The de novo transcriptome and its analysis of the worldwide vegetable pest, *Delia antiqua* (Diptera: Anthomyiidae)

Yujuan Zhang, Youjin Hao, Fengling Si, Shuang Ren, Ganyu Hu, Li Shen, Bin Chen*

Institute of Entomology and Molecular Biology, College of Life Sciences, Chongqing Normal University, Chongqing, P.R. China.

*Correspondence E-mail: c_bin@hotmail.com

Running title: transcriptome and its analysis of *Delia antiqua*

Abstract

The onion maggot *Delia antiqua* is a major insect pest of cultivated vegetable especially the onion and a good model to investigate the molecular mechanisms of diapause. To better understand the biology and the diapause mechanism of the insect pest species, *D. antiqua* transcriptome were sequenced using Illumina paired-end sequencing technology. Approximate 54 million reads were obtained, trimmed, and assembled into 29,659 unigenes, with an average length of 607 bp and an N50 of 818 bp. Among them, 21,605 (72.8%) unigenes were annotated in the public databases. All unigenes were then compared against *Drosophila melanogaster* and *Anopheles*.
*gambiae.* Codon usage bias was analyzed and 332 Simple Sequence Repeats (SSRs) were detected in this organism. The data presented the most comprehensive transcriptomic resource currently available for *D. antiqua*, and will facilitate the study of genetics, genomics and diapause, and further pest control of *D. antiqua*.

**Keywords**: Onion maggot, High-Throughput RNA Sequencing, De novo assemble, Codon usage bias, SSR

### Introduction

The onion maggot *Delia antiqua* is a major insect pest of cultivated vegetable especially onions, and is widely distributed in the northern hemisphere. It can be induced into summer and winter diapauses, both happening at the pupal stage and just after the head evagination [1]. This characteristics make it a good model to investigate the molecular mechanisms of pupal diapause [2]. To date, there have been no efforts to sequence the complete transcriptome of *D. antiqua*. The research on this species is limited by a very small amount of genomic data. As of February 2014, only 245 expressed sequence tags (ESTs) have been deposited in GenBank. To better understand the biological basis of diapause and *D. antiqua*, there is a need to explore the genomic biology in the *D. antiqua*.

Advances in Next-Generation Sequencing (NGS) and assembly algorithms have rapidly promoted the development of next generation RNA sequencing (RNA-seq), which can explore the entire transcriptome in a selected tissue and species of interest and generates quantitative expression scores for each transcript [3]. This transcriptome analysis will likely replace large-scale microarray approaches [4; 5] due to its lower cost and greater sequence yield that allows
the measurement of transcriptome composition, quantitatively surveys RNA expression patterns [3], and addresses comparative genomic-level questions and allows the development of molecular markers [6]. In the past several years, RNA-seq has been used on practically any genome, from bacteria, archaea, and lower eukaryotes to higher eukaryotes. It is particularly effective when a reference genome is not available[7].

The main objective of our study was to construct a reference transcriptome of *D. antiqua* for future genetic and genomic studies of this species. Here, we present a comprehensive analysis of the de novo transcriptome sequencing results for cDNA samples derived from eggs, larvae, pupae, and adults of *D. antiqua* (which covers every stage of development) by using the Illumina Hiseq2000 sequencing platform. As a result, a total of 29,659 unigenes were assembled and identified; among them 21,050 were annotated. Based on the transcriptome, the codon usage bias was analyzed, and the Simple Sequence Repeats (SSRs) were detected. To our knowledge, this is the first report on of the complete transcriptome and transcriptome characteristics of *D. antiqua*. This new dataset will provide a useful resource for future study of genetics and genomics of this species.

**Results and Discussion**

**Illumina sequencing and assembly**

In order to achieve an overall understanding of the *D. antiqua* transcriptome, a mixed cDNA sample obtaining from diverse developmental stages of this species was prepared and sequenced using the Illumina Hiseq2000 sequencing platform. Each sequenced sample yielded a 2*50-nt independent reads from either end of a cDNA fragment. We obtained a total of 53.81 million raw reads. After the removal of
raw reads that only had adaptor fragments, ambiguous and low-quality reads, 51.50 million (4.63Gnt, 95.7% of the raw reads) clean reads with a Q20 percentage of 98.2%, GC content of 39.03%, and unknown nucleotide ‘N’ of 0.00% remained. An overview of the sequencing, assembling and functional annotation results is presented in Table 1.

