1 **Reconstruction and functional annotation of** 2 ***Ascosphaera apis* full-length transcriptome via PacBio** 3 **single-molecule long-read sequencing**

4 Dafu Chen ^{1,†}, Yu Du ^{1,†}, Xiaoxue Fan ¹, Zhiwei Zhu ¹, Haibin Jiang ¹, Jie Wang ¹,
5 Yuanchan Fan ¹, Huazhi Chen ¹, Dingding Zhou ¹, Cuiling Xiong ¹, Yanzhen Zheng ¹,
6 Xijian Xu ², Qun Luo ², Rui Guo ^{1,*}

7 ¹ College of Bee Science, Fujian Agriculture and Forestry University, Fuzhou
8 350002, China

9 ² Jiangxi Province Institute of Apiculture, Nanchang, Jiangxi 330201, China

10 [†] These authors contributed equally to this work.

11 * Correspondence author:

12 E-mail address: ruiguo@fafu.edu.cn;

13 Tel: +86-0591-87640197; Fax: +86-0591-87640197

14

15

16

17

18

19

20

21

22

23

24

25

26

27 **Abstract:**

28 *Ascosphaera apis* is a widespread fungal pathogen of honeybee larvae that results
 29 in chalkbrood disease, leading to heavy losses for the beekeeping industry in China and
 30 many other countries. This work was aimed at generating a full-length transcriptome of
 31 *A. apis* using PacBio single-molecule real-time (SMRT) sequencing. Here, more than
 32 23.97 Gb of clean reads was generated from long-read sequencing of *A. apis* mecyliia,
 33 including 464,043 circular consensus sequences (CCS) and 394,142 full-length
 34 non-chimeric (FLNC) reads. In total, we identified 174,095 high-confidence transcripts
 35 covering 5141 known genes with an average length of 2728 bp. We also discovered
 36 2405 genic loci and 11,623 isoforms that have not been annotated yet within the current
 37 reference genome. Additionally, 16,049, 10,682, 4520 and 7253 of the discovered
 38 transcripts have annotations in the Non-redundant protein (Nr), Clusters of Eukaryotic
 39 Orthologous Groups (KOG), Gene Ontology (GO), and Kyoto Encyclopedia of Genes
 40 and Genomes (KEGG) databases. Moreover, 1205 long non-coding RNAs (lncRNAs)
 41 were identified, which have less exons, shorter exon and intron lengths, shorter
 42 transcript lengths, lower GC percent, lower expression levels, and fewer alternative
 43 splicing (AS) events, compared with protein-coding transcripts. A total of 253 members
 44 from 17 transcription factor (TF) families were identified from our transcript datasets.
 45 Finally, the expression of *A. apis* isoforms was validated using a molecular approach.
 46 Overall, this is the first report of a full-length transcriptome of entomogenous fungi
 47 including *A. apis*. Our data offer a comprehensive set of reference transcripts and hence
 48 contributes to improving the genome annotation and transcriptomic study of *A. apis*.

49
 50 **Keywords:** *Ascosphaera apis*; full-length transcriptome; PacBio; chalkbrood;
 51 honeybee

52

53

54

55

56 **1. Introduction**

Chalkbrood is a widespread disease of the honeybee caused by *Ascosphaera apis* (Maassen ex Claussen) Olive and Spiltoir [1-2], an entomopathogenic fungus that exclusively infects western honeybee larvae. Recently, *A. apis* was reported to infect the larvae of eastern honeybee drones and workers [3]. This brood disease weakens colony productivity and honey production by lowering the number of newly emerged bees and, under certain circumstances, may result in colony losses [4].

The transcriptome can provide the information associated with the number and variety of intracellular genes and uncover the physiological and biochemical processes at a molecular level [5]. To date, an array of technologies has been developed and applied for transcriptome sequencing. Among these, short-read sequencing (i.e., Illumina and Ion Torrent) has become a useful tool for precisely analyzing RNA transcripts and gene expression levels [6-7]. However, most second-generation sequencing (also known as next-generation sequencing (NGS)) platforms offer a read-length shorter than the typical length of a eukaryotic mRNA, including a methylated cap at the 5' end and poly-A at the 3' end. To overcome the limitation of short-read sequences, single-molecule real-time (SMRT) sequencing (Pacific Biosciences of California, Inc., CA, USA) was developed, which can produce kilobase-sized sequencing reads, thus eliminating the need for sequence assembly [8-9]. For example, the average read length of PacBio SMRT sequencing is around 10 kb and the subread length can reach up to 35 kb [9]. The full-length transcriptome based on long reads can be used for the exploration and functional characterization of genes, the collection of large-scale long-read transcripts with complete coding sequences, and the identification of gene families [10-11]. However, the technology has a high sequencing-error rate (~15%) when compared to Illumina sequencing (~1%); and it can not currently be directly used to quantify gene expression [12-13]. Fortunately, the limitations of SMRT can be algorithmically improved and corrected by short and high-accuracy sequencing reads [14-15]. Hence, hybrid data derived from SMRT and NGS can offer high-quality and more complete assemblies for genome and transcriptome studies [16-17].

The genome of *A. apis* was published in 2006 with a total size of 20.31 Mb [18]. This version of the reference genome (AAP 1.0) is composed of 8092 contigs which are further assembled into 1627 scaffolds [18]; however, it is yet to be fully assembled into complete chromosomes. Transcriptome analysis is a powerful tool for uncovering the relationships between genotypes and phenotypes, leading to a better understanding of

the underlying pathways and genetic mechanisms controlling cell growth, development, immune defense, and so forth [19-21]. Our group previously *de novo* assembled and annotated a transcriptome of *A. apis* using short reads from NGS [22]. Based on this reference transcriptome, we further investigated the transcriptomic alteration and pathogenesis of *A. apis* during the infection process of two different bee species, *Apis mellifera ligustica* and *Apis cerana cerana* [23-24]. To provide a high-quality transcriptome of *A. apis*, in this work, the *A. apis* mycelia were subjected to third-generation sequencing (TGS) using the PacBio Sequel™ system (PacBio, Menlo Park, CA, USA). In parallel, Illumina paired short RNA reads generated separately from *A. apis* mycelia were used to support the SMRT data. Functional annotation of the transcriptome was performed followed by prediction and analysis of long non-coding RNAs (lncRNAs) and transcription factors (TFs). Overall, to the best of our knowledge, this is the first documentation of PacBio-based transcriptomic data of fungi including *A. apis*.

