# Reference transcriptome data in silkworm *Bombyx mori*

Kakeru Yokoi[1,2,*,**], Takuya Tsubota[3,*], Jianqiang Sun[2], Akiya Jouraku[1], Hideki Sezutsu[3], Hidemasa Bono[4]

1 Insect Genome Research and Engineering Unit, Division of Applied Genetics, Institute of Agrobiological Sciences (NIAS), National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba, Ibaraki 305-8634, Japan
2 Research Center for Agricultural Information Technology (RCAIT), National Agriculture and Food Research Organization (NARO), Kintetsu Kasumigaseki Building Kasumigaseki 3-5-1 Chiyoda-ku, Tokyo 100-0013, Japan
3 Transgenic Silkworm Research Unit, Division of Biotechnology, Institute of Agrobiological Sciences (NIAS), National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba, Ibaraki 305-8634, Japan
4 Database Center for Life Science (DBCLS), Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan
*These authors equally contributed to this work.
**Corresponding author: Kakeru Yokoi

Email adresses
Kakeru Yokoi: yokoi123@affrc.go.jp
Takuya Tsubota: tsubota@affrc.go.jp
Sun Jianqiang: j.sun@affrc.go.jp
Akiya Jouraku: joraku@affrc.go.jp
Hideki Sedutsu: hsezutsu@affrc.go.jp
Hidemasa Bono: bono@dbcls.rois.ac.jp

## Abstract

The silkworm *Bombyx mori* has long been used in the silk industry and utilized in studies of physiology, genetics, molecular biology, and pathology. We recently reported high quality reference genome data for *B. mori.* In the present study, we constructed a

reference transcriptome data set using the reference genome data and RNA-seq data of 10 tissues from P50T strain larvae. Reference transcriptome data contained 51,926 transcripts, with 39,619 having one or more coding sequence region. The abundance of each transcript in the 10 tissues as well as 5 tissues from other strain larvae was estimated, and hierarchical clustering was performed. The results obtained showed that data on abundance were highly reproducible and there here is a little difference of transcriptome abundance between the two strain larvae. New isoforms of silk-related genes were searched for in the reference transcriptomes, and the longest isoform of *sericin-1* possessing a long exon was identified. We also extracted transcripts that were strongly expressed in one or more parts of the silk glands. An enrichment analysis performed using the functional annotation data of the transcripts provided novel insights into the functions of the silk gland parts. Therefore, the reference transcriptome data set obtained has extended *B. mori* genomic and transcriptomic insights and may contribute to advances in entomologic and vertebrate research, including that on humans.

# Introduction

The silkworm *Bombyx mori* is a lepidopteran insect that has been utilized in studies of physiology, genetics, molecular biology, and pathology. Functional analyses of genes related to hormone synthesis/degradation, pheromone reception, larval marking formation, and virus resistance have been performed using this silkworm (Tan et al., 2005; Ito et al., 2008; Sakurai et al., 2011; Daimon et al., 2015; Kondo et al., 2017), and the findings obtained have contributed to the promotion of insect science. The silkworm has the ability to produce large amounts of silk proteins, which is one of the most prominent characteristics of this species. Silk proteins are mainly composed of the fibrous protein Fibroin and aqueous protein Sericin and are produced in the larval tissue silk gland (SG) (Yukuhiro et al., 2016). A transgenic technique has been applied to the silkworm (Tamura et al., 2000), and has enabled the production of a large amount of recombinant proteins through the introduction of transgenes, which are overexpressed in the SG (Tatematsu et al., 2010). Through this method, it is possible to utilize the silkworm as a significant bioreactor.

Based on its biological and industrial importance, the whole genome sequence data of the silkworm was reported in 2004 from two research groups (Mita el al., 2004; Xia et al., 2004), and was the first lepidopteran whole genome data. Silkworm whole genome data was updated in 2008 from an international research group (International Silkworm Genome Consortium 2008). Several data related to the silkworm genome

68 have since become available (e.g. microarray-based gene expression profiles in
69 multiple tissues, BAC-based linkage map full-length cDNA data and of *B. mori* in
70 KaikoBase) (Xia et al., 2007; Yamamoto et al., 2008; Suetsugu et al., 2013). These
71 findings have strongly promoted studies on *B. mori* and other lepidopteran insects in
72 the past decade.
73 In 2019, we reported a new and high-quality genome silkworm reference
74 genome assembly the silkworm p50T strain using PacBio long-read and Illumina short-
75 read sequencers (Kawamoto et al., 2019). The new genome assembly consists of 696
76 scaffolds with N50 of 16.8 M and only 30 gaps, and a new gene model based on this
77 sequence was predicted. These data are expected to be utilized in silkworm and
78 Lepidopteran research. Reference transcriptome data using this genome sequence
79 and the predicted gene set and transcriptome profile of important tissues significantly
80 contribute to advances in silkworm and Lepidopteran research. In the present study,
81 we constructed a reference transcript sequence data set using the RNA-seq data of 10
82 important tissues from silkworm larvae and reference genome data for improving the
83 predicted gene set data of Kawamoto et al. (2019) (Fig. 1). We also showed
84 comprehensive expression data for 10 different organs. These results will contribute
85 to further advances in silkworm as well as entomological and vertebrate research,
86 including that on humans.

# 87 Results

88 Reference transcriptome data

89 Total RNAs were extracted from the fat body (FB), midgut (MG), Malpighian
90 tubules (MT), testis (TT), anterior silk gland (ASG), anterior part of the middle silk gland
91 (MSG-A), middle part of the middle silk gland (MSG-M), posterior part of the middle silk
92 gland (MSG-P) and posterior silk gland (PSG) of one male P50T larva and from the
93 ovary (OV) of a female larva. Extraction was repeated three times. Total RNAs were
94 sequenced and thirty sets of sequenced data were used as RNA-seq data. We
95 obtained “reference transcriptome data” by using the reference genome, gene model
96 data (Kawamoto et al., 2019), and RNA-seq data. Transcriptome data contains 51,926
97 transcripts in 24236 loci (Supporting data 1). The previously constructed gene model
98 data contained 16,880 genes in 16,845 loci, while our reference transcriptome data
99 contains new transcripts and loci (Fig. 2A). Among 51,926 transcripts, 7,704 transcripts
100 belonged to new loci while 27,342 transcripts are newly identified isoforms of which loci
101 was already identified in Kawamoto et al., (2019) (Fig. 2B). These results suggest that
102 our reference transcriptome data extend gene model data and contain many