These clean data were assembled de novo by Trinity, producing 29,659 unigenes (longer than 200nt) (18Mnt), with an average length of 607nt and a N50 of 818nt (Table 1). Of these, 4,507 (15.2%) could be classified into distinct clusters, 25,152 (84.8%) were distinct singletons, 11,705 unigenes (39.5%) were longer than 500nt, and 5530 unigenes (18.6%) were longer than 1000nt. The length distributions of unigenes are shown in Additional File 1.

To test the assembly quality, 267 D. antiqua available cDNA (245 ESTs downloaded and 31 cDNA cloned in our lab) sequences were used as queries to blast all unigenes generated in our study with a stringent E-value of 1e-7. As a result, 227 queries were searched out with 82.2% (227/276) coverage.

**Functional annotation**

BLASTX alignments were conducted between the predicted protein sequences and several public databases, including NCBI Non-redundant protein database (Nr), NCBI Non-redundant nucleotide database (Nt), Swiss-Prot, Gene Ontology (GO) database, Cluster of Orthologous Groups (COG) database, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, with an E-value threshold of 1e-5. The results indicate that out of 29,659 unigenes, a total of 21,050 (71.0%) unigenes were annotated against Nr, 10,058 (33.9%) against Nt, 15,886 (53.6%) against Swiss-Prot, 13,238 (44.6%) against GO database, 6,600 (19%) against COG database, and 14,147 (43%) against the KEGG database obtained annotations.
(Table 1). Altogether, BLAST searches against Nr, Nt, Swiss-Prot, GO, COG and KEGG databases showed that a total of 21,605 (72.8% of 29,659 unigenes) identified unigenes could be annotated with known biological functions.

After Nr database annotation, results including the E-value distribution, identity distribution and species distribution were then analyzed [8] (Figure 1). For the E-value distribution of the predicted proteins, the top hits indicated that 41.2% of the mapped sequences had a significant similarity with a stringent threshold of less than 1e-45, and 58.8% of the similar sequences ranged from 1e-5 to 1e-45 (Figure 1A). For the similarity distribution...

![Figure 1](image)

**Figure. 1 NR classification of *D. antiqua* unigenes.** A) E-value distribution; B) Similarity distribution; C) Species distribution.
distribution of the predicted proteins, most of the sequences (53.5%) had a similarity higher than 60%, and 22.1% of the sequences had a similarity higher than 80% (Figure 1B). The species distribution showed that *D. antiqua* genes had the greatest number of matches with *Drosophila* genes. Among them, 11.5% of the unigenes had top matches and first hit against the sequences of *Dr. virilis* and 11.2% with sequences from *Dr. willistoni* and 9.5% with sequences from *Glossina morsitans*, followed by other species within *Drosophila*, including *Dr. mojavensis* (9.0%), *Dr. grimshawi* (7.8%), *Dr. ananassae* (7.5%), and *Dr. pseudoobscura* (6.6%), and 36.5% with sequences from other species (Figure 1C).

GO terms were assigned to *D. antiqua* unigenes for functional categorization, 13,238 unigenes were categorized into 60 subcategories belonging to 3 main categories, including biological process (25), molecular function (17), and cellular component (18) after assignment (Table 1). In the category of biological processes, the dominant GO terms were grouped into cellular processes and metabolic processes. Among the 18 subcategories of cellular components, those assignments were mostly given to cell part, and organelle. Within the molecular function category, there was a high-percentage of genes with binding, catalytic activity and transporter activity (Figure 2). These GO annotations revealed that diverse structural, regulatory, metabolic and transporter proteins were encoded by expressed genes in *D. antiqua*.

COG is a database which classifies gene products into different clusters of orthologous groups. As a result of the search of 29,659 unigenes against the
Figure 2 GO function classification of *D. antiqua* unigenes.

