Title: Spatial and temporal distribution of genome divergence among California populations of *Aedes aegypti*.

Running head title (60 chars w/ spaces): Population genomics of California *Aedes aegypti*

Yoosook Lee[1*], Travis C. Collier[2], William R. Conner[3], Mark J. Hanemaaijer[4], Chelsea T. Smartt[5], Montgomery Slatkin[6], Joanna C. Chiu[7], Gregory C. Lanzaro[8], F. Steve Mulligan, III[9], Anthony J. Cornel[10]

1. Yoosook Lee, University of California - Davis, Davis, California, USA
2. Travis C. Collier, University of California - Davis, Davis, California, USA
3. William R. Conner, University of California - Davis, Davis, California, USA
4. Mark J. Hanemaaijer, University of California - Davis, Davis, California, USA
5. Chelsea T. Smartt, University of Florida Florida Medical Entomology Laboratory, Vero Beach, FL, USA
6. Montgomery Slatkin, University of California - Berkeley, California, USA
7. Joanna C. Chiu, University of California - Davis, Davis, California, USA
8. Gregory C. Lanzaro, University of California - Davis, Davis, California, USA
9. F. Steve Mulligan, III, Consolidated Mosquito Abatement District, Clovis, California, USA
10. Anthony J. Cornel, University of California - Davis, Davis, California, USA

Corresponding Author: Yoosook Lee

yoslee@ucdavis.edu

530-752-7333

Keywords: *Aedes aegypti*, population genomics, California

Abstract (250 words)

In the summer of 2013, *Aedes aegypti* Linnaeus was first detected in three cities in central California (Clovis, Madera and Menlo Park). It has now been detected in multiple locations in central and southern CA as far south as San Diego and Imperial Counties. Here we analyzed 26 *Ae. aegypti* genomes from multiple locations throughout California to determine if these populations originated from a single or multiple introductions. We also compared *Ae. aegypti* populations in Clovis at two time points, 2013 and 2016, to describe changes at the genome level that might suggest adaptation to its new environment. We identified 3 distinct genetic clusters loosely corresponding to different regions within California. Populations at all five southern California sites are grouped into a single genetic cluster. However, all three clusters occur in parapatry in the Central Valley. We also found two major mitochondrial lineages with an estimated divergence time of 55,700 - 557,000 years ago. Genome-wide comparisons indicate extensive differentiation between genetic clusters. A potential selective sweep within the Clovis population was also detected. We believe these observations support recent multiple introductions of *Ae. aegypti* into California.

Author Summary

The mosquito species *Aedes aegypti*, introduced in 2013, has now been detected in multiple locations throughout California. Our genome analyses identified 3 distinct population groups loosely corresponding to different regions within California. Genome-wide comparisons indicate extensive differentiation between genetic clusters. Samples collected from Clovis in two different years (2013 and 2016) reveal genomic signatures of potential selection. Here, we also first describe mitochondrial lineages of *Ae. aegypti* that diverged around 55,700 – 557,000 years ago. Our analysis suggests that founding populations were polymorphic for two mitochondrial lineages with one or the other lost in the various extant populations. These observations support recent multiple introductions of *Ae. aegypti* into California.

Introduction

Aedes aegypti Linnaeus (Diptera: Culicidae) transmits multiple human viral pathogens including Yellow Fever, Dengue, Chikungunya and Zika viruses. There are no cures for the diseases caused by these pathogens and there is a vaccine for only one: the Yellow Fever virus. At this time, with the exception of Yellow Fever, protecting the public depends exclusively on vector control.

California (CA) had no known established local populations of *Ae. aegypti* prior to the summer of 2013 when it was first detected in three cities in central California: Clovis, Madera and Menlo Park [1, 2]. In the spring and summer of the following year, this mosquito was again found in the same three California locations and for the first time in additional communities in central California and further south in San Diego County (Figure 1). These observations provided the first indication that *Ae. aegypti* is capable of overwintering in California and had become established in multiple locations. Towards the end of 2016, *Ae. aegypti* were collected from a number of additional locations in both central and southern CA, indicating that *Ae. aegypti* has now spread throughout the state (Figure 1).

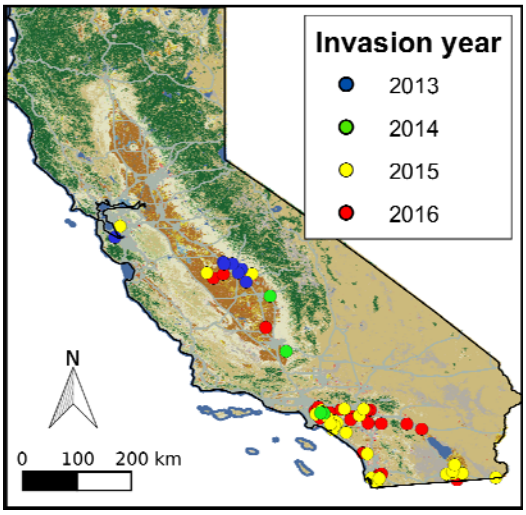


Figure 1. History of *Ae. aegypti* invasion in California (data from CalSurv Maps [3]). USGS National Land Cover Database (NLCD) 2011 Land Cover product available at https://www.mrlc.gov/nlcd2011.php was used as background . QGIS version 2.18 was used to create this map.

67 There is only a single previous study on population genetics/genomics of CA *Aedes aegypti* [1]. Gloria-
68 Soria et al. [1] analyzed three CA *Ae. aegypti* populations and concluded that the likely source of
69 introduction was New Orleans or Houston area. Its genetic similarity indicated by the Structure
70 assignment profile based on microsatellite genotypes, however, cannot rule out other regions such as
71 Florida and Amacuzac in southern Mexico. This stresses the importance of having a good representative
72 reference populations for source identification [1].
73 *Aedes aegypti* has a short flight range; mostly within 200m from their breeding source [4], but it is
74 exquisitely adapted to hitchhiking in transport vehicles [5]. Since its introduction into California, we
75 hypothesize that *Ae. aegypti* currently exists as a group of genetically distinct subpopulations with
76 limited gene flow between them. If true this hypothesis suggests that elimination of individual
77 subpopulations is feasible and that sustainable "*Ae. aegypti*-free zones" may be established and
78 maintained. For example, Exeter and Menlo Park subpopulations were successfully eliminated in 2014
79 and 2016 respectively (personal communications from mosquito control district managers). To test our
80 hypothesis, we used whole genome sequence data for 26 individuals collected from 11 locations
81 throughout CA, plus 2 from Florida.
82
83 **Materials and Methods**
84 Mosquito Collection
85 Adult female *Ae. aegypti* were collected by personnel from Mosquito Abatement Districts in Fresno, San
86 Diego, and Orange Counties from 13 towns or cities (Figure 2A and Table 1). All collections in private
87 properties were conducted after obtaining permission from residents and/or owner. Mosquito samples
88 were individually stored in 80% ethanol prior to DNA extraction.
89