2. Results

2.1. PacBio SMRT sequencing and error correction of long reads

The workflow of the current work is presented in **Figure 1**. To obtain a representative full-length transcriptome for *A. apis*, the mycelia of *A. apis* were sequenced using PacBio Sequel system, and a total of 13,302,489 subreads (about 23.97 Gb) were yielded from the long-read sequencing, with an average read length of 1802 bp and an N50 of 3077 bp. To provide more accurate sequence information, circular consensus sequences (CCS) were generated from subreads that passed at least once time through the insert, and 464,043 CCS with a mean length of 2970 bp were gained (**Figure 2A**). By detecting the sequences, 402,415 were identified as being full-length (containing a 5' primer, 3' primer and the poly-A tail) and 394,142 were identified as being full-length non-chimeric (FLNC) reads with low artificial concatemers (**Figure 2B, Table 1**). The mean length of the FLNC reads was 2820 bp (**Figure 2B, Table 1**). FLNC reads with similar sequences were clustered together using the Iterative Clustering for Error Correction (ICE) algorithm, and 182,165 unpolished consensus isoforms with a mean length of 2701 bp were obtained (**Figure 2C**). In total, 121,776 high-quality isoforms and 58,307 low-quality isoforms were gained after polishing these unpolished consensus isoforms with the Quiver algorithm. Further, the aforementioned low-quality isoforms were corrected using the NGS short

reads with Proovread software, resulting in significant improvement in sequence accuracy (**Figure 3**). Finally, we identified 174,095 corrected isoforms with a mean read length of 2728 bp and an N50 length of 3543 bp (**Table 2**).

2.2. Reconstruction of *A. apis* full-length transcriptome

Here, 84.9% of the FLNC reads were mapped to the *A. apis* genome (**Figure 4A**). In addition, all 174,095 corrected isoforms were aligned against the reference genome of *A. apis*, and 168,740 (96.92%) reads were mapped to the reference genome, including 165,206 (94.89%) unique mapped reads and 3534 (2.03%) multiple mapped reads (**Figure 4B**); 84,734 (48.67%) reads were mapped to the positive strand of the genome, while 80,472 (46.22%) reads were mapped to the opposite strand of the genome (**Figure 4B**). A total of 17,195 isoforms from 5141 genetic loci were mapped to the *A. apis* genome gene set (AAP 1.0), which contains 6442 isoforms from 6442 genetic loci. We identified 3167 known isoforms in reference genome, 2405 novel genic loci from unannotated regions and 11,623 new isoforms from various exons.

2.3. Functional annotation of the full-length transcriptome of *A. apis*

Functional annotations of the non-redundant transcripts were determined by searching in the public databases. The results showed that 16,049 (93.34%), 10,682 (62.12%), 4520 (26.29%) and 7253 (42.18%) of the 17,195 isoforms could be found in the NCBI non-redundant protein sequences (Nr), Clusters of euKaryotic Orthologous Groups (KOG), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases, respectively (**Figure 5 A-C**). In addition, the transcripts had the highest number of hits to the *A. apis* (11,441 hits, 88.83%) proteins, followed by *Blastomyces dermatitidis* (118 hits, 0.92%) and *Histoplasma capsulatum* (97 hits, 0.75%) proteins (**Figure 5 D**).

2.4. LncRNA and TF identification

LncRNAs have been reported to play vital regulatory roles in a wide range of biological processes [25]. The number of lncRNAs predicted by each prediction method is presented in **Figure 6A**. In total, 1205 high-confidence lncRNAs were identified with an average length of about 912 bp. LncRNAs were classified into five groups according to their biogenesis positions relative to the protein-coding genes of

155 AAP 1.0 annotations: 19.00% (229) of them were generated from intergenic regions,
156 0.58% (7) from the intronic regions, 15.44% (186) from the sense strand, 18.59%
157 (224) from the antisense strand and 24.65% (297) belong to bidirectional lncRNA
158 (**Figure 6B**). In addition, the majority (72.37%) of the lncRNAs were single exons,
159 and this percentage was obviously higher than that of mRNAs (20.51%) (**Figure 6C**).
160 We additionally observed that when compared with protein-coding transcripts,
161 non-coding transcripts had fewer exons, shorter exon and intron lengths, shorter
162 transcript lengths, lower GC percentages, lower expression levels, and fewer
163 comparison of alternative splicing (AS) events (**Figure 6D-I**), which are similar to
164 findings in other species [26-29].

165 TFs are key components involved in the transcriptional regulatory system in
166 various animals, plants, and insects [12,30]. A total of 253 members from 17 TF
167 families were identified from our transcript datasets. The top 10 TF families were
168 C2H2 (89), bHLH (29), bZIP (29), HB-other (23), HSF (22), MYB_related (19), C3H
169 (13), GATA (8), M-type (5), and NF-YA (3) (**Figure 7**). However, considering the lack
170 of reports associated with *A. apis* TFs, more evidence is needed to confirm our
171 prediction.

172 2.5. Molecular validation of *A. apis* isoforms

173 In this work, 16 isoforms were randomly selected for RT-PCR to confirm the
174 expression of novel isoforms. As shown in **Figure 8A**, the signal fragments were
175 successfully detected by agarose gel electrophoresis. In addition, one of these
176 fragments was subjected to molecular cloning and Sanger sequencing, the result
177 further validated the reliability of *A. apis* isoforms (**Figure 8B**).

178 3. Discussion

179 *A. apis* is a widespread fungal pathogen of the honeybee, but its molecular and
180 omics study is lagging due to a lack of high-quality genome and transcriptome data.
181 Though the genome of *A. apis* was published as early as 2006 [18], the functional
182 annotation was not available until 2016. Transcriptome construction and annotation,
183 particularly for species without a reference genome or a complete genome, has greatly
184 improved with the development and revolution of sequencing techniques and plays a
185 critical role in gene discovery, genomic signature exploration, and genome annotation

[31-32]. Our group previously sequenced the *A. apis*-infected honeybee larval guts using the Illumina HiSeq platform, and *de novo* assembled and annotated a transcriptome using the short reads from *A. apis* [22]. Transcriptome analysis is a powerful tool for unraveling the relationships between genotype and phenotype, allowing a better understanding of the underlying pathways and molecular mechanisms regulating metabolism, growth and development, and the immune system [21,33-35]. Based on the previously assembled transcriptome, we conducted a further comprehensive transcriptomic investigation of *A. apis* infecting larvae from the western honeybee and eastern honeybee [23-24]. However, it still remains challenging to reliably assemble full-length from the short reads, and such transcripts are essential to explore post-transcriptional processes, such as AS and alternative polyadenylation (APA) events.

