A High-Quality Genome Assembly of SMRT Sequences Reveals Long-Range Haplotype Structure in the Diploid Mosquito *Aedes aegypti*

Sarah B Kingan[1], Paul Peluso[1], Richard Hall[1], Ben Matthews[2], Leslie Vosshall[2], and Jonas Korlach[1]
1. PacBio, 1380 Willow Rd, Menlo Park, CA 94025. 2. HHMI-Rockefeller University,1230 York Ave, New York, NY 10065

Summary

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, is resolved using a combination of the FALCON-Unzip assembler, genome annotation, coverage depth, and pairwise nucleotide alignments.

Strain Preparation

- 3 generations single-pair mating with same father each generation
- Founding strain: Liverpool
- Max ploidy = 4N
- Genomic DNA extracted with MagAttract Kit (Qiagen)

80 male pupae
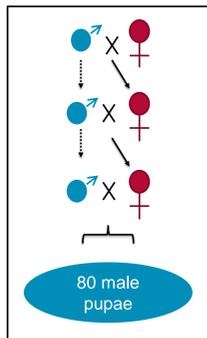
Library Preparation and Sequencing

- Three SMRTbell libraries constructed with 20 and 30 kb size selection using BluePippin.
- Libraries run on PacBio RS II using P6-C4 chemistry for 6 hour movies
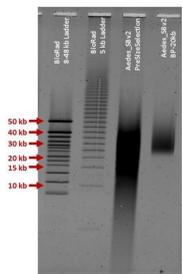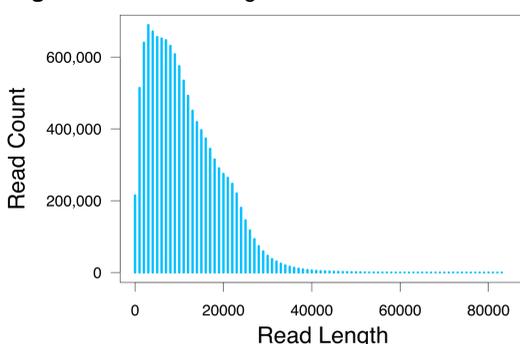

Figure 1. Size selection of a 20 kb library with BluePippin system

PacBio RS II SMRT Cells	177
Total Raw Sequence	140 Gb
Genome Coverage	110 fold
Subread N50	13,733 bp

Table 1. Sequencing Results

Figure 2. Subread length distribution

FALCON-Unzip Diploid Assembly

Contig Set	Primary Contigs	Associated Haplotigs	Current Reference (L3.31)[1]
Total Length	1.45 Gb	0.59 Gb	1.38 Gb
Contig Number	3,462	4,328	36,204
Contig N50	1.43 Mb	0.38 Mb	0.083 Mb

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

Identification of Additional Haplotigs

- Genome was annotated with conserved, single-copy genes from BUSCO[3] arthropod dataset (N=2675).
- Shorter members of pairs of primary contigs with duplicated BUSCO genes and reduced raw read coverage were recategorized as haplotigs.
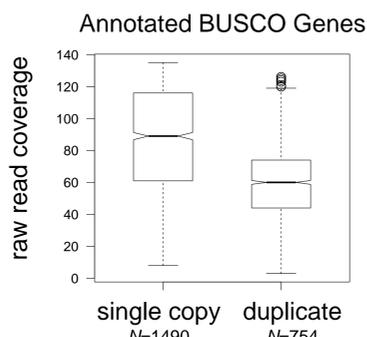

Annotated BUSCO Genes

Figure 3. Reduced coverage in windows around for duplicate BUSCO genes compared to single copy genes is consistent with haploid read coverage.