90 **Table 1**. Sample used for genome analysis. Raw reads indicate the number of reads obtained per
91 sample. Mapped reads indicate the number of reads mapped to the AaegL4 reference genome. Mean
92 coverage indicate the average depth of reads of a given base pair in the AaegL4 reference genome.
93 Mean mt coverage indicate the average depth of reads in the mitochondrial genome. Median insert size
94 is the median length of insert sequences (length of library fragment minus sequencing adapters).

Sample	Collection Location	Collection Period	latitude	longitude	raw reads (million)	mapped reads (million)	mean coverage (X)	Mean mt coverage (X)	Median insert size (bp)
Ae15SAND001	San Diego	Aug. 2015	32.56	-117.05	107.5	105.4	9.4	1139.2	381
Ae15SAND002	San Diego	Aug. 2015	32.56	-117.05	138.0	135.2	11.9	1168.1	356
Ae15SAND015	San Diego	Aug. 2015	32.56	-117.05	97.7	95.6	8.2	864.4	339
Ae15IMPE002	Brawley	2015	32.97	-115.53	97.9	95.8	8.4	1722.8	379
Ae15GARG021	Garden Grove	Oct. 2015	33.76	-117.92	155.4	150.6	13.3	1961.3	389
Ae15GARG026	Garden Grove	Oct. 2015	33.76	-117.92	79.5	77.6	6.9	347.9	349
Ae15MISV029	Mission Viejo	Oct. 2015	33.61	-117.67	146.7	143.2	12.5	1196.0	389
Ae15MISV036	Mission Viejo	Oct. 2015	33.61	-117.67	74.6	72.8	6.5	1266.4	420
Ae15COMF3.001	Commerce	Sep. 2015	34.01	-118.16	96.8	94.3	8.0	839.2	371
Ae14EXET135	Exeter	Sep. 2014	36.30	-119.16	74.1	72.4	6.2	1420.1	394
Ae14EXET165	Exeter	Sep. 2014	36.30	-119.16	94.4	92.3	7.9	3201.5	368
Ae15SANG073	Sanger	Sep. 2015	36.73	-119.51	56.8	55.2	4.7	412.5	400
Ae13CLOV028	Clovis	Jul. 2013	36.81	-119.67	101.9	100.2	8.6	1500.0	256
Ae13CLOV030	Clovis	Jul. 2013	36.81	-119.67	107.5	105.5	8.9	1511.1	272
Ae13CLOV032	Clovis	Jul. 2013	36.81	-119.67	99.0	97.5	8.5	2488.1	249
Ae16CLOV002	Clovis	Apr. 2016	36.49	-119.40	183.6	179.3	15.6	5051.4	408
Ae16CLOV003	Clovis	Apr. 2016	36.49	-119.40	79.9	77.9	6.9	2593.6	418
Ae16CLOV007	Clovis	Apr. 2016	36.48	-119.40	222.5	219.3	18.8	753.3	366
Ae16CLOV008	Clovis	Apr. 2016	36.48	-119.40	98.9	96.5	8.2	217.2	371
Ae15FRES008.05	Fresno	Aug. 2015	36.83	-119.86	28.5	27.9	2.3	782.5	348
Ae15FRES017	Fresno	Oct. 2015	36.83	-119.86	91.0	88.9	7.6	434.9	358
Ae15MADE001	Madera	Jul. 2015	36.94	-120.08	67.1	65.6	5.5	967.6	383
Ae15MADE015	Madera	Aug.2015	36.94	-120.08	85.8	83.4	7.1	1369.0	370
Ae14SAMC001	Menlo Park	2014	37.44	-122.20	58.6	57.1	4.9	1104.1	416
Ae15SAMC001	Menlo Park	Apr. 2015	37.44	-122.20	44.2	28.5	2.3	3208.5	236
Ae15SAMC004	Menlo Park	May. 2015	37.44	-122.20	92.1	90.0	7.6	215.0	293
Ae16KEYW001	Key West, FL	2016	24.56	-81.78	93.1	91.0	7.9	936.4	361
Ae16VERO001	Vero Beach, FL	2016	29.65	-80.38	40.5	39.5	3.2	1256.2	372

