Published in final edited form as:
Physiol Entomol. 2013 June ; 38(2): 173–181. doi:10.1111/phen.12016.

Transcriptome sequencing as a platform to elucidate molecular components of the diapause response in the Asian tiger mosquito, *Aedes albopictus*

MONICA F. POELCHAU[1], **JULIE A. REYNOLDS**[2], **DAVID L. DENLINGER**[2], **CHRISTINE G. ELSIK**[3], and **PETER A. ARMBRUSTER**[1]

[1]Department of Biology, Georgetown University, Washington, DC, U.S.A.

[2]Department of Entomology, Ohio State University, Columbus, Ohio, U.S.A.

[3]Divisons of Animal and Plant Sciences, S134-D Animal Sciences Research Center, University of Missouri, Columbia, Missouri, U.S.A.

Abstract

Diapause has long been recognized as a crucial ecological adaptation to spatio-temporal environmental variation. More recently, rapid evolution of the diapause response has been implicated in response to contemporary global warming and during the range expansion of invasive species. Although the molecular regulation of diapause remains largely unresolved, rapidly emerging next-generation sequencing (NGS) technologies provide exciting opportunities to address this longstanding question. Herein, a new assembly from life-history stages relevant to diapause in the Asian tiger mosquito, *Aedes albopictus* (Skuse) is presented, along with unique methods for the analysis of NGS data and transcriptome assembly. A digital normalization procedure that significantly reduces computational resources required for transcriptome assembly is evaluated. Additionally, a method for protein reference-based and genomic reference-based merged assembly of 454 and Illumina reads is described. Finally, a gene ontology analysis is presented, which creates a platform to identify physiological processes associated with diapause. Taken together, these methods provide valuable tools for analyzing the transcriptional underpinnings of many complex phenotypes, including diapause, and provide a basis for determining the molecular regulation of diapause in *Ae. albopictus*.

Keywords

Aedes albopictus; diapause; next-generation sequencing; transcriptome

Introduction

The annual rotation of the earth around the sun gives rise to various forms of recurring seasonal environmental variation that affect most life forms on Earth. Conspicuous examples include the annual arrival of winter in temperate habitats, wet and dry seasons in tropical habitats, and biotic interactions that can severely impact opportunities for survival and reproduction of a wide range of organisms. Photoperiodic diapause is a crucial ecological adaptation that allows many insects to cope with recurring seasonal environmental variation by using day length (photoperiod) as a token cue to initiate

Correspondence: Peter A. Armbruster, , Department of Biology, Georgetown University, 37th and O Streets NW, Washington, DC, U.S.A. Tel.: + 1 202 687 2567; paa9@georgetown.edu.

physiological changes that prepare the insect for persistence under adverse conditions. In temperate habitats, the seasonal timing of diapause initiation exhibits remarkably consistent geographic trends across latitudinal gradients (Anderwartha, 1952; Danks, 1987). The rapid evolution of diapause timing has been implicated in response to contemporary global warming (Bradshaw & Holzapfel, 2001), and also during range expansion of invasive species (Bean *et al.*, 2012; Urbanksi *et al.*, 2012). Thus, studies of diapause address both basic questions concerning organismal adaptation to spatio-temporal environmental heterogeneity, as well as more applied questions such as anticipating biological responses to rapid ongoing global warming and the evolution of invasive species during range expansion.

Despite the well-established ecological significance of photoperiodic diapause, the molecular regulation of this crucial adaptation remains largely unresolved. In part, this is because diapause is a physiologically complex process that involves multiple regulatory hierarchies across the trajectory from diapause preparation, to diapause maintenance and eventually termination (Kostal, 2006). A number of common physiological themes are associated with diapause in a diverse range of insects, including reduced metabolism and developmental arrest (Tauber *et al.*, 1986), increased cold and desiccation resistance (Rinehart *et al.*, 2007; Benoit, 2010), and increased nutrient storage (Hahn & Denlinger, 2011). However, current understanding of the molecular regulation of these physiological processes during diapause is mostly based on studies of one or several genes in a diverse range of insect taxa. The extent to which common molecular regulatory mechanisms underlie the consistent physiological themes of diapause remains unclear. Thus, elucidation of the molecular regulation of diapause across the trajectory from preparation to termination remains an elusive goal with important implications for a wide range of both basic and applied questions in insect science.

Rapid advances in next-generation sequencing (NGS) technology are opening up exciting opportunities for elucidating the molecular basis of complex phenotypes, even in non-model organisms for which complete genome sequences are not available (Davey *et al.*, 2011; Martin & Wang, 2011). This is an especially exciting opportunity for diapause research, since many insect species in which diapause has been studied thoroughly from an ecological, physiological, or evolutionary perspective do not currently have a complete genome sequence available (e.g., Wyeomyia smithii, Sarcophaga crassipalpis, Culex pipiens, Aedes albopictus). Nevertheless, many challenges remain in applying NGS to elucidate the molecular basis of diapause. The amount of data from a single NGS experiment can be overwhelming, requiring substantial computational resources for analysis. Additionally, working with a non-model organism for which a genome sequence is not available requires creative approaches to assembly, annotation and estimation of differential gene expression by read mapping.

In this paper, methods are described for the development of transcriptomic resources from multiple NGS diapause experiments in the Asian tiger mosquito, *Aedes albopictus* (Skuse). A digital normalization procedure that significantly reduces computational resources required for assembly of hundreds of millions of Illumina reads is evaluated. Additionally, a method is described for protein reference-based and genomic reference-based merged assembly of 454 and Illumina reads. The resulting merged assembly leads to increased gene discovery and annotation. Finally, results of a gene ontology (GO) analysis are presented which establish a basis for identifying physiological processes associated with diapause. The results of the GO analysis are discussed in relation to previous results on the transcriptional basis of diapause in *Ae. albopictus*. These methods illustrate a valuable set of analytical tools that provide a basis for determining the molecular regulation of diapause in *Ae. albopictus*, as well as the transcriptional underpinnings of many other complex phenotypes.