PacBio SMRT sequencing provides better completeness to the sequencing of both the 5' and 3' ends of cDNA molecules; thus, it is a superior strategy for the direct generation of a comprehensive transcriptome with precise AS isoforms and novel genes [36-37]. Yi et al. identified 33,300 full-length transcripts (transcript N50 of 5234 bp) of *Misgurnus anguillicaudatus* based on SMRT sequencing, and constructed a transcriptome by performing functional annotations of the non-redundant transcripts with public databases [38]. By sequencing mixed samples of *Agasicles hygrophila* eggs, larvae, pupae, and adults using PacBio SMRT, Jia and colleagues constructed a transcriptome composed of 28,982 full-length transcripts (transcript N50 of 2331 bp) [5]. Here, we employed PacBio SMRT technology for whole-transcriptome profiling in *A. apis* mycelia. A total of ~23.97 Gb of subreads were generated, including 464,043 CCS (mean length of 2970 bp) and 394,142 FLNC reads (mean length of 2820 bp). After removing redundant sequences from 182,165 high-quality isoforms, 174,095 transcripts (mean length of 2728 bp) and 5141 genes (mean length of 4087 bp) were obtained, which is much better than the 42,609 assembled unigenes (mean length of 966 bp) recorded in our previous work [22]. Meanwhile, the non-assembled transcripts from SMRT sequencing (transcript N50 of 3543 bp) were much longer than the assembled transcripts from Illumina sequencing (unigene N50 of 1550 bp) [22]. Additionally, the expression of 16 *A. apis* isoforms was verified by RT-PCR. Among these transcripts, one was further confirmed by Sanger sequencing (**Figure 8**). The results indicated that full-length transcripts can be recovered using SMRT sequencing. Additionally, 5141 genes were detected by SMRT with a mean length of 4087 bp,

which is 934 bp larger in size than that in the reference genome. About 13,853 isoforms were found to carry a completed ORF, further showing the long-read property of PacBio SMRT sequencing. The human transcriptome was revealed to be much more complex than previously expected owing to the application of TGS technology, which identified an array of novel isoforms that had not yet been annotated [9]. Similar findings were reported in some other species such as pigs [39], rabbits [26], switchgrass [13], red clover [30], and so forth. In the present study, we identified 2405 novel genic loci from unannotated regions and 11,623 new isoforms from various exons in the draft genome based on SMRT data, suggestive of a more complex transcriptome of *A. apis*. These data not only enrich the transcriptional information of the draft genome sequence but could also be used in functional studies of important genes in further research.

Here, 17,195 transcripts were annotated in four functional databases including the Nr, KOG, GO, and KEGG databases. Genomic sequencing clearly suggested that most of genes specifying the key biological functions are shared by all eukaryotes [40]. In this study, KOG database annotations showed that majority of transcripts (3,370, 31.55%) were enriched in the function of general function prediction only. The *A. apis* transcripts were annotated to various subcategories such as metabolic process, cellular process, cell, cell part, binding, and catalytic activity in the three categories based on the GO database annotations. Additionally, KEGG database annotations demonstrated that these transcripts were annotated to as many as 87 material and energy metabolism-related pathways such as biosynthesis of secondary metabolites and oxidative phosphorylation, genetic information processing-related pathways, such as RNA transport and spliceosome, cellular processes-related pathways such as cell cycle and endocytosis; environmental information processing-related pathways, such as Mitogen-activated protein kinase (MAPK) signaling pathway and ATP-binding cassette (ABC) transporters, and an organismal systems-related pathway (longevity regulating pathway). These results suggest that the transcripts of *A. apis* are associated with the abovementioned functions. Collectively, this high-quality transcriptome of *A. apis* could be used as a reference in some cases.

In a previous study, we identified 379 lncRNAs in the mixed samples of *A. apis* mycelia and spores, including 242 antisense lncRNAs, 123 intergenic lncRNAs, one intronic lncRNA and 13 sense lncRNAs, based on NGS and bioinformatics [41]. In this work, 1205 *A. apis* lncRNAs with an average length of 912 bp were identified on the

254 basis of PacBio SMRT sequencing. These high-quality lncRNAs were much longer
255 than previous ones [41], indicative of the advantage of SMRT sequencing in mining
256 lncRNAs at the transcriptome scale. In addition, the *A. apis* lncRNAs identified in this
257 study have fewer exons, shorter exon and intron lengths, shorter transcript lengths,
258 lower expression levels, and less AS events compared with protein-coding transcripts,
259 which are similar to our previous findings [41]. Collectively, our data provide
260 enrichment for the lncRNA reservoir of *A. apis*, but also enlarge the ncRNA database
261 of the fungal kingdom. It should be noted that the occurrences of false positives of
262 non-coding transcripts can not be absolutely excluded since this conclusion just
263 depends on the computational approach of a homologous search against reference
264 protein databases, thus the functions in *A. apis* require further experimental evidence.

265 **4. Materials and Methods**

266 *4.1. Preparation of A. apis mycelia samples*

267 *A. apis* was previously isolated from a fresh chalkbrood mummy of *A. m. ligustica*
268 larvae [41] and kept at the Honeybee Protection Laboratory of the College of Bee
269 Science at Fujian Agriculture and Forestry University.

270 *A. apis* was cultured at 33±0.5 °C on plates of Potato-Dextrose Agar (PDA)
271 medium according to the method developed by Jensen et al. [42]. One week after
272 culturing, mycelia (shown in **Figure 1**) were harvested and purified as previously
273 described [42], and then immediately frozen in liquid nitrogen and stored at -80 °C.

274 *4.2. Library construction and SMRT sequencing*

275 Firstly, the total RNA was extracted by grinding *A. apis* mycelia in TRIzol reagent
276 (Thermo Fisher, Shanghai, China) on dry ice and processed following the protocol
277 provided by the manufacturer. The integrity of the RNA was determined with the
278 Agilent 2100 Bioanalyzer and agarose gel electrophoresis. The purity and
279 concentration of the RNA were determined with the Nanodrop
280 micro-spectrophotometer (Thermo Fisher, Shanghai, China). Secondly, mRNA was
281 enriched by Oligo (dT) magnetic beads, followed by reverse transcription of the
282 enriched mRNA into cDNA using Clontech SMARTer PCR cDNA Synthesis Kit
283 (Takara, Shiga, Japan). PCR cycle optimization was used to determine the optimal
284 amplification cycle number for the downstream large-scale PCR reactions. Then the

285 optimized cycle number was used to generate double-stranded cDNA. Thirdly, >4kb
286 size selection was performed using the BluePippinTM Size-Selection System (Select
287 science, Corston, UK) and mixed equally with the no-size-selection cDNA. Fourthly,
288 large-scale PCR was performed for the next SMRT bell library construction; cDNAs
289 were DNA damage repaired, end repaired, and ligated to sequencing adapters. Finally,
290 the SMRT bell template was annealed to sequencing primer and bound to polymerase
291 followed by sequencing on the PacBio Sequel platform using P6-C4 chemistry with 10
292 h movies by Gene Denovo Biotechnology Co. (Guangzhou, China).