95

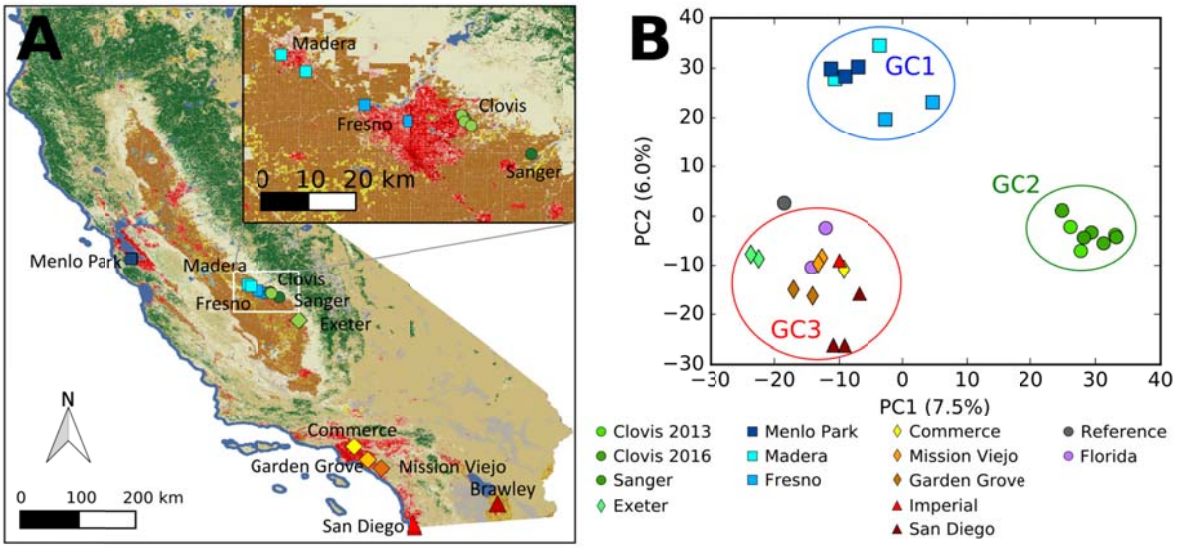


Figure 2. A: Geographic origin of *Ae. aegypti* used in this study. Different colors were used to denote four regions: (1) blue for central coast area (Menlo Park), (2) green for central valley area (Madera, Fresno, Clovis, Sanger and Exeter), (3) yellow for greater Los Angeles area (Commerce, Garden Grove, and Mission Viejo) and (4) red for San Diego and Imperial County area (San Diego, Brawley). B: Principal Component Analysis based on 28,057 strict-filtered (=no missing data) biallelic SNPs. Three genetic clusters (GC) are evident, labeled as GC1, GC2 and GC3. Reference, indicated in grey, refers to genotype inferred from *AaegL4* reference sequence [6]. QGIS version 2.18 was used to create this map.

<u>Whole genome sequencing</u>

Genomic DNA was extracted using the protocol described in [7] and [8]. DNA concentrations for each sample were measured using a high sensitivity DNA kit on the Qubit instrument (Life Technologies). A genomic DNA library was constructed for each individual mosquito using 25ng DNA, Qiaseq FX 96 (Qiagen, Valencia, CA), and Ampure SPRI beads (Beckman) according to the library construction protocol described in [8]. Library concentrations were measured using Qubit (Life Technologies). Libraries were sequenced using HiSeq 4000 at the UC Davis DNA Technologies Core.

113 Sequencing Analysis

114 Demultiplexed raw reads were trimmed using Trimmomatic [9] version 0.36 and mapped to the *Aaeg*L4
115 reference genome [6] using BWA-MEM [10] version 0.7.15. Chromosome 2 of AaegL4 was too large for
116 some bioinformatics tools. To resolve this issue, we subdivided the chromosome 2 sequence in half at a
117 position where a stretch of uncalled bases are located (around 240Mbp position) and labeled as the
118 chromosome segments 2a and 2b. Freebayes [11] version 1.0.1 was used for variant calling without
119 population prior. Following the recommendation by Crawford and Lazarro [12], we used minimum depth
120 of 8 to call variant for each individual. The neighbor-joining algorithm implemented in PHYLIP [13]
121 version 3.696 was used for constructing a phylogenetic trees. F_{ST} and Principal Component Analysis
122 (PCA) analyses was done in Python version 3.4.3 using the scikit-allel module version 0.21.1 [14].
123

124 Mitochondrial genome sequences were mapped to the reference sequence of *Ae. aegypti* (Genbank
125 accession number = NC_010241 [15]), an inbred sub-strain (LVPib12) of the Liverpool strain used for
126 reference genome assembly. Freebayes version 1.0.1 was used for mitochondrial variant calling
127 assuming single ploidy. Mitochondrial coverage was on average 221 times of nuclear genome coverage
128 (Table 1). Thus forcing a single ploidy in variant calls will minimize the impact of variants introduced by
129 nuclear copies of the mitochondrial genome. The Vcf2fasta program [16] was used to extract
130 mitogenome sequences from the vcf file to fasta format. Geneious version 10.1.3 was used for
131 mitogenome alignments. Mitogenome reference sequences of *Culex quinquefasciatus* (Genbank
132 accession number = HQ724617), *Aedes notoscriptus* (KM676219), *Aedes vigilax* (KP995260), and *Aedes*
133 *albopictus* (NC_006817) were obtained from GenBank. Sequences for the thirteen mitochondrial
134 protein-coding genes in *Ae. aegypti* were obtained from GenBank [17]. The corresponding genes in the
135 28 *A. aegypti* mitochondrial genomes included in this study were identified with BLAST, extracted, and
136 concatenated. The resulting sequence was partitioned by codon position and used to generate a

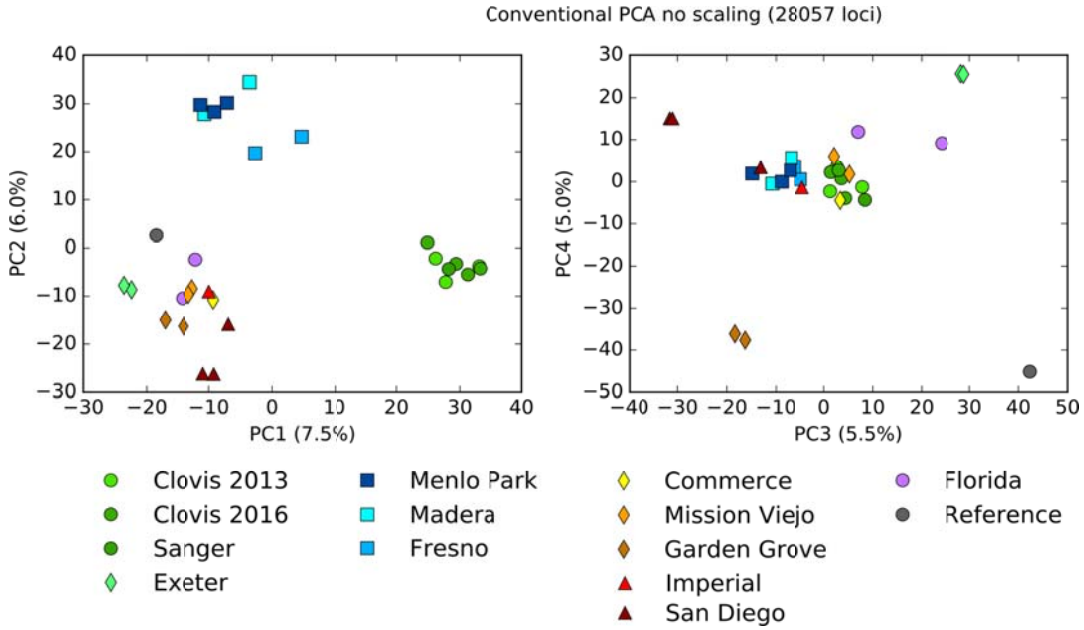
137 chronogram with a Bayesian phylogenetic inference program, RevBayes version 1.0.3 [18]. We
138 attempted to partition by gene and codon positions, but many of the resulting partitions lacked
139 sufficient information to estimate the model parameters. For divergence time estimation, we used a
140 strict clock model. Two independent runs were performed. The parameter point estimates for the
141 second run were all within the 95% confidence interval for the first, demonstrating that results of the
142 two analyses are consistent. Continuous parameters were checked for convergence and adequate
143 mixing with Tracer 1.6 [19]. Roughly 100 million years of divergence time between *Cx. quinquefasciatus*
144 and *Ae. aegypti* [20] were used to calculate divergence time from relative node ages obtained from
145 RevBayes analysis.