Materials and methods

Study organism

Aedes albopictus females typically oviposit desiccation-resistant eggs above the water line in a variety of natural (e.g., tree holes) or artificial (e.g., tyres) containers. Once embryological development is complete, a non-diapause pharate larva inside the chorion of the egg persists in a quiescent state and typically hatches within 1 h of submersion in water with reduced oxygen content. In contrast, a diapause pharate larva is refractory to hatching stimuli for up to several months. In temperate populations, exposure of the maternal pupa and adult to short day lengths induces diapause (Wang, 1966; Mori *et al.*, 1981). RNA was sequenced from pharate larvae (eggs) at three time points after oviposition in order to assemble a trascriptome of gene expression across the trajectory from the early to late developmental arrest phases of diapause and quiescence.

Insect rearing and RNA generation

A laboratory F_{13} *Ae. albopictus* strain collected from Manassas, Vancouver, U.S.A. was reared at 21 °C, ca. 80% relative humidity and a long-day photoperiod (LD 16:8 h) as described previously (Armbruster & Hutchinson, 2002; Armbruster & Conn, 2006). Upon pupation, mosquitoes were divided into eight replicate cohorts of approx. 100 mosquitoes per cohort, each in a ca. 9.5-L cage. Four cohorts (biological replicates) were maintained under a diapause-inducing photoperiod (D; LD 8:16 h) and four cohorts were maintained under a non-diapause-inducing photoperiod (ND; LD 16:8 h). Females were allowed to blood-feed on a human host 9-16 days after eclosion, and again 7 and 14 days later, so that eggs could be collected over multiple gonotrophic cycles. Three days after the first blood feeding, a small brown jar half-filled with *ca.* 50 mL of deionized water and lined with unbleached seed germination paper was placed into each cage 6-7 h after lights-on to stimulate oviposition. The seed germination paper containing eggs was removed after 24 h and this procedure was repeated for twenty-six consecutive days. Egg papers were gently air-dried 72 h after collection, and then stored at *ca.* 80% relative humidity on a LD 8:16 h photoperiod until further treatment; short-day photoperiods experienced at the egg stage do not result in diapause in *Ae. albopictus* (Mori *et al.*, 1981). A subset of eggs from each replicate cohort was reserved for diapause incidence and measurements (see below). The remainder was divided and allowed to develop for 11, 21, and 40 days post-oviposition. Upon completion of the specified development period, individual egg samples were carefully brushed into 2-mL microcentrifuge tubes, snap-frozen in liquid nitrogen, and stored at -80 °C. After collection of all samples, eggs were pooled according to treatment (D, ND) and development period (11, 21 and 41 days post-ovoposition), ground in TRI® Reagent (Sigma Aldrich, St. Louis, Missouri), and RNA was extracted according to manufacturer's instructions. DNA was removed from each sample with Turbo-DNA free (Applied Biosystems/Ambion, Austin, Texas). RNA quality was assessed on an RNA chip (Bioanalyzer 2100, Agilent Technologies, Santa Clara, California). Three of the four biological replicates from each treatment and development period were chosen for sequencing based on RNA quality and quantity, with the exception of 40 days post-oviposition pharate larvae reared under ND conditions, for which only two biological replicates showed sufficient RNA quality for high-throughput sequencing. This resulted in 17 sequenced RNA libraries (see below). Incubator malfunction caused some 40d eggs to experience a *ca.* 4 °C fluctuation on three consecutive days. Eggs scheduled for snap-freezing on these days were discarded so that all eggs had at least 24 h to recover. Because ND and D eggs were stored together, temperature fluctuations should not result in systematic differences in gene expression between ND and D treatments.

For diapause incidence measurements, eggs ranging from 2-4 weeks of age were hatched, the number of hatched larvae recorded, and the egg papers re-dried. This procedure was repeated after 7 and 14 days. After the last hatch, eggs were counted and bleached as described by Trpis (1970) to record the number of embryonated but unhatched (= diapause) eggs. Diapause incidence was calculated as - (# embryonated unhatched eggs) / (# hatched eggs + # embryonated unhatched eggs) (Urbanski *et al.*, 2012). Percentage embryonation was calculated as (# embryonated unhatched eggs + # hatched eggs) / total # eggs. Counts were pooled across collection dates within each replicate for the final calculations.

Sequencing

Illumina paired-end mRNA-Seq library construction and sequencing was performed by The University of Maryland Genomics Institute following the TruSeq RNA sample preparation guide (v2). The 17 libraries were bar-coded (Wong *et al.*, 2013) according to manufacturer's instructions and sequenced on three flow-cell lanes on an Illumina HiSeq 2000 high-throughput sequencer, where a proportion of each library was sequenced on each lane (average fragment size: 320 bp; average read length: 101 bp). Raw reads are available in NCBI's short read archive under BioProject accession PRJNA187045.