Figure 3 COG classification of *D. antiqua* unigenes.
COG database for orthologous genes, 6,600 *D. antiqua* unigenes were classified into 25 functional categories. The four largest categories were: 1) “General function prediction only” (2703, 40.1%); 2) “Transcription” (1431, 21.7%); 3) “Translation, ribosomal structure and biogenesis” (1308, 19.8%); 4) “Carbohydrate transport and metabolism” (1209, 18.3%). “Defense mechanisms” (109, 1.7%), “RNA processing and modification” (88, 1.3%), “Extracellular structures” (48, 0.7%) and “Nuclear structure” (3, 0.04%) represented the smallest groups (Figure 3).

Based on the KEGG database, the potential involvement of the assembled unigenes in biological pathways was annotated with corresponding Enzyme Commission (EC) numbers and pathways, thus provides inner-cell metabolic pathways and functions of gene products of unigenes. After searching 29,659 assembled unigenes against KEGG, a total of 14,147 unigenes were assigned to 257 KEGG pathways (Table 1). The pathways most represented by unique sequences were “metabolic pathways” with 1,935 unigenes (13.7%) that were associated with basic metabolic functions, followed by “Focal adhesion” (503, 3.6%), “pathways in cancer” (501, 3.5%), “RNA transport” (445, 3.1%) and “Regulation of actin cytoskeleton” (409, 2.9%). These functional annotations provided a basis for exploring specific biological processes, functions, subcellular localization, and pathways of gene products in *D. antiqua* research.

**Comparative analysis with D. antiqua and other Dipteran genomes**

In order to further understand the relationship between *D. antiqua* and other two Dipteran model species, *Dr. melanogaster* and *An. gambiae*, and to identify unigenes that might be unique to *D. antiqua*, BLASTX was used for the comparisons of *D. antiqua* to *Dr.*
*melanogaster* and *An. gambiae* with a cut-off E-value \( \leq 1 \times 10^{-5} \), because good BLAST hits are more easily obtained with amino acid sequences than with nucleotide sequences. Comparisons of unigenes to amino acid sequences of *Dr. melanogaster* and *An. gambiae* (E-value <1e-5 in BLASTX) showed that 17,922 (60.4%) *D. antiqua* unigenes had similarity hits to *Dr. melanogaster* peptides and 14,969 (50.5%) unigenes had similarity hits to *An. gambiae* peptides. Among these aligned sequences, 14,787 unigenes had similarities with both *Dr. melanogaster* and *An. gambiae*, while 3,135 (10.57%) unigenes only had similarity hits to *Dr. melanogaster* and 182 (0.61%) unigenes only had similarity hits to *An. gambiae* (Table 2 and Figure 4A). A total of 11,555 (38.96%) *D. antiqua* unigenes did not match any of *Dr. melanogaster* and *An. gambiae* sequences were assumed to be *D. antiqua* unique (Table 2 and Figure 4A).

After similarity search, GO functional classification were used to compare two groups of *D. antiqua* unigenes, one including shared homologs with *Dr. melanogaster* and *An. gambiae* and the other assumed to be unique to *D. antiqua* (Figure 4B). In all, among 14,787 shared homologs, there were 11,271 unigenes which were assigned with one or more GO terms. The GO analysis showed that for biological processes, genes involved in cellular processes and metabolic processes were highly represented. For molecular functions, binding activity was the most represented GO term, followed by catalytic activity. Regarding cellular components, the most represented categories were “cell” and “cell part”. Only 644 of 11,555 unigenes predicted to be unique to *D. antiqua* were annotated by GO analysis. This low annotation percent is probably due to the relatively small fraction of *D. antiqua* gene information available in public
databases, especially compared with *D. melanogaster* and *An. gambiae*.

Figure 4 *D. antiqua* unigenes similarity comparison with *Dr. melanogaster* and *An. gambiae* and functional classification by GO analysis. (A) similarity search of *D. antiqua* sequences against *Dr. melanogaster* and *An. Gambiae*; (B) Functional classification of *D. antiqua* unigenes with and without homologs with *Dr. melanogaster* and *An. gambiae*. 
There are some differences between annotated shared homologs and *D. antiqua* unique unigenes, such as GO term “synapse part” in cellular component; “auxiliary transport protein”, “metallochaperone”, “nutrient reservoir” and “protein tag” in molecular functions; and “cell killing” and “viral reproduction” in biological process annotation. On the one hand, we postulate that differences in the function allocation of the *D. antiqua* unigenes with sequence similarity hits to *D. melanogaster* and *An. gambiae* contribute to the divergence of *D. antiqua* from other insects. On the other hand, *D. antiqua* is famous for its diapause. The large bulk of predicted unigenes unique to *D. antiqua* represent a valuable resource to explore *D. antiqua* gene diversity and provide a basis for finding the genes involved in the specific physiological processes of diapause.