146 **Results**

147 We sequenced 26 individual *Ae. aegypti* genomes from California and two from Florida with median
148 genome coverage of 8.5X per sample (Table 1). Sequence data is available through NCBI Sequence Read
149 Archive (Study accession number: SRP106694). Data is also available through UC Davis PopI
150 OpenProjects page: https://popi.ucdavis.edu/PopulationData/OpenProjects/AedesGenomes/.
151 Principal Components Analysis revealed 3 distinct genetic clusters (GC1, GC2 and GC3), denoted in
152 different colors in Figure 2B. The genetic cluster designated GC1 included the northernmost site near
153 the coast in Menlo Park, San Mateo County and extends south including populations at Madera and
154 Fresno. Populations in Clovis and Sanger in central California form the GC2 cluster. The GC3 cluster
155 includes all five sites in southern California and extends north to include Exeter in Tulare County. The
156 samples from Florida also fall into this cluster. Overall the distribution of the three *Ae. aegypti* genetic
157 clusters have a parapatric distribution with the three groups converging in the center of the Central
158 Valley. We did not observe evidence that any of the groups occur in sympatry, but given our relatively
159 small sample size we cannot rule this out. Other Principal components indicates additional population

160 subdivisions further dividing Mission Viejo, Garden Grove, Exeter and Florida samples from the rest

161 groups (Figure S1).



162

163 **Figure S1**. Principal Component Analysis based on 28,057 strict-filtered (=no missing data) biallelic SNPs.

164 Reference, indicated in grey, refers to genotype inferred from *AaegL4* reference sequence [6]. QGIS

165 version 2.18 was used to create this map.

166

167 Examination of our sequence data indicates genome-wide differentiation among the three clusters

168 (Figure 3A -C). GC1 and GC2 were the most diverged (genome-wide average F_{ST} = 0.164). Genetic

169 distance between GC3 and the other clusters was intermediate (Figure 3C & Figure 3D). Reduction in F_{ST}

170 in the middle of Chromosome 1 and 2 were evident especially in the GC1 vs. GC3 comparison (Figure

171 3C). Whether this region corresponds to specific chromosomal features, such as centromeres, remains

172 to be determined. Overall SNP density was lowest in Chromosome 1, which contains the sex

173 determining locus in *Ae. aegypti*.