Assembly and annotation

Read cleaning. Reads with matches to the NCBI UniVec Core database, which contains common vector, adapter, linker and primer contaminants (ftp://ftp.ncbi.nih.gov/pub/UniVec/; accessed August 10 2010), as well as *Ae. albopictus* rRNA sequence (GenBank # L22060.1), and Illumina multiplexed, paired-end mRNA-Seq adapters were identified using sshaha2 (Ning *et al.*, 2001) and removed along with their read pair. Match criteria for removal were set at 95% identity (rRNA, UniVec) or 90% (adapters), and a Smith-Waterman score larger than 18 (UniVec). The first 15 bp of each read was trimmed because Illumina libraries frequently display low k-mer diversity in the first 12-15 bp of each read, suggesting non-random priming during the Illumina mRNA-Seq library preparation (see Poelchau *et al.*, 2013). The remaining reads were cleaned using SolexaQA (v. 1.10) (Cox *et al.*, 2010) by trimming regions with masked contaminant sequence or a phred score equivalent of less than 20 and removing reads shorter than 25 bp.

Digital normalization and reference-free assembly. A novel “digital normalization” technique (Brown *et al.*, 2012) was used to reduce the computational resources required for contig assembly. The great sequencing depth and large sequencing error rate of many NGS experiments results in large and error-prone datasets that are difficult and computationally expensive to assemble. To circumvent this problem, the “digital normalization” method identifies all k-mers (short DNA-sequences) of a particular length in a dataset, and eliminates all reads that contain a k-mer over a given abundance. This method effectively removes the majority of erroneous k-mers, while keeping almost all real k-mers in the dataset, thereby drastically reducing the number of redundant reads and the computational resources required for assembly (Brown *et al.*, 2012). One round of digital normalization was performed on the cleaned reads using parameters almost identical to Brown *et al.* (2012) (using k-mer size 20, and a coverage cut-off of 20; the “x” parameter, which determines memory usage, was set to 1e^{10}). The program was downloaded from https://github.com/ged-lab/khmer/tree/2012-paper-diginorm.

The reduced read set was assembled into contigs with Velvet (Zerbino & Birney, 2008), using k-mers from 19 to 59 in 10 k-mer-intervals and a coverage cut-off of 10. The resulting contigs were then further assembled by Oases, which clusters contigs generated by Velvet into loci, and uses the paired-end information to generate transcript isoforms (Schulz *et al.*, 2012). Assemblies from separate k-mers were merged using CD-HIT-EST (Li & Godzik, 2006), which grouped sequences with 99% identity over the entire sequence length.

To evaluate the efficacy of the digital normalization technique for this dataset, a comparison of test assemblies using a normalized and non-normalized subset of the data was performed. Specifically, assemblies were constructed as outlined above for the D treatment at 11 days post-oviposition using either a digitally normalized or complete set of reads. The first measure of assembly quality was the number of contigs, where a smaller number generally indicates lower contig redundancy, and thus an increased assembly quality. Second, assembly completeness was evaluated by mapping the non-normalized read set back to each assembly using bowtie (version 0.12.7) (Langmead *et al.*, 2009) based on the rationale that more of the original reads would map back to the more complete assembly. Third, contig length was compared between the two assemblies, where longer contigs likely represent more completely reconstructed transcripts.

Previous Data. Previously, transcriptomes were generated from *Ae. albopictus* oocyte (Poelchau *et al.*, 2011) and embryonic (Poelchau *et al.*, 2013) tissue. To generate a comprehensive *Ae. albopictus* transcriptome assembly, the present assembly was merged with the previous assemblies using a reference-based approach, similar to Poelchau *et al.* (2013) and Surget-Groba & Montoya-Burgos (2010). Assembled contigs and unassembled reads from 454 sequencing of oocyte tissue were derived from Poelchau *et al.* (2011). Contigs from Illumina reads generated from embryonic tissue (Poelchau *et al.*, 2013) and assembled in Velvet (Zerbino & Birney, 2008) and Oases (Schulz *et al.*, 2012) followed by CD-HIT-EST (Li & Godzik, 2006) were also added. Unassembled Illumina reads from the embryo tissue were not included because the short reads of Illumina sequencing limits the utility of these data for the current merged assembly.

Protein reference-based assembly

A non-redundant dipteran protein set was generated from Ae. aegypti, Culex quinquefasciatus, Anopheles gambiae, and *Drosophila melanogaster* protein sequences, downloaded as ortholog groups from OrthoDB v.4 (Waterhouse *et al.*, 2011). For each ortholog group, a single protein sequence from the taxonomically closest organism to *Ae. albopictus* was retained (*Ae. aegypti*, then *Cx. quinquefasciatus*, then *An. gambiae*, then *D. melanogaster*). The final reference set, which comprised 21,066 proteins, maximized protein sequence diversity and eliminated redundancy, thereby leading to increased computational efficiency.

Unassembled reads from the oocytes and all contigs were aligned to the dipteran protein set using fastx (Pearson *et al.*, 2007). The alignment with the lowest e-value ≤ $1e^{-6}$ was retained for further analysis. Contigs assigned to the same reference protein and at least 95% identity of the overlapping sequence were merged in cap3 (Huang & Madan, 1999). The annotations of the re-assembled contigs were verified by again aligning them to the reference protein set in fastx; only alignments with >70% identity to the best matching reference protein were used in the final, annotated assembly. Chimeric contigs were identified by searching for additional alignments outside of the primary alignment that were within 80% of the primary alignment's percentage identity. Contigs that met these criteria were considered chimeric and discarded.

Genomic reference-based assembly

Contigs and oocyte reads that did not align to the dipteran protein set described above were matched to *Ae. aegypti* genomic supercontigs (Nene *et al.*, 2007) using *blastn* (e-value < $1e^{-6}$), and then aligned in *exonerate* (Slater & Birney, 2005) (parameters were est2genome --

softmaskquery --bestn 1 --dnahspdropoff 0). At least 72.8 % identity of alignments was required for further analysis. Contigs were merged in cap3 if they overlapped and re-aligned, as above, retaining only contigs with >70% identity in their re-alignments. If contigs spanned multiple gene models, they were considered chimeric and discarded. Contigs that matched within 1 kb up- or down-stream of annotated genes over >90% of their length were annotated as potential UTRs for those genes. The final, annotated assembly included contigs that matched both annotated and un-annotated regions of the genome.

Functional characterization of the transcriptome

Gene ontology (GO) and GO-Slim (Ashburner *et al.*, 2000) categories were downloaded for each reference gene model from BioMart (Haider *et al.*, 2009). The numbers and composition of each group were then tallied. GO-Slim categories were investigated in addition to GO categories in order to organize genes into broader functional groups whose biological significance in the context of diapause can be more easily summarized.

Results

Diapause incidence

Diapause incidence ranged from 87.5 to 100% in biological replicates of the diapause-inducing photoperiod treatments (D). Percentage embryonation ranged from 82.9 to 98.9% across all replicates, with an average of 90.7%.

Assembly and annotation

The final assembly, and a spreadsheet with annotations for each gene model, is provided at http://www.albopictusexpression.org/?q=data.

Read cleaning and digital normalization of sequences from diapause and non-diapause eggs

Illumina sequencing of 17 RNA libraries from D and ND eggs (pharate larvae) produced 648,339,954 reads (324,169,167 pairs), of which 602,178,150 (92.9%) remained after quality control. Digital normalization of the quality-filtered reads reduced the total number of reads to 35,833,461 (5.5% of the original reads; Fig. 1), substantially reducing the computational resources necessary to complete the assembly (memory usage scales roughly linearly with the number of reads used: http://listserver.ebi.ac.uk/pipermail/velvet-users/2009-June/000359.html). The digitally normalized assembly was performed in much less time and required less memory (digitally-normalized assembly, *ca.* 4.5 h, 11 Gb maximum memory usage; non-normalized assembly, *ca.* 43 h, 62 Gb maximum memory usage).

Digitally normalized and non-normalized test assemblies on sequences from the day 11 D sample showed that the normalized assembly out-performed the non-normalized assembly in most aspects (Table 1). In the digitally normalized assembly 19% fewer contigs were generated and 10% more reads mapped back to the assembly. However, the normalized assembly had a 9% shorter average contig length. Because the substantially reduced computational requirements, smaller contig number, and greater number of mapped reads indicated more advantages than disadvantages to the digital normalization procedure, the digitally normalized full dataset was used in the analysis.

Assembly metrics - reference-free assembly from diapause and non-diapause eggs

Velvet assembly (Zerbino & Birney, 2008) of sequences from D and ND eggs (pharate larvae) followed by Oases (Schulz *et al.*, 2012) yielded 311,071 contigs, which were reduced to 176,502 non-redundant contigs using CD-HIT-EST (Table 2). Combining this assembly with the oocyte (Poelchau *et al.* 2011) and embryo (Poelchau *et al.* 2013) reference-free assemblies resulted in 627,154 contigs (Fig. 1).

Assembly metrics: protein-reference assembly

Of all reference-free contigs described above, 43.3% aligned to the dipteran protein set, generating 61,624 quality-filtered, re-assembled contigs after cap3 re-assembly (Table 2; Fig. 1). This substantially reduced the number of redundant contigs annotated as a given gene model. 4,254 re-assembled contigs were considered chimeric and discarded. 12,139 gene models were identified in the protein-based assembly (Fig. 1).

Assembly metrics: genomic-reference assembly

A number, 24.7%, of reference-free contigs were used in the genomic cap3 re-assembly resulting in 28,079 quality-filtered, re-assembled contigs that matched to annotated regions of the *Ae. aegypti* genome (25,603) or potential UTRs (2,476). An additional 28,460 re-assembled contigs matched to un-annotated regions of the genome. Of the reassembled contigs 2,311 were chimeric, and therefore discarded. Re-assembled contigs in the genome-based assembly recovered 7,629 gene models, of which 1,122 were not found in the protein-based assembly.

Assembly metrics: unaligned contigs

239,091 reference-free contigs remained that did not meet any of the alignment criteria, and therefore were not re-assembled, or included in the annotated assembly (Fig. 1). An inspection of these remaining sequences revealed that 9,777 aligned to the UniProt database (*blastx*, e-value < $1e^{-3}$), and 57,327 sequences derived from the pharate larval assembly aligned to un-annotated sequences from previous assemblies (Poelchau *et al.,* 2011, 2013; *blastn*, e-value < $1e^{-6}$; data not shown). These sequences could thus include rapidly diverged sequence specific to *Ae. albopictus*, and viral or bacterial sequences. The remaining sequences will be aligned to the *Ae. albopictus* genome sequence, once it is completed, for further annotation.

Both protein and genomic reference assemblies had high percentage identity to gene models, similar to the previous *Ae. albopictus* transcriptome assembly containing only reads from oocyte and embryo stages (median percentage identity, protein: 91.5% , genomic: 83.5%; Fig. 2). A substantial percentage of the annotated contigs' length was contained in the sequence alignment (median percentage of contig in the alignment, protein: 69.1%, genomic: 83.1%; Fig. 2). Similarly, the proportion of reference sequence covered by contigs in protein alignments was high (median reference coverage: 85.3%; Fig. 2). Due to the nature of sequence alignments using the genomic reference, genomic reference coverage was correspondingly low (median reference coverage: 12.4%; Fig. 2), which suggests that the majority of many genomic alignments occurred outside of, or adjacent to, annotated sequence on the *Ae. aegypti* genome. It is also likely that many un-annotated UTR regions of the *Ae. aegypti* genome were contained in the alignments, or it may be possible that *Ae. albopictus* genes have different gene boundaries than their *Ae. aegypti* homologs.