**Codon usage bias**

Codon usage bias is a phenomenon that unequal usage frequency of specific codon than other synonymous codons during the translation of genes. It provides a useful tool in functional genomic research and helps understand the physiological, biochemical and molecular mechanism of a gene or genome [9]. Factors influence the extent of codon usage bias including expression level, GC content, codon position, gene length, environmental stress and population size [10].

In our study, a total of 20,578 predicted open reading frames (ORFs) longer than 150nt in *D. antiqua* transcriptome were used in the codon usage bias analyses. Total counts of codons and the relative synonymous codon usage (RSCU) for these sequences were calculated. Average RSCU values showed that the seven most frequently used codons in *D. antiqua* are: CGU, UUA, GGU, UUG, GAU, AAU, and GCU. The seven
seldom used codons in *D. antiqua* are: GGG, CGG, CUC, GCG, AGG, GAC and AAC, without considering stop codons (Additional File 2). The results are very similar to the results obtained from Dipteran insects based on the genome-specific frequencies of the codons [11].

The average GC content of 20,578 ORFs was 39.14% and average GC content of all 29,659 unigenes in *D. antiqua* was 39.03%. These values are different from the GC content in the *An. gambiae* (55.8%) and *Dr. melanogaster* (53.9%) genomes [12]. Different species possess different GC content at the third codon position (GC3), which largely decide genome base composition. Our results showed the average GC3 was 28.36% in *D. antiqua* (Additional File 3).

The effective Number of Codons (Nc) quantifies how far the codon usage of a gene departs from equal usage of synonymous codons [13] and is a measure of codon usage biases in genes and genomes that ranges from 20 (maximal bias) to 61 (unbiased) [14; 15]. *D. antiqua* shows a medium degree of codon usage bias, as measured by the mean Nc value (45.06) (Additional Files 3). Nc plot (plot of Nc versus GC3) showing us relationship between codon usage bias and GC3, is widely used to study the codon usage variation among genes in different genomes [16]. In Nc plot, if the values of Nc fall on the continuous curve between Nc and GC3, then the codon usage variation among the genes was only determined by variation in GC3 content. In our study, most of the genes have an Nc value lower than expected on the curve, and fall within a restricted cloud, at GC3 between 0.056 and 0.928, and Nc values between 33.97 and 56.21 (Additional Files 4). The Nc plot showed here implied that the codon usage of a large number gene in *D. antiqua* are subject to other factors.
Gene expression level can heavily influence codon usage [10]. For instance, genes encoding ribosomal proteins are known to be highly expressed and tend to have more biased usage than lowly expressed genes [17]. Significant positive correlation between gene expression and codon usage bias has been studied on microarray data in many species [11]. To test the correlation between gene expression and codon bias, Codon Adaptation Index (CAI)[18] values and gene expression levels ($\log_{10} \text{FPKM}$) for each ORF (totally 20,578 ORFs) were compared. A significantly positive correlation between the CAI and gene expression levels was observed (Pearson Correlation: $r = 0.126$, $P = 1.28e^{-73}$) (Additional Files 3 and 5).

In this study, we calculated total codon usages and identified most frequently, seldom used codons. The influences of GC3 and expression level on codon usage bias in $D.\ antiqua$ were also tested. A large number of genes have an $Nc$ value lower than the expected value located on curve in $Nc$ plot, implying that the codon usage of many $D.\ antiqua$ genes are not just determined by GC3. A significant correlation between expression levels and codon usage was observed in $D.\ antiqua$. Knowledge of the codon usage pattern obtained in $D.\ antiqua$ help us understanding the mechanisms of codon usage bias and improving exogenous genes expression in future transgenic manipulation and thus serves as useful tool in functional genomic research in $D.\ antiqua$.