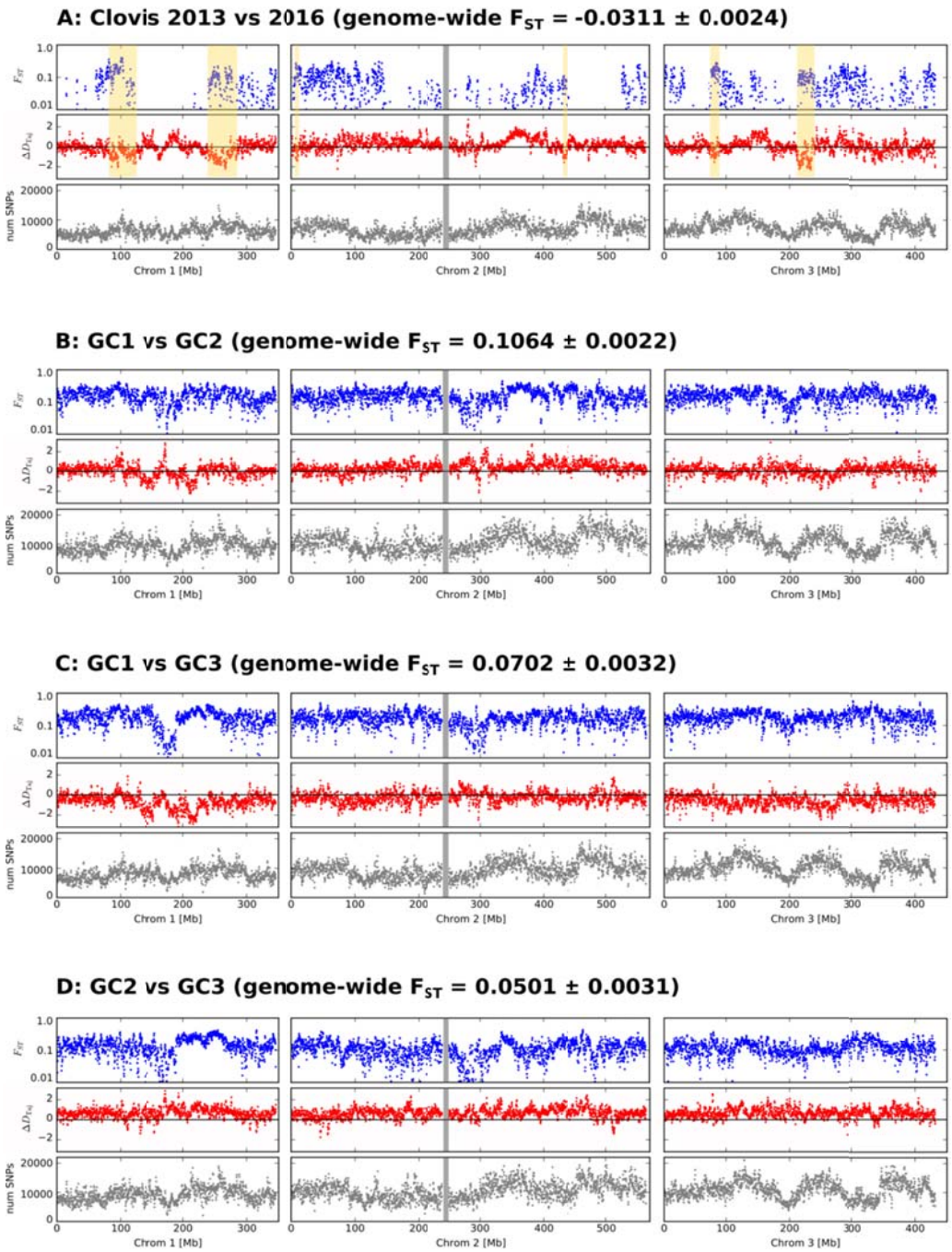


Figure 3. Genome-wide comparison of *Ae. aegypti* populations revealing broad highly differentiated (F_{ST}>0.1) genomic regions. Hudson F_{ST} [21] is marked in blue dots. Relative nucleotide diversity change is marked in red. The SNP density is marked in gray. A: Clovis 2013 (N=3) vs 2016 (N=4). Regions of potential selective sweep are highlighted in yellow. B: GC1 vs GC2. C: GC1 vs GC3. D: GC2 vs GC3. The gray bar around the 250 Mbp region of Chromosome 2 is an artificial break.

Overall differentiation between Clovis 2013 and 2016 samples (all in GC1) was relatively small (F_{ST} = -0.0311). However, a whole genome scan of Hudson F_{ST} [21] values with 500 kbp windows indicates that there were a fair number of specific regions with high F_{ST} values (>0.1) (Figure 3A) between samples from 2013 and 2016. Relative Difference in nucleotide diversity (ΔD = (*Tajima's* D_{2013}- *Tajima's* D_{2016}))/*Tajima's* D_{2013}) between 2013 and 2016 samples showed a substantial reduction in nucleotide diversity over time (highlighted in yellow shades in Figure 3A). These highly differentiated regions with reduced nucleotide diversity may indicate genomic regions under selection presumably as the founding population adapts to local environmental conditions [22].

Geographic distances were not correlated with genetic distance (Linear model p-value = 0.728, Mantel test p-value = 0.38; Figure 4). For instance, the genetic distance measured using F_{ST} between Fresno and Clovis (17 km apart) was as high as that between Clovis and Mission Viejo (over 400 km apart) (Figure 2).

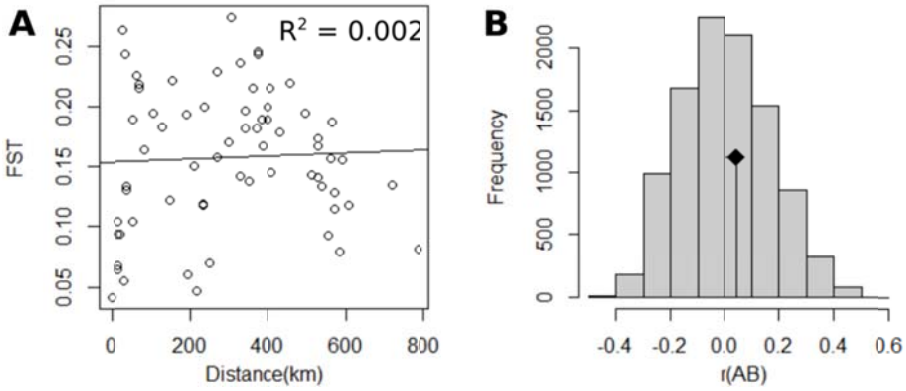


Figure 4. A: Relationship between geographic distance and genetic distance (Hudson F_{ST} [21]). No significant correlation between geographic and genetic distances (linear model slope indicated in black line; p value = 0.728). B: Mantel test result with 10,000 replicates (p value = 0.38). Black line indicates the average correlation value between geographic and genetic distances (=0.0437).

198 Differentiation among populations within GC1 (F_{ST} =-0.063 ~ -0.031) and GC3 (F_{ST} =-0.054 ~ -0.007) was

199 minimal (Table 2). However, differentiation within GC2 was high (F_{ST} =0.074 ~ 0.274) suggesting potential

200 substructure within GC2 (Figure 5).

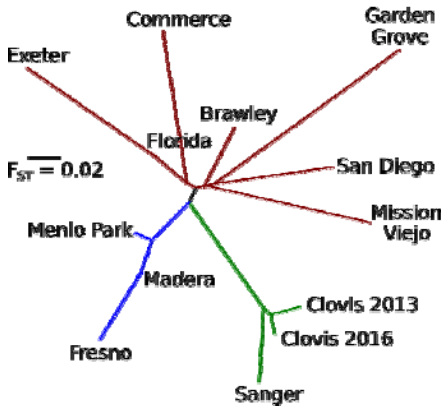
201

202 **Table 2.** Pairwise F_{ST} values based on nuclear genome SNP data. Colors match genetic clusters described

203 in Figure 2B.

F_{ST}	Clovis 2013	Clovis 2016	Sanger	Fresno	Madera	Menlo Park	Exeter	Commerce	Brawley	Garden Grove	Mission Viejo	San Diego	Florida
Clovis 2013	-												
Clovis 2016	-0.031	-											
Sanger	-0.063	-0.050	-										
Fresno	0.059	0.072	0.005	-									
Madera	0.132	0.129	0.100	-0.054	-								
Menlo Park	0.111	0.108	0.058	-0.007	-0.025	-							
Exeter	0.214	0.218	0.187	0.147	0.194	0.164	-						
Commerce	0.181	0.196	0.138	0.115	0.181	0.132	0.229	-					
Brawley	0.113	0.127	0.091	0.057	0.116	0.073	0.194	0.157	-				
Garden Grove	0.245	0.243	0.214	0.175	0.215	0.182	0.274	0.244	0.199	-			
Mission Viejo	0.189	0.199	0.165	0.125	0.178	0.148	0.236	0.226	0.149	0.264	-		
San Diego	0.166	0.172	0.142	0.119	0.156	0.130	0.219	0.193	0.122	0.221	0.182	-	
Florida	0.115	0.122	0.077	0.051	0.099	0.070	0.098	0.117	0.074	0.187	0.142	0.121	-

204



205

206 **Figure 5.** Neighbor-joining tree based on pairwise nuclear genome-wide $F_{ST.}$ Populations within GC1

207 designated in green, GC2 in red and GC3 in blue.