293 *4.3. Illumina short-read sequencing*

294 (1) The total RNA was isolated from *A. apis* mycelia using a Trizol Kit (Thermo
295 Fisher, Shanghai, China). (2) Oligo (dT) primers were used to isolate poly-A mRNA,
296 followed by fragmentation and reverse transcription with random primers (Qiagen,
297 Hilden, Germany). Second-strand cDNAs were synthesized using RNase H and DNA
298 polymerase I. The double-strand cDNAs were then purified using the QiaQuick PCR
299 extraction kit (Qiagen, Hilden, Germany). (3) After agarose gel electrophoresis, the
300 required fragments were purified using a DNA extraction kit (Qiagen, Hilden,
301 Germany) and then enriched via PCR amplification in a total volume of 50 µL
302 containing 3 µL of NEB Next USER Enzyme (NEB, Ipswich, USA), 25 µL of NEB
303 Next High-Fidelity PCR Master Mix (2×) (NEB, Ipswich, USA), 1 µL of Universal
304 PCR Primer (25 mmol) (NEB, Ipswich, USA), and 1 µL of Index (X) Primer (25 mmol)
305 (NEB, Ipswich, USA). The reaction conditions were as follows: 98 °C for 30 s,
306 followed by 13 cycles of 98 °C for 10 s and 65 °C for 75 s, and 65 °C for 5 s. (4) The
307 amplified fragments were sequenced on the Illumina HiSeqTM 4000 platform (Illumina,
308 San Diego, USA) by Gene Denovo Biotechnology Co. (Guangzhou, China) following
309 the manufacturer's protocols.

310 *4.4. Processing of SMRT reads and error correction*

311 The raw sequencing reads of cDNA libraries were classified and clustered into
312 transcript consensus using the SMRT Link v5.0.1 pipeline [43] supported by Pacific
313 Biosciences. Briefly, CCS reads were extracted out of subreads of the BAM file with a
314 minimum full pass of 1 and a minimum read score of 0.65. Subsequently, CCS reads
315 were classified into FLNC, non-full-length (nFL), chimeras, and short reads based on
316 cDNA primers and the poly-A tail signal. Reads shorter than 50 bp were discarded.

317 Next, the FLNC reads were clustered by ICE software to generate the cluster consensus
318 isoforms.

319 Two strategies were employed to improve the accuracy of PacBio reads: (1) The
320 nFL reads were used to polish the above obtained cluster consensus isoforms by Quiver
321 software to obtain the FL-polished high-quality consensus sequences (accuracy $\geq$ 99%).
322 (2) The low-quality isoforms were further corrected using Illumina short reads obtained
323 from the same samples using the LoRDEC tool (version 0.8) [44]. The pipeline of the
324 SMRT sequencing data process is shown in **Figure 1**.

325 4.5. Mapping of PacBio data to reference genome

326 The corrected high quality consensus sequences were then mapped to the
327 reference genome of *A. apis* (AAP 1.0) using Genomic Mapping and Alignment
328 Program (GMAP) [45], and redundant transcripts were collapsed with minimum
329 identity of 95% and a minimum coverage of 99%. The finally obtained isoforms were
330 compared with the reference genome annotation and classified into three groups:
331 known isoforms, novel isoforms, and new isoforms.

332 4.6. Functional annotation of transcripts

333 Transcripts were aligned against the NCBI Nr database
334 (<http://www.ncbi.nlm.nih.gov>), KOG database (<http://www.ncbi.nlm.nih.gov/KOG>),
335 and KEGG database (<http://www.genome.jp/kegg>) with the BLASTx program
336 (<http://www.ncbi.nlm.nih.gov/BLAST/>) at an E-value threshold of 1×10^{-5} to evaluate
337 the sequence similarity with genes of other species. GO annotation was analyzed by
338 Blast2GO software [46] with the Nr annotation results of isoforms. Isoforms ranking as
339 having the highest 20 scores and no shorter than 33 High-scoring Segment Pair (HSP)
340 hits were selected for the Blast2GO analysis. Functional classification of isoforms was
341 then performed using WEGO software [47].

342 4.7. ORF prediction

343 The ORFs were detected by using ANGEL [48] software for the isoform
344 sequences to obtain the coding sequences (CDS), protein sequences, and (Untranslated
345 region) UTR sequences.

346 4.8. Prediction and analysis of lncRNAs

CNCI (version 2) [49] and CPC [49] (<http://cpc.cbi.pku.edu.cn/>) were used to evaluate the protein-coding potential of novel isoforms and new isoforms by default parameters. Meanwhile, isoforms were mapped to the SwissProt database to assess protein annotation. The intersection of both non protein-coding potential results and non-protein annotation results were regarded as candidate lncRNAs. To better annotate lncRNAs at the evolution level, Infernal [50] (<http://eddylib.org/infernal/>) was used to assess the secondary structures and sequence conservation of lncRNAs. Cuffcompare was used to select the different types of lncRNAs including lincRNA, intronic lncRNA, and anti-sense lncRNA. Fragments per kilobase per million fragments mapped (FPKM) of both lncRNAs and mRNAs were calculated using StringTie (1.3.1). The transcript lengths, exon numbers and lengths, intron lengths, GC content, expression levels, and AS event numbers of lncRNAs were compared with those of mRNAs.

4.9. TF analysis

Protein coding sequences of isoforms were aligned by hmmscan to Plant TFdb (<http://planttfdb.cbi.pku.edu.cn/>) to predict TF families.

4.10. RT-PCR validation of isoforms

Sixteen isoforms of *A. apis* (Isoform000014, Isoform000018, Isoform000019, Isoform000021, Isoform000027, Isoform000028, Isoform000029, Isoform000042, Isoform000047, Isoform000063, Isoform000066, Isoform000085, Isoform000094, Isoform000113, Isoform000127 and Isoform0000365) were randomly selected for RT-PCR validation. Specific forward and reverse primers (presented in Table S1) were designed using DNAMAN software on the basis of the corresponding transcript sequences. One microgram of total RNA of *A. apis* mycelia was reverse transcribed to cDNA using the RevertAid First Strand cDNA Synthesis Kit (TaKaRa, China) and Oligo dT primers. PCR amplification was conducted on a T100 thermo cycler (BIO-RAD) using Premix (TaKaRa, China) under the following conditions: pre-denaturation step at 94 °C for 5 min; 34 amplification cycles of denaturation at 94 °C for 30 s, annealing at 60 °C for 30 s, and elongation at 72 °C for 1 min; this was followed by a final elongation step at 72 °C for 10 min. The PCR products were monitored on 1.8% agarose gel through electrophoresis with Genecolor (Gene-Bio,

China) staining. The fragment amplified from Isoform000014 was purified and cloned to pMD-19T vector (Takara, China) followed by Sanger sequencing.