**SSR Discovery**

In the $D.\ antiqua$ transcriptome, 4,637 unigene sequences longer than 1 Kb
were used for SSRs identification. The result is that 315 sequences containing a total of 352 and 94 kinds of SSRs were identified, with 37 of the sequences containing more than 1 SSR (Table 3). The frequency of SSR in the *D. antiqua* transcriptome was 1 per 14.7 kilobases. The most abundant repeat motif was the tri-nucleotide, accounting for 82.7%, followed by the di-nucleotide repeat motif (14.7%), and tetra-nucleotide (2.5%) repeat units (Table 4). We also calculated the frequencies of SSRs with different numbers of tandem repeats (Table 4). SSRs with five tandem repeats (69.9%) were the most common, followed by six tandem repeats (21.6%), seven tandem repeats (5.4%), eight tandem repeats (1.1%), and four tandem repeats (1.7%). A detailed list of SSRs identified is shown in Additional File 6.

SSR markers are useful tool for assessing genetic variation and relationships in genetic mapping studies [19]. SSR markers developed from transcriptome data are easily and cheaper when compared with traditional isolation of genomic DNA-derived SSRs, because large-scale transcriptome sequencing programs based on NGS methods produced large amounts of sequence data with cheap price. Moreover, because mainly occur in the protein-coding regions of annotated unigenes, transcriptome-based SSRs are more benefit for identifying associations with functional genes and thus with phenotypes [20].

Efficient identifying of transcriptome-based SSRs have been reported in many organisms, such as in *Ma Bamboo* [21] and *Spodoptera exigua* [22], but no studies have been reported in *D. antiqua*. The unigenes obtained from *D. antiqua* have provided a good resource for SSRs mining. Different software (such as SSRIT, SPUTNIK, and SSRPrimer) [23; 24; 25] and parameter-setting used in detection may produce minor influences on the
efficiency of detection. In our study, 352 SSRs were identified and the tri-nucleotide repeat motif is the most abundant form of SSR repeat, consistent with the results from another insect species, *Sp. exigua* [22]. The SSR markers identified from *D. antiqua* transcriptome will serve as potential marker assists in genome mapping and identification of important functional genes of this species.

**Conclusions**

The onion maggot *Delia antiqua* is a major insect pest of cultivated vegetable and a good model for summer and winter diapauses studies. But the research of this species was restricted by the availability of research data and scientific resources. In order to establish a genomic resource, we used the Illumina Hiseq2000 sequencing platform to sequence *D. antiqua* transcriptome and produced 29,659 assembled unigenes with 21,050 (71%) obtaining annotation. This study dramatically increased the number of genes from the *D. antiqua* and will facilitate future genomic level studies in this species. To our knowledge, our results represent approximately 120-fold more genes than all *D. antiqua* genes deposited in GenBank (as of December, 2012). Certain numbers of them showed homology with other Dipteran genomes, of significance in evolutionary studies of Diptera. Other members that did not match any of *Dr. melanogaster* and *An. gambiae* sequences were assumed to be *D. antiqua* unique, offering a valuable resource for gene diversity research in *D. antiqua*. Research of the codon usage bias in *D. antiqua* help us understanding the mechanisms of codon usage and helpful for exogenous genes expression in this species. Furthermore, the transcriptome-based SSR marker identified in *D. antiqua* will help identification of pest-related genes and
contribute to genome mapping and pest control. We believe that results obtained from this study will play as a useful genomic resource to accelerate exploring the molecular mechanism of pest adaptation, the diapause mechanism and functional genomics in this important species.

Materials and methods

Insect rearing and RNA extraction

The non-diapausing colony (ND) of *D. antiqua* was maintained in the Institute of Entomology and Molecular Biology, Chongqing Normal University, China at 20 ± 0.2 °C under 50-70% relative humidity (RH) with a 16L:8D photocycle as previously described [26]. Larvae were raised at 25 ± 0.5 °C under 50-70% RH with 16L:8D to induce summer diapause (SD). Newly formed puparia were used for experiments. They were kept under the same conditions as the larvae until day 15 after pupariation and then transferred to 16 °C and 16L:8D to trigger diapause termination. For winter diapause (WD) induction, the conditions were 15 ± 0.5 °C, 50-70% RH with a 12L:12D through larvae and pupae. Ten samples were taken and then immediately frozen in liquid nitrogen for later RNA extraction. These 10 samples include ND eggs (from every incubation day), ND larvae (from every larval instar), ND pupae (female and male, from the stage equal to diapause development), ND adults (female and male, from different developmental stages), 3 SD pupae samples (female and male, sampled separately at diapause initiation, maintenance and termination), and 3 WD pupae samples (female and male, sampled separately at diapause initiation, maintenance and termination) of *D. antiqua*.

