, and conducted a discussion on the conserved structural features of these proteins. Another noteworthy aspect of our study involves the confirmation of microsyntenic relationships surrounding the sericin gene cluster. We observed these relationships among *E. kuehniella*, *G. mellonella*, and *Amyelois transitella*. This insight into the gene expansion observed in *G. mellonella* adds valuable knowledge to the field of silk research. In Chapter 2, we have provided a detailed description of a previously misannotated silk component gene, *P150/ser6*, found in *B. mori*. Prior to our study, this gene had been incorrectly annotated as two separate genes encoding uncharacterized proteins in both GenBank and SilkBase. We proposed a new model for *P150/ser6*, and verified its specific expression in the middle silk glands of *B. mori*. Previously, the P150/ser6 protein of *B. mori* has been identified in both the non-cocoon silk and the inner cocoon layer of the silk. We observed a similar expression pattern for P150/ser6 homologs in *B. mori*, *E. kuehniella* and *G. mellonella*. However, we also detected variations in the intensity of the P150/ser6 protein within silk cocoons. This discrepancy suggests that the utilization of the P150/ser6 protein during silk synthesis may be influenced by additional genetic factors.

One of the ongoing projects in this laboratory revolves around the characterization of natural silk across various lepidopteran and trichopteran species. As our dataset continues to expand, we encounter two significant challenges that demand immediate attention.

Firstly, it is imperative to address the effective management of our sequence database to conduct comparative studies. Currently, while we do document the sources of the data and the methods employed in generating the sequences, this information is not integrated seamlessly. This disjointed documentation often results in confusion, particularly in the nomenclature of

genes. Furthermore, sequences generated on different platforms frequently contain invalid characters and exhibit varying line endings, which subsequently lead to errors in their utilization. The unprocessed headers within FASTA files also pose a potential source of errors due to the presence of invalid characters, exceeding program limits, or encountering duplicates. Many tools have been created to sanitize sequence data (Waldmann et al., 2014; Foley et al., 2019); however, there is a pressing need to establish a unified standard to streamline future utilization.

Secondly, in comparison to the extensive sequence data, we find ourselves lacking a complete evaluation of silk properties. To address this gap in our research, it is essential to systematically describe a wide range of silk properties, including structural, mechanical, thermal, hydration, optical, electrical, biodegradability, and more. By accumulating more data on these properties, we can enhance our understanding of the intricate relationship between the protein sequences and the resulting silk properties. A compelling example of a resource that has already made significant strides in this direction is the Spider Silkome Database (https://spider-silkome.org). This comprehensive repository houses both sequence and property data for over 1,000 spider species (Arakawa et al., 2022). By adopting a systematic approach, we can enhance our ability to draw meaningful conclusions about silk properties across moths and caddisflies.

Our extensive sequence dataset serves as a valuable resource for our collaborators in the biomaterial field, enabling them to explore and develop new materials with diverse applications. Simultaneously, our dataset holds substantial significance for evolutionary biologists who are keen on untangling the origins of silk production (“phenotype”) from the wealth of sequence information we've gathered (“genotype”) in Lepidoptera. To tackle this question, it is necessary to adopt a more sophisticated strategy that broadens our search for homologous genes. Beyond traditional approaches, such as sequence alignment-based and microsynteny information-based methods, there is a growing interest in utilizing protein structure-based techniques for homology detection. Historically, structure alignment programs relied heavily on a priori knowledge of solved protein structures. With the advancements in machine learning, however, we now have access to highly accurate protein structure prediction tools like AlphaFold (Jumper et al., 2021) and ESMFold (Lin et al., 2023). Additionally, emerging structure-aware aligners like PRotein Ortholog Search Tool (PROST) (Kilinc et al., 2023) and TM-Vec/DeepBLAST (Hamamsy et al., 2023) are gaining

prominence as alternative methods to identify remote homologs. Incorporating these tools into our research toolkit is an important consideration for future investigations. Another important consideration is the inclusion of as many species as possible in our search for an all-inclusive understanding. Accessing high-quality genome data through repositories like NCBI or the Darwin Tree of Life Data Portal (Blaxter et al., 2022) has become increasingly convenient. However, it's crucial to recognize that to gain a holistic and accurate perspective, we should not restrict ourselves to data from only a handful of organisms. This approach allows us to capture variations, adaptations, and unique features at different taxonomic levels that might be missed when relying solely on a limited set of species.

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–95. doi: 10.1126/science.287.5461.2185.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Arakawa, K., Kono, N., Malay, A. D., Tateishi, A., Ifuku, N., Masunaga, H., et al. (2022). 1000 spider silkomes: Linking sequences to silk physical properties. *Sci. Adv.* 8, 1–14. doi: 10.1126/sciadv.abo6043.

Baalsrud, H. T., Tørresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., et al. (2018). De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Mol. Biol. Evol.* 35, 593–606. doi: 10.1093/molbev/msx311.

Badet, T., and Croll, D. (2020). The rise and fall of genes: origins and functions of plant pathogen pangenomes. *Curr. Opin. Plant Biol.* 56, 65–73. doi: 10.1016/j.pbi.2020.04.009.

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20, 159. doi: 10.1186/s13059-019-1774-4.

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6, 914–920. doi: 10.1038/s41477-020-0733-0.

Bell, J. R., Bohan, D. A., Shaw, E. M., and Weyman, G. S. (2005). Ballooning dispersal using silk: world fauna, phylogenies, genetics and models. *Bull. Entomol. Res.* 95, 69–114. doi: 10.1079/BER2004350.

Blaxter, M., Mieszkowska, N., Di Palma, F., Holland, P., Durbin, R., Richards, T., et al. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl. Acad. Sci.* 119, 1–7. doi: 10.1073/pnas.2115642118.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a

platform for investigating biology. *Science* 282, 2012–8. doi: 10.1126/science.282.5396.2012.

CABI (2023). *Ephestia kuehniella* (Mediterranean flour moth). *CABI Compend.* doi: 10.1079/cabicompendium.21412.

Caresche, L. A., and Wapshere, A. J. (1975). Biology and host specificity of the Chondrilla root moth *Bradyrrhoa gilveolella* (Treitschke) (Lepidoptera, Phycitidae). *Bull. Entomol. Res.* 65, 171–185. doi: 10.1017/S0007485300005885.

Caspari, E., and Gottlieb, F. J. (1959). On a modifier of the gene a in *Ephestia kuehniella*. *Z. Vererbungsl.* 90, 263–72. doi: 10.1007/BF00888759.

Chen, L., DeVries, A. L., and Cheng, C.-H. C. (1997a). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl. Acad. Sci.* 94, 3817–3822. doi: 10.1073/pnas.94.8.3817.

Chen, L., DeVries, A. L., and Cheng, C.-H. C. (1997b). Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci.* 94, 3811–3816. doi: 10.1073/pnas.94.8.3811.

Chen, Q., Zobel, J., and Verspoor, K. (2017). Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database* 2017, baw163. doi: 10.1093/database/baw163.

Chen, X., Wang, Y., Wang, Y., Li, Q., Liang, X., Wang, G., et al. (2022). Ectopic expression of sericin enables efficient production of ancient silk with structural changes in silkworm. *Nat. Commun.* 13, 6295. doi: 10.1038/s41467-022-34128-5.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562. doi: 10.1038/nature01262.

Collin, M. A., Mita, K., Sehnal, F., and Hayashi, C. Y. (2010). Molecular Evolution of Lepidopteran Silk Proteins: Insights from the Ghost Moth, *Hepialus californicus*. *J. Mol. Evol.* 70, 519–529. doi: 10.1007/s00239-010-9349-8.

Damas, J., Corbo, M., Kim, J., Turner-Maier, J., Farré, M., Larkin, D. M., et al. (2022). Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc. Natl. Acad.*

Sci. 119, 1–12. doi: 10.1073/pnas.2209139119.

Das, G., Shin, H. S., Campos, E. V. R., Fraceto, L. F., del Pilar Rodriguez-Torres, M., Mariano, K. C. F., et al. (2021). Sericin based nanoformulations: a comprehensive review on molecular mechanisms of interaction with organisms to biological applications. *J. Nanobiotechnology* 19, 1–22. doi: 10.1186/s12951-021-00774-y.

Ding, W., Baumdicker, F., and Neher, R. A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids Res.* 46, e5–e5. doi: 10.1093/nar/gkx977.

Dong, Z., Guo, K., Zhang, X., Zhang, T., Zhang, Y., Ma, S., et al. (2019). Identification of *Bombyx mori* sericin 4 protein as a new biological adhesive. *Int. J. Biol. Macromol.* 132, 1121–1130. doi: 10.1016/j.ijbiomac.2019.03.166.

Dong, Z., Song, Q., Zhang, Y., Chen, S., Zhang, X., Zhao, P., et al. (2016). Structure, evolution, and expression of antimicrobial silk proteins, seroins in Lepidoptera. *Insect Biochem. Mol. Biol.* 75, 24–31. doi: 10.1016/j.ibmb.2016.05.005.

Ensembl (2023). Synteny. *Ensembl*. Available at: http://www.ensembl.org/Help/Permalink?url=http%3A%2F%2FJul2023.archive.ensembl.org [Accessed September 22, 2023].

Eom, J., Park, S., Jin, H.-J., and Kwak, H. W. (2020). Multiscale Hybridization of Natural Silk–Nanocellulose Fibrous Composites With Exceptional Mechanical Properties. *Front. Mater.* 7, 1–12. doi: 10.3389/fmats.2020.00098.

Foley, G., Sützl, L., D'Cunha, S. A., Gillam, E. M. J., and Bodén, M. (2019). SeqScrub: a web tool for automatic cleaning and annotation of FASTA file headers for bioinformatic applications. *Biotechniques* 67, 50–54. doi: 10.2144/btn-2018-0188.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 546, 563–7. doi: 10.1126/science.274.5287.546.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49.

Goudey, B., Geard, N., Verspoor, K., and Zobel, J. (2022). Propagation, detection and correction of errors using the sequence database network. *Brief. Bioinform.* 23, 1–12.

doi: 10.1093/bib/bbac416.

Guo, K., Zhang, X., Zhao, D., Qin, L., Jiang, W., Hu, W., et al. (2022). Identification and characterization of sericin5 reveals non-cocoon silk sericin components with high β-sheet content and adhesive strength. *Acta Biomater.* 150, 96–110. doi: 10.1016/j.actbio.2022.07.021.

Guo, P. C., Dong, Z., Xiao, L., Li, T., Zhang, Y., He, H., et al. (2015). Silk gland-specific proteinase inhibitor serpin16 from the *Bombyx mori* shows cysteine proteinase inhibitory activity. *Biochem. Biophys. Res. Commun.* 457, 31–36. doi: 10.1016/j.bbrc.2014.12.056.

Hamamsy, T., Morton, J. T., Blackwell, R., Berenberg, D., Carriero, N., Gligorijevic, V., et al. (2023). Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.* provisiona. doi: 10.1038/s41587-023-01917-2.

Heckenhauer, J., Stewart, R. J., Ríos-Touma, B., Powell, A., Dorji, T., Frandsen, P. B., et al. (2023). Characterization of the primary structure of the major silk gene, h-fibroin, across caddisfly (Trichoptera) suborders. *iScience* 26, 107253. doi: 10.1016/j.isci.2023.107253.

Holland, C., Numata, K., Rnjak-Kovacina, J., and Seib, F. P. (2019). The Biomedical Use of Silk: Past, Present, Future. *Adv. Healthc. Mater.* 8. doi: 10.1002/adhm.201800465.

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. doi: 10.1074/jbc.M006897200.

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11, 431. doi: 10.1186/1471-2105-11-431.

Julien, E., Coulon-Bublex, M., Garel, A., Royer, C., Chavancy, G., Prudhomme, J.-C., et al. (2005). “Silk Gland Development and Regulation of Silk Protein Genes,” in *Comprehensive Molecular Insect Science* (Elsevier), 369–384. doi: 10.1016/B0-44-451924-6/00022-3.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:

10.1038/s41586-021-03819-2.

Kawamoto, M., Jouraku, A., Toyoda, A., Yokoi, K., Minakuchi, Y., Katsuma, S., et al. (2019). High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* 107, 53–62. doi: 10.1016/j.ibmb.2019.02.002.

Kilinc, M., Jia, K., and Jernigan, R. L. (2023). Improved global protein homolog detection with major gains in function identification. *Proc. Natl. Acad. Sci.* 120, 1–9. doi: 10.1073/pnas.2211823120.

Kirilenko, B. M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., et al. (2023). Integrating gene annotation with orthology inference at scale. *Science (80-.).* 380. doi: 10.1126/science.abn3107.

Kludkiewicz, B., Kucerova, L., Konikova, T., Strnad, H., Hradilova, M., Zaloudikova, A., et al. (2019). The expansion of genes encoding soluble silk components in the greater wax moth, *Galleria mellonella*. *Insect Biochem. Mol. Biol.* 106, 28–38. doi: 10.1016/j.ibmb.2018.11.003.

Kmet, P., Kucerova, L., Sehadova, H., Chia-hsiang Wu, B., Wu, Y.-L., and Zurovec, M. (2023). Identification of silk components in the bombycoid moth *Andraca theae* (Endromidae) reveals three fibroin subunits resembling those of Bombycidae and Sphingidae. *J. Insect Physiol.* 147, 104523. doi: 10.1016/j.jinsphys.2023.104523.

Koh, L.-D., Cheng, Y., Teng, C.-P., Khin, Y.-W., Loh, X.-J., Tee, S.-Y., et al. (2015). Structures, mechanical properties and applications of silk fibroin materials. *Prog. Polym. Sci.* 46, 86–110. doi: 10.1016/j.progpolymsci.2015.02.001.

Kucerova, L., Zurovec, M., Kludkiewicz, B., Hradilova, M., Strnad, H., and Sehnal, F. (2019). Modular structure, sequence diversification and appropriate nomenclature of seroins produced in the silk glands of Lepidoptera. *Sci. Rep.* 9, 3797. doi: 10.1038/s41598-019-40401-3.

Künstner, A., Busch, H., Hartmann, E., and Traut, W. (2022). Data on draft genomes and transcriptomes from females and males of the flour moth, *Ephestia kuehniella*. *Data Br.* 42, 108140. doi: 10.1016/j.dib.2022.108140.

Li, Y., Zhao, P., Liu, H., Guo, X., He, H., Zhu, R., et al. (2015). TIL-type protease inhibitors

may be used as targeted resistance factors to enhance silkworm defenses against invasive fungi. *Insect Biochem. Mol. Biol.* 57, 11–19. doi: 10.1016/j.ibmb.2014.11.006.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (80-.).* 379, 1123–1130. doi: 10.1126/science.ade2574.

Liu, J., Shi, L., Deng, Y., Zou, M., Cai, B., Song, Y., et al. (2022). Silk sericin-based materials for biomedical applications. *Biomaterials* 287, 121638. doi: 10.1016/j.biomaterials.2022.121638.

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., et al. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* 11, 1–20. doi: 10.7554/eLife.78526.

Lucas, F., Shaw, J. T. B., and Smith, S. G. (1957). Amino-Acid Composition of the Silk of Chrysopa Egg-stalks. *Nature* 179, 906–907. doi: 10.1038/179906a0.

Ma, Y., Zeng, W., Ba, Y., Luo, Q., Ou, Y., Liu, R., et al. (2022). A single-cell transcriptomic atlas characterizes the silk-producing organ in the silkworm. *Nat. Commun.* 13, 3316. doi: 10.1038/s41467-022-31003-1.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., et al. (2004). The genome sequence of silkworm, *Bombyx mori*. *DNA Res.* 11, 27–35. doi: 10.1093/dnares/11.1.27.

Mori, K., Tanaka, K., Kikuchi, Y., Waga, M., Waga, S., and Mizuno, S. (1995). Production of a Chimeric Fibroin Light-chain Polypeptide in a Fibroin Secretion-deficient Naked Pupa Mutant of the Silkworm *Bombyx mori*. *J. Mol. Biol.* 251, 217–228. doi: 10.1006/jmbi.1995.0429.

Offord, C., Vollrath, F., and Holland, C. (2016). Environmental effects on the construction and physical properties of *Bombyx mori* cocoons. *J. Mater. Sci.* 51, 10863–10872. doi: 10.1007/s10853-016-0298-5.

Passarge, E., Horsthemke, B., and Farber, R. A. (1999). Incorrect use of the term synteny. *Nat. Genet.* 23, 387–387. doi: 10.1038/70486.

Perdrix-Gillot, S. (1979). DNA synthesis and endomitoses in the giant nuclei of the silkgland

of *Bombyx mori*. *Biochimie* 61, 171–204. doi: 10.1016/S0300-9084(79)80066-8.

Pevzner, P., and Tesler, G. (2003). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13, 37–45. doi: 10.1101/gr.757503.

Reizabal, A., Costa, C. M., Pérez-Álvarez, L., Vilas-Vilela, J. L., and Lanceros-Méndez, S. (2023). Silk Fibroin as Sustainable Advanced Material: Material Properties and Characteristics, Processing, and Applications. *Adv. Funct. Mater.* 33. doi: 10.1002/adfm.202210764.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* 12, 85–94. doi: 10.1093/protein/12.2.85.

Rouhova, L., Kludkiewicz, B., Sehadova, H., Sery, M., Kucerova, L., Konik, P., et al. (2021). Silk of the common clothes moth, *Tineola bisselliella*, a cosmopolitan pest belonging to the basal ditrysian moth line. *Insect Biochem. Mol. Biol.* 130, 103527. doi: 10.1016/j.ibmb.2021.103527.

Rouhová, L., Sehadová, H., Pauchová, L., Hradilová, M., Žurovcová, M., Šerý, M., et al. (2022). Using the multi-omics approach to reveal the silk composition in *Plectrocnemia conspersa*. *Front. Mol. Biosci.* 9, 1–14. doi: 10.3389/fmolb.2022.945239.

Sehnal, F., and Žurovec, M. (2004). Construction of Silk Fiber Core in Lepidoptera. *Biomacromolecules* 5, 666–674. doi: 10.1021/bm0344046.

Siengchin, S. (2023). A review on lightweight materials for defence applications: Present and future developments. *Def. Technol.* 24, 1–17. doi: 10.1016/j.dt.2023.02.025.

Simakov, O., Bredeson, J., Berkoff, K., Marletaz, F., Mitros, T., Schultz, D. T., et al. (2022). Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* 8. doi: 10.1126/sciadv.abi5884.

Singh, C. P., Vaishna, R. L., Kakkar, A., Arunkumar, K. P., and Nagaraju, J. (2014). Characterization of antiviral and antibacterial activity of *Bombyx mori* seroin proteins. *Cell. Microbiol.* 16, 1354–1365. doi: 10.1111/cmi.12294.

Song, H., Lin, K., Hu, J., and Pang, E. (2018). An Updated Functional Annotation of Protein-Coding Genes in the Cucumber Genome. *Front. Plant Sci.* 9, 1–14. doi: 10.3389/fpls.2018.00325.

Stoler, N., and Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* 3, 1–9. doi: 10.1093/nargab/lqab019.

Takei, F., Kikuchi, Y., Kikuchi, A., Mizuno, S., and Shimura, K. (1987). Further evidence for importance of the subunit combination of silk fibroin in its efficient secretion from the posterior silk gland cells. *J. Cell Biol.* 105, 175–180. doi: 10.1083/jcb.105.1.175.

Tanaka, K., Inoue, S., and Mizuno, S. (1999). Hydrophobic interaction of P25, containing Asn-linked oligosaccharide chains, with the H-L complex of silk fibroin produced by *Bombyx mori*. *Insect Biochem. Mol. Biol.* 29, 269–276. doi: 10.1016/S0965-1748(98)00135-0.

Tanaka, K., and Mizuno, S. (2001). Homologues of fibroin L-chain and P25 of *Bombyx mori* are present in *Dendrolimus spectabilis* and *Papilio xuthus* but not detectable in *Antheraea yamamai*. *Insect Biochem. Mol. Biol.* 31, 665–677. doi: 10.1016/S0965-1748(00)00173-9.

Tettelin, H., and Medini, D. (2020). *The Pangenome*. , eds. H. Tettelin and D. Medini Cham: Springer International Publishing doi: 10.1007/978-3-030-38281-0.

The International Silkworm Genome Consortium (2008). The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1036–1045. doi: 10.1016/j.ibmb.2008.11.004.

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

Underhill, A. P. (1997). Current issues in Chinese Neolithic archaeology. *J. World Prehistory* 11, 103–160. doi: 10.1007/BF02221203.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681. doi: 10.1016/j.tig.2018.05.008.

Vergara, I. A., and Chen, N. (2010). Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics* 11, 516. doi: 10.1186/1471-2164-11-516.

Visser, S., Voleníková, A., Nguyen, P., Verhulst, E. C., and Marec, F. (2021). A conserved

role of the duplicated Masculinizer gene in sex determination of the Mediterranean flour moth, *Ephestia kuehniella*. *PLoS Genet.* 17, e1009420. doi: 10.1371/journal.pgen.1009420.

Volenikova, A., Nguyen, P., Davey, P., Sehadova, H., Kludkiewicz, B., Koutecky, P., 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. doi: 10.1038/s42003-022-04240-9.

Waldmann, J., Gerken, J., Hankeln, W., Schweer, T., and Glöckner, F. O. (2014). FastaValidator: an open-source Java library to parse and validate FASTA formatted sequences. *BMC Res. Notes* 7, 365. doi: 10.1186/1756-0500-7-365.

Wang, X., Li, Y., Liu, Q., Xia, Q., and Zhao, P. (2017). Proteome profile of spinneret from the silkworm, *Bombyx mori*. *Proteomics* 17, 1600301. doi: 10.1002/pmic.201600301.