4.11. Data availability

The raw data produced from PacBio SMRT sequencing and Illumina sequencing in this work were submitted to NCBI SRA database under BioProject numbers: PRJNA557811 and PRJNA560452.

5. Conclusions

Taken together, this work, for the first time, proposed the full-length transcriptome of *A. apis* using PacBio SMRT sequencing, providing a basis for further exploration of gene structures such as AS and APA. Moreover, the annotation of the *A. apis* gene set could improve the reference genome annotation and facilitate deeper understanding of the complexity of the *A. apis* genome and transcriptome.

Author Contributions: Conceptualization, D.C. and R.G. designed this study. R.G., Y.D., X.F., Z.Z., J.W., H.J., Y.F., H.C., D.Z., X.X., Q.L., C.X. and Y.Z. conducted laboratory work. R.G. and Y.D. performed bioinformatic analysis. D.C., R.G. and Y.D. supervised the work and contributed to preparation of the manuscript. All authors read and approved the final manuscript.

Funding: This research was supported by the Earmarked Fund for China Agriculture Research System (No. CARS-44-KXJ7), the Science and Technology Planning Project of Fujian Province (No. 2018J05042), the Teaching and Scientific Research Fund of Education Department of Fujian Province (No. JAT170158), the Outstanding Scientific Research Manpower Fund of Fujian Agriculture and Forestry University (No. xjq201814), and the Scientific and Technical Innovation Fund of Fujian Agriculture and Forestry University (No. CXZX2017342, No. CXZX2017343).

Acknowledgments: We thank all editors and reviewers for their invaluable comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ABC	ATP-binding cassette
APA	Alternative polyadenylation
AS	Alternative splicing
CCS	Circular consensus sequence
CDS	Coding sequences

CNCI	Coding-Non-Coding Index
CPC	Coding potential calculator
FLNC	Full-length non-chimeric
FPKM	Fragments per kilobase per million fragments mapped
GMAP	Genomic Mapping and Alignment Program
GO	Gene Ontology
HSP	High-scoring Segment Pair
ICE	Iterative Clustering for Error Correction
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOG	Clusters of euKaryotic Orthologous Groups
LncRNA	Long non-coding RNA
MAPK	Mitogen-activated protein kinase
nFL	Non-full-length
NGS	Next-generation sequencing
Nr	NCBI non-redundant protein
ORF	Open reading frame
PDA	Potato-Dextrose Agar
SMRT	Single molecule real time
TF	Transcription factor
TGS	Third-generation sequencing
UTR	Untranslated region

406

407 **References**

- 408 1. Spiltoir, C.F. Life cycle of *Ascosphaera apis*. *Am J Bot.* **1955**, 42, 501-518, doi:
409 10.2307/2438686
- 410 2. Spiltoir, C.F.; Olive L.S. A reclassification of the genus *Pericyctis* Betts.
411 *Mycologia.* **1955**, 47, 238-244, doi: 10.2307/3755414
- 412 3. Chen, D.F.; Guo, R.; Xiong, C.L.; Zheng, Y.Z.; Hou, C.S.; Fu, Z.M.
413 Morphological and molecular identification of chalkbrood disease pathogen
414 *Ascosphaera apis* in *Apis cerana cerana*. *J. Apic. Res.* **2018**, 57, doi:
415 10.1080/00218839.2018.1475943
- 416 4. Evison, S.E. Chalkbrood: epidemiological perspectives from the host-parasite
417 relationship. *Chalkbrood disease in honey bees.* **2015**, 10, 65-70, doi:
418 10.1016/j.cois.2015.04.015
- 419 5. Dong, J.; Wang, Y.X.; Liu, Y.H.; Hu, J.; Guo, Y.Q.; Gao, L.L.; Ma, R.Y. SMRT
420 sequencing of full-length transcriptome of fea beetle *Agasicles hygrophila*
421 (Selman and Vogt). *Sci. Rep.* **2018**, 8, 2197, doi: 10.1038/s41598-018-20181-y
- 422 6. Ugrappa, N.; Zhong, W.; Karl, W.; Chong, S.; Debasish, R.; Mark, G.; Michael,
423 S.; The transcriptional landscape of the yeast genome defined by RNA sequencing.
424 *Science.* **2008**, 320, 1344-1349, doi: 10.1126/science.1158441.

- 425 7. Djebali, S.; Davis, C.A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.;
426 Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; et al. Landscape of transcription
427 in human cells. *Nature*. **2012**, *489*, 101-108, doi: 10.1038/nature11233
- 428 8. Au, K.F.; Sebastiano, V.; Afshar, P.T.; Durruthy, J.D.; Lee, L.; Williams, B.A.;
429 van Bakel, H.; Schadt, E.E.; Reijo-Pera, R.A.; Underwood, J.G.; et al.
430 Characterization of the human ESC transcriptome by hybrid sequencing. *Proc.*
431 *Nati. Acad. Sci. USA*. **2013**, *110*, E4821–E4830, doi: 10.1073/pnas.1320101110
- 432 9. Sharon, D.; Tilgner, H.; Grubert, F.; Snyder, M. A single-molecule long-read
433 survey of the human transcriptome. *Nat. Biotechnol.* **2013**, *31*, 1009-1014, doi:
434 10.1038/nbt.2705
- 435 10. Koren, S.; Schatz, M.C.; Walenz, B.P.; Martin, J.; Howard, J.T.; Ganapathy, G.;
436 Wang, Z.; Rasko, D.A.; McCombie, W.R.; Jarvis, E.D.; et al. Hybrid error
437 correction and *de novo* assembly of single-molecule sequencing reads. *Nat.*
438 *Biotechnol.* **2012**, *30*, 693-700, doi: 10.1038/nbt.2280
- 439 11. Treutlein, B.; Gokce, O.; Quake, S.R.; Südhof, T.C. Cartography of neurexin
440 alternative splicing mapped by single-molecule long-read mRNA sequencing.
441 *Proc. Natl. Acad. Sci. USA*. **2014**, *111*, E1291–E1299, doi:
442 10.1073/pnas.1403244111
- 443 12. Luo, Y.; Ding, N.; Shi, X.; Wu, Y.; Wang, R.; Pei, L.; Xu, R.; Cheng, S.; Lian, Y.;
444 Gao, J.; et al. Generation and comparative analysis of full-length transcriptomes in
445 sweetpotato and its putative wild ancestor *I. trifida*, **2017**, doi: 10.1101/112425
- 446 13. Zuo, C.M.; Matthew, B.; Avinash, S.; Rita, C.K.; Govindarajan, K.R.; Ivone, T.J.;
447 Li, G.F.; Wang, M.; David, D.; Kerrie, R.; et al. Revealing the transcriptomic
448 complexity of switchgrass by PacBio long-read sequencing. *Biotechnol Biofuels*.
449 **2018**, *11*, 170, doi: 10.1186/s13068-018-1167-z
- 450 14. Li, Q.; Li, Y.; Song, J.; Xu, H.; Xu, J.; Zhu, Y.; Li, X.; Gao, H.; Dong, L.; Qian, J.;
451 et al. High-accuracy *de novo* assembly and SNP detection of chloroplast genomes
452 using a SMRT circular consensus sequencing strategy. *New Phytol.* **2014**, *204*,
453 1041-1049, doi: 10.1111/nph.12966
- 454 15. Hack, T.; Hedrich, R.; Schultz, J.; Förster, F.; Proovread: large-scale
455 high-accuracy pacbio correction through iterative short read consensus.
456 *Bioinformatics*. **2014**, *30*, 3004-3011, doi: 10.1093/bioinformatics/btu392
- 457 16. Huddleston, J.; Ranade, S.; Malig, M.; Antonacci, F.; Chaisson, M.; Hon, L.;
458 Sudmant, P.H.; Graves, T.A.; Alkan, C.; Dennis, M.Y.; et al. Reconstructing