Total RNA was separately extracted from the 10 samples using TRIizol Reagent (Invitrogen, Carlsbad, CA, USA) following the manufacturer’s
protocol. To eliminate genomic DNA, the RNA samples were treated with RNase-Free DNase I according to manufacturer’s protocol (Qiagen, USA). The RNA integrity was confirmed using the Agilent 2100 Bioanalyzer with a minimum integrity number value of 7. Thirteen percent amounts of total RNA each from ND eggs, ND larvae and ND adults samples, and 8.7% of total RNA from each other 7 sample were pooled together for cDNA preparation.

**mRNA purification, and cDNA synthesis and Illumina sequencing preparation and sequencing**

Beads with Oligo(dT) were used to isolate poly(A) mRNA after total RNA extraction. Fragmentation buffer was added for cutting mRNA into short fragments. Taking these short fragments as templates, random hexamer-primer was used to synthesize the first-strand cDNA. The second-strand cDNA was synthesized using buffer, dNTPs, RNaseH and DNA polymerase I. Short fragments were purified with QiaQuick PCR extraction kit (Qiagen, USA) and resolved with EB buffer for end reparation and tailing A. After that, the short fragments were connected with sequencing adapters. And, after the agarose gel electrophoresis, the ligated products were purified and amplified with PCR to create the final cDNA library. The cDNA library was sequenced by Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China, using Illumina HiSeq™ 2000, according to manufacturer’s instructions.

**De novo transcriptome assembly**

The raw reads produced from sequencing machines were cleaned by removing adapter sequences, ambiguous reads (reads with unknown nucleotides ‘N’ larger than 5%) and low-quality sequences (reads with more than 10% Q<20 bases). The quality reads were assembled into unigenes using short reads assembling program - Trinity [27].
Reads that contain a certain length of overlap area were first joined to form longer fragments, which are called contigs without gaps. Then the reads are mapped back to contigs; with paired-end reads it is able to detect contigs from the same transcript as well as the distances between these contigs. Next, Trinity connects the contigs, and obtains sequences that no longer can be extended. Such sequences are defined as unigenes. The assembled sequences less than 200nt were deleted. At last, unigenes were divided into two classes by gene family clustering. One is clusters where several unigenes with a similarity higher than 70% are classified into one cluster with the prefix CL. And the other is singletons with the prefix is Unigene. FPKM for samples were calculated to show the expression quantity, thus avoiding the influence of sequencing length and difference [28]. Each FPKM was log_{10} transformed.

**Functional annotation**

The generated unigenes larger than 200nt were searched against Nr, Swiss-Prot, KEGG, COG and GO databases using BLASTX alignment (E-value ≤ 1e-5), and against Nt by BLASTN (E-value ≤ 1e-5). The best aligning results were used to decide sequence direction and the coding sequence (CDS) of unigenes, respectively. If results of different databases conflict with each other, a priority order of Nr, Nt, Swiss-Prot, KEGG and COG should be followed when deciding sequence direction of unigenes. ESTScan [29] was used to predict the sequence direction and CDS when unigenes were unaligned to any of the databases. GO annotations of the unigenes were determined by the Blast2GO program [30]. After obtaining GO annotation for each unigene, WEGO software [31] was used to display GO functional classification.

**Characterization of ORFs and Codon Usage**
The open reading frames (ORFs) in each unigene sequence were predicted by search against protein databases using BLASTX (E-value ≤1e-5) in the following order: Nr, SwissProt, KEGG, COG. Sequences having hits in the former database will not go to the next round search against a later database. The coding regions were then extracted according to the best BLASTX match with a custom perl script. CodonW (http://codonw.sourceforge.net/) [32] were used for analyzed RSCU and GC3 of ORFs (≥150nt). Perl scripts were used to process the output files of CodonW. We used SPSS 13.0 statistics software (SPSS Inc., Chicago, IL) for correlation analysis.