103 unidentified transcripts and loci. To annotate transcripts, we predicted the coding
104 sequence region (CDS) and amino acid sequences (Supporting data 2) against all
105 transcripts. We found that 39,619 transcripts and 16,632 loci had at least one CDS.
106 The predicted amino acid sequences were used for gene functional annotations by a
107 homology search against human and *Drosophila* gene sets (Supporting data 3).
108
109 Estimating the abundance of the reference transcriptome in multiple tissues
110 We calculated the abundance of each transcript in ten tissues: FB, MG, MT, TT,
111 OV, ASG, MSG-A, MSG-M, MSG-P, PSG, TT, and OV. In comparisons of our
112 calculated results and evaluations of the reliability of our expression analysis, we
113 additionally quantified transcript expression by using public RNA-seq data that were
114 sequenced from the FB, MG, MT, TT, and SG of the *B. mori* o751 strain (Kikuchi et al.,
115 2017; Ichino et al., 2018; Kobayashi et al., 2019). To distinguish between our RNA-seq
116 data and public data, we referred to the FB, MG, MT, TT, and SG of the o751 strain as
117 BN_FB, BM_MG, BN_ MT, BM_TT, and BM_SG, respectively. Abundance values are
118 shown as transcripts per million (tpm) and listed in Supporting data 4.
119 Hierarchical clustering was performed for comparisons of transcriptome
120 abundance between multiple tissues (Fig. 3). The results obtained showed that, except
121 for MSG_P and MSG_M, all samples were clearly grouped by tissue types. Moreover,
122 clusters of samples that were the same tissues, but extracted from different strain
123 larvae, namely, FB and BN_FB, MT and BN_MT, MG and BN_MG, and TT and
124 BN_TT, were placed in adjacent positions, of which samples were obtained from
125 different studies as well as different strains. These results suggest that the data
126 obtained on abundance were highly reproducible and there were a little difference in
127 transcriptome profiles between P50T and o751 larvae. These results also indicated
128 that there were slight individual genetic differences between the larval samples used
129 and a marginal artificial effect, demonstrating that our abundance data may be used as
130 reference expression data for silkworm larvae.
131 Transcripts of *sericin*, *fibroin*, and *fibrohexamerin* genes
132 The *sericin*, *fibroin*, and *fibrohexamerin* (*fhx*) genes are known to play
133 important roles in silk synthesis in *B. mori*. Our transcriptome data contains several
134 isoforms of *sericin*, *fibroin*, *and fhx* (Table 1). A detailed sequence analysis was
135 performed to elucidate isoform structures. A previous analysis of *sericin-1* revealed that
136 this gene is composed of 9 exons, with exons 3-6 being under the selection of
137 alternative splicing (Garel et al., 1997; Yukuhiro et al., 2016). Within these exons, exon
138 6 is responsible for the most abundant component of the sericin protein (Sericin M),

139 which is ~6500 bp in length and encodes a serine-rich repetitive sequence (Yukuhiro et
140 al., 2016). However, the structure of this exon has not yet been elucidated in detail. We
141 herein found that exon 6 of MSTRG.2477.1, the longest *sericin-1* isoform identified in
142 the present study, had a length of 6234 bp (from 2,552,212 to 2,558,445 bp in
143 chromosome 11, see Supporting data 2 and Fig. 4A) In the present study, we newly
144 identified a *sericin-1* isoform that contained exon 6 with 6234 bp. We speculate that
145 this isoform corresponds to the full-length or nearly full-length transcript of *sericin-1*.
146 The product from this exon is enriched with serine residues and also has abundant
147 residues of glycine, asparagine and threonine (Supplemental Fig. 1).
148 Sericin-3 is another major silk protein that has a relatively soft texture (Takasu
149 et al., 2006; 2007). In gene model data, there is a frame shift in the predicted amino
150 acid sequence (KWMTBOMO08464), presumably due to the 73-bp deletion present in
151 exon 3. In the present study, we found that reference transcriptome data
152 (MSTRG.2595.1) provided an accurate gene structure. Sericin-4 is another recently
153 identified sericin protein that is composed of 34 exons (Dong et al., 2019). In gene
154 model data, it is split into three genes (KWMTBOMO06324, KWMTBOMO06325 and
155 KWMTBOMO06326) whereas our reference transcriptome data represent an exact
156 structure (MSTRG.2610.1, Fig. 4B).
157 In contrast, a better structure is provided by gene model data for the fibroin
158 heavy chain (*h-fib*); this gene encodes a protein with a large and highly repetitive
159 sequence (Zhou et al., 2000) and although there is a small deletion in the repeat motifs
160 for both models, the deletion length is shorter for gene model data (32 amino acid
161 deletion for gene model data and 223 for reference transcriptome data). Regarding
162 other silk genes (*sericin-2*, *fibroin light chain* (*l-fib*), and *fhx*), both models provide exact
163 structures (data not shown).

164 Transcript abundance in the silk gland

165 As described above, silk is synthesized in the SG. While the role of each SG
166 part in silk synthesis is known, the underlying molecular and genetic mechanisms
167 remain unclear. Therefore, the genes or transcripts that are strongly expressed in each
168 SG part need to be identified in order to elucidate these mechanisms. We searched for
169 transcripts that showed values of more than 30 transcripts per kilobase million (tpm) in
170 the five SG parts. The results obtained are shown in Fig. 5. The numbers of transcripts
171 that were strongly expressed in only ASG, MSG_A, MSG_M, MSG_P, and PSG were
172 351, 180, 99, 71, and 100, respectively, while more than 1,000 transcripts were
173 strongly expressed in all parts of the SG.

By using the annotation data of the strongly expressed transcripts, a functional enrichment analysis (FEA) was performed using the transcripts strongly expressed in each part of the SG to predict their role. In the enrichment analysis, we utilized the annotation results against the human gene set. Fig. 6A shows the results of FEA using the annotation of transcripts strongly expressed in MSG_P, MSG_M specific plus both in MSG_P and MSG_M. The reason for utilizing MSG_P, MSG_M specific plus MSG_P and MSG_M classes is that the samples of MSG_P and MSG_M in Fig. 3 did not form different clusters, suggesting that both tissues have the same functions. The highly enriched function of the category (-log(P) > 10) was “Metabolism of RNA”, while the moderately enriched functions (6 < -log(P) < 10) were “ncRNA metabolic process”, “regulation of mRNA processing”, “HIV Infection”, and “Asparagine N-linked glycosylation”. In MSG_A, the moderately enriched function was “Metabolism of vitamins and cofactors”, while the highly enriched functions of the category were not found (Fig. 6B). In ASG, the highly enriched functions of the category were “carbohydrate metabolic process” and “Transport of small molecules” (Fig. 6C). The moderately enriched functions were “anion transport”, “Glycolysis/Gluconeogenesis”, “Ascorbate and aldarate metabolism”, and “Metabolism of carbohydrates”. In PSG, the moderately enriched function was “tRNA modification”, while the highly enriched function of the category was not found (Fig. 6D).

