Long-Read Assembly of the *Aedes aegypti* Aag2 Cell Line Genome Resolves Ancient Endogenous Viral Elements

Matthew G. Seetin[1], Mark Kunitomi[2], Steve Oh[1], Cheryl Heiner[1], Ellen Paxinos[1], and Raul Andino[2]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2]Department of Microbiology and Immunology, University of California, San Francisco, CA.

Abstract

Transmission of arboviruses such as Dengue and Zika Viruses by *Aedes aegypti* causes debilitating disease across the globe[1]. Disease in humans can include severe acute symptoms such as hemorrhagic fever and organ failure, but it is unclear why mosquitoes tolerate high titers of virus in a persistent infection[2].

Recent publications highlighted the integration of genetic material from non-retroviral RNA viruses into the genome of the host during infection that relies upon endogenous retro-transcriptase activity from transposons[3,4]. These endogenous viral elements (EVEs) found in the genome are predicted to be ancient, and at least some EVEs are under purifying selection, suggesting they are beneficial to the host[5].

To characterize EVE biogenesis in a tractable system, we sequenced the *Ae. aegypti* cell line, Aag2, to 58-fold coverage and present a *de novo* assembly of the genome. The assembly contains 1.7 Gb of genomic and 255 Mb of alternative haplotype-specific sequence, consisting of contigs with a N50 of 1.4 Mb, considerably longer than other assemblies of the Aedes genus[6].

The Aag2 genome is highly repetitive, most of which is classified as transposable elements. We identify EVEs in the genome homologous to a range of extant viruses, many of which cluster in these regions of repetitive DNA. The contiguous assembly allows for more comprehensive identification of the transposable elements and EVEs that are most likely to be lost in assemblies lacking the read length of SMRT Sequencing.

Aedes aegypti Aag2 Assembly

	Reference[6]	This Work
Sample	LVP Strain	Aag2 Cell line
Sequencing	Sanger	PacBio
Assembled size	1.38 Gb	1.72 Gb
Gap length	73 Mb	0
Contigs	36204	3752
Contig N50	0.082 Mb	1.42 Mb

Table 1. Assembly statistics. Assembly was performed with FALCON v 0.4.1 on 58-fold coverage of PacBio data, P6/C4 chemistry, 15.5 kb subread N50

The Highly Repetitive Aag2 Genome

Repeat Class	Number of Elements	Total Length (Mb)	Percent of genome
SINE	93046	18.7	0.95
LINE	550030	318.8	16.1
LTR	461766	196.4	9.91
DNA	1580634	442.8	22.4
Unclassified	1096069	369.5	18.7
Total	3781095	1346.1	68.0

Table 2. Identification of repetitive elements. Analysis was performed with RepeatMasker[7] to identify Short Interspersed Nuclear Elements (SINEs), Long Interspersed Nuclear Elements (LINEs), Long Terminal Repeats (LTRs), DNA transposons (DNA), and other kinds of repetitive elements.

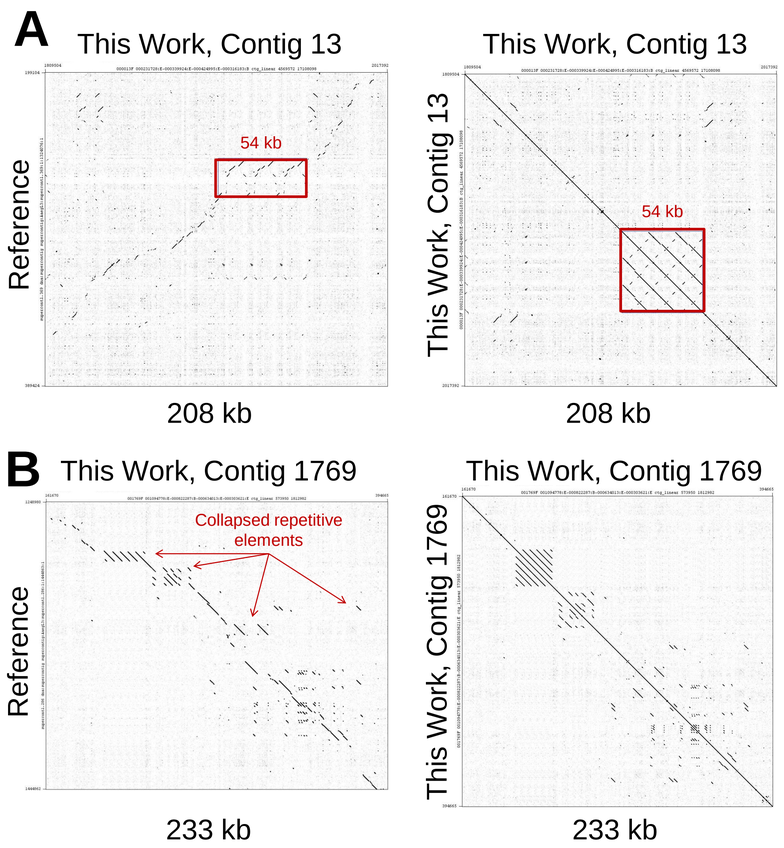

Figure 1. Dot plot comparisons of this assembly with the prior reference. (A) Plot of a portion of contig 13 from this work against the corresponding portion of supercontig 369 from the prior reference (left) and against itself (right). The red box highlights a 54 kb element of nested repeats poorly resolved in the reference but that is fully resolved in this work. (B) Plot of contig 1769 from this work against supercontig 269 from the prior reference and itself. Many repetitive elements that are missing, collapsed, or fragmented in the prior reference are resolved.

EVE Identification

	Reference[6]	This Work
EVEs Identified	188	417
Viral Families	5+	6+

Table 3. EVE statistics. EVEs were identified by translating candidate portions of the genome in 3 frames and aligning to known viral protein sequences.

EVEs are Found in Repeat-Rich Regions

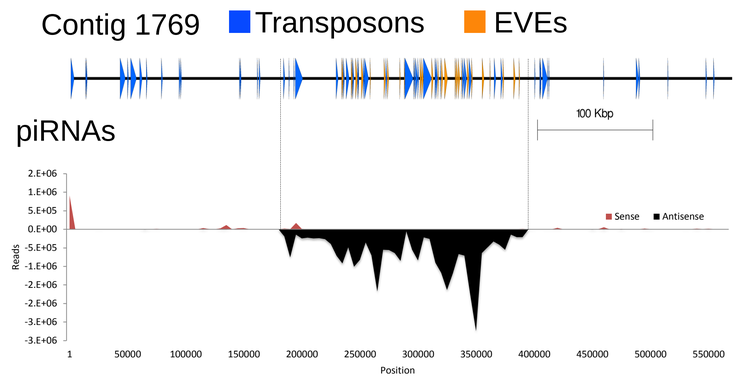

Figure 2. Annotation of contig 1769 from this work with transposons and EVEs, and the corresponding piwi-interacting RNA (piRNA) expression in the same region observed in RNA-seq data. EVEs integrate into the genome by way of transposon machinery and thus are going to be found with the repetitive elements from these transposons that confound short-read assembly. That these EVEs can subsequently serve as templates for anti-sense piRNA production suggests a mechanism for transgenerational immunity and for viral tolerance even while competent to spread infection.

An Active EVE

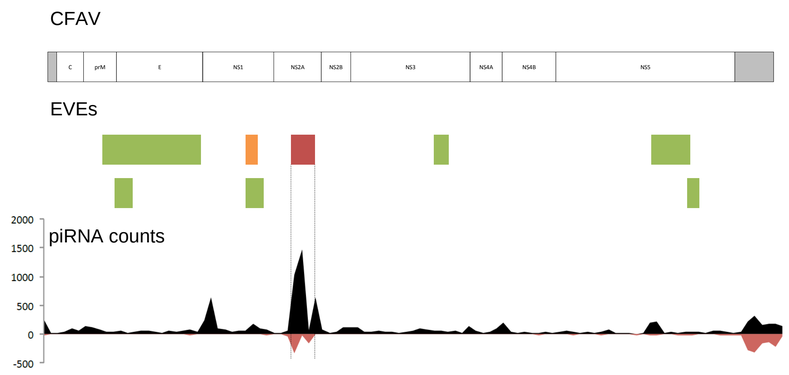

Figure 3. Several EVEs found in this assembly align to the Cell Fusing Agent Virus (CFAV) genome with >60% (green), >70% (orange), or >90% (red) sequence identity. Aligning candidate piRNAs from RNA-seq data against the CFAV genome reveals the highest piRNA expression from the EVE with the highest sequence identity to CFAV.

References

[1] Bhatt, S., et al. “The global distribution and burden of dengue.” *Nature* **496:**504–507 (2013).

[2] Olson, K. E. and C.D. Blair. “Arbovirus-mosquito interactions: RNAi pathway.” *Curr. Opin. Virol.* **15:**119–126 (2015).

[3] Goic, B., et al. “RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila.” *Nat. Immunol.* **14:**396–403 (2013).

[4] Horie, M., et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463:**84–87 (2010).

[5] Fort, P., et al. “Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality.” *Mol. Biol. Evol.* **29:**381–390 (2012).

[6] Nene, V., et al. “Genome sequence of Aedes aegypti, a major arbovirus vector.” *Science* **316**(5832):1718-23 (2007).

[7] Smit, A.F.A., R. Hubley & P. Green. RepeatMasker at http://repeatmasker.org