# Systematic functional annotation workflow for insects

Hidemasa Bono [1,2*], Takuma Sakamoto[3,4], Takeya Kasukawa[5] and Hiroko Tabunoki[3,4]

[1]Laboratory of Bio-DX, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan; bonohu@hiroshima-u.ac.jp

[2]Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan

[3]Institute of Global Innovation Research, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo, 183-8509; tsakamoto@go.tuat.ac.jp (T.S.); h_tabuno@cc.tuat.ac.jp (H.T.)

[4]Department of Science of Biological Production, Graduate School of Agriculture, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo, 183-8509, Japan

[5]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; takeya.kasukawa@riken.jp

*Correspondence: bonohu@hiroshima-u.ac.jp; Tel.: +81-82-424-4013

**Abstract**

**Background:** Next generation sequencing has revolutionized entomological study, rendering it possible to analyze the genomes and transcriptomes of non-model insects. However, use of this technology is often limited to obtaining nucleotide sequences of target or related genes, with many of the acquired sequences remaining unused because other available sequences are not sufficiently annotated.

**Results:** To address this issue, we have developed a functional annotation workflow for transcriptome-sequenced insects to determine transcript descriptions, which represents a significant improvement over the previous method (functional annotation pipeline for insects). The developed workflow attempts to annotate not only the protein sequences obtained from transcriptome analysis but also the ncRNA sequences obtained simultaneously. In addition, the workflow integrates the expression level information obtained from transcriptome sequencing for application as functional annotation information. Using the workflow, functional annotation was performed on the sequences obtained from transcriptome sequencing of stick insect (*Entoria okinawaensis*) and silkworm (*Bombyx mori*), yielding richer functional annotation information than that obtained in our previous study.

**Conclusions:** The improved workflow allows more comprehensive exploitation of transcriptome data and is applicable to other insects because the workflow has been openly developed on GitHub.

**Keywords:** Functional annotation, RNA sequencing, Transcriptome assembly, Stick insect, Silkworm

## Background

Genome sequencing is becoming common practice in insect research. As of March 2022, the genomes of nearly 2,800 insect species have been decoded and registered in the genome section of NCBI Datasets (https://www.ncbi.nlm.nih.gov/datasets/genomes/?taxon=50557). In addition, long-read technology is further accelerating the pace of insect genome sequencing [1].

However, some insects have genomes larger than that of humans, further complicating the difficult process of whole genome sequencing. As an alternative, transcriptome sequencing using next generation sequencing technology, also termed RNA sequencing (RNA-Seq), provides a powerful tool for evaluating non-model species such as large-genome-size insects [2]. In particular, this strategy can efficiently identify tens of thousands of possible genes in a specific tissue by assembling tens of millions of reads. The sequences are then assembled for the identification of transcriptional units. Systematic sequence similarity searches against non-redundant sequence sets and BLAST2GO searches [3] are often employed for the functional inference of such transcriptomes to be utilized in gene set enrichment analyses. Moreover, a new method termed Seq2Fun that does not require transcriptome *de novo* assembly was recently developed for rapid functional profiling of RNA-Seq data for non-model organisms [4]. Nevertheless, the output of such analyses is reliant upon the comprehensiveness of the available datasets and their functional annotation.

Traditionally, reference data have been maintained as research community databases (DBs) such as FlyBase for fruit flies [5]. In turn, such communities spearheaded the development of Gene Ontology (GO) [6], which provides machine-interpretable annotations incorporating human-readable descriptions. These annotations for genomic research, including GO, are provided by Ensembl and cover a variety of vertebrates [7]. For insects, Ensembl Metazoa in Ensembl Genomes integrates the community annotations of individual insects [8].

77 Nevertheless, it is unable to keep pace with the increase of insect genome sequencing including
78 currently available genome and transcriptome data.

79 Concurrently, we developed a functional annotation pipeline in non-model insects for
80 microarray-based transcriptome analysis of a unique mutant of the silkworm *Bombyx mori op*
81 [9]. Subsequently, we also utilized this pipeline for comparative studies with humans and flies
82 [10,11]. In particular, by assigning human gene identifiers (IDs) to the silkworm genes, we
83 could facilitate the reconstruction of pathway DBs, which were originally developed for the
84 analysis of model organisms [9,12]. However, such prior studies have almost exclusively been
85 limited to protein coding sequences, as functional annotation pipelines for transcriptome
86 sequences were originally designed for mammalian transcriptomes [13] produced by
87 sequencing cDNAs including expressed sequence tags (ESTs) using Sanger sequencers in the
88 Functional Annotation of Mammalian genomes (FANTOM) project [14,15]. Although
89 functional annotation for non-coding RNAs (ncRNA) was considered in the pipeline for
90 FANTOM3 [16], this feature was ultimately not incorporated because the reference
91 information and biological knowledge regarding ncRNAs were not yet sufficient [17].

92 As transcriptome sequencing becomes popular, many groups run pipelines of their own,
93 with the information regarding the transcription units from various studies being reported on a
94 study-by-study basis. Moreover, the transcript assemblies are often not registered in a public
95 DB even though it is recommended that such data be deposited in the Transcriptome Shotgun
96 Assembly (TSA) sequence database under the International Nucleotide Sequence Database
97 Collaboration (INSDC) [18,19], as the data need to be annotated uniformly to facilitate reuse in
98 other analyses. However, the functional information of the currently registered transcription
99 units is not sufficiently rich to support functional analysis and almost never annotated once
100 archived even if it is available. Thus, the functional annotation of transcriptomes in a manner

101 that can be integrated with datasets from other groups including the data in public DBs is
102 urgently required.

103 Therefore, in this study we evaluated the functional annotation of transcripts from
104 assembled transcriptomes to establish a functional annotation workflow for insect
105 transcriptomes by updating the previously developed functional annotation pipeline for
106 silkworm. Then, we developed a systematic functional annotation workflow for insect research
107 termed "Fanflow4Insects". Fanflow4Insects enables the automatic annotation of sequences to
108 exploit the exponential increase in genome and transcriptome information. Finally,
109 Fanflow4Insects was tested using not only the reference transcriptome of silkworm (*Bombyx*
110 *mori*) but also transcriptomes of the Japanese stick insect (*Entoria okinawaensis*) including the
111 newly sequenced transcriptome of the fat body.