208

209 Mitogenome sequence analysis revealed two major mitochondrial lineages within *Ae. aegypti* (Figure 6).

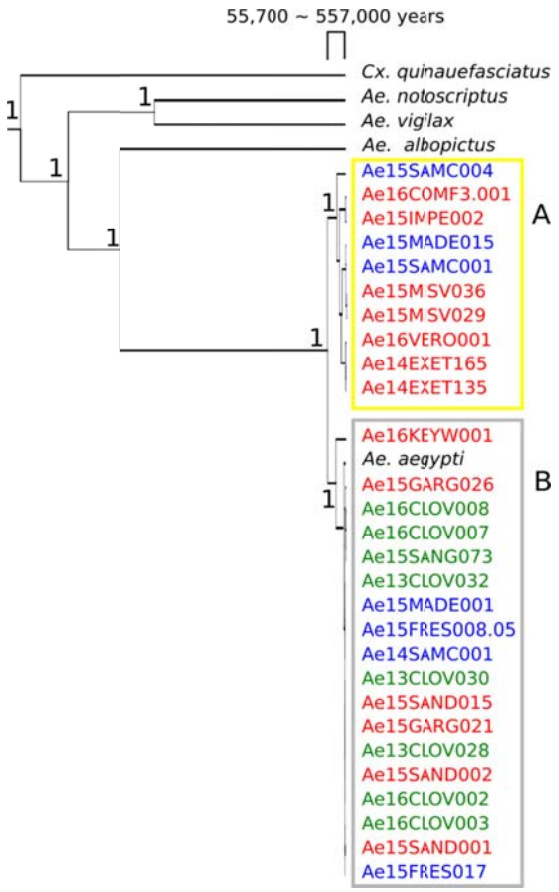
210 Lineage B includes the *Ae. aegypti* reference sequence. The mitochondrial lineages did not correspond

211 well with nuclear genome clusters as shown in Figure 2. Locations where mixed lineages were found

212 include Menlo Park and Madera. One sample from Vero Beach, Florida corresponded to lineage A and

213 the sample from Key West, Florida corresponded to lineage B, also indicating the possibility of mixed

214 mitochondrial lineages occurring in Florida.



215

216 **Figure 6.** Tree based on mitochondrial genome sequence generated from the RevBayes program. *Ae.*

217 *aegypti* sequences produced from this study are noted in capital letters. Green letters indicate samples

218 belonging to the GC1 cluster from Figure 2B; red for GC2; blue for GC3. Two major *Ae. aegypti*

219 mitochondrial lineages were labeled as A (yellow box) and B (gray box). The two mitochondrial lineages

220 do not correspond to the nuclear genome clusters. All major branches are 100% supported in our

221 simulation (node posterior probability = 1.00). Divergence time estimates were based on 28 individual
222 *Ae. aegypti* mitochondrial genome data under the assumption of 1-10 times nuclear genome mutation
223 rate and 5 generations per year.
224
225 Our RevBayes estimate of the substitution rate was 0.144 substitutions per site. If we assume 100 Myr
226 of divergence time between *Cx. quinquefasciatus* and *Ae. aegypti* [20] and 5 generations per year, this
227 equates to a mutation rate of 2.88 x 10^{-10} substitutions/site/generation. This value is 10 times slower
228 than the nuclear genome mutation rates reported from *Drosophila melanogaster* (2.8 x 10^{-9}
229 substitutions/site/generation [23]) or *Heliconius melpomene* (2.9 x 10^{-9} substitutions/site/generation,
230 [24]). Even with the lower estimates of 50 Myr between *Cx. quinquefasciatus* and *Ae. aegypti* [25], the
231 mutation rate (5.77 x 10^{-10}) is lower than the nuclear genome substitution rate. This directly contradicts
232 the notion that mitochondrial genomes have a higher (1-10 times) substitution rate than nuclear
233 genomes [26, 27]. Dating using *Cx. quinquefasciatus* is likely inadequate due to saturation of
234 substitutions. In other words, multiple substitutions occurred at a single site resulting in
235 underestimation of mutation rates. Existing models cannot adequately handle cases with multiple
236 substitutions per site. Therefore, the divergence time estimates for the two lineages of *Ae. aegypti* (~5.6
237 Myr) may not be accurate (Supplement Figure S2). The substitution rate estimates only using our *Ae.*
238 *aegypti* data was an average of 0.0079 substitutions/site. If we assume the mitochondrial mutation rate
239 is 1-10 times higher than the nuclear genome mutation rate as indicated in Keightley et al. (2015) (mean
240 = 2.85 x 10^{-9}) and 5 generations per year, then the divergence time of the two *Ae. aegypti* mitochondrial
241 lineages would fall between 55,700 ~ 557,000 years. Because gene annotation data for *AaegL4* is not
242 available, we did not calculate the nuclear genome mutation rate.