The number of gene models identified increased from 11,505 in the oocyte assembly (Poelchau *et al.*, 2011), to 12,345 in the merged embryo and oocyte assembly (Poelchau *et al.*, 2013), to 13,261 in the final merged assembly from all three experiments reported here. This result affirms that new life-history stages included in the assembly increase the assembly's gene content, likely due to the addition of genes with unique, stage-specific expression. A total of 17,356 gene models have been annotated in the *Ae. aegypti* genome sequence (Assembly AaegL1.3, Vector Base: https://www.vectorbase.org/organisms/aedes-aegypti/liverpool-lvp/AaegL1.3, accessed December 30, 2012), suggesting that a large proportion of the genes from the *Ae. albopictus* genome are represented in the current full assembly presented here.

Functional characterization

The new assembly added 125 new GO categories and 9 new GO-Slim categories relative to the previous assembly (a full list of GO categories is available on the website, http://www.albopictusexpression.org/?q=data). The new GO-Slim categories added a diversity of functions that likely contribute to physiological and developmental processes during the pharate larval stage in *Ae. albopictus*. Several of these new categories were related to cellular structure and growth (cilium, cell proliferation, cell junction organization, pigmentation), translation (structural constituent of ribosome), and metabolism (protein transporter activity, enzyme regulator activity, GTPase activity, nitrogen cycle metabolic process). Many genes in the full assembly had functions with known relevance to diapause, including 231 genes related to lipid metabolism and 165 related to stress response (Table 3). This demonstrates that this assembly represents a rich resource for potential diapause-related genes.

Discussion

Temperate populations of *Ae. albopictus* survive through winter by entering diapause as a pharate larva inside the chorion of the egg (Mori *et al.*, 1981; Wang, 1966). Diapause eggs are more cold resistant (Hawley *et al.*, 1987), desiccation resistant (Sota & Mogi, 1992; Urbanski *et al.*, 2010), are larger and harbour greater lipid reserves (Reynolds *et al.*, 2012), relative to non-diapause eggs. In addition to enhancing overwinter survival, all of these properties are likely to contribute to the ability of *Ae. albopictus* eggs to survive long distance transport, and therefore may be important factors contributing the world-wide spread of this invasive mosquito. Diapause timing has evolved rapidly among populations across the invasive U.S.A. range of *Ae. albopictus* in just a 20-year period (Urbanski *et al.*, 2012), further emphasizing the central importance of diapause to invasion and range expansion of this mosquito.

Rapidly advancing NGS technologies and analysis tools provide exciting opportunities to determine the molecular regulation of complex life-history adaptations such as diapause. For example, new techniques are emerging that facilitate transcriptome assembly, which can require prohibitively large computational resources (Brown *et al.* 2012). The digitally normalized transcriptome assembly presented here demonstrates that such techniques can be used to successfully assemble large Illumina datasets with significant advantages relative to non-normalized data (Table 1). Additionally, previous assemblies of the *Ae. albopictus* transcriptome included sequence data from pre-diapause and non-diapause oocytes (Poelchau *et al.*, 2011) and embryos (Poelchau *et al.*, 2013), encompassing only some of the life history stages that a diapause-destined mosquito experiences. In the present assembly, transcriptome data from multiple pharate larval stages is added in order to generate a comprehensive transcriptome encompassing all diapause stages and the non-diapause counterparts. As anticipated, sequence data from pharate larval stages in the present, comprehensive, assembly increases both the number of gene models in the assembly and the number of functional groups based on Gene Ontology assignments.

The reference-based assembly technique presented in the present study reduces contig redundancy and increases contig length relative to reference-free assemblies (Table 2), while creating contigs with high-confidence annotations to reference gene sets from other organisms, and identifying potential novel UTR sites. Similar to Poelchau *et al.* (2013), the protein-based reference assembly had longer contigs than the genome-based reference assembly (Table 2). However, the genome-based reference assembly added 1,122 additional gene models to the assembly, as well as potential UTRs, which will be of future utility in identifying regulatory regions. Therefore, this “hybrid” transcriptome assembly method can be a useful strategy for other taxa that are closely related to an organism with an available genome sequence.

Previous studies have identified several physiological themes that appear to be shared across the diapause response of multiple insect species. Diapausing insects demonstrate up-regulation of stress-response genes (Denlinger *et al*., 2005; Rinehart *et al*., 2007), increased lipid synthesis and storage (Reynolds & Hand, 2009; Robich & Denlinger, 2005; Sim & Denlinger, 2009a), changes in insulin signalling (Hahn & Denlinger, 2011; Sim & Denlinger, 2008, 2009b; Tatar *et al*., 2001; Williams *et al*., 2006), shifts in metabolism (Kukal *et al*., 1991; Michaud & Denlinger, 2007; Ragland *et al*., 2010; Hahn & Denlinger, 2011), and changes in patterns of cell cycle arrest (Tammariello & Denlinger, 1998; Tammariello, 2001; Reynolds & Hand, 2009). Analysis of diapause-related expression changes of genes relevant to these physiological processes can illuminate the specific mechanisms involved in diapause-driven physiological change (e.g., Sim & Denlinger 2008, 2009a; Ragland *et al*., 2010, 2011; Reynolds *et al*., 2012). In the present assembly, many genes in potentially diapause-relevant pathways, such as lipid metabolism and stress response, are identified (Table 3). One approach to using this resource to investigate the transcriptional regulation of diapause-associated physiological processes is to identify candidate genes that can then be screened by quantitative RT-PCR for differential expression (e.g., Reynolds *et al*., 2012) or functionally evaluated using RNAi knockdown experiments (Sim & Denlinger, 2008, 2009b). Additionally, the original reads from the transcriptome sequencing can be mapped back to the transcriptome assembly to quantify gene expression (e.g., Poelchau *et al*., 2011, 2013). Thus, the *Ae. albopictus* diapause transcriptome presented here represents a comprehensive foundation to elucidate the molecular basis of key traits underpinning geographic adaptation and invasion success in this invasive mosquito.