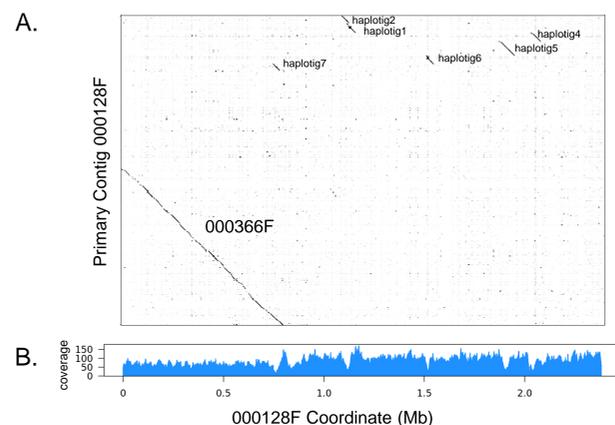

Figure 4. **Raw read coverage is reduced in regions where haplotigs align to primary contig.**

(**A**). Dot plot[5] of primary contig 000128F, its associated haplotigs, and an additional *homologous primary contig*, 000366F, which shares two BUSCO genes with 000128F. (**B**). Raw read coverage across primary contig 000128F is reduced by half in regions with multiple assembled haplotypes.

Contig	Length (Mb)	Proportion Aligned	P-distance	Mean (s.d.) Read Coverage
000128F	2.38	21.9%	NA	80 (24)
haplotig1	0.0336	100%	0.32%	21 (14)
haplotig2	0.0492	81.2%	0.42%	50 (35)
haplotig4	0.0420	100%	0.50%	59 (27)
haplotig5	0.0703	100%	0.40%	51 (27)
haplotig6	0.0413	99.1%	0.36%	48 (27)
haplotig7	0.0316	100%	0.92%	10 (8)
000366F	1.26	21.0%	1.79%	61 (13)

Table 3. Contig 000366F is recategorized as haplotig of 000128F. Nucmer[4] alignments show it is more divergent than the other haplotigs identified by FALCON-Unzip.

Genome Completeness and Quality

Contig Set	Primary + Haplotig	Primary Contigs	Associated Haplotigs	Current Reference
Complete	87%	81%	53%	85%
Duplicated	32%	5.4%	11%	11%
Fragmented	10%	10%	8.7%	11%
Missing	2.0%	8.1%	37%	2.1%

Table 4. BUSCOv1 analysis with arthropod dataset.

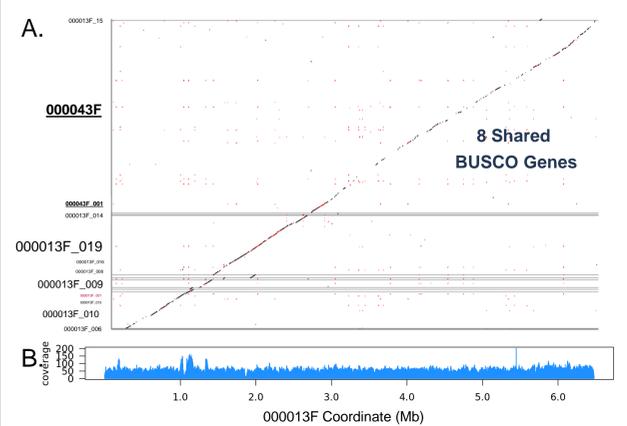## Long-range Haplotype Phasing

8 Shared BUSCO Genes

Figure 5. Divergent haplotypes share genes but align over less than 40% of their length.

(**A**). Dot plot[6] of primary contig 000013F, its *homologous primary contig* 000043F, and their associated haplotigs. (**B**). Raw read coverage across primary contig 000013F is reduced across regions of haplotype phasing.

Conclusion

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

References

1. Nene et al. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*, 316(5832), 1718-1723.
2. Chin et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12),1050-105
3. Simão et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31(19), 3210-3212.
4. Kurtz et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology* 5(2), R12.
5. Krumsiek et al. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*. 23(8),1026-1028.
6. Nattestad and Schatz (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*. 32(19), 3021-3023.