# Discussion

In the present study, we obtained RNA-seq data on ten tissues of *B. mori* on the 3rd day of fifth instar larvae from the P50T strain. Using RNA-seq data and new reference genome data (Kawamoto et al., 2019), we constructed reference transcriptome data. Our transcriptome data contained new loci and isoforms, thereby updating the reference genomic and transcriptome data of *B. mori*. The reference transcriptome consists of 51,926 transcripts in 24,236 loci (16,632 loci have coding genes), and 39,619 transcripts contain single or multiple CDS. In the mouse reference data set (GRCm38.p6), there are 52332 loci (22,480 coding genes, 16,324 non-coding genes, and 13,528 pseudogenes) and 142,333 transcripts (http://asia.ensembl.org/Mus_musculus/Info/Annotation). In the human reference data set (GRCh38.p12), there are 63,048 loci (20,454 coding genes, 23,940 non-coding genes, and 15,204 pseudogenes) and 226,950 transcripts (http://asia.ensembl.org/Homo_sapiens/Info/Annotation). In *Drosophila melanogaster* (BDGP6.22), there are 17,753 loci (13,931 coding genes, 3,513 non-coding genes, and 309 pseudogenes) and 34,802 transcripts

209 (http://asia.ensembl.org/Drosophila_melanogaster/Info/Annotation). In Zebrafish
210 (GRCz11), there are 32,506 loci (25,592 coding genes, 6,599 non-coding genes, and
211 315 pseudogenes) and 59,878 transcripts
212 (http://asia.ensembl.org/Danio_rerio/Info/Annotation). In consideration of the basic
213 status of the reference data of these model species, our transcriptome data is not
214 unusual. It suggests that transcriptome data cover the majority of actual transcripts.
215 We estimated transcriptome abundance in multiple tissues plus several
216 tissues of other strain larvae. Transcriptome abundance in the tissues MG, TT, MT,
217 and FB did not markedly differ between the P50T and o751 strains. These results
218 suggest that these tissues at this stage did not contribute to phenotypic differences
219 between the two strains. To elucidate the underlying genetic mechanisms for
220 phenotypic differences, the RNA-seq data and transcriptome data of other stages are
221 needed. On the other hand, transcriptome abundance in MSG_M and MSG_P samples
222 was not divided into two independent clusters, suggesting that both parts have similar
223 roles in this stage, while MSG_A has distinct roles from the other parts.
224 We searched for new or previously unidentified isoforms of the *sericin, fibroin*,
225 *and fhx* genes in reference transcriptome data. While new or previously unidentified
226 isoforms of *sericin-2*, *l-fib*, *h-fib*, and *fhx* were not found, the long or structured isoforms
227 of *Sericin-1, Sericin-3*, and *Sericin-4* were identified in the reference transcriptome. The
228 longest isoform of *Sericin-3* (MSTRG.2595.1) possessed slightly more extensive
229 nucleotide sequences than that of KWMTBOMO08464, in which 73 bases of exon 3
230 were deleted, resulting in the prediction of incorrect ORF. The predicted amino acid
231 sequences of KWMTBOMO08464 were not similar to the sericin-3 amino acid
232 sequence in UniProtKB (ID: A8CEQ1), while that of MSTRG.2595.1 was similar.
233 Therefore, our transcriptome data provide more precise gene predictions. In the case
234 of *Sericin-4*, which was recently identified (Dong et al., 2019), we found a longer
235 transcript in our transcriptome data than the gene model reported by Dong et al.
236 (2019), which may contribute to the further characterization of sericin-4. The new
237 isoform of *Sericin-1* contains CDS that code glycine-, asparagine-, and threonine-rich
238 regions. It was not possible to elucidate the sequences of CDS because they were very
239 repetitive. Using long-read sequencers, repetitive sequences have been accurately
240 elucidated in the new reference genome. We consider our reference transcriptome
241 data to have significantly improved gene model data.
242 We searched for strongly expressed transcripts in one or more SG parts.
243 While more than 1,000 transcripts were strongly and ubiquitously expressed in the SG,
244 801 transcripts were strongly expressed in single parts of the SG. FEA with annotation

245 data on these transcripts in each part of the SG, except for the categories of MSG_A,
246 MSG_M specific plus MSG_A and MSG_M, was performed. The FEA results for
247 MSG_A, MSG_M specific plus MSG_A and MSG_M showed that these parts have
248 roles in coding or non-coding RNA processing. Some functional descriptions of these
249 ontologies are related to “splice variant processing”. Some isoforms of *sericin-1* (IDs of
250 MSTRG.2477.1 - MSTRG.2477.16, and KWMTBOMO06216.mrna1) were strongly
251 expressed in MSG_A and MSG_M. Moreover, the FEA result contained “Asparagine
252 N-linked glycosylation”. These results suggest that the splice variant processing of
253 *sericin-1* and asparagine processing of the sericin-1 protein, which possesses many
254 asparagine residues, occurred in MSG_A and MSG_M. The FEA results for ASG
255 suggested that ASG produced large amounts of energy via carbohydrate metabolic
256 processes. Silk proteins are mainly synthesized in PSG and MSG. After several
257 processes, matured silk protein, which is a large complex, is exported and released
258 through ASG (Takiya et al., 2016). Therefore, the strong expression of “carbohydrate
259 metabolic”-related transcripts may contribute to the export of silk protein. Since there is
260 moderate ontology for MSG_A and PSG, we cannot predict the roles of these parts.
261 In the present study, we performed RNA-seq on multiple tissues of *B. mori* and
262 constructed reference transcriptome data. The reference transcriptome data
263 constructed using RNA-seq data and new reference genome data contained
264 unidentified loci and isoforms, including a long and almost complete *sericin-1* isoform,
265 which are not present in the gene model data based on a reference new genome
266 (Kawamoto et al., 2019). Moreover, comprehensive transcriptome abundance and
267 annotation data will contribute to elucidating the functions of SG parts previously not
268 proven. The transcript data obtained herein will lead to advances in entomologic and
269 vertebrate research, including that on humans (Tabunoki et al., 2016).

# 270 Methods

271 <u>Silkworm rearing, RNA extraction, and sequencing</u>

272 The silkworm P50T (*daizo*) strain was reared on an artificial diet (Nihon Nosan
273 Kogyo, Yokohama, Japan) at 25°C under a 12-hour light/dark photoperiod. Tissues of
274 the SG, FB, MG, MT, TT, and OV were dissected on the 3$^{rd}$ day of fifth instar larvae.
275 The SG was further subdivided into ASG, MSG-A, MSG-M, MSG-P, and PSG. Each
276 tissue was dissected from one individual, except for OV, and three biological replicates
277 were obtained and analyzed separately. Tissues were homogenized using ISOGEN
278 (NIPPON GENE, Tokyo, Japan) and the SV Total RNA Isolation System (Promega,

279 Madison, WI) was used for RNA extraction. Extracted total RNA samples were
280 sequenced by Illumina Novaseq6000 (Macrogen Japan Corp., Kyoto, Japan).
281 Construction of reference transcription data and estimation of the expression of
282 each transcript
283 The raw RNA-seq data of 30 samples were trimmed by Trimmomatic-version
284 0.36 (Bolger et al., 2014). The trimmed RNA-seq data of each tissue were mapping to
285 the new reference genome with the new gene model (Kawamoto et al., 2019) by Hisat2
286 version 2.1.0. Each mapped data were assembled to transcriptome data by stringtie
287 version 1.3.3 (Pertea et al., 2016). The 30 transcriptome data sets were merged to one
288 transcriptome data set referred to as "a reference transcriptome" by the stringtie.
289 gffcompare v0.10.6 was used (https://ccb.jhu.edu/software/stringtie/gffcompare.shtml)
290 for comparisons with the reference transcriptome and gene set of Kawamoto et al.
291 (2019).
292 To estimate the expression of the reference transcriptome in 30 samples, the
293 raw fastq data of each sample and reference transcript data were used with Kallisto
294 ver0.44.0 (Bray et al., 2016). In comparisons of transcriptome data, the raw RNA-seq
295 data of multiple tissues in *B. mori* strain o751 from the Sequence Read Archive (SRA)
296 and reference transcript data were used: the accession numbers of raw RNA-seq data
297 are DRA005094, DRA005878 and DRA005094 (Kikuchi et al., 2017; Ichino et al.,
298 2018; Kobayashi et al., 2019).
299 We used TIBCO Spotfire Desktop (v7.6.0) software with the "Better World"
300 program license (TIBCO, Inc., Palo Alto, CA; http://spotfire.tibco.com/better-world-
301 donation-program/) for the classification of differentially expressed samples in silkworm
302 tissues in hierarchical clustering using Ward's method.
303 Annotation for the reference transcriptome and functional enrichment analysis
304 Transcoder (v5.5.0) was used to identify coding regions within transcript
305 sequences and convert transcript sequences to amino acid sequences
306 (https://transdecoder.github.io/).
307 Transcriptome sequence sets were compared at the amino acid sequence level by the
308 successive execution of the blastp program in the NCBI BLAST software package
309 (v2.9.0+) with default parameters and an E-value cut-off of 1e-10 (Altschul et al.,1997).
310 Regarding the reference database sets to be blasted, human and fruit fly (*D.*
311 *melanogaster*) protein datasets in the Ensembl database (v.97) were used because the
312 sequences of these organisms were functionally well-annotated and amenable to
313 computational methods, such as a pathway analysis (Tabunoki et al., 2013). The
314 names of top-hit genes in the human and fruit fly datasets were annotated to *B. mori*