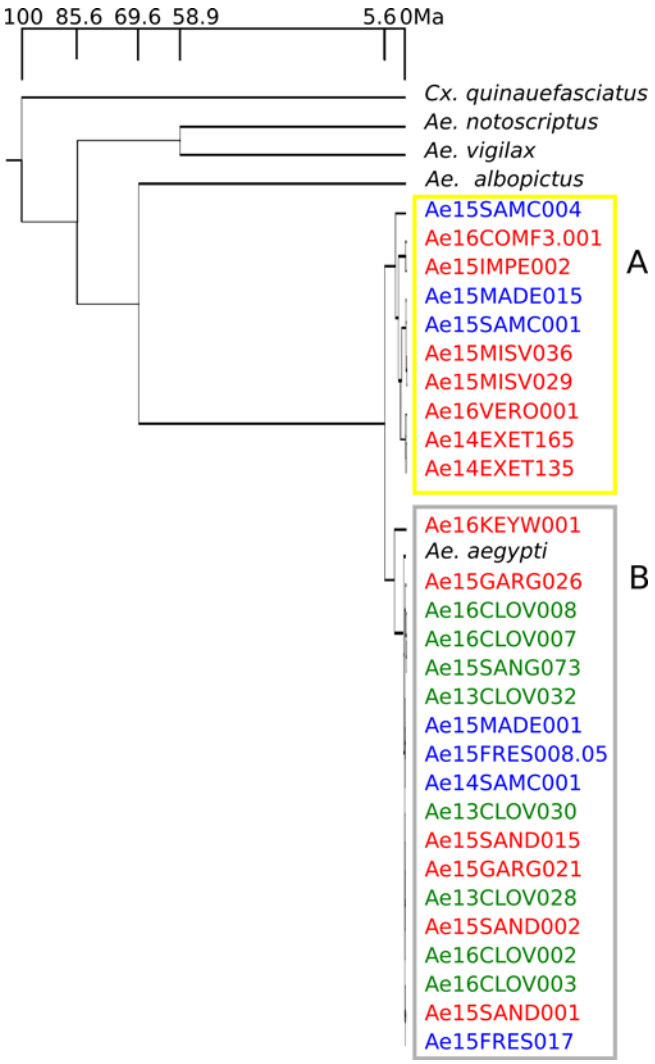


Figure S1. Tree based on mitochondrial genome sequence generated from RevBayes program. Relative age of nodes were translated into absolute divergence time using the *Cx. quinquefasciatus* and *Ae. aegypti* divergence time estimates of ~100 million years [20]. *Ae. aegypti* sequences produced from this study are noted in capital letters. Green letters indicate samples belongs to the GC1 cluster from Figure 2; red for GC2; blue for GC3. Two major *Ae. aegypti* lineages were labeled as A (yellow box) and B (gray box). At face value, *Culex* v. *Aedes aegypti* had 0.144 subs/site averaged across all three positions. Assuming the 100 million years divergence time between the two species and 5 generations per year, it gives 2.88 x 10-10 subs/site/gen. For comparison, the nuclear genome mutation rate from two insect species (*Drosophila melanogaster* and *Heliconius melpomene*) were estimated to be 2.8 ~ 2.9 x 10-9

253 subs/site/generation [23, 24]. If we assume the *Ae. aegypti* nuclear genome is similar to other insects,
254 this observation contradicts many previous reports showing mitochondrial substitution rates are higher
255 than nuclear genome substitution rates [26-28]. It is likely that multiple substitutions occurred at a
256 single site resulting in underestimation of mutation rates.
257
258 **Discussion**
259 We identified three major nuclear genome clusters within California *Ae. aegypti* roughly corresponding
260 to their geographic origin (Figure 2). However, geographic proximity did not correlate with genetic
261 similarity (Figure 4) corroborating the conventional wisdom that this mosquito species is limited in its
262 ability to disperse, relying on hitchhiking on human transport to move through the environment [5, 29].
263 Mitochondrial data also suggested two major lineages within California populations of *Ae. aegypti.*
264 However, their genealogy differs from the nuclear genome genealogy (Figures 5 and 6). This suggests
265 that the three founding populations (GC1-3) were polymorphic for the two mitochondrial lineages, but
266 one or the other lineage was lost due to founder effects in each CA population. For example, GC1 only
267 includes Lineage B (Figure 6). The incongruence between nuclear and mitochondrial gene genealogies
268 could also be due to different evolutionary rates between different loci producing differing topologies
269 [26-28]. It is possible that mitochondrial lineages capture historic divergence events, while nuclear
270 genome divergence reflects relatively recent divergence. Linkage disequilibrium decays rapidly in
271 mosquito genome as seen in *Anopheles arabiensis* [30]. Thus any contact between two distinct *Ae.*
272 *aegypti* populations may have resulted in relatively recent gene flow homogenizing populations within a
273 locality. It is possible that mitochondrial pseudogenes in the nuclear genome may influence the
274 mitochondrial variant [31, 32]. We used the option to call single ploidy genome to call mitochondrial
275 variants of each individual knowing that mitochondrial genome coverages are 10-100 times greater than

276 nuclear genome coverage and used and thus variant call will largely influenced by the abundant
277 mitochondrial copies.
278 The two insect species - *Drosophila melanogaster* and *Heliconius melpomene* - have similar estimated
279 nuclear genome mutation rates (2.8 ~ 2.9 x 10^{-9} substitution/generation). Whether *Ae. aegypti* has a
280 mutation rate similar to them remains to be determined. If we assume that *Culex quinquefasciatus* and
281 *Ae. aegypti* diverged 50-100 million years ago [20, 33], our mitochondrial mutation rate estimation is 5-
282 10 times slower than known insect nuclear genome mutation rates [23, 24](Figure S1). This contradicts
283 the general notion that mitochondrial genomes evolve faster than nuclear genomes [26-28]. Therefore,
284 it is more likely that multiple mutations occurred at a single mitochondrial locus, resulting in
285 underestimation of our mutation rate when we used *Culex quinquefasciatus* as an outgroup.
286 Consequently we rejected the divergence time estimate of 5.6 million years between two *Ae. aegypti*
287 mitochondrial lineages as shown in Figure S1. Further study of mutation rates in *Ae. aegypti* and/or
288 divergence time between more closely related taxa would illuminate a more accurate evolutionary
289 history.
290 Our data support the hypothesis that *Ae. aegypti* in CA currently exists as multiple, isolated populations.
291 High (> 0.1) genetic distance measured by F_{ST} as well as genome-wide differentiation patterns (Figure 3)
292 support multiple introductions into California from genetically distinct source populations as the most
293 plausible history of this invasion. We are investigating highly informative SNPs for the identification of
294 different genetic clusters (Figure 2B) in order to develop an affordable (<$10/sample) multiplex SNP
295 genotyping assay which can be used for a population-level analysis of *Ae. aegypti* in California. Examples
296 of informative SNPs distinguishing three genetic clusters are presented in Table 3. Successful
297 development of a genetic cluster identification assay will be useful in future comparative virus vector
298 competence studies [34] and examination of traits critical to environmental suitability such as drought
299 tolerance, egg overwintering and insecticide resistance.