- complex regions of genomes using long-read sequencing technology. *Genome Res*,
2014, 24, 688-696, doi: 10.1101/gr.168450.113
17. Xu, Z.; Peters, R.J.; Weirather, J.; Luo, H.; Liao, B.; Zhang, X.; Zhu, Y.; Ji, A.;
Zhang, B.; Hu, S.; et al. Full-length transcriptome sequences and splice variants
obtained by a combination of sequencing platforms applied to different root tissues
of *Salvia miltiorrhiza*, and tanshinone biosynthesis. *Plant J.* **2015**, 82, 951–961,
doi: 10.1111/tpj.12865
18. Qin, X.; Evans, J.D.; Aronstein, K.A.; Murray, K.D.; Weinstock, G.M. Genome
sequences of the honey bee pathogens *Paenibacillus larvae* and *Ascosphaera apis*.
Insect Mol. Biol. **2006**, 15, 715-718, doi: 10.1111/j.1365-2583.2006.00694.x
19. Qian, X.; Ba, Y.; Zhuang, Q.; Zhong, G. RNA-Seq technology and its application
in fish transcriptomics. *OMICS*. **2013**, 18, 98, doi:10.1089/omi.2013.0110
20. Chen, D.F.; Guo, R.; Xu, X.J.; Xiong, C.L.; Liang, Q.; Zheng, Y.Z.; Luo, J.; Zhang,
Z.N.; Huang, Z.J.; Kumar, D.; et al. Uncovering the immune responses of *Apis*
mellifera ligustica larval gut to *Ascosphaera apis* infection utilizing transcriptome
sequencing. *Gene*. **2017**, 621, 40-50, doi : 10.1016/j.gene.2017.04.022
21. Guo, R.; Chen, D.F.; Diao, Q.Y.; Xiong, C.L.; Zheng, Y.Z.; Hou, C.S.
Transcriptomic investigation of immune responses of the *Apis cerana cerana*
larval gut infected by *Ascosphaera apis*. *J. Invertebr. Pathol.* **2019**, 166, 107210,
doi: 10.1016/j.jip.2019.107210
22. Zhang, Z.N.; Xiong, C.L.; Xu, X.J.; Huang, Z.J.; Zheng, Y.Z.; Luo, Q.; Liu, M.; Li,
W.D.; Tong, X.Y.; Zhang, Q.; et al. *De novo* assembly of a reference transcriptome
and development of SSR markers for *Ascosphaera apis*. *Acta Entomol. Sin.* **2017**,
60, 34-44 doi: 10.16380/j.kcxb.2017.001.05
23. Chen, D.F.; Guo, R.; Xiong, C.L.; Liang, Q.; Zheng, Y.Z.; XU, X.J.; Huang, Z.J.;
Zhang, Z.N.; Zhang, L.; LI, W.D.; Tong, X.Y.; XI, W.J.; Transcriptomic analysis
of *Ascosphaera apis* stressing larval gut of *Apis mellifera ligustica*
(Hymenoptera: Apidae). *Acta Entomol. Sin.* **2017**, 60, 401-411, doi:
10.16380/j.kcxb.2017.04.005
24. Guo, R.; Chen, D.F.; Huang, Z.J.; Liang, Q.; Xiong, C.L.; Xu, X.J.; Zheng, Y.Z.;
Zhang, Z.N.; Xie, Y.L.; Tong, X.Y.; et al. Transcriptome analysis of *Ascosphaera*
apis stressing larval gut of *Apis cerana cerana*. *Acta Microbiol. Sin.* **2017**, 57,
1865-1878, doi: 10.13343/j.cnki.wsxb.20160551