112

113

114 **Results**

115 **Functional annotation workflow for insects**

116 We first evaluated rules for the functional annotation of transcripts. In our previous studies, we
117 only focused on translated protein sequences [11,20]. In the present study, we also aimed to
118 annotate sequences without coding regions using ncRNA references, and quantified expression
119 information from RNA-Seq as well as protein coding sequences ascertained based on sequence
120 similarity. For transcripts with no sequence similarity, quantified expression values were used
121 for annotation (Fig. 1). The functional annotation workflow described below, being developed
122 especially for insects, was termed "Fanflow4Insects".

123 *Functional annotation of coding sequences*

124 Using predicted protein sequences translated from each assembled transcriptome, we
125 performed systematic sequence similarity searches against the predicted protein sequence

126 datasets of functionally well-annotated organisms, including human (*Homo sapiens*), mouse
127 (*Mus musculus*), nematode worm (*Caenorhabditis elegans*), and fruit fly (*Drosophila*
128 *melanogaster*). Because local alignment is not appropriate for evaluating the overall
129 percentage similarity between protein sequences, we adopted global alignment as an
130 alternative as insects are far distant from well-annotated reference organisms. The program
131 GGSEARCH in the FASTA package (https://fasta.bioch.virginia.edu/) was used for
132 implementation of the global alignment search, and organism specific homology searches by
133 GGSEARCH were executed against the four functionally well-annotated organisms described
134 above followed by a search against UniProtKB [21]. A corresponding functional description of
135 the top hit in the reference organism or UniProtKB was annotated as "homolog" (Fig. 2).

136 If no homologs could be found, the search result of HMMSCAN against Pfam was
137 referenced [22] (Fig. 2). According to the protein domain hits, protein domain information was
138 annotated if present (example: "zinc finger containing protein"); else, just "hypothetical
139 protein" was used. The sources of the reference DBs used in the functional annotation are
140 summarized in Table 1.

141 **Table 1** Source of reference databases.

| **Category** | **Name of resource** | **URL** |
|---|---|---|
| Protein sequences | Ensembl | https://ftp.ensembl.org/pub/release-105/fasta/homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz[1] |
| | UniProtKB | https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz |
| Non-coding RNA | Ensembl | http://ftp.ensembl.org/pub/release-105/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz[1] |

| | | |
|---|---|---|
| sequences | EnsemblGenomes | http://ftp.ensemblgenomes.org/pub/release-52/metazoa/fasta/drosophila_melanogaster/ncrna/Drosophila_melanogaster.BDGP6.32.ncrna.fa.gz[1] |
| Protein and RNA domain | Pfam | http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam35.0/Pfam-A.hmm.gz |
| | Rfam | http://ftp.ebi.ac.uk/pub/databases/Rfam/14.7/Rfam.cm.gz |

[1] Only the URL for the reference data of the typical organism is listed.

*Functional annotation of non-coding sequences*

For sequences in which TransDecoder was unable to find any protein coding region (i.e., > 100 amino acids (aa) in length by default), functional information of the ncRNAs were annotated based on nucleotide Basic Local Alignment Search Tool (BLAST) search against ncRNA reference sequences (Fig. 2). For insect functional annotation, reference ncRNA sequences of the four functionally well-annotated organisms (*H. sapiens*, *M. musculus*, *C. elegans*, and *D. melanogaster*) were used in Fanflow4Insects (Table 1).

If no functional information could be assigned, then the RNA structure and sequence similarity against Rfam [23] were assessed using the CMSCAN program in the Infernal package [24] (Fig. 2). If any information regarding the sequence was located, the RNA-related information in DB hits was annotated; else, the sequence was annotated as an "unclassifiable transcript".

*Functional annotation by quantified expression information*

Transcript expression was quantified based on the assembled transcriptome and RNA-Seq raw reads. The matrix of gene expression was generated by pasting columns. Using the expression matrix, we were able to annotate gene expression features for all genes.

161 Although the threshold for the judgment remains to be calibrated, genes with specific
162 expression were then annotated (Fig. 3). If a transcript is expressed in all tissues, it was
163 annotated as "constitutive expression". Alternatively, if expressed only in e.g. the fat body, it
164 was annotated as "fat body-specific expression". Practical examples of annotation based on
165 expression information are described in the following section.
166

167 **Case study for Fanflow4Insects**

168 Practical applications of Fanflow4Insects for the Japanese stick insect *Entoria okinawaensis*
169 and silkworm *Bombyx mori* are described below.
170

171 *Functional annotation of the Japanese stick insect transcriptome using Fanflow4Insects*

172 In addition to three biological replicates of an *E. okinawaensis* midgut sample, which we
173 previously reported and deposited in SRA under DRA007226 (DRR148118, DRR148119, and
174 DRR148120) [20], three biological replicate reads of the *E. okinawaensis* fat body
175 (DRR346171, DRR346172 and DRR346173) were used for functional annotation via
176 Fanflow4Insects in the present study.

177 Of 311,357 transcripts, 18.5% (57,732 transcripts) could be translated into 68,162
178 predicted protein sequences with length > 100 aa. These protein coding sequences were
179 annotated based on systematic sequence similarity and protein domain searches (Table 2). In
180 brief, 55,068 predicted protein sequences (80.8%) had homologs, among which 58,876
181 (86.4%) had some sequence similarity information including protein domains. A total of 9,286
182 predicted protein sequences could not be assigned to any functional information based on
183 sequence similarity and domain searches at the protein level and thus were annotated as a
184 "hypothetical protein".
185

186 **Table 2** Protein-level annotation for *E. okinawaensis*.

| Annotation category | Annotation level | Number of annotations | Cumulative number | Percentage |
|---|---|---|---|---|
| Protein homolog based on the top hit | Human or mouse homolog | 44,351 | 44,351 | 65.1 |
| | *C. elegans* homolog | 3,739 | 48,090 | 70.6 |
| | *D. melanogaster* homolog | 2,349 | 50,439 | 74.0 |
| | Homolog found in UniProtKB | 4,629 | 55,068 | 80.8 |
| No protein homolog | Protein domain | 3,808 | 58,876 | 86.4 |
| | Hypothetical protein | 9,286 | 68,162 | 100 |

187

188 The remaining 253,625 sequences comprised transcripts without protein coding regions. Of
189 these, only 185 sequences had sequence similarity to known ncRNA sequences. Moreover, 264
190 sequences could be assigned based on a CMSCAN search against Rfam DB, whereas 253,176
191 sequences were annotated as an "unclassifiable transcript". These functional annotation results
192 from sequence information for *E. okinawaensis* are summarized graphically in Fig. S1.

193 For the hypothetical protein and the unclassifiable transcripts with no sequence similarity,
194 quantified expression values were used for annotation. Functional information could be
195 inferred and annotated based on the expression matrix generated from sequence counts (Fig. 1).
196 Table 3 summarizes the numbers of sequences annotated as "hypothetical protein" and
197 "unclassifiable transcript" and the corresponding expression annotations. Only a small fraction
198 of sequences, 2.7% (156) of "hypothetical proteins" and 1.5% (3,860) of "unclassifiable
199 transcripts", had no expression information (Table 3).

200 **Table 3** Functional annotation based on expression for *E. okinawaensis*.

| Annotation level | All (311,357) | Hypothetical protein | Unclassifiable transcript |
|---|---|---|---|

| | | (5,699) | (253,176) |
|---|---|---|---|
| Fat body-specific expression | 48,675 | 622 | 42,790 |
| Midgut-specific expression | 28,918 | 315 | 25,005 |
| Constitutive expression | 228,103 | 4,606 | 181,521 |
| Not expressed | 5,661 | 156 | 3,860 |

201

202 The assembled transcriptomes and the estimated transcript abundances were deposited to
203 the TSA Sequence database under accession IDs: ICSG01000001–ICSG01311357 and the
204 Genomic Expression Archive (GEA) under accession ID: E-GEAD-476. Full functional
205 annotation and transcript descriptions generated by Fanflow4Insects were deposited to figshare
206 (DOI: 10.6084/m9.figshare.19368110) to be reused in stick insect research.

207

208 *Functional annotation of silkworm using Fanflow4Insects*

209 We also evaluated the reference transcriptome data of silkworm *B. mori* as a lepidopteran
210 insect with biological and industrial importance [11]. Corresponding functional annotation is
211 available from the Life Science Database Archive (DOI: 10.18908/lsdba.nbdc02443-002);
212 however, only BLAST top hits for human and fly are listed. We therefore applied
213 Fanflow4Insects to silkworm using the assembled transcriptomes (TSA IDs: ICPK01000001–
214 ICSG01051926) and estimated abundance of transcripts (GEA ID: E-GEAD-315). Of 51,926
215 transcripts, 78.1% (40,555 transcripts) could be translated into 45,719 predicted protein
216 sequences of length > 100 aa. These protein coding sequences were annotated via systematic
217 sequence similarity and protein domain searches, with 38,107 predicted protein sequences
218 (83.4%) able to be annotated with homologs (Table 4). Moreover, protein domains could be
219 assigned to 5,398 predicted protein sequences for the remaining 7,612 predicted protein

220 sequences without homologs, resulting in only 2,214 predicted protein sequences (4.8%) being
221 annotated as a "hypothetical protein" (Table 4).

222 **Table 4** Protein-level annotation for *B. mori*.

| Annotation category | Annotation level | Number of annotations | Cumulative number | Percentage |
|---|---|---|---|---|
| Protein homolog from the top hit | Human or mouse homolog | 31,354 | 31,354 | 68.6 |
| | *C. elegans* homolog | 1,752 | 33,106 | 72.4 |
| | *D. melanogaster* homolog | 2,113 | 35,219 | 77.0 |
| | Homolog found in UniProtKB | 2,888 | 38,107 | 83.4 |
| No protein homolog | Protein domain | 5,398 | 43,505 | 95.2 |
| | Hypothetical protein | 2,214 | 45,719 | 100 |

223

224 The remaining 11,371 sequences comprised transcripts without protein coding regions. Of
225 these, only 45 sequences had sequence similarity to known ncRNA sequenced. Moreover, 300
226 sequences could be assigned based on CMSCAN search against Rfam DB. Finally, 11,026
227 sequences were annotated as an "unclassifiable transcript". These functional annotation results
228 from sequence information for *B. mori* are summarized graphically in Fig. S2.

229 The estimated abundances of transcripts in the expression matrix (GEA ID: E-GEAD-315)
230 were also used for annotation. Table 5 summarizes the numbers of sequences annotated as
231 "hypothetical protein" and "unclassifiable transcript", together with the corresponding
232 expression annotations. Unlike the results for *E. okinawaensis* in which data from only two
233 tissues were available, the abundance of transcripts in six tissues, notated as transcripts per
234 million (TPM), was available for *B. mori*. A small number of sequences were judged to be

235 tissue-specific because only the genes with the expression in specific tissues (TPM > 0) and no
236 expression (TPM = 0) in other tissues were considered to be tissue specific. In summary, 3.0%
237 (44) of 1,486 transcripts carrying the "hypothetical protein" annotation and 5.4% (693) of
238 12,845 annotated as an "unclassifiable transcript" could be annotated with strong expression
239 features, although the function of those transcripts could not be inferred based only on
240 sequence similarity information (Table 5).

241

242 **Table 5** Functional annotation of *B. mori* transcripts based on expression.

| Annotation level | All (51,927) | Hypothetical protein (1,486) | Unclassifiable transcript (12,845) |
|---|---|---|---|
| Fat body-specific expression | 39 | 1 | 11 |
| Midgut-specific expression | 108 | 0 | 29 |
| Malpighian tubule-specific expression | 83 | 3 | 14 |
| Silk gland-specific expression | 365 | 12 | 92 |
| Testis-specific expression | 861 | 19 | 492 |
| Ovary-specific expression | 179 | 9 | 55 |
| Constitutive expression | 7,825 | 114 | 1,205 |
| No expression | 609 | 24 | 125 |

243

244 The functional annotation described above including the transcript descriptions generated
245 by Fanflow4Insects along with the expression annotation has been deposited to figshare (DOI:
246 10.6084/m9.figshare.19368137) for use in silkworm research.

247

248 *Comparison of ncRNAs between Japanese stick insect and silkworm*

249 A comparative analysis of ncRNA transcripts was performed to utilize the output of
250 Fanflow4Insects. Notably, the functional annotation to ncRNA references (human and fruit
251 fly) by Fanflow4Insects provided transcript IDs for *E. okinawaensis* and *B. mori* along with
252 corresponding ncRNA IDs from reference organisms for use in this comparison (Fig. 4).

253 Whereas few transcripts were functionally annotated based on existing ncRNAs in the
254 public database, several functionally annotated transcripts were identified in the intersection of
255 *E. okinawaensis* and *B. mori*; the corresponding transcript IDs in the reference organisms
256 (human and fruit fly) are listed in Table 6. We also evaluated the corresponding functional
257 annotation derived from the expression of these transcripts, revealing that most transcripts
258 exhibit constitutive expression whereas several show midgut-specific expression in *E.*
259 *okinawaensis*; in contrast, all transcripts are constitutively expressed in *B. mori*.

260 **Table 6** ncRNA transcripts expressed both in *E. okinawaensis* and *B. mori*.

| | **Transcript ID** | **Gene name** | **Gene description** |
|---|---|---|---|
| Annotation from fruit fly | FBtr0100888 | mt:lrRNA | mitochondrial large ribosomal RNA |
| | FBtr0345722 | asRNA:CR45330 | antisense RNA:CR45330 |
| | FBtr0346876 | 28SrRNA:CR45837 | 28S ribosomal RNA:CR45837 |
| | FBtr0346877 | pre-rRNA:CR45846 | ribosomal RNA primary transcript:CR45846 |
| | FBtr0346881 | pre-rRNA:CR45847 | ribosomal RNA primary transcript:CR45847 |
| | FBtr0346882 | 18SrRNA:CR45841 | 18S ribosomal RNA:CR45841 |
| Annotation from | ENST00000450451 | | novel transcript |
| | ENST00000501016 | | novel transcript |

| | | | |
|---|---|---|---|
| human | ENST00000518947 | HOXA-AS3 | HOXA cluster antisense RNA 3 [Source:HGNC Symbol;Acc:HGNC:43748] |
| | ENST00000547387 | | novel transcript, antisense to TUBA1B |
| | ENST00000618978 | U2 | U2 spliceosomal RNA [Source:RFAM;Acc:RF00004] |
| | ENST00000623543 | | novel transcript, antisense to TUBA8 |
| | ENST00000631211 | | novel transcript, similar to YY1 associated myogenesis RNA 1 YAM1 |
| | ENST00000638356 | | novel transcript, antisense to ATP4A |

## Discussion

Next-generation sequencing has facilitated decoding of the genomes of numerous insects along with their transcript sequences. However, the biological interpretation of these sequences remains as a primary bottleneck of transcriptome analysis. An important first step is functional annotation, which serves as e.g. an important clue for selecting genome editing targets. Toward this end, we have had a long-term interest in providing functional annotation, starting with the annotation of enzymatic genes using the Gene Function Identification Tool (GFIT) in bacteria [25]. More recently, we have contributed to the functional annotation pipeline in the FANTOM project to provide the scientific community with accurate functional annotation of mouse cDNA [14–16].

Currently, BLAST2GO is often used for functional annotation of non-model organisms [3]; however, BLAST2GO is a commercial application and considerable delay exists before results are obtained. Recently, Seq2fun has been used to predict the function of the total set of

genes encoded in a genome at very high speed [4]. However, Seq2fun does not generate sequences for each gene, making it unusable for applications such as individual transcriptome analysis. Thus, it is not possible to trace how the functional annotation was assigned to each gene using Seq2fun. We consider that traceable functional annotation supported by concrete evidence is crucial in biological studies. In addition, these tools cannot be customized to the circumstances of each organism. To address these issues, we therefore developed Fanflow4Insects.

In our previous study for annotating the *E. okinawaensis* midgut transcriptome, we performed a systematic sequence similarity search using translated protein sequences against the predicted protein sequences of well-annotated model organisms using a locally installed NCBI protein BLAST program [20], which provided sufficient speeds to obtain results in a reasonable time. However, although BLAST reports the local alignment of the query sequence and DB hit sequence [30], the local alignment is not appropriate for the present purposes because overall sequence identity and sequence similarity between protein sequences cannot be determined. Therefore, in the present study we re-evaluated the program to be applied for the project. The program FASTY in the FASTA package (https://fasta.bioch.virginia.edu/), previously utilized in the FANTOM project [13,17], compares a DNA sequence to a protein sequence DB, translating the DNA sequence in three forward (or reverse) frames and allowing frameshifts. However, despite dramatic improvements in computation technology, FASTY still requires considerable computing power and is difficult to be used for practical functional annotation. Considering that next-generation sequencing technology using "sequence by synthesis" with Illumina sequencers is much more accurate as compared to capillary sequencing using the Sanger method, and nucleotide deletions are less likely to occur in the final sequences, we proposed to utilize global rather than local alignment in the present study. Global alignment is also more appropriate because several well-conserved alignment

301 fragments are apparent upon the alignment of insects and model organisms while insects are far
302 distant from the well-annotated reference organisms. We found that the GGSEARCH program
303 in the FASTA package is most suitable for the implementation of global alignment searches in
304 a local installation environment. In particular, GGSEARCH calculates sequence identity and
305 similarity, which allow biologists to intuitively evaluate the homology between two sequences,
306 using global alignment while providing similar speed to BLAST. However, because of the
307 evolutionary distance between insects and well-annotated model organisms, we set a lower
308 E-value threshold of 0.1 in the systematic sequence similarity search by GGSEARCH against
309 genes from well-annotated organisms (human, mouse, *C. elegans*, and *D. melanogaster*)
310 followed by UniProtKB [21] (Fig. 2).

311 In practice, more than 80% of protein sequences for *E. okinawaensis* and *B. mori* could be
312 assigned homologs using GGSEARCH (Table 2 and 4). In our previous study, we showed that
313 approximately 58% of *B. mori* genes had human homologs [10]. In the present study, we
314 showed that 68.6% of *B. mori* protein sequences could be assigned human or mouse homologs
315 (Table 4). We believe that the increase in percentage directly reflects advances in our
316 knowledge of reference sequences.

317 Although many DBs have been developed for protein domains, Pfam [22], which describes
318 protein domains using the Hidden Markov Model (HMM), is often used. We therefore included
319 the program HMMSCAN in the HMMER package (https://hmmer.org/), available as a
320 program to search Pfam, in the workflow. Notably, we obtained functional information for
321 numerous sequences for which no sequence similarity to the reference protein sequence was
322 identified using protein domain assignment. In particular, some protein domain information
323 could be assigned for 3,808 sequences in *E. okinawaensis* and 5,398 sequences in *B. mori*
324 (Table 2 and 4), which allows the function of the sequence to be inferred despite the lack of
325 similarity to a known protein sequence. Overall, this strategy allowed 95.2% of protein

326 sequences in *B. mori* to be annotated with only 4.8% (2,214 sequences) being left unannotated.
327 These findings indicate that sequence similarity and protein domain searches with proper
328 reference databases (Table 1) can serve as a powerful tool for functional annotation.

329 The accumulation of ncRNA reference sequences in model organisms facilitates
330 sequence-level comparisons; thus, we incorporated ncRNAs incorporated into the workflow.
331 Toward this end, we utilized nucleotide BLAST, as sequences are not as conserved at the
332 nucleotide sequence level as at the protein sequence level. Therefore, although few
333 assignments were obtained, they were useful to infer functionality. In particular, knowledge
334 regarding RNA domains has been accumulated in a DB called Rfam [23], with RNA domains
335 able to be searched using the program CMSCAN in the Infernal package [24]; thus, this
336 program was also incorporated into the workflow. We anticipate that although only a few
337 sequences were annotated from ncRNA DB and Rfam for *E. okinawaensis* and *B. mori* (Figs.
338 S1 and S2), additional sequences will be annotated in the future as ncRNA reference sequences
339 and RNA domain information become more complete.

340 Performing functional annotation based on gene expression is challenging, encompassing
341 numerous problems such as how to determine the threshold and the inability to detect weak
342 expression unless the RNA-Seq reads are sufficiently deep. However, gene expression can
343 provide useful information for inferring the function of the sequence. Toward this end,
344 information from derived libraries has been previously used in the analysis of ESTs [14–16].
345 More recently such information has been replaced by quantified expression values derived
346 from RNA-Seq, which can be used for functional annotation together with the level of
347 expression in other tissues. Accordingly, we also integrated quantified expression values for
348 providing functional annotation. In the present study, sequences that were expressed only in a
349 specific tissue were annotated as showing tissue-specific expression (Table 3 and 5). Notably,
350 as quantified expression data were available for only two tissues in *E. okinawaensis*, multiple

351 sequences were judged as tissue specific. This allowed us to add functional annotation based on
352 expression to sequences previously annotated as "hypothetical protein" or "unclassifiable
353 transcript" (Table 3). In comparison, as data for six tissues were available in *B. mori* only few
354 sequences exhibited tissue specific expression and could accordingly be assigned functional
355 annotation (Table 5). Moreover, the functional annotation to ncRNA references in
356 Fanflow4Insects allowed the comparative analysis of ncRNA transcripts (Fig. 4). In
357 conjunction with the expression-based functional annotation, the comparative use of
358 Fanflow4Insects data could therefore readily identify candidate genes for future evaluation of
359 e.g. insects with distinct phenotypes (Table 6). These results indicated that even sequences for
360 which no information was obtained from sequence similarity were able to provide useful
361 functional information.

362 For enrichment analysis, it is especially important to render the data not only
363 human-readable with regard to transcript description but also machine-readable using GO and
364 other functional annotations. Although originally developed by fly geneticists, GO is now
365 widely used in various model organisms [6]. In our previous study for annotating the *E.*
366 *okinawaensis* midgut transcriptome, enrichment analyses were performed by assigning
367 transcripts of *E. okinawaensis* to those of *D. melanogaster* because no original GO annotations
368 were available for *E. okinawaensis* [20]. The next target is to annotate GO terms to newly
369 sequenced organism and utilize those for enrichment analysis. Currently, GO annotation
370 applies only for protein coding sequences; however, we speculate that annotation using a
371 controlled vocabulary including that of GO will be incorporated for functional annotation of
372 both proteins and transcripts because it affords machine-readability. As the genome sequences
373 of more organisms become available, functional annotation of the genes encoded therein will
374 be required together with GO and protein domain annotation to e.g. play an active role in
375 selecting which genes to edit for genome editing.

Following the Findable, Accessible, Interoperable, Re-usable (FAIR) principle [26], raw RNA-Seq data and hopefully assembled transcriptome sequences should be registered in the public DB as well as to reproduce data analysis. Our data regarding RNA-Seq reads, transcriptome assembly, and quantified expression values are archived in the public DBs Sequence Read Archive (SRA), TSA, and GEA, respectively. A workflow that integrates these elements is very important and therefore should be re-usable with proper interoperability. All scripts to run the program described in the present study are available on Github: https://github.com/bonohu/SAQE/. The program is not fully automated as it is expected to be customized for actual use. In particular, gene expression information will vary with each insect species and data acquisition situation. Moreover, the calculation of correspondence relationships with closely related insects for which sequence information is already available will also vary depending on the situation. Such construction of individual workflows according to the available information represents the key to digital transformation.

## Conclusions

We have developed a functional annotation workflow for insects, termed “Fanflow4Insects”. Fanflow4Insects has been openly developed on GitHub, and is freely accessible. Fanflow4Insects was tested for functional annotation of the Japanese stick insect *E. okinawaensis* and silkworm *B. mori*. In conjunction with functional annotation derived from expression, the data from Fanflow4Insects can be applicable to the comparative study of insects with distinct phenotypes.

## Methods

401 **Programs comprising the workflow**

402 All scripts that form the workflow described below are freely accessible from the Systematic
403 Analysis for Quantification of Everything (SAQE) code repository at GitHub
404 (https://github.com/bonohu/SAQE) [27] along with the detailed parameters.

405 For transcriptomes without reference genome sequence, we utilized Trinity (v2.12.0) with
406 paired-end parameters in the Docker environment (trinityrnaseq/trinityrnaseq:2.12.0) to
407 construct *de novo* transcriptome assemblies [28].

408 Protein coding sequences within transcriptome sequences produced by *de novo* assembly
409 or by mapping the reference genome sequence were predicted using TransDecoder (v5.2.0)
410 [29]. The predicted and then translated protein sequences were subsequently compared via
411 successive execution of the GGSEARCH program (ggsearch36; v36.3.8g) with '-d1 -m 10 -E
412 0.1' parameters in the FASTA package (https://fasta.bioch.virginia.edu/) against the protein
413 reference sequence DBs described in Table 1. The protein domain search program (hmmscan)
414 in the HMMER package (https://hmmer.org/) was used for the predicted protein sequences.
415 Pfam (v35.0) was used as the protein domain DB [22].

416 The transcriptome sequences were then used for annotation as ncRNA sequences.
417 Nucleotide BLAST (blastn; v2.6.0+) [30] was used to detect sequence similarity to known
418 ncRNA sequences in Ensembl (human) [7] and Ensembl Genomes (fly) [8] DBs. The program
419 (cmscan) in the Infernal: inference of the RNA alignments package (v1.1.4) [24], which was
420 developed for searching nucleotide sequence DBs for RNA structure and sequence similarities,
421 was used to annotate transcript sequences against Rfam (v14.7), utilized as the RNA family DB
422 [23].

423 A program (align_and_estimate_abundance.pl) in the Docker container for Trinity
424 (trinityrnaseq/trinityrnaseq:2.11.0) was used to estimate the abundance of RNA-Seq reads [28].

425 Salmon (v1.0.0) [31] was selected to quantify the transcripts to produce the expression matrix
426 for each organism.
427
428 **Insects and sample collection**
429 The stick insects *E. okinawaensis* were obtained from Amami-Ohshima in Kagoshima, Japan,
430 in 2011 and Ishigaki Island in Okinawa, Japan, in 2013, as described in our previous study [20].
431 Three biological replicates of the fat body were dissected from adult *E. okinawaensis*. These
432 tissues were stored at −80 °C until use. The fat body samples were weighed, homogenized with
433 lysis buffer from a PureLink® RNA extraction kit (Thermo Fisher Scientific Inc., Valencia,
434 CA, USA), and then centrifuged at 13,000 × *g* for 10 min. The supernatants were then collected
435 and processed for RNA purification according to the manufacturer's instructions. Purified total
436 RNA (1 μg) samples were processed for RNA sequencing.
437
438 **RNA sequencing**
439 RNA quality was assessed using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA,
440 USA). Libraries for cDNA sequencing were constructed using the Illumina TruSeq v2 kit
441 (Illumina Inc., San Diego, CA, USA) according to the manufacturer's protocol, and 100 bp
442 paired-end sequencing was then performed on the HiSeq 2500 platform (Illumina). Using Trim
443 Galore! (v0.4.4; https://github.com/FelixKrueger/TrimGalore), low-quality bases and adapter
444 sequences were trimmed, and the data quality was visualized and then confirmed. The raw
445 RNA sequence reads for biological triplicates of *E. okinawaensis* fat body samples were
446 deposited in the Sequence Read Archive (SRA) under accession ID: DRA013458
447 (DRR346171, DRR346172, and DRR346173).
448
449

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and materials

The developed workflow and original scripts necessary to replicate the study are available at GitHub: https://github.com/bonohu/SAQE/. The RNA sequencing reads reported in this article are available in the Sequence Read Archive (SRA) under accession ID DRA013458 (DRR346171, DRR346172, and DRR346173). The assembled transcriptome sequences are available in the Transcriptome Shotgun Assembly (TSA) sequence database under accession ID ICSG01 (ICSG01000001–ICSG01311357), and the estimated abundance of transcripts is available from the Genomic Expression Archive (GEA) under accession ID E-GEAD-476.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This research was supported by JSPS KAKENHI Grant 21K19118 to HB. This work was also supported by the Center of Innovation for Bio-Digital Transformation (BioDX), an open innovation platform for industry-academia co-creation (COI-NEXT), Japan Science and Technology Agency (JST, COI-NEXT, JPMJPF2010) to HB, and the ROIS-DS-JOINT (004RP2017) to HT. The APC was funded by JSPS KAKENHI Grant 21K19118. The funders

had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Authors' contributions

HB, TK and HT conceptualized the study; HB developed the methodology and software; HB and TK performed validation; HB performed the formal analysis; TS and HB conducted the investigations; TS, HT, and HB provided resources; HB performed data curation and original draft preparation of the manuscript; HB, TS, HT, and TK reviewed and edited the manuscript; HB performed data visualization, supervised the study, and provided project administration; and HB and HT oversaw funding acquisition. All authors have read and agreed to the published version of the manuscript.

## Acknowledgments

The authors wish to thank Dr. Kakeru Yokoi for helpful discussion. Computations were performed on the computers at Hiroshima University Genome Editing Innovation Center.

## Supplementary Information

The online version contains supplementary material available.

**Additional file 1:** Fig. S1: Graphical summary of functional annotation from the sequence information for *E. okinawaensis*. Fig. S2: Graphical summary of functional annotation from the sequence information for *B. mori*.

## References

498 1. Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, et al. Long
499 reads are revolutionizing 20 years of insect genome sequencing. Genome Biol. Evol.
500 2021;13:evab138; doi:10.1093/gbe/evab138.
501 2. Oppenheim SJ, Baker RH, Simon S, DeSalle R. We can't all be supermodels: the value
502 of comparative transcriptomics to the study of non-model insects. Insect Mol. Biol.
503 2015;24:139-54; doi:10.1111/imb.12154.
504 3. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A
505 universal tool for annotation, visualization and analysis in functional genomics research.
506 Bioinformatics. 2005;21:3674-6; doi:10.1093/bioinformatics/bti610.
507 4. Liu P, Ewald J, Galvez JH, Head J, Crump D, Bourque G, et al. Ultrafast functional
508 profiling of RNA-Seq data for nonmodel organisms. Genome Res. 2021;31:713-20;
509 doi:10.1101/gr.269894.120.
510 5. Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, et al.
511 FlyBase: Updates to the *Drosophila melanogaster* knowledge base. Nucleic Acids Res.
512 2021;49:D899-907; doi:10.1093/nar/gkaa1026.
513 6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene
514 ontology: tool for the unification of biology. Nat. Genet. 2000;25:25-9; doi:10.1038/75556.
515 7. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al.
516 Ensembl 2022. Nucleic Acids Res. 2022;50:D988-95; doi:10.1093/nar/gkab1049.