Wang, Z., Zhang, Y., Zhang, J., Huang, L., Liu, J., Li, Y., et al. (2014). Exploring natural silk protein sericin for regenerative medicine: an injectable, photoluminescent, cell-adhesive 3D hydrogel. *Sci. Rep.* 4, 7064. doi: 10.1038/srep07064.

Wu, B. C., Sauman, I., Maaroufi, H. O., Zaloudikova, A., Zurovcova, M., Kludkiewicz, B., et al. (2022). Characterization of silk genes in *Ephestia kuehniella* and *Galleria mellonella* revealed duplication of sericin genes and highly divergent sequences encoding fibroin heavy chains. *Front. Mol. Biosci.* 9, 1–16. doi: 10.3389/fmolb.2022.1023381.

Wu, B. C., Zabelina, V., Zurovcova, M., and Zurovec, M. (2023). Unravelling the complexity of silk sericins: *P150/sericin* 6 is a new silk gene in *Bombyx mori*. *bioRxiv*, 2023.09.22.558982. doi: 10.1101/2023.09.22.558982.

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., et al. (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306, 1937–40. doi: 10.1126/science.1102210.

Ye, X., Zhao, S., Wu, M., Ruan, J., Tang, X., Wang, X., et al. (2021). Role of sericin 1 in the immune system of silkworms revealed by transcriptomic and proteomic analyses after gene knockout. *FEBS Open Bio* 11, 2304–2318. doi: 10.1002/2211-5463.13239.

Yonemura, N., Mita, K., Tamura, T., and Sehnal, F. (2009). Conservation of Silk Genes in

Trichoptera and Lepidoptera. *J. Mol. Evol.* 68, 641–653. doi: 10.1007/s00239-009-9234-5.

Zabelina, V., Takasu, Y., Sehadova, H., Yonemura, N., Nakajima, K., Sezutsu, H., et al. (2021). Mutation in *Bombyx mori* fibrohexamerin (P25) gene causes reorganization of rough endoplasmic reticulum in posterior silk gland cells and alters morphology of fibroin secretory globules in the silk gland lumen. *Insect Biochem. Mol. Biol.* 135, 103607. doi: 10.1016/j.ibmb.2021.103607.

Zhang, X., Guo, K., Dong, Z., Chen, Z., Zhu, H., Zhang, Y., et al. (2020a). Kunitz-type protease inhibitor BmSPI51 plays an antifungal role in the silkworm cocoon. *Insect Biochem. Mol. Biol.* 116, 103258. doi: 10.1016/j.ibmb.2019.103258.

Zhang, Y., Tang, M., Dong, Z., Zhao, D., An, L., Zhu, H., et al. (2020b). Synthesis, secretion, and antifungal mechanism of a phosphatidylethanolamine-binding protein from the silk gland of the silkworm *Bombyx mori*. *Int. J. Biol. Macromol.* 149, 1000–1007. doi: 10.1016/j.ijbiomac.2020.01.310.

Zhao, T., Zwaenepoel, A., Xue, J.-Y., Kao, S.-M., Li, Z., Schranz, M. E., et al. (2021). Whole-genome microsynteny-based phylogeny of angiosperms. *Nat. Commun.* 12, 3498. doi: 10.1038/s41467-021-23665-0.

Zheng, X. H., Lu, F., Wang, Z.-Y., Zhong, F., Hoover, J., and Mural, R. (2005). Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics* 21, 703–710. doi: 10.1093/bioinformatics/bti045.

Zhou, C. Z., Confalonieri, F., Medina, N., Zivanovic, Y., Esnault, C., Yang, T., et al. (2000). Fine organization of *Bombyx mori* fibroin heavy chain gene. *Nucleic Acids Res.* 28, 2413–9. doi: 10.1093/nar/28.12.2413.

Zurovec, M., Yang, C., Kodrík, D., and Sehnal, F. (1998). Identification of a Novel Type of Silk Protein and Regulation of Its Expression. *J. Biol. Chem.* 273, 15423–15428. doi: 10.1074/jbc.273.25.15423.

bulah@entu.cas.cz

Comparative Analysis of Silk Proteins and Discovery of Novel Sericin Gene in Lepidopteran Moths

Ph.D. Thesis

University of South Bohemia in České Budějovice

Faculty of Science

Branišovská 1760

CZ-37005 České Budějovice, Czech Republic

Phone: +420 387 776 201

www.prf.jcu.cz, e-mail: sekret-fpr@prf.jcu.cz