Table 3. Example of informative SNPs distinguishing genetic clusters GC1, GC2 and GC3 (Figure 2). SNP id is the genomic coordinates in reference to the AaegL3 genome as presented in VectorBase [35]. The first decimal number before the colon (:) indicates supercontig number, the 2nd number between colons indicates coordinate within supercontig, and the third and last number after the 2nd colon indicates the strand (1 for positive strand and -1 for negative strand). The reference and variant nucleotide is based on the positive strand of the AaegL4 genome. NC stands for no call. Reference (AaegL3) homozygotes are marked in light blue, heterozygotes in yellow and variant homozygotes are marked in dark blue cells. * indicates SNPs located with an intron. Remaining SNPs are located within intergenic regions.

	reference	A	T	A	T	T	T	G	A	C	G	A	G
	variant	T	G	T	A	A	C	A	T	T	T	T	A
		GC1 vs GC2				**GC1 vs GC3**				**GC2 vs GC3**			
	sample	1.93:1735550:-1	1.151:1532211:1	1.185:1005919:1*	1.301:1307877:-1	1.404:229495:1*	1.482:650583:1	1.14:1026176:-1*	1.240:1496129:1*	1.508:771166:-1	1.32:1151270:-1	1.43:2941963:1	1.240:1496129:1
GC1	Ae13CLOV028	TT	GG	TT	AA	AA	CC	AA	TT	CT	TT	AA	GG
GC1	Ae13CLOV030	TT	GG	TT	AA	AA	CC	AA	TT	CT	GT	AA	GA
GC1	Ae13CLOV032	TT	GG	TT	AA	AA	CC	AA	TT	TT	GG	AA	GG
GC1	Ae15SANG073	TT	GG	TT	AA	AA	CC	AA	TT	CC	NC	AA	GG
GC1	Ae16CLOV002	TT	GG	TT	AA	AA	CC	AA	TT	TT	GG	AA	GG
GC1	Ae16CLOV003	TT	GG	TT	AA	AA	CC	AA	TT	CT	GG	AA	GG
GC1	Ae16CLOV007	NC	GG	TT	AA	AA	CC	AA	TT	TT	GG	AA	GG
GC1	Ae16CLOV008	TT	GG	TT	AA	AA	CC	AA	TT	CC	TT	AA	GG
GC2	Ae14EXET135	AA	TT	AA	TT	TT	TC	GG	AT	CC	TT	AA	GG
GC2	Ae14EXET165	AA	TT	AA	TT	TT	TC	GG	TT	CC	TT	AA	GG
GC2	Ae15GARG021	AA	TT	AA	TT	TT	TT	GG	AT	CC	TT	AA	GG
GC2	Ae15GARG026	AA	TT	AA	TT	TA	TT	NC	TT	CC	TT	AA	GG
GC2	Ae15IMPE002	AA	TT	AA	TT	TA	TT	GG	TT	CC	TT	AA	GG
GC2	Ae15MISV029	AA	TT	AA	TT	TT	TT	GG	AA	CC	TT	AA	GG
GC2	Ae15MISV036	AA	TT	AA	TT	TT	TT	GA	TT	CC	TT	AA	GG
GC2	Ae15SAND001	AA	TT	AA	TT	TA	CC	GG	AA	CC	TT	AA	GG
GC2	Ae15SAND002	AA	TT	AA	TT	TA	CC	GG	AA	CC	TT	AA	GG
GC2	Ae15SAND015	AA	TT	AA	TT	TA	CC	AA	AT	CC	TT	AT	GG
GC2	Ae16COMF3.001	AA	TT	AA	TT	TA	TT	AA	TT	CC	TT	AA	GG
GC3	Ae14SAMC001	NC	NC	AA	TT	TT	TT	GG	AA	NC	GG	TT	AA
GC3	Ae15FRES008.05	NC	NC	NC	TT	NC	NC	GG	AA	TT	GG	NC	AA
GC3	Ae15FRES017	NC	TG	AA	TA	TT	TT	GG	AA	TT	GG	TT	AA
GC3	Ae15MADE001	NC	NC	AT	TT	TT	TT	GG	AA	TT	GG	TT	AA
GC3	Ae15MADE015	NC	TG	AT	TA	TT	TT	GG	AA	TT	GG	TT	AA
GC3	Ae15SAMC001	NC	NC	AA	NC	TT	TT	GG	AA	TT	GG	TT	AA
GC3	Ae15SAMC004	NC	TG	AA	TT	TT	TT	GG	AA	TT	GG	TT	AA

310 We are also investigating the option of generating SNP genotypes that are compatible with the *Ae.*
311 *aegypti* SNP chip dataset [36]. The SNP positions provided in [36] were based on the *AaegL1* genome
312 assembly [15], and efforts to convert those *Aaeg* L1 genomic coordinates to the positions in the latest
313 reference genome *AaegL4* [6] are currently underway.
314
315 Microsatellite data from [1] indicated that San Mateo (=Menlo Park), Madera and Fresno samples were
316 genetically similar to samples from the southeastern USA which includes samples from Vaca Key and
317 Miami, Florida. Our results are inconsistent with this relationship. Our analysis placed the three central
318 California populations (Menlo Park, Madera and Fresno) in a group (GC3) distinct from the group
319 containing the Florida samples which also included populations from southern California (GC2, Figure 2).
320 Investigating a larger number of samples may provide a more complete picture of the genetics of *Ae.*
321 *aegypti* populations in California; a clearer description of its invasion history and spread in the state; and
322 provide information useful to its control. Annotation of the *AaegL4* assembly, when available, will also
323 open the door for investigations using our genome data.
324
325 **Acknowledgments**
326 We thank Orange County Mosquito and Vector Control District team (Mr. Michael Hearst, District
327 Manager) and Community Health Division of the Department of Environmental Health (Ms. Rebecca
328 Lafreniere, Chief and Ms. Elizabeth Pozzebon, Director) for sharing mosquito specimens. We also thank
329 Dr. Chris Barker for sharing specimens from Commerce. We thank Youki Yamasaki, Allison Chang and
330 Parker Houston for assisting DNA extraction and library preparations for this study. Special thanks to Dr.
331 Bradley Main for his comment to the manuscript. This project was funded by the UC Davis Bridge
332 Funding Program, School of Veterinary Medicine Vector-borