Aedes aegypti is a tropical and subtropical mosquito vector for Zika, yellow fever, dengue fever, and chikungunya. We describe the first diploid assembly of an insect genome, using SMRT Sequencing and the open-source assembler FALCON-Unzip. This assembly has high contiguity (contig N50 1.3 Mb), is more complete than previous assemblies (Length 1.45 Gb with 87% BUSCO genes complete), and is high quality (mean base >QV30 after polishing). Long-range haplotype structure, in some cases encompassing more than 4 Mb of extremely divergent homologous sequence with dramatic differences in coding sequence content, is resolved using a combination of the FALCON-Unzip assembler, genome annotation, coverage depth, and pairwise nucleotide alignments.

Assembly Results

<table>
<thead>
<tr>
<th>Contig Set</th>
<th>Primary Contig</th>
<th>Associated Haplotig</th>
<th>Current Reference (L3.31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Length</td>
<td>1.45 Gb</td>
<td>0.59 Gb</td>
<td>1.38 Gb</td>
</tr>
<tr>
<td>Contig Number</td>
<td>3,462</td>
<td>4,328</td>
<td>36,204</td>
</tr>
<tr>
<td>Contig N50</td>
<td>1.43 Mb</td>
<td>0.38 Mb</td>
<td>0.083 Mb</td>
</tr>
</tbody>
</table>

Table 2. Assembly Statistics. FALCON-Unzip1 (v0.7.0) used for assembly followed by genome polishing with Arrow in SMRT Link.

Identifying Divergent Haplotypes using Gene Annotations

- Genome was annotated with conserved, single-copy genes from BUSCO3 arthropod dataset (N=2675)
- Shorter members of pairs of primary contigs with duplicated BUSCO genes and reduced raw read coverage were recategorized as haplotigs4

<table>
<thead>
<tr>
<th>Annotated BUSCO Genes</th>
</tr>
</thead>
</table>

![Figure 3. Reduced coverage in windows around duplicate BUSCO genes compared to single copy genes is consistent with haploid read coverage.](image)

Phased Assembly Identifies Dramatic Allelic Differences

- Long-range haplotype phasing spans 4Mb of contigs 000013F and 000043F, which align over ~40% of their length but share 8 BUSCO genes
- Region contains heterozygous premature stop codon in AAEL005110, a DNA repair protein

| Figure 5. (A) CDS and peptide alignment of AAEL005110 showing intact 302aa protein (above) and truncated 166aa allele (below). (B) Location of heterozygous variant in dot plot between contigs. |

Conclusion

- FALCON-Unzip can efficiently assemble long-range phased haplotypes in heterozygous non-model organisms, elucidating allelic differences between parental chromosomes
- Annotation with BUSCO genes is a simple and powerful way to identify divergent homologous genomic regions, in conjunction with read depth data

References

5. Kruyswijk et al. (2007) Rapport a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics, 23(8), 1028-1028

Acknowledgements

The authors would like to thank Jason Chin, David Rank, Greg Conception, and Matt Seo in Pacific as well as members of the Aedes Genome Working Group, including Adam Philliply and Sergey Koren, for helpful advice and thoughtful discussion.