- 492 25. Atianand, M.K.; Fitzgerald, K.A. Long non-coding RNAs and control of gene
493 expression in the immune system. *Trends Mol Med.* **2014**, *20*, 623-631, doi:
494 10.1016/j.molmed.2014.09.002
- 495 26. Chen, S.Y.; Deng, F.; Jia, X.; Li C.; Lai, S.J. A transcriptome atlas of rabbit
496 revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* **2017**, *7*,
497 7648, doi: 10.1038/s41598-017-08138-z
- 498 27. Pauli, A.; Valen E.; Lin, M.F.; Garber, M.; Vastenhouw, N.L.; Levin, J.Z.; Fan,
499 L.; Sandelin, A.; Rinn, J.L.; Regev, A.; et al. Systematic identification of long
500 noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **2012**,
501 *22*, 577-91, doi: 10.1101/gr.133009.111
- 502 28. Li, L.; Eichten, S.R.; Shimizu, R.; Petsch, K.; Yeh, C.T.; Wu, W.; Chettoor, A.M.;
503 Givan, S.A.; Cole, R.A.; Fowler, J.E.; et al. Genome-wide discovery and
504 characterization of maize long non-coding RNAs. *Genome Biol.* **2014**, *15*, R40,
505 doi: 10.1186/gb-2014-15-2-r40
- 506 29. Guo, R.; Chen, D.F.; Xiong, C.L.; Hou, C.S.; Zheng, Y.Z.; Fu, Z.M.; Liang, Q.;
507 Diao, Q.Y.; Zhang, L.; Wang, H.Q.; et al. First identification of long non-coding
508 RNAs in fungal parasite *Nosema ceranae*. *Apidologie.* **2018**, *49*, 660-670, doi:
509 10.1007/s13592-018-0593-z
- 510 30. Chao, Y.H.; Yuan, J.B.; Li, S.F.; Jia, S.Q.; Han, L.B.; Xu, L.X. Analysis of
511 Transcripts and splice isoforms in Red Clover (*Trifolium pratense* L.) by
512 single-molecule long-read sequencing. *BMC Plant Biol.* **2018**, *18*, 300, doi:
513 10.1186/s12870-018-1534-8
- 514 31. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van, Baren,
515 M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and
516 quantification by RNA-seq reveals unannotated transcripts and isoform switching
517 during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511-5, doi: 10.1038/nbt.1621
- 518 32. Haas, B.J.; Papanicolaou, Y.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.;
519 Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. *De novo* transcript sequence
520 reconstruction from RNA-seq using the trinity platform for reference generation
521 and analysis. *Nat. Protoc.* **2013**, *8*, 1494-512 doi: 10.1038/nprot.2013.084
- 522 33. Song, L.Y.; Rong, Li.; Chen, S.L.; Chen, H.F.; Zhang, C.J.; Chen, L.M.; Hao,
523 Q.N.; Shan, Z.H.; Yang, Z.L.; Qiu, D.Z.; et al. RNA-Seq analysis of differential
524 gene expression responding to different rhizobium strains in Soybean (*Glycine*
525 *max*) Roots. *Front. Plant Sci.* **2016**, *7*, 721, doi: 10.3389/fpls.2016.00721

- 526 34. Guo, R.; Chen, D.; Chen, H.; Fu, Z.; Xiong, C.; Hou, C.; Zheng, Y.; Guo, Y.;
527 Wang, H.; Du, Y.; et al. Systematic investigation of circular RNAs in *Ascosphaera*
528 *apis*, a fungal pathogen of honeybee larvae. *Gene*. **2018**, 678, 17-22. doi:
529 10.1016/j.gene.2018.07.076
- 530 35. Chen, D.F.; Chen, H.Z.; Du, Y.; Zhou, D.D.; Geng, S.H.; Wang, H.P.; Wan, J.Q.;
531 Xiong, C.L.; Zheng, Y.Z.; Guo, R.; Genome-wide Identification of long
532 non-coding RNAs and their regulatory networks involved in *Apis mellifera*
533 *ligustica* Response to *Nosema ceranae* infection. *Insects*. **2019**, 10, doi:
534 10.3390/insects10080245
- 535 36. Chuang, T.J.; Chen, Y.J.; Chen, C.Y.; Mai T.L.; Wang, Y.D.; Yeh, C.S.; Yang,
536 M.Y.; Hsiao, Y.T.; Chang T.H.; Kuo T.C.; et al. Integrative transcriptome
537 sequencing reveals extensive alternative trans-splicing and *cis*-backsplicing in
538 human cells. *Nucleic Acids Res*. **2018**, 46, 3671-3691, doi: 10.1093/nar/gky032.
- 539 37. Filichkin, S.A.; Hamilton, M.; Dharmawardhana P.D.; Singh, S.K.; Sullivan. C.;
540 Ben-Hur. A.; Reddy, A.S.N.; Jaiswal, P.; Abiotic stresses modulate landscape of
541 poplar transcriptome via alternative splicing, differential intron retention, and
542 isoform ratio switching. *Front. Plant Sci*. **2018**, 9, 5, doi:
543 10.3389/fpls.2018.00005
- 544 38. Yi, S.; Zhou, X.; Li, J.; Zhang, M.; Luo, S. Full-length transcriptome of *Misgurnus*
545 *anguillicaudatus* provides insights into evolution of genus *Misgurnus*. *Sci. Rep*.
546 **2018**, 8, 11699, doi: 10.1038/s41598-018-29991-6
- 547 39. Li, Y.; Fang, C.; Fu, Y.; Hu, A.; Li, C.; Zou, C.; Li, X.; Zhao, S.; Zhang, C.; Li, C.
548 A survey of transcriptome complexity in *Sus scrofa* using single-molecule
549 long-read sequencing. *DNA Res*. **2018**, 25, 421-437, doi: 10.1093/dnares/dsy014
- 550 40. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.;
551 Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: tool for
552 the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. **2000**, 25,
553 25-29, doi: 10.1038/75556
- 554 41. Guo, R.; Chen, D.; Xiong, C.; Hou, C.; Zheng, Y.; Fu, Z.; Diao, Q.; Zhang, L.;
555 Wang, H.; Hou, Z.; et al. Identification of long non-coding RNAs in the
556 chalkbrood disease pathogen *Ascosphaera apis*. *J. Invertebr. Pathol*. **2018**, 156,
557 1-5, doi: 10.1016/j.jip.2018.06.001

- 558 42. Jensen, A.B.; Aronstein, K. Flores, J.M.; Vojvodic, S.; Palacio, M.A.; Spivak, M.
559 Standard methods for fungal brood disease research. *J. Apic. Res.* **2013**, *52*,
560 doi:10.3896/IBRA.1.52.1.13
- 561 43. Gordon, S.P.; Tseng, E.; Salamov, A.; Zhang, J.; Meng, X.; Zhao, Z.; Kang, D.;
562 Underwood, J.; Grigoriev, I.V.; Figueroa, M. Widespread polycistronic transcripts
563 in fungi revealed by single-molecule mRNA sequencing. *PLoS One.* **2015**, *10*,
564 e0132628, doi: 10.1371/journal.pone.0132628
- 565 44. Salmela, L.; Rivals, E. LoRDEC: accurate and efficient long read error correction.
566 *Bioinformatics.* **2014**, *30*, 3506-3514, doi: 10.1093/bioinformatics/btu538
- 567 45. Wu, T.D.; Watanabe, C.K.; GMAP: a genomic mapping and alignment program
568 for mRNA and EST sequences. *Bioinformatics.* **2005**, *21*, 1859-1875, doi:
569 10.1093/bioinformatics/bti310
- 570 46. Conesa, A.S.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M.
571 Blast2GO: a universal tool for annotation, visualization and analysis in functional
572 genomics research. *Bioinformatics.* **2005**, *21*, 3674-3676, doi:
573 10.1093/bioinformatics/bti610
- 574 47. Ye, J.; Fang, L.; Zheng, H.K.; Zhang, Y.; Chen, J.; Zhang, Z.J.; Wang, J.; Li, S.T.;
575 Li, R.Q.; Lars, B.; et al. WEGO: a web tool for plotting GO annotations. *Nucleic*
576 *Acids Res.* **2006**, *34*, W293-W297, doi: 10.1093/nar/gkl031
- 577 48. Shimizu, K.; Adachi, J.; Muraoka, Y. ANGLE: a sequencing errors resistant
578 program for predicting protein coding regions in unfinished cDNA. *J. Bioinform.*
579 *Comput. Biol.* **2006**, *4*, 649-664, doi: 10.1142/S0219720006002260
- 580 49. Sun, L.; Luo, H.; Bu, D.; Zhao, G.; Yu, K.; Zhang, C.; Liu, Y.; Chen, R.; Zhao, Y.
581 Utilizing sequence intrinsic composition to classify protein-coding and long
582 non-coding transcripts. *Nucleic Acids Res.* **2013**, *41*, e166, doi:
583 10.1093/nar/gkt646
- 584 50. Nawrocki, E.P.; Eddy, S.R.; Infernal 1.1: 100-fold faster RNA homology searches.
585 *Bioinformatics.* **2013**, *29*, 2933-2935, doi: 10.1093/bioinformatics/btt509