NC and CAI were calculated by DAMBE[33], using Delia antiqua ribosomal proteins codon usage table as the reference when calculated CAI.

**SSR Detection**

SSRs Identification Tool (SSRIT, http://www.gramene.org/db/markers/ssritool) [23] were used for detection of SSRs on unigenes of D. antiqua longer than 1kb. The parameters were designed for identifying perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motifs, with minimum thresholds of six, five, four, four and four repeats, respectively. Mononucleotide repeats were not considered due to the possibility of the Illumina homopolymer sequencing problem associated with this technology.

**Accession code**

The clean reads produced in this study have been deposited at DDBJ/EMBL/GenBank Short Read Archive under project number PRJNA208983, BioSample number SAMN02208942 and accession code SRR916227. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GAWI00000000.
The version described in this paper is the first version, GAWI01000000.

**Additional Files**

Additional File 1. Length distribution of unigenes in *D. antiqua*.

Additional File 2. Total codon usage and codon usage bias in the *D. antiqua* transcriptome.

Additional File 3. The ID and FPKM of unigenes, and the CAI, Nc and GC3 of ORFs in each unigene in *D. antiqua*.

Additional File 4. A plot of Nc versus GC3 (Nc-plot) for *D. antiqua* ORFs. The pink-dotted curve represents the expected curve between GC3 and Nc under random codon usage. A blue dot each indicates corresponding ORFs of each unigene.

Additional File 5. Relationship between CAI (Codon Adaptation Index) and expression level (Log10 (FPKM)) of all *D. antiqua* transcriptome unigenes.

Additional File 6. SSR identification of *D. antiqua* unigenes.

**Abbreviations**

CAI: Codon Adaptation Index; CDS: coding sequence; COG: Cluster of Orthologous Groups; FPKM: Fragment Per Kilobase per Million reads; GC3: GC content at the third codon position; GO: Gene Ontology; KEGG: the Kyoto Encyclopedia of Genes and Genomes; Nc: effective Number of Codons; Nc plot: A plot of Nc versus GC3; NGS: Next-Generation Sequencing; Nr: NCBI Non-redundant protein database; Nt: NCBI Non-redundant nucleotide database; ORFs: open reading frames; RNA-seq: next generation RNA sequencing; RSCU: relative synonymous codon usage; SSRs: simple sequence repeats.

**Competing interests**

The authors declare that they have no
competing interests.

Author contributions
YZ performed data analysis and drafted the manuscript. YH, FS, SR, GH and LS participated in sample collecting and experiments. BC conceived and designed the study, and joined in data analysis and manuscript drafting. All authors read and approved the final manuscript.

Acknowledgements
We appreciate the technical support for Illumina sequencing and initial data analysis provided by Beijing Genome Institute at Shenzhen, China. This work was supported by National Natural Science Foundation of China (Nos 31372265 and 31200947), IAEA Coordinated Research Project (No 18268/R0), National Institute of Health (R01 AI095184), Key Scientific and Technological Project of Chongqing (CSTC2012GG-YJJSB80002) and Par-Eu Scholars Program.

Author details
Institute of Entomology and Molecular Biology, College of Life Sciences, Chongqing Normal University, Chongqing, P.R. China.

References


Table 1: Statistics of RNA-seq based sequencing, assembling and functional annotation for *D. antiqua*.

<table>
<thead>
<tr>
<th>Sequencing results</th>
<th>Assembling results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of total raw reads</td>
<td>Number of unigenes</td>
</tr>
<tr>
<td>53,814,782</td>
<td>29,659 (4,507 into distinct clusters; 25,152 singletons)</td>
</tr>
<tr>
<td>Number of total clean reads</td>
<td>Total length (nt) of total unigenes</td>
</tr>
<tr>
<td>51,497,228</td>
<td>18,008,540</td>
</tr>
<tr>
<td>Number of total clean nucleotides (nt)</td>
<td>Mean length (nt) of total unigenes</td>
</tr>
<tr>
<td>4,634,750,520</td>
<td>607</td>
</tr>
<tr>
<td>Q20 percentage of total clean reads</td>
<td></td>
</tr>
<tr>
<td>98.24%</td>
<td></td>
</tr>
<tr>
<td>GC percentage of total clean nucleotides</td>
<td></td>
</tr>
<tr>
<td>39.03%</td>
<td></td>
</tr>
<tr>
<td>N percentage of total clean nucleotides</td>
<td></td>
</tr>
<tr>
<td>0.00%</td>
<td></td>
</tr>
</tbody>
</table>
N50 (nt) of total unigenes 818

### Annotation
- Unigenes with Nr database 21,050 (71.0%)
- Unigenes with Nt database 10,058 (33.9%)
- Unigenes with Swiss-Prot database 15,886 (53.6%)
- Unigenes with KEGG database 14,147 (47.7%), 257 pathways
- Unigenes with COG database 6,600 (22.3%), 25 functional categories
- Unigenes with GO database 13,238 (44.6%), 60 subcategories belonging to 3 main categories
  - Biological process 25 sub-categories
  - Cellular component 18 sub-categories
  - Molecular function 17 sub-categories
- Total unigenes annotated 21,605 (72.8% of 29,659 unigenes)

### Table 2 Comparative analysis between *D. antiqua* and other dipteran genomes using BLASTX with a cut-off E-value of 1E-5.

<table>
<thead>
<tr>
<th></th>
<th><em>Drosophila melanogaster</em></th>
<th><em>Anopheles gambiae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of a.a. sequences</td>
<td>27,538</td>
<td>14,324</td>
</tr>
<tr>
<td>Number of sequences with a.a. &gt; 50</td>
<td>27,410</td>
<td>14,296</td>
</tr>
<tr>
<td>Genome sequence version</td>
<td>r5.47</td>
<td>AgamP3.6</td>
</tr>
<tr>
<td>Source of genome sequence</td>
<td>Flybase</td>
<td>Vectorbase</td>
</tr>
<tr>
<td>With hits to</td>
<td>17,922 (60.43%)</td>
<td>14,969 (50.47%)</td>
</tr>
<tr>
<td>Only with hits to</td>
<td>3,135 (10.57%)</td>
<td>182 (0.61%)</td>
</tr>
<tr>
<td>With hits to both</td>
<td>14,787 (49.86%)</td>
<td></td>
</tr>
<tr>
<td>With no hits to both</td>
<td>11,555 (38.96%)</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3 Features of SSRs identified in the *D. antiqua* transcriptome

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of examined unigenes</td>
<td>29,659</td>
</tr>
<tr>
<td>Number of unigenes longer than 1 kb</td>
<td>4,637</td>
</tr>
<tr>
<td>Total nucleotides screened (knt)</td>
<td>7,423</td>
</tr>
<tr>
<td>Number of unigenes containing SSRs</td>
<td>315</td>
</tr>
<tr>
<td>Number of identified SSRs</td>
<td>352</td>
</tr>
<tr>
<td>Kinds of identified SSRs</td>
<td>94</td>
</tr>
<tr>
<td>Number of unigenes containing more than 1 SSRs</td>
<td>37</td>
</tr>
<tr>
<td>Frequency of SSR in transcriptome</td>
<td>1/14.7 Kb</td>
</tr>
</tbody>
</table>
## Table 4 Frequency of SSRs in *D. antiqua* transcriptome

<table>
<thead>
<tr>
<th>Number of nucleotides</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>&gt;10</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Di</td>
<td>-</td>
<td>-</td>
<td>40</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>52</td>
<td>14.7%</td>
</tr>
<tr>
<td>Tri</td>
<td>-</td>
<td>245</td>
<td>34</td>
<td>9</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>291</td>
<td>82.7%</td>
</tr>
<tr>
<td>Tetra</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9</td>
<td>2.5%</td>
</tr>
<tr>
<td>Penta</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hexa</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>246</td>
<td>76</td>
<td>19</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>352</td>
<td></td>
</tr>
</tbody>
</table>

% 1.7% 69.9% 21.6% 5.4% 1.1% 0.3% 0 0