Acknowledgments

We thank Shin Goto, Jim Hardie and Kyeong-Yeoll Lee for the opportunity to contribute to this special issue and to participate in the symposium on New Complexities in the Regulation of Insect Diapause and Cold Hardiness at the 2012 International Congress of Entomology in Daegu, South Korea. This work was supported by the National Institutes of Health grant 5R21AI081041-02 and Georgetown University.

References

Andrewartha HG. Diapause in relation to the ecology of insects. Biological Reviews. 1952; 27:50–107.

Armbruster PA, Conn JE. Geographic variation of larval growth in North American *Aedes albopictus* (Diptera: Culicidae). Annals of the Entomological Society of America. 2006; 99:1234–1243.

Armbruster PA, Hutchinson RA. Pupal mass and wing length as indicators of fecundity in *Aedes albopictus and Aedes geniculatus* (Diptera : Culicidae). Journal of Medical Entomology. 2002; 39:699–704. [PubMed: 12144308]

Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]

Bean DW, Dalin P, Dudley TL. Evolution of critical day length for diapause induction enables range expansion of *Diorhabda carinulata*, a biological control agent against tamarisk (Tamarix spp.). Evolutionary Applications. 2012; 5:511–523. [PubMed: 22949926]

Benoit, JB. Water management by dormant insects: comparisons between dehydration resistance during summer aestivation and winter diapause.. In: Navas, CA.; Caralho, JE., editors. Aesivation: Molecular and Physiological Aspects. Springer-Verlag; Berlin: 2010. p. 209-229.

Bradshaw WE, Holzapfel CM. Genetic shift in photoperiodic response correlated with global warming. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:14509–14511. [PubMed: 11698659]

Brown CT, Howe A, Zhang Q, et al. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802. 2012

Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 2010; 11:485. [PubMed: 20875133]

Danks, HV. Insect dormancy: An ecological perspective. Biological Survey of Canada (Terrestrial Arthropods); Ottowa, Canada: 1987.

Davey JW, Hohenlohe PA, Etter PD, et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics. 2011; 12:499–510.

Denlinger, DL.; Yocum, GD.; Rinehart, JL. Hormonal control of diapause. Gilbert, L.; Iatrou, K.; Gill, S., editors. Elsevier Press; Amsterdam, The Netherlands: 2005. p. 615-650.

Hahn DA, Denlinger DL. Energetics of insect diapause. Annual Review of Entomology. 2011; 56:103–121.

Haider S, Ballester B, Smedley D, et al. BioMart Central Portal-unified access to biological data. Nucleic Acids Research. 2009; 37:W23–W27. [PubMed: 19420058]

Hawley WA, Reiter P, Copeland RS, et al. *Aedes albopictus* in North America: Probable introduction in used tires from northern Asia. Science. 1987; 236:1114–1116. [PubMed: 3576225]

Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Research. 1999; 9:868–877. [PubMed: 10508846]

Kostal V. Eco-physiological phases of insect diapause. Journal of Insect Physiology. 2006; 52:113–127. [PubMed: 16332347]

Kukal O, Denlinger DL, Lee RE. Developmental and metabolic changes induced by anoxia in diapausing and nondiapausing flesh fly pupae. Journal of Comparative Physiology B-Biochemical Systemic and Environmental Physiology. 1991; 160:683–689.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10:R25. [PubMed: 19261174]

Li WZ, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22:1658–1659. [PubMed: 16731699]

Martin JA, Wang Z. Next-generation transcriptome assembly. Nature Reviews Genetics. 2011; 12:671–682.

Michaud MR, Denlinger DL. Shifts in the carbohydrate, polyol, and amino acid pools during rapid cold-hardening and diapause-associated cold-hardening in flesh flies (*Sarcophaga crassipalpis*): a metabolomic comparison. Journal of Comparative Physiology B-Biochemical Systemic and Environmental Physiology. 2007; 177:753–763.

Mori A, Oda T, Wada Y. Studies on the egg diapause and overwintering of *Aedes albopictus* in Nagasaki. Tropical Medicine. 1981; 23:79–90.

Nene V, Wortman JR, Lawson D, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science. 2007; 316:1718–1723. [PubMed: 17510324]

Ning ZM, Cox AJ, Mullikin JC. SSAHA: A fast search method for large DNA databases. Genome Research. 2001; 11:1725–1729. [PubMed: 11591649]

Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography. 2007; 34:102–117.

Poelchau MF, Reynolds JA, Denlinger DL, et al. A de novo transcriptome of the Asian tiger mosquito, *Aedes albopictus*, to identify candidate transcripts for diapause preparation. BMC Genomics. 2011; 12:619. [PubMed: 22185595]

Poelchau MF, Reynolds JA, Denlinger DL, et al. Deep sequencing reveals complex mechanisms of diapause preparation in the invasive mosquito, *Aedes albopictus*. Proceedings of the Royal Society of London B. 2013 in press.

Ragland GJ, Denlinger DL, Hahn DA. Mechanisms of suspended animation are revealed by transcript profiling of diapause in the flesh fly. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:14909–14914. [PubMed: 20668242]

Ragland GJ, Egan SP, Feder JL, et al. Developmental trajectories of gene expression reveal cadidates for diapause termination: a key life-history transition in the apple maggot fly Rhagoletis pomonella. Journal of Experimental Biology. 2011; 214:3948–3959. [PubMed: 22071185]

Reynolds JA, Hand SC. Embryonic diapause highlighted by differential expression of mRNAs for ecdysteroidogenesis, transcription and lipid sparing in the cricket *Allonemobius socius*. Journal of Experimental Biology. 2009; 212:2074–2083.

Reynolds JA, Poelchau MF, Rahman Z, et al. Transcript profiling reveals mechanisms for lipid conservation during diapause in the mosquito, *Aedes albopictus*. Journal of Insect Physiology. 2012; 58:966–973. [PubMed: 22579567]

Rinehart JP, Li A, Yocum GD, et al. Up-regulation of heat shock proteins is essential for cold survival during insect diapause. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:11130–11137. [PubMed: 17522254]

Robich RM, Denlinger DL. Diapause in the mosquito *Culex pipiens* evokes a metabolic switch from blood feeding to sugar gluttony. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:15912–15917. [PubMed: 16247003]

Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012; 28:1086–1092. [PubMed: 22368243]

Sim C, Denlinger DL. Insulin signaling and FOXO regulate the overwintering diapause of the mosquito *Culex pipiens*. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:6777–6781. [PubMed: 18448677]

Sim C, Denlinger DL. A shut-down in expression of an insulin-like peptide, ILP-1, halts ovarian maturation during the overwintering diapause of the mosquito *Culex pipiens*. Insect Molecular Biology. 2009a; 18:325–332. [PubMed: 19523064]

Sim C, Denlinger DL. Transcription profiling and regulation of fat metabolism genes in diapausing adults of the mosquito *Culex pipiens*. Physiological Genomics. 2009b; 39:202–209. [PubMed: 19706691]

Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005; 6:31. [PubMed: 15713233]

Sota T, Mogi M. Survival time and resistance to desication of diapause and non-diapause eggs of temperate *Aedes* (Stegomyia) mosquitoes. Entomologia Experimentalis et Applicata. 1992; 63:155–161.

Surget-Groba Y, Montoya-Burgos J. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Research. 2010; 20:1432–1440. [PubMed: 20693479]

Tammariello, SP. Regulation of the cell cycle during diapause. Denlinger, DL.; Giebultowicz, J.; Saunders, DS., editors. Elsevier Science, B.V; Amsterdam, The Netherlands: 2001. p. 173-183.

Tammariello SP, Denlinger DL. G0/G1 cell cycle arrest in the brain of *Sarcophaga crassipalpis* during pupal diapause and the expression pattern of the cell cycle regulator, proliferating cell nuclear antigen. Insect Biochemistry and Molecular Biology. 1998; 28:83–89. [PubMed: 9639874]

Tatar M, Kopelman A, Epstein D, et al. A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. Science. 2001; 292:107–110. [PubMed: 11292875]

Tauber, MJ.; Tauber, CA.; Masaki, S. Seasonal adaptations of insects. Oxford University Press; New York, New York: 1986.

Trpis M. A new bleaching and decalcifying method for general use in zoology. Canadian Journal of Zoology. 1970; 48:892–893.

Urbanski JM, Benoit JB, Michaud MR, et al. The molecular basis of increased desiccation resistance during diapause in the Asian tiger mosquito, *Aedes albopictus*. Proceedings of the Royal Society of London B. 2010; 277:2683–2692.

Urbanski J, Mogi M, O'Donnell D, et al. Rapid adaptive evolution of photoperiodic response during invasion and range expansion across a climatic gradient. The American Naturalist. 2012; 179:490–500.

Wang RL. Observations on the influence of photoperiod on egg diapause in *Aedes albopictus* Skuse. Acta Entomologica Sinica. 1966; 15:75–77.

Waterhouse RM, Zdobnov EM, Tegenfeldt F, et al. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Research. 2011; 39:D283–D288. [PubMed: 20972218]

Williams KD, Busto M, Suster ML, et al. Natural variation in *Drosophila melanogaster* diapause due to the insulin-regulated PI3-kinase. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:15911–15915. [PubMed: 17043223]

Wong, K.; Jin, Y.; Moqtaderi, Z. Multiplex Illumina sequencing using DNA barcoding.. In: Ausebel, FM., et al., editors. Current Protocols in Molecular Biology. John Wiley & Sons; New York, New York: 2013. p. 7.11.1-7.11.11.

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008; 18:821–829. [PubMed: 18349386]

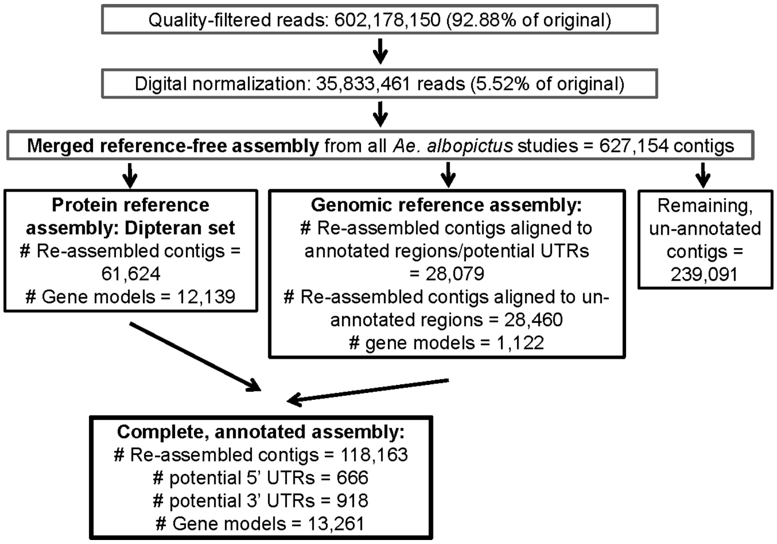
Quality-filtered reads: 602,178,150 (92.88% of original)
Digital normalization: 35,833,461 reads (5.52% of original)
Merged reference-free assembly from all *Ae. albopictus* studies = 627,154 contigs
Protein reference assembly: Dipteran set
# Re-assembled contigs = 61,624
# Gene models = 12,139
Genomic reference assembly:
# Re-assembled contigs aligned to annotated regions/potential UTRs = 28,079
# Re-assembled contigs aligned to un-annotated regions = 28,460
# gene models = 1,122
Remaining, un-annotated contigs = 239,091
Complete, annotated assembly:
# Re-assembled contigs = 118,163
# potential 5' UTRs = 666
# potential 3' UTRs = 918
# Gene models = 13,261

Figure 1.

Steps of the assembly and the number of reads, contigs and gene models resulting from each step of Illumina sequencing from RNA libraries of the Asian tiger mosquito *Aedes albopictus*.

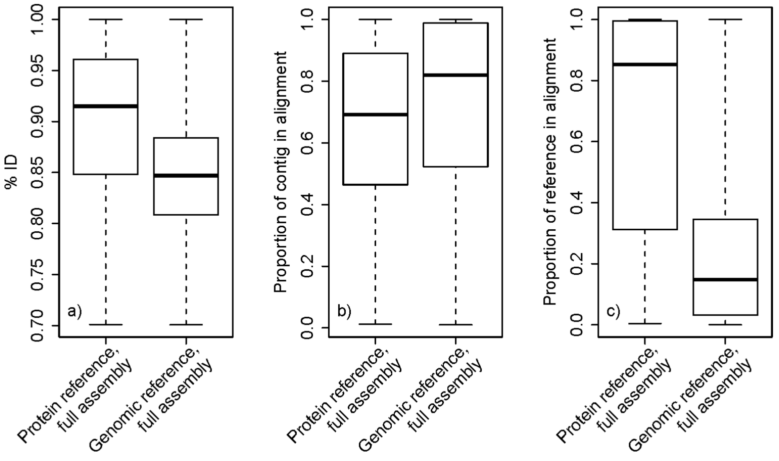

Figure 2.

Percentage identity (a), proportion of contig in the alignment (b), and proportion of reference in the alignment (c) of re-assembled contigs of the Asian tiger mosquito *Aedes albopictus* resulting from alignments to the protein sequence and genomic references.

Table 1

Quality metrics for digitally-normalized and non-normalized test assemblies of next generation sequencing reads from the Asian tiger mosquito *Aedes albopictus* using the library from diapause (D) conditions at 11 days post-oviposition.

	Number of reads	**Number of contigs**	**Minimum contig length**	**Average contig length**	**Median contig length**	**Maximum contig length**	**Mean % paired reads mapped to assembly**
Without normalization	111,574,867	93,622	94	1,042.3	619	14,481	68.6
Digital normalization	17,351,298	76,018	96	947.2	553	14,480	75.3

Table 2

Quality metrics for consecutive stages of assembly of next generation sequencing data from the Asian tiger mosquito *Aedes albopictus*.

	Number contigs length	Mean contig length	Median contig length	Maximum contig length	Average % GC
Pharate larval assembly, Oases	311,071	975.9	518	23,266	44.10
Pharate larval assembly, merged by cdhit-est	176,502	896.8	490	23,266	43.50
All contigs in merged assembly	627,154	885.8	502	23,934	45.27
Protein-reference assembly	61,624	1,961.0	1,388	25,247	49.07
Genomic-reference assembly	28,079	840.2	513	20,598	44.46
Complete annotated assembly	118,163	1,426.2	806	25,247	46.53

Table 3

Gene Ontology categories ("GO-Slim") of all gene models represented in the full assembly of Illumina sequencing reads from the Asian tiger mosquito *Aedes albopictus*. The number of gene models assigned to the top 25 categories are listed; the remainder are combined.

GOSlim GOA Accession No.	Description	No. of gene models[1]
GO:0008150	biological_process	5768
GO:0005575	cellular_component	4299
GO:0005623	cell	3995
GO:0005622	intracellular	2727
GO:0043226	organelle	1493
GO:0034641	cellular nitrogen compound metabolic process	1063
GO:0006810	transport	993
GO:0005634	nucleus	892
GO:0009058	biosynthetic process	878
GO:0005737	cytoplasm	795
GO:0044281	small molecule metabolic process	699
GO:0007165	signal transduction	624
GO:0006464	cellular protein modification process	479
GO:0043234	protein complex	433
GO:0009056	catabolic process	421
GO:0055085	transmembrane transport	401
GO:0005975	carbohydrate metabolic process	354
GO:0005576	extracellular region	282
GO:0006412	translation	276
GO:0006629	lipid metabolic process	231
GO:0006259	DNA metabolic process	197
GO:0034655	nucleobase-containing compound catabolic process	193
GO:0005840	ribosome	175
GO:0006950	response to stress	165
GO:0005856	cytoskeleton	152
NA	other categories, combined	2215
NA	No ontology information	6248

[1] The combined number of gene models in each category exceeds the total number of gene models in the assembly because gene models can be assigned to multiple GO-Slim categories.