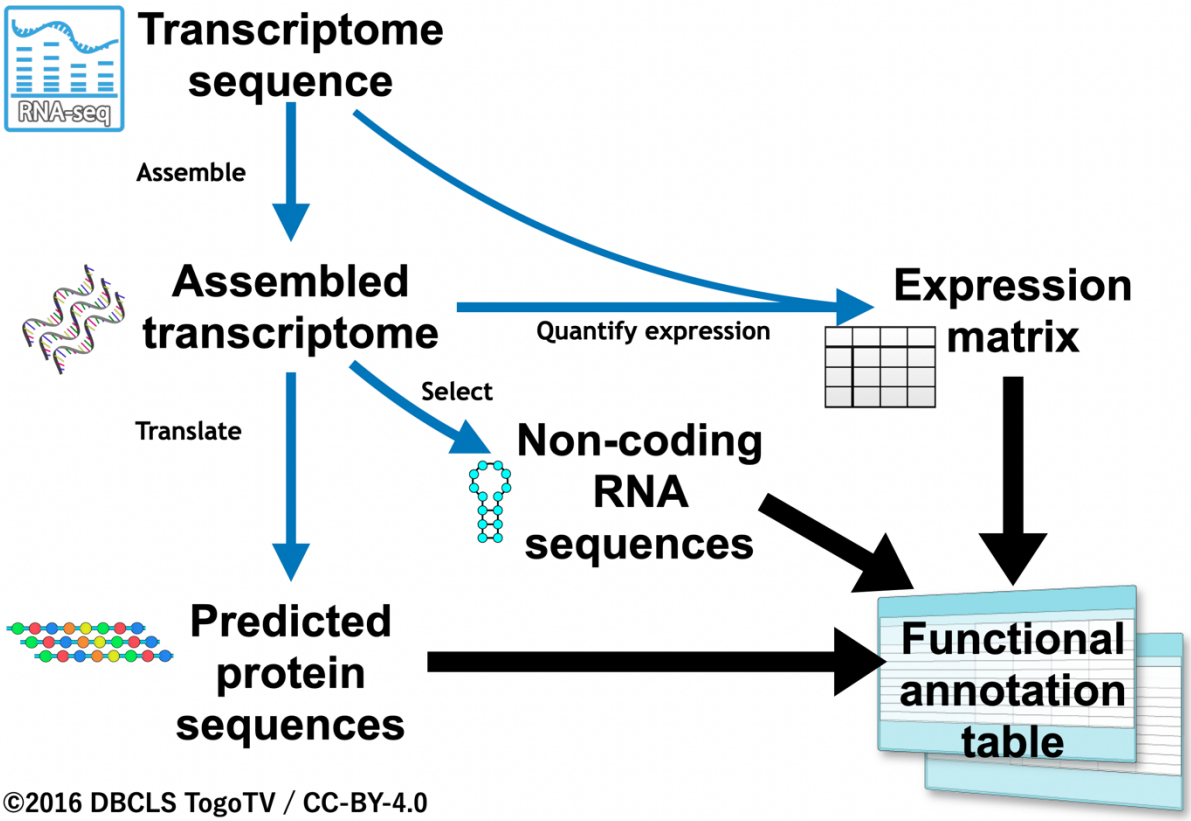
517 8. Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, et al. Ensembl
518 Genomes 2022: An expanding genome resource for non-vertebrates. Nucleic Acids Res.
519 2022;50:D996-1003; doi:10.1093/nar/gkab1007.
520 9. Tabunoki H, Ono H, Ode H, Ishikawa K, Kawana N, Banno Y, et al. Identification of
521 key uric acid synthesis pathway in a unique mutant silkworm *Bombyx mori* model of
522 Parkinson's disease. PLoS One. 2013;8:e69130; doi:10.1371/journal.pone.0069130.
523 10. Tabunoki H, Bono H, Ito K, Yokoyama T. Can the silkworm (*Bombyx mori*) be used as
524 a human disease model? Drug Discov. Ther. 2016;10:3-8; doi:10.5582/ddt.2016.01011.
525 11. Yokoi K, Tsubota T, Jouraku A, Sezutsu H, Bono H. Reference transcriptome data in
526 silkworm *Bombyx mori*. Insects. 2021;12:519; doi:10.3390/insects12060519.
527 12. Nojima Y, Bono H, Yokoyama T, Iwabuchi K, Sato R, Arai K, et al. Superoxide
528 dismutase down-regulation and the oxidative stress is required to initiate pupation in *Bombyx*
529 *mori*. Sci. Rep. 2019;9:14693;doi:10.1038/s41598-019-51163-3.
530 13. Kasukawa T, Furuno M, Nikaido I, Bono H, Hume DA, Bult C, et al. Development and
531 evaluation of an automated annotation pipeline and cDNA annotation system. Genome Res.
532 2003;13:1542-51: doi:10.1101/gr.992803.
533 14. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, et al. Functional
534 annotation of a full-length mouse cDNA collection. Nature. 2001;409:685-90;
535 doi:10.1038/35055500.

536 15. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the
537 mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature.
538 2002;420:563-73; doi:10.1038/nature01266.
539 16. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The
540 transcriptional landscape of the mammalian genome. Science. 2005;309:1559-63;
541 doi:10.1126/science.1112014.
542 17. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engström PG, et al. Transcript
543 annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet.
544 2006;2:e62; doi:10.1371/journal.pgen.0020062.
545 18. Blaxter M, Danchin A, Savakis B, Fukami-Kobayashi K, Kurokawa K, Sugano S, et al.
546 Reminder to deposit DNA sequences. Science. 2016;352:780; doi:10.1126/science.aaf7672.
547 19. Salzberg SL. Databases: Reminder to deposit DNA sequences. Nature. 2016;533:179;
548 doi:10.1038/533179a.
549 20. Sakamoto T, Sasaki S, Yamaguchi N, Nakano M, Sato H, Iwabuchi K, et al. De novo
550 transcriptome analysis for examination of the nutrition metabolic system related to the
551 evolutionary process through which stick insects gain the ability of flight (Phasmatodea).
552 BMC Res. Notes. 2021;14:182; doi:10.1186/s13104-021-05600-0.
553 21. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic
554 Acids Res. 2021;49:D480-9; doi:10.1093/nar/gkaa1100.