transcripts utilizing Ensembl Biomart (Kinsella et al., 2011) and Spotfire Desktop software under TIBCO Spotfire's "Better World" program license (TIBCO Software, Inc., Palo Alto, CA, USA) (https://spotfire.tibco.com/better-world-donation-program/ ). Functional enrichment analyses were performed using the metascape portal site with annotation results against the human gene set (URL: http://metascape.org/gp/index.html#/main/step1, Zhou et al., 2019).

Investigation of gene structures of *sericin*, *fibroin*, and *fhx*

In investigations on the *sericin*, *fibroin*, and *fhx* gene structures, we visualized the positions of the new gene set and reference transcript data in the new reference genome (Kawamoto et al., 2019) using the Integrative genomics viewer (IGV) (James et al., 2011). In the gene model data set, *sericin-1* corresponded to KWMTBOMO06216, *sericin-2* KWMTBOMO06334, *sericin-3* KWMTBOMO06311, *sericin-4* KWMTBOMO06324-06326, fibroin heavy chain (*h-fib*) KWMTBOMO15365, fibroin light chain *l-fib* KWMTBOMO08464, and *fhx* KWMTBOMO01001. The structures of these models were compared visually with our new reference transcriptome data. The several isoforms identified are listed in Table 1. We performed sequence alignment using gene model sequence data and public sequences deposited in the NCBI database.

# Data Availability

The RNA-seq reads supporting the conclusions of this study are available in the SRA with accession number DRA008737 (The accession number of RNA-seq data of each sample is shown in Table 2A).
Assembled transcriptome sequences are available at the Transcriptome Shotgun Assembly (TSA) database under accession IDs ICPK01000001-ICPK01051926.
The estimated abundance of transcripts is available from the Gene Expression Archive (GEA) in DDBJ under accession ID E-GEAD-315.
Supporting data are available in The Life Science Database Archive. The title in the Archive is "KAIKO - Metadata of reference transcriptome data" (DOI:10.18908/lsdba.nbdc02443-000.V001).

# References

Altschul, S.F. et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research 25:3389-3402 (1997). DOI: 10.1093/nar/25.17.3389

348 Bolger, A. M. et al. Trimmomatic: A flexible trimmer for Illumina Sequence Data.
349 Bioinformatics, btu170 (2014). DOI: 10.1093/bioinformatics/btu170
350 Bray, N. L. et al. Near-optimal probabilistic RNA-seq quantification. Nature
351 Biotechnology 34:525–527 (2016). DOI: 10.1038/nbt.3519
352 Daimon, T. et al. Knockout silkworms reveal a dispensable role for juvenile hormones
353 in holometabolous life cycle. Proceedings of the National Academy of Sciences of
354 the United States of America 112 E4226-E4235 (2015). DOI:
355 10.1073/pnas.1506645112
356 Dong, Z. et al. Identification of *Bombyx mori* sericin 4 protein as a new biological
357 adhesive. International Journal of Biological Macromolecules 132:1121-1130 (2019).
358 DOI: 10.1016/j.ijbiomac.2019.03.166
359 Garel, A. et al. Structure and organization of the *Bombyx mori* sericin 1 gene and of the
360 sericins 1 deduced from the sequence of the Ser 1B cDNA. Insect Biochemistry and
361 Molecular Biology 27:469–477 (1997). DOI: 10.1016/S0965-1748(97)00022-2
362 Ichino, F. et al. Construction of a simple evaluation system for the intestinal absorption
363 of an orally administered medicine using *Bombyx mori* larvae. Drug Discoveries and
364 Therapeutics 12:7-15 (2018). DOI: 10.5582/ddt.2018.01004
365 International Silkworm Genome Consortium. The genome of a lepidopteran model
366 insect, the silkworm Bombyx mori. 38(12):1036-45 (2008). DOI:
367 10.1016/j.ibmb.2008.11.004.
368 Ito, K. et al. Deletion of a gene encoding an amino acid transporter in the midgut
369 membrane causes resistance to a *Bombyx* parvo-like virus. Proceedings of the
370 National Academy of Sciences of the United States of America 105 7523-7527
371 (2008). DOI: 10.1073/pnas.0711841105
372 James, T. et al. Integrative genomics viewer. Nature Biotechnology 29:24–26 (2011).
373 DOI: 10.1038/nbt.1754
374 Kawamoto, M. et al. High-quality genome assembly of the silkworm, Bombyx mori.
375 Insect Biochemistry and Molecular Biology 107:53-62 (2019). DOI:
376 10.1016/j.ibmb.2019.02.002
377 Kikuchi, A. et al. Identification of functional enolase genes of the silkworm *Bombyx mori*
378 from public databases with a combination of dry and wet bench processes. BMC
379 Genomics 18: 83 (2017). DOI: 10.1186/s12864-016-3455-y
380 Kinsella, R.J. et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space.
381 Database (Oxford) 2011:bar030 (2011). DOI: 10.1093/database/bar030