586
587
588
589
590
591
592

593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623

Figure legends

Figure 1. Workflow for fungal mycelia preparation, Iso-seq, data processing and bioinformatic analysis.

Figure 2. Length distribution of PacBio SMRT sequencing. (A) Number and length distribution of CCS. (B) Number and length distribution of FLNC reads. (C) Number and length distribution of corrected isoforms.

Figure 3. Accuracy of low-quality isoforms after correction with Illumina short reads.

Figure 4. Mapping of SMRT reads (A) and corrected isoforms (B) to *A. apis* genome.

Figure 5. Function annotation of corrected isoforms. (A) KOG classifications of transcripts. (B) KEGG pathways enriched by transcripts. (C) Distribution of GO terms for all annotated transcripts in biological process, cellular component and molecular function. (D) RefSeq Nr Homologous species distribution diagram of transcripts.

Figure 6. Identification of *A. apis* lncRNAs. (A) Venn diagram of lncRNAs predicted by Coding Potential Calculator (CPC), Coding-Non-Coding Index (CNCI), and pfam methods. (B) Proportions of different types of lncRNAs. (C) Comparison of exon number between lncRNAs and mRNAs. (D) Comparison of exon length between lncRNAs and mRNAs. (E) Comparison of intron length between lncRNAs and mRNAs. (F) Comparison of GC content between lncRNAs and mRNAs. (G) Comparison of transcript length between lncRNAs and mRNAs. (H) Comparison of expression level between lncRNAs and mRNAs. (I) AS events between lncRNAs and mRNAs.

Figure 7. Identification of TFs in *A. apis* mycelia. The number and family of TFs were predicted by SMRT.

Figure 8. RT-PCR and Sanger sequencing validation of *A. apis* isoforms. (A) Agarose gel electrophoresis of RT-PCR products from 16 *A. apis* isoforms; Lane 1: Isoform000014; Lane 2: Isoform000021; Lane 3: Isoform000027; Lane 4: Isoform000036; Lane 5: Isoform000042; Lane 6: Isoform000085; Lane 7: Isoform000094; Lane 8: Isoform000113; Lane 9: Isoform000127; Lane 10: Isoform000018; Lane 11: Isoform000019; Lane 12: Isoform000028; Lane 13: Isoform000029; Lane 14: Isoform000047; Lane 15: Isoform000063; Lane 16: Isoform000066; Lane M: DNA marker. (B) Sanger sequencing of amplified fragment from Isoform000014.

656

Table 1. PacBio SMRT sequencing output statistics.

cDNA size	Number
Number of five prime reads	434,570
Number of three prime reads	439,279
Number of poly-A reads	431,206
Number of filtered short reads	3865
Number of non-FL reads	57,763

Number of FL reads	402,415
Number of FLNC reads	394,142
Mean FL non-chimeric read length	2820

658

659

Table 2. Overview of *A. apis* isoforms after correction.

	Aam
Number of total reads	174,095
Length of total reads (bp)	474,928,820
Maximum length of total reads (bp)	13,808
Minimum length of total reads (bp)	50
Average length of total reads (bp)	2728
N50 length of total reads (bp)	3543
GC content of total reads (%)	49.00

660

661

662

663

664

665

666

667

668

669

670

Table S1. Primers used in the present study.

Isoform		Sequence
Isoform000014	F	CATCAAGTCCACTGCCATC
	R	ACACAGACACCAGAAACGC
Isoform000021	F	GATTCCCACCTCCGATAAG
	R	GAACTGGTCAACACCGACA
Isoform000027	F	CATCTCCCATCACCATAGG
	R	GCAACGGGCACTTATTTG

Isoform000036	F	GGGTTTCAGTTTCCTCGGAT
	R	AGACGATGTCACCATAGCG
Isoform000042	F	CTACTGCTGCTGCTGCTACT
	R	AACCCACTGATGTGCCTATT
Isoform000085	F	CATCACCATCCACACCATAG
	R	CAGGAGCATTTAGGCGATTA
Isoform000094	F	TCTACCAAACGACTTCCTCG
	R	CCCATTCTTCCTTTCTATGCTC
Isoform000113	F	GGCACCATTGTTTCAGTCAG
	R	GGTCTAAGGCACTTCACGAC
Isoform000127	F	CGGAGTCTCTGGTGTTCATCG
	R	ACCTATCGGGAACCTGGAT
Isoform000018	F	CATCACCAGCGTCAACATC
	R	CAGAGAAGAAACCACGGATAG
Isoform000019	F	CGGCACTCCTGAGATTCCTA
	R	TCATCCTCTGCGACATCGT
Isoform000028	F	GCATAGCGTTGTCACTTACG
	R	GGACTTCCAGCCAGTATTGTTA
Isoform000029	F	TGGTAGATTGGACGCTAACG
	R	TCTGGTGAGTGTTCAGCCTT
Isoform000047	F	GGGTTACCATTAGCAGCGT
	R	CAAGACAATCCACCAGAGG
Isoform000063	F	CATCCTTCATCATCTGGCA
	R	CAGCCGTGTCAACTTGTCTA
Isoform000066	F	GCATACTGTCAAGGGTCACA
	R	TTCGTTGCCTGTAACTTCG

671

672

673

674

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

787
788
789
790
791
792
793
794
795
796
797
798
799
800