555 22. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al.
556 Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49:D412-9;
557 doi:10.1093/nar/gkaa913.
558
559 20. Bono H. Meta-analysis of oxidative transcriptomes in insects. Antioxidants (Basel).
560 2021;10:345; doi:10.3390/antiox10030345.
561 21. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
562 transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol.
563 2011;29:644-52; doi:10.1038/nbt.1883.
564 22. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo
565 transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference
566 generation and analysis. Nat. Protoc. 2013;8:1494-512; doi:10.1038/nprot.2013.084.
567 24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped
568 BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic
569 Acids Res. 1997;25:3389-402; doi:10.1093/nar/25.17.3389.
570 25. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches.
571 Bioinformatics. 2013;29:2933-5; doi:10.1093/bioinformatics/btt509.
572 26. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M,
573 et al. Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. Nucleic
574 Acids Res. 2021;49:D192-200; doi:10.1093/nar/gkaa1047.

575 27. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and
576 bias-aware quantification of transcript expression. Nat. Methods. 2017;14:417-9;
577 doi:10.1038/nmeth.4197.
578 30. Bono H, Ogata H, Goto S, Kanehisa M. Reconstruction of amino acid biosynthesis
579 pathways from the complete genome sequence. Genome Res. 1998;8:203-10;
580 doi:10.1101/gr.8.3.203.
581 31. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al.
582 The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data.
583 2016;3:160018; doi:10.1038/sdata.2016.18
584

585




586

587 **Fig. 1** Overview of the annotation workflow, Fanflow4Insects.

588

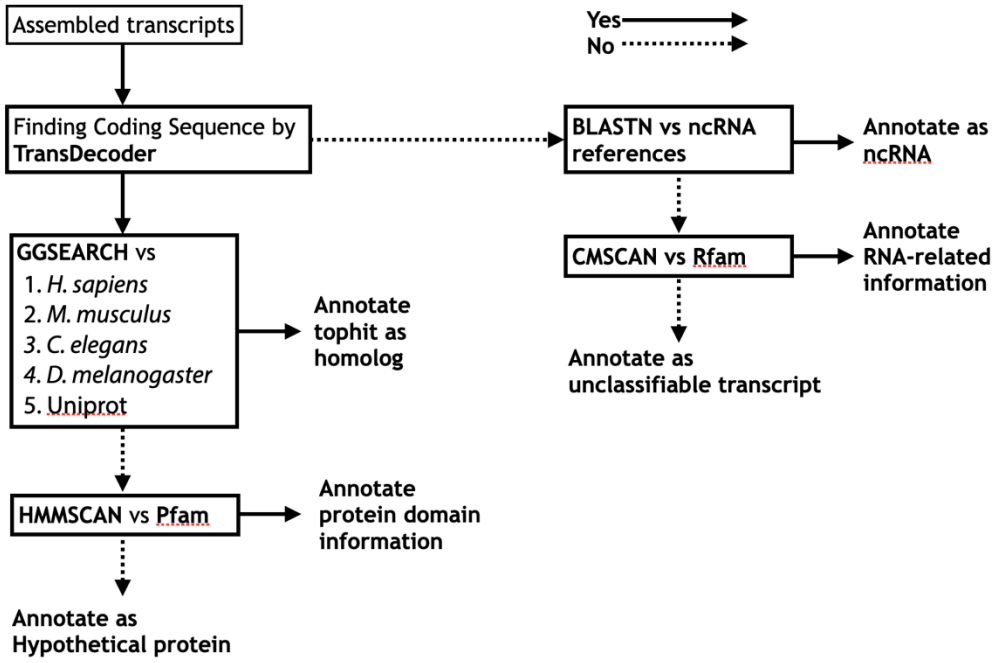
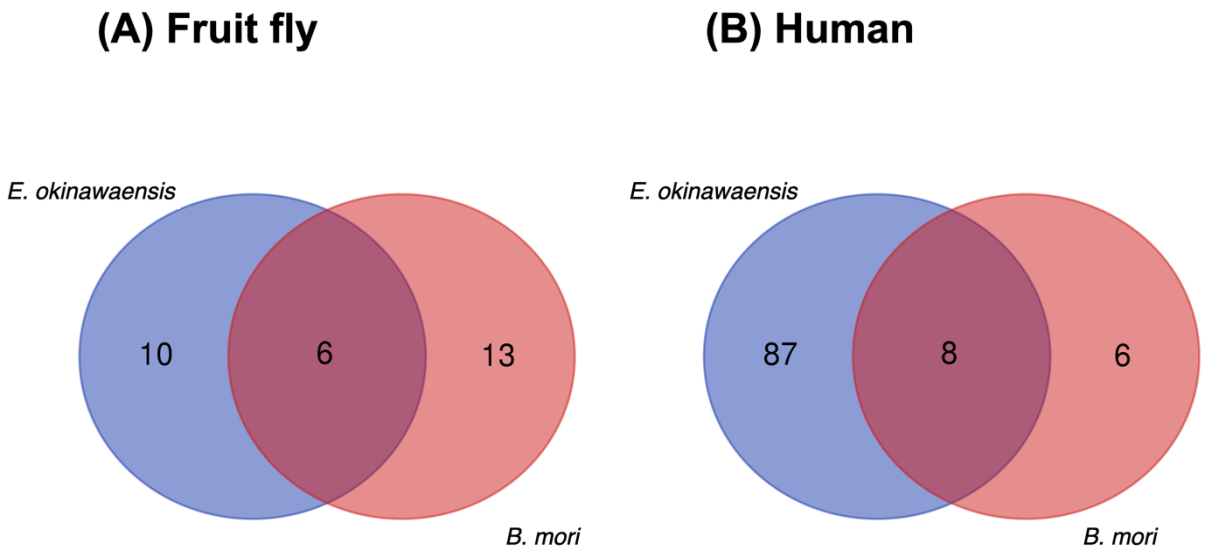



**Fig. 2** Annotation workflow for transcript description from sequence information. Left half for protein sequence level annotation; right half for nucleotide sequence level annotation

| | fat body | midgut | |
|---|---|---|---|
| transcript1 | 2 | 1 | ← constitutive expression |
| transcript2 | 0 | 2 | ← midgut specific expression |
| transcript3 | 3 | 0 | ← fat body specific expression |
| transcript4 | 0 | 0 | ← not expressed |
| ... | ... | ... | |

**Fig. 3** Annotation workflow for transcript descriptions derived from expression information.

602




603

604 **Fig. 4** Comparison of ncRNA transcripts between *E. okinawaensis* and *B. mori*

605