382 Kobayashi, Y. et al. Comparative analysis of seven types of superoxide dismutases for
383 their ability to respond to oxidative stress in *Bombyx mori*. Scientific Reports 9, 2170
384 (2019). DOI: 10.1038/s41598-018-38384-8
385 Kondo, Y. et al. Toll ligand Spätzle3 controls melanization in the stripe pattern
386 formation in caterpillars. Proceedings of the National Academy of Sciences of the
387 United States of America 114 8336-8341 (2017). DOI: 10.1073/pnas.1707896114
388 Mita, K. et al. The genome sequence of silkworm, *Bombyx mori.* DNA Research
389 29;11(1):27-35 (2004). DOI:10.1093/dnares/11.1.27
390 Pertea, M. et al. Transcript-level expression analysis of RNA-seq experiments with
391 HISAT, StringTie and Ballgown. Nature Protocols 11:1650-1667 (2016).
392 DOI:10.1038/nprot.2016.095
393 Sakurai, T. et al. A single sex pheromone receptor determines chemical response
394 specificity of sexual behavior in the silkmoth *Bombyx mori*. PLoS Genetics
395 7:e1002115 (2011). DOI 10.1371/journal.pgen.1002115
396 Suetsugu, Y. et al. Large scale full-Length cDNA sequencing reveals a unique
397 genomic landscape in a lepidopteran model insect, *Bombyx mori*. G3 3(9):1481-
398 1492 (2013). DOI:10.1534/g3.113.006239
399 Takiya, S. et al. Regulation of Silk Genes by Hox and Homeodomain Proteins in the
400 Terminal Differentiated Silk Gland of the Silkworm *Bombyx mori*. Journal of
401 Developmental Biology 4(2):19 (2016). DOI:10.3390/jdb4020019
402 Tabunoki, H. et al. Can the silkworm (Bombyx mori) be used as a human disease
403 model? Drug Discoveries and Therapeutics 10:3-8 (2016).
404 DOI:10.5582/ddt.2016.01011
405 Takasu, Y. et al. The silk sericin component with low crystallinity. Sanshi-Konchu
406 Biotec 75:133–139 (2006) DOI: 10.11416/konchubiotec.75.133
407 Takasu, Y. et al. Identification and characterization of a novel sericin gene expressed
408 in the anterior middle silk gland of the silkworm *Bombyx mori*. Insect Biochemistry
409 and Molecular Biology 37:1234–1240 (2007) DOI: 10.1016/j.ibmb.2007.07.009
410 Tamura, T. et al. Germline transformation of the silkworm *Bombyx mori* L. using a
411 piggyBac transposon-derived vector. Nature Biotechnology 18:81–84 (2000)
412 DOI:10.1038/71978
413 Tan, A. et al. Precocious metamorphosis in transgenic silkworms overexpressing
414 juvenile hormone esterase. Proceedings of the National Academy of Sciences of the
415 United States of America. 102:11751-11756. (2005) DOI:
416 10.1073/pnas.0500954102

Tatematsu, K. et al. Construction of a binary transgenic gene expression system for recombinant protein production in the middle silk gland of the silkworm *Bombyx mori*. Transgenic Research, 19:473-487 (2010). DOI:10.1038/71978

Xia, Q. et al. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 10;306(5703):1937-40 (2004). DOI:10.1126/science.1102210

Xia, Q. et al. Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, *Bombyx mori*. Genome Biology 8(8):R162 (2007). DOI:10.1186/gb-2007-8-8-r162

Yamatomo, K. et al. A BAC-based integrated linkage map of the silkworm Bombyx mori. Genome Biology. 9:R21 (2008). DOI:10.1186/gb-2008-9-1-r21

Yukuhiro, K. et al. Insect silks and cocoons: structural and molecular aspects. Extracellular composite matrices in Arthropods (eds. E. Cohen & B. Moussian), pp.515-555. Springer, Cham. (2016) DOI: 10.1007/978-3-319-40740-1_18

Zhou, C. et al. Fine organization of *Bombyx mori* fibroin heavy chain gene. Nucleic Acids Research 28:2413–2419 (2000). DOI: 10.1093/nar/28.12.2413

Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications 10(1):1523 (2019). DOI:10.1038/s41467-019-09234-6

# Funding

This work was supported by the National Bioscience Database Center of the Japan Science and Technology Agency (JST) to HB.

This work was supported by the Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP), “Technologies for Smart Bio-industry and Agriculture” (funding agency: Bio-oriented Technology Research Advancement Institution, NARO) to KY, TT, JS, and HS.

This work was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Research Project for Sericultural Revolution) to KY, TT, and HS.

# Acknowledgments

The computing resource was partly provided by the super computer system at the National Institute of Genetics (NIG), Research Organization of Information and Systems (ROIS), Japan.

# Author contributions

450 Conceived and designed the experiments: K.Y., T.T., and H.S.
451 Performed the experiments: T.T.
452 Contributed reagents/materials/analysis tools: H.S.
453 Analyzed the data: K.Y., J.S., A.J., and H.B.
454 Contributed to the writing of the manuscript under draft version: K.Y., T.T., and H.B.
455 All authors discussed the data and helped with manuscript preparation. K.Y.
456 supervised the project.
457 All authors read and approved the final manuscript.

## 458 Competing interests

459 The authors declare no conflicts of interest.

## 460 Figures

461 Fig. 1 Workflow of the data analysis performed in the present study. To obtain
462 reference transcriptome sequences, Fastq data of 10 tissues from $5^{th}$ instar larvae
463 were mapped to the new reference genome (Kawamoto et al., 2019). Kallisto software
464 was used to estimate the expression abundance of each transcript in these tissues
465 plus other *B. mori* samples of which RNA-seq data were obtained from a public
466 database (Accession numbers are listed in Table 2B). We performed a Blast search
467 against human and *Drosophila* genome data to perform functional annotations of the
468 reference transcriptome. Insect, human, database image, and sequencer drawings
469 (http://togotv.dbcls.jp/ja/pics.html) are licensed at
470 (http://creativecommons.org/licenses/by/4.0/deed.en).

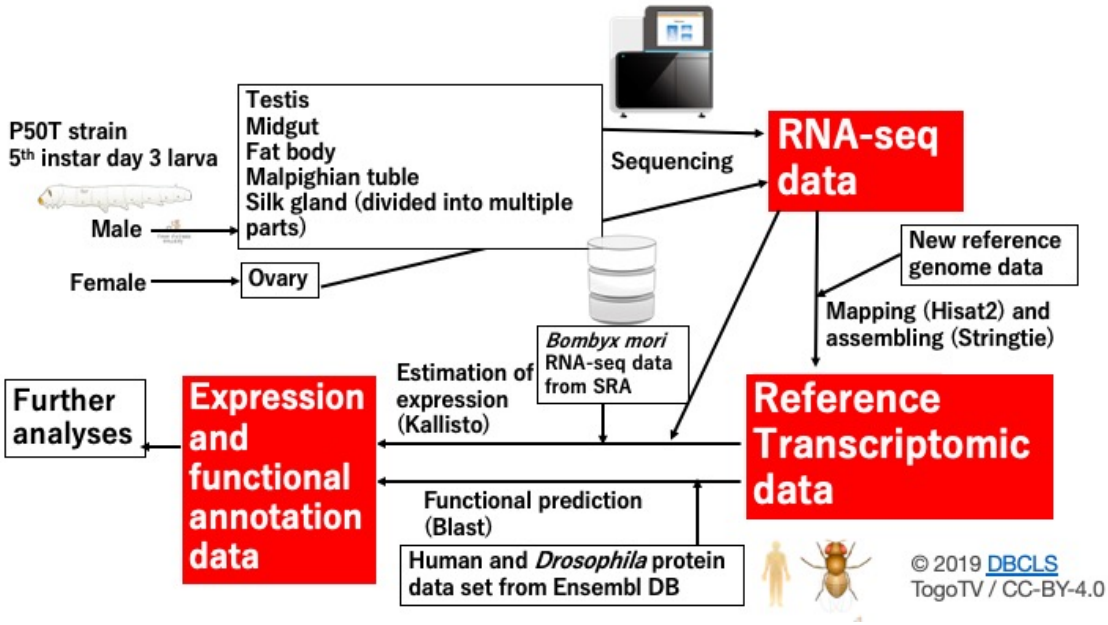



Fig. 2 Basal characteristics of the reference transcriptome. (A) Comparison of gene model data (Kawamoto et al., 2019) and the reference transcriptome data of the present study. The number of loci and transcripts are shown. These numbers were calculated from gff files of the two data sets. (B) Classification of 51,926 transcripts. Each transcript was classified into three categories, and the numbers of the three categories are shown in a pie chart. Definitions of the three categories were described in the main text.

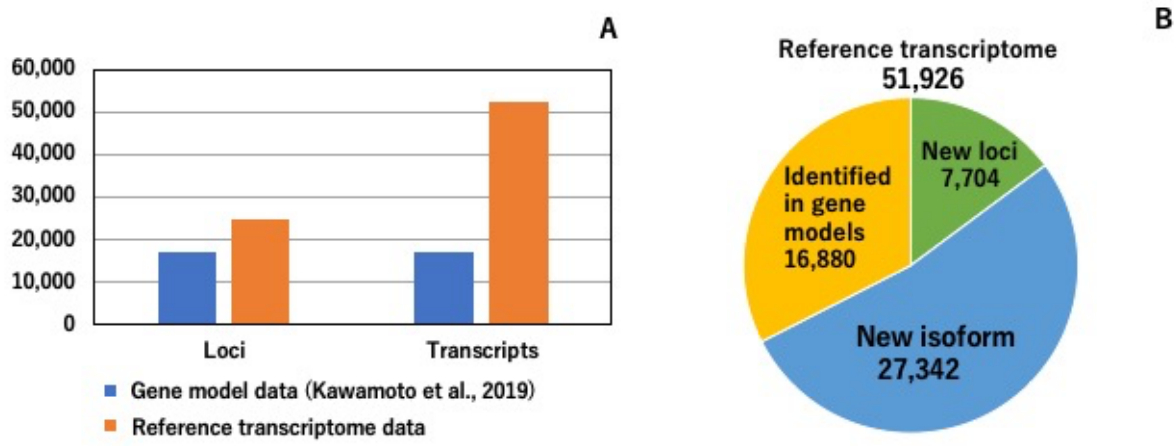



Fig. 3 Hierarchical clustering of expression data in 45 samples. Hierarchical clustering was performed using transcriptome expression data (tpm values). Abbreviations of the samples are shown and described in the main text. The numbers added to the

abbreviations mean biological replicates.

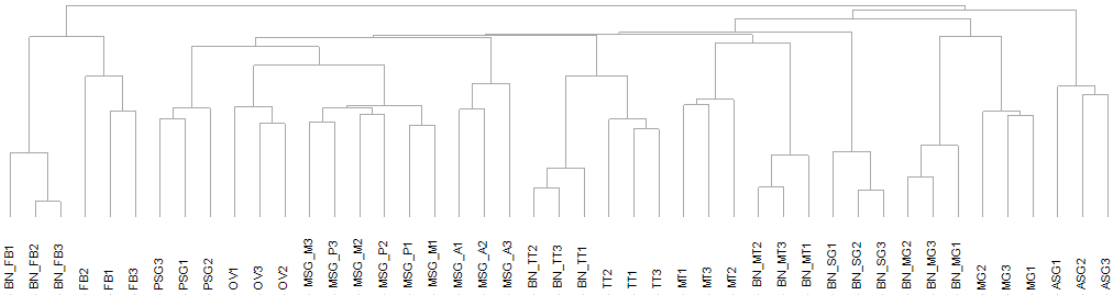
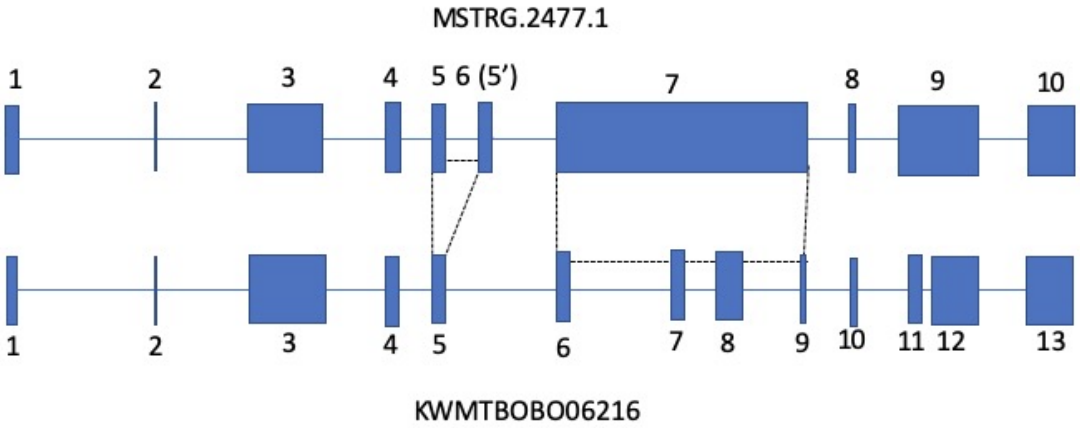
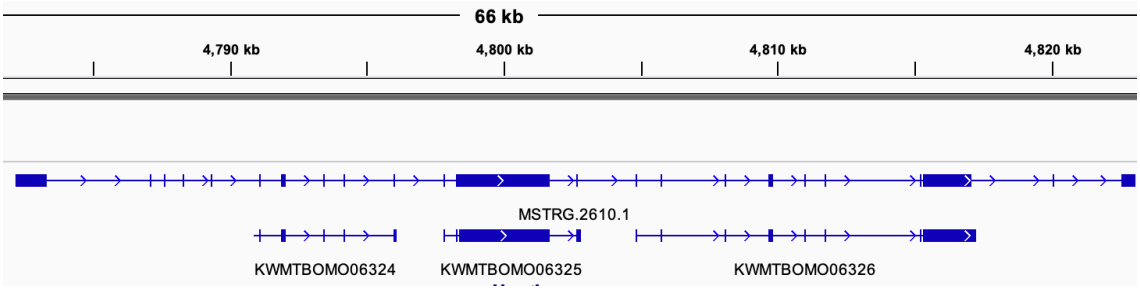


