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



Sarah B Kingan¹, Paul Peluso¹, Richard Hall¹, Ben Matthews², Leslie Vosshall², and Jonas Korlach¹ 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.

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-Unzip² (v0.7.0)

Genome Completeness and Quality

laplotig	Primary Contigs	Associated Haplotigs	Current Reference
87%	81%	53%	85%
32%	5.4%	11%	11%
10%	10%	8.7%	11%
	<pre>Imary + laplotig 87% 32% 10%</pre>	Imary + aplotig Primary Contigs 87% 81% 32% 5.4% 10% 10%	Imary 1 aplotigPrimary ContigsAssociated Haplotigs87%81%53%32%5.4%11%10%8.7%

Strain Preparation

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



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 BUSCO³ 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

2.0

coverage.

Miss	sing	2.0%	8.1%	37%	2.1%

 Table 4. BUSCOv1 analysis with arthropod dataset.

Long-range Haplotype Phasing



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

(A). Dot plot⁶ 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.

with MagAttract Kit (Qiagen)



Α.

B.

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



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

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

Table 1.	Sequencing	Results

			haplotig2 haploti	g1		haplo	tig4
		∖ haplotig7		۲ hap	olotig6	_napiolig5	
28F			2012년 1월 2012년 1월 2012년 1월 2012년 1월 2012년 1월 28년 1월 28년 1월 2012년 1월 28년 1월 28년 1월 2012년				
012							
00							
tig		(a) A set of the se					
.uo	 A second sec second second sec	general (Balanda Karpan) (Balanda) Station (Balanda Karpan) (Balanda Station (Balanda Karpan)		n - Sanagi (Mirilano) Silaati Angeleti (Miren) Ti Sulating Hiring)			na na sina Sana sina Sana sina
U >							
ົ່ງສຸບ	0003	866F					
rin							
۵.							
			<u>, , , , , , , , , , , , , , , , , , , </u>		ne kon na se s		

000128F Coordinate (Mb)

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

(A). Dot plot⁵ 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						

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) <u>Genome sequence of Aedes aegypti, a major</u> <u>arbovirus vector</u>. *Science*, 316(5832), 1718-1723.
- 2. Chin et al. (2016) <u>Phased diploid genome assembly with single-</u> <u>molecule real-time sequencing</u>. *Nature Methods*. 13(12),1050-105
- 3. Simão et al. (2015) BUSCO: assessing genome assembly and

Figure 2. Subread length distribution



Table 3. Contig 000366F is recategorized as haplotig of 000128F. Nucmer⁴ alignments show it is more divergent than the other haplotigs identified by FALCON-Unzip.

annotation completeness with single-copy orthologs. *Bioinformatics*. 31(19), 3210-3212.

- 4. Kurtz et al. (2004) <u>Versatile and open software for comparing large</u> <u>genomes.</u> *Genome Biology* 5(2), R12.
- 5. Krumsiek et al. (2007) <u>Gepard: a rapid and sensitive tool for creating</u> <u>dotplots on genome scale.</u> *Bioinformatics*. 23(8),1026-1028.

 Nattestad and Schatz (2016) <u>Assemblytics: a web analytics tool for the</u> <u>detection of variants from an assembly.</u> *Bioinformatics*. 32(19), 3021-3023.

Acknowledgements

The authors would like to thank Jason Chin, David Rank, Greg Concepcion, and Matt Seetin at PacBio as well as members of the *Aedes* Genome Working Group, including Adam Phillippy and Sergey Koren, for helpful advice and thoughtful discussion.

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2017 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. Fragment Analyzer is a trademark of Advanced Analytical Technologies. All other trademarks are the sole property of their respective owners.