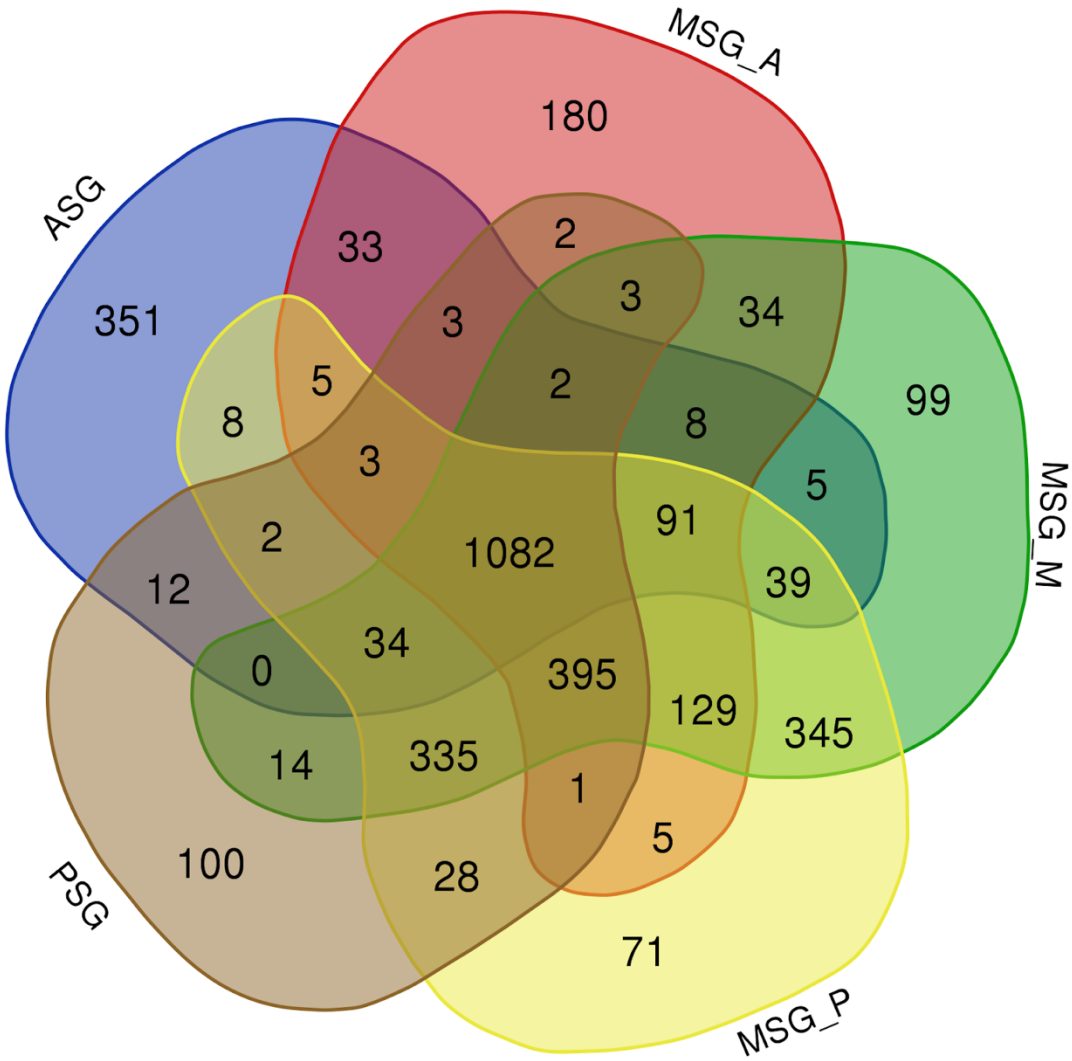

Fig. 4 Longest isoforms of *sericin-1* and *sericin-4* in the reference transcriptome. (A) A schematic drawing showing the exon positions of isoform MTSRG.2477.1 (longest isoform of *sericin-1*) and a gene model of *sericin-1* (Kawamoto et al., 2019). Squares indicate exons with exon numbers in the gff file (Supporting data 1). Exon 6 of MSTRG.2477.1 corresponds to exon 5' and exons 7-10 correspond to exons 6-9 in KWMTBOMO06216 (each group of exons are connected with dashed lines).




(B) Exon positions of isoform MTSRG.2477.1 (the longest isoform of *sericin-4*) and the gene model of *sericin-4* (KWMTBOMO06325, KWMTBOMO06325, and KWMTBOMO06326) in the new reference genome is shown by IGV. The scale above the transcript indicates the location of chromosome 11.

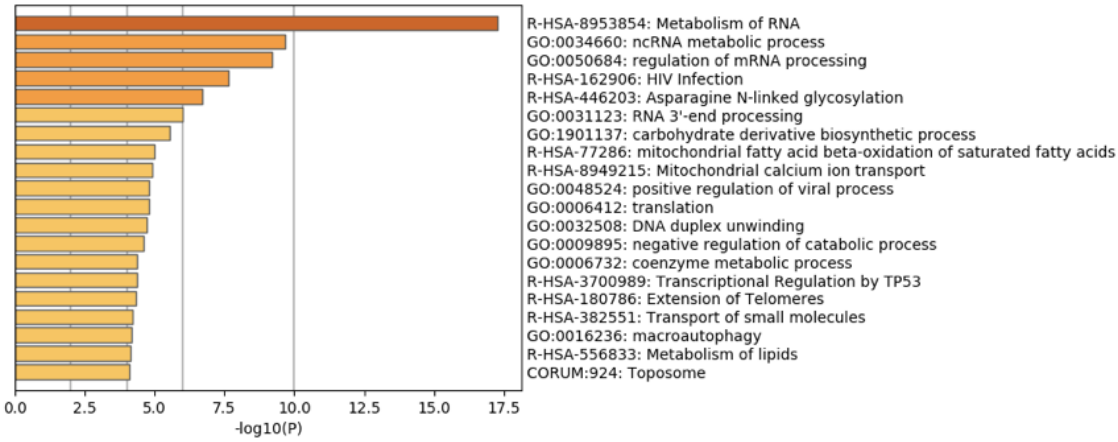


498 Fig. 5 Strongly expressed transcripts within the silk gland. The numbers in the Venn
499 diagrams indicate the number of transcripts of which values of tpm were more than 30
500 in the corresponding silk gland parts.

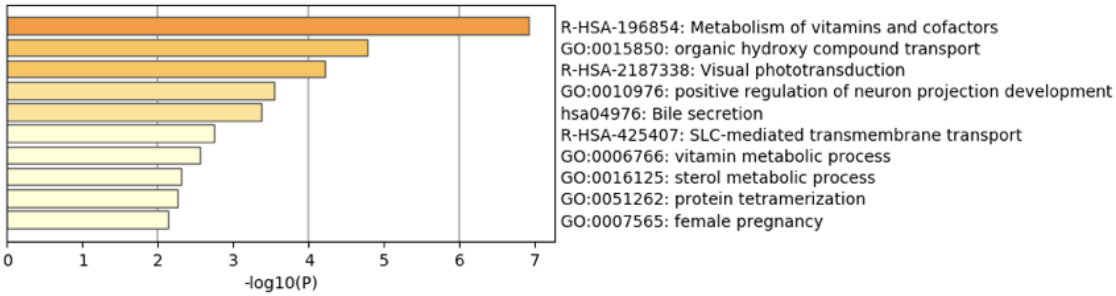



501

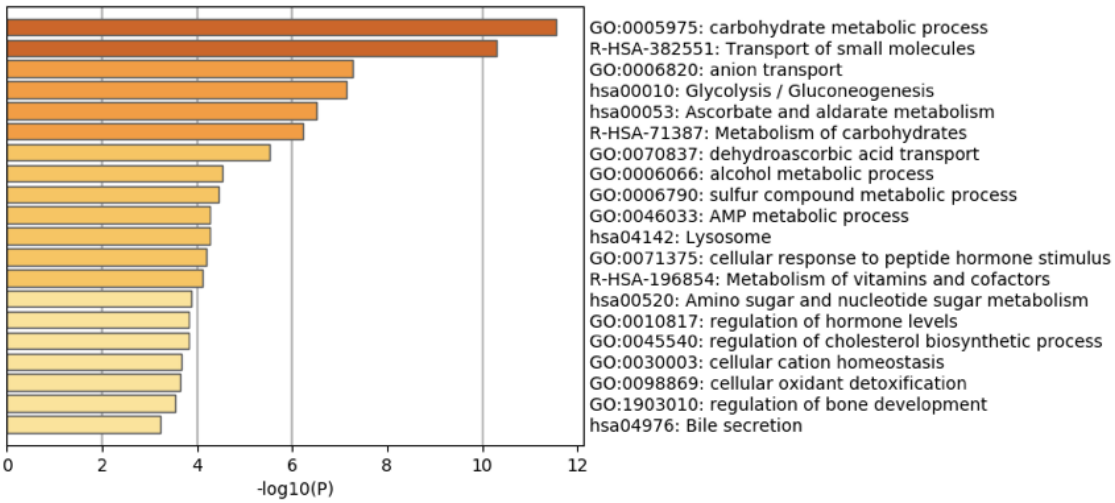
502 Fig. 6 Results of the enrichment analysis by Metascape. Using annotation data against
503 the human gene set of the reference transcripts expressed in one or several parts of
504 the silk gland (Fig. 5), an enrichment analysis was performed (numbers in brackets
505 after the silk gland parts indicate the numbers of transcripts). -log10 (P) represents -
506 log10 (P-value). For example, -log10 (P)=5 represents P-value= $10^{-5}$ (A) Transcripts
507 showing tpm > 30 in MGM_M (99),MGM_P (71), and MGM_M and MGM_P (345).



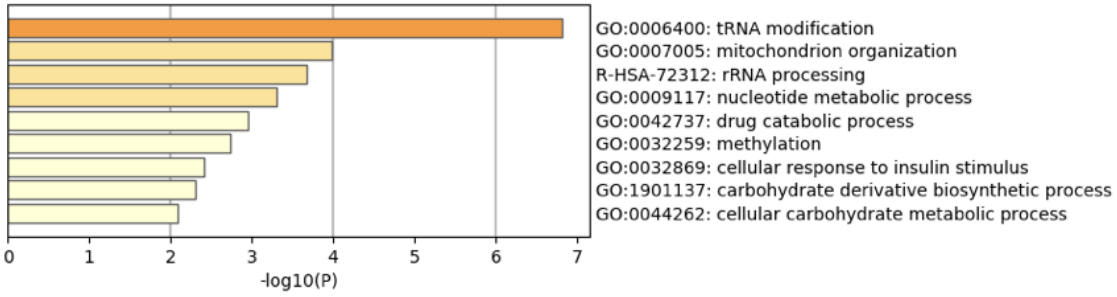

(B) Transcripts showing tpm > 30 in MSG_A (180).




(C) Transcripts showing tpm > 30 in ASG (351).




(D) Transcripts showing tpm > 30 in PSG (100).




# Table

Table 1

*Sericin, fibroin*, and *fibrohexamerin* genes and isoform IDs

| Gene name | Gene model ID in Kawamoto et al., 2019 | Isoform IDs in Supporting data 1 (GenBank IDs) |
|---|---|---|
| *sericin-1* | KWMTBOMO06216 | MSTRG.2477.1 - MSTRG.2477.16, KWMTBOMO06216.mrna1 (ICPK01006046 -ICPK01006062) |
| *sericin-2* | KWMTBOMO06334 | MSTRG.2627.1, MSTRG.2627.2, KWMTBOMO06334.mrna1 (ICPK01006484 -ICPK01006486) |
| *sericin-3* | KWMTBOMO06311 | MSTRG.2595.1 - MSTRG.2595.9, KWMTBOMO06311.mrna1 (ICPK01006421 -ICPK01006430) |
| *sericin-4* | KWMTBOMO06324-06326 | MSTRG.2610.1 (ICPK01006453) |
| fibroin heavy chain (*h-fib*) | KWMTBOMO15365 | MSTRG.14927.1 - MSTRG.14927.23, KWMTBOMO15365.mrna1 (ICPK01035046 -ICPK01035068) |

| | | |
|---|---|---|
| fibroin light chain (*l-fib*) | KWMTBOMO08464 | MSTRG.5511.1, KWMTBOMO08464.mrna1 (ICPK01013031 -ICPK01013032) |
| *fibrohexamerin* | KWMTBOMO01001 | |

519

520 Table 2

521 A. Samples for RNA-seq and run accession IDs

| Sample | SRA Run ID | Sex |
|---|---|---|
| Anterior silk gland (ASG) | DRR186474, DRR186475, DRR186476 | Male |
| Anterior part of the middle silk gland (MSG_A) | DRR186477, DRR186478, DRR186479 | Male |
| Middle part of the middle silk gland (MSG_M) | DRR186480, DRR186481, DRR186482 | Male |
| Posterior part of the middle silk gland (MSG_P) | DRR186483, DRR186484, DRR186485 | Male |
| Posterior silk gland (PSG) | DRR186486, DRR186487, DRR186488 | Male |
| Fat body (FB) | DRR186489, DRR186490, DRR186491 | Male |
| Midgut (MG) | DRR186492, DRR186493, DRR186494 | Male |
| Malpighian tubules (MT) | DRR186495, DRR186496, DRR186497 | Male |
| Testis (TT) | DRR186498, DRR186499, DRR186500 | Male |

| | | |
|---|---|---|
| Ovary (OV) | DRR186501, DRR186502, DRR186503 | Female |

522

523 B. RNA-seq data from SRA

| Sample | SRA Run ID | Reference |
|---|---|---|
| Testis | DRR068893, DRR068894, DRR068895 | Kikuchi et al. 2017 |
| Fat body | DRR095105, DRR095106, DRR095107 | Kobayashi et al. 2019 |
| Midgut | DRR095108, DRR095109, DRR095110 | Ichino et al. 2018 |
| Malpighian tubules | DRR095111, DRR095112, DRR095113 | Kobayashi et al. 2019 |
| Silk gland | DRR095114, DRR095115, DRR095116 | Kobayashi et al. 2019 |

524

525 # Supporting data

526 All supporting data are available in The Life Science Database Archive
527 (https://dbarchive.biosciencedbc.jp/index-e.html).
528
529 Supporting data 1
530 Metadata of reference transcriptome data
531 URL:https://togodb.biosciencedbc.jp/db/kaiko_trascnript_data
532 DOI:10.18908/lsdba.nbdc02443-001.V001
533
534 Supporting data 2
535 Predicted amino acid sequences of the reference transcriptome
536 DOI:10.18908/lsdba.nbdc02443-004.V001
537
538 Supporting data 3
539 Annotations of each transcript (blast against human and *Drosophila* gene sets)

URL:https://togodb.biosciencedbc.jp/db/kaiko_annotation_human_drosophila_data
DOI:10.18908/lsdba.nbdc02443-003.V001

Supporting data 4
Expression data of each transcript in multiple tissues
URL:https://togodb.biosciencedbc.jp/db/kaiko_transcript_tpm_data
DOI:10.18908/lsdba.nbdc02443-002.V001

# Supplemental Figure

Supplemental Fig. 1
Predicted amino acid sequences of the longest *sericin1* isoforms (MSTRG.2477.1.p1 ). Glycine, asparagine and threonine residues are colored in red. The region of exon7 is underlined.
DOI:10.6084/m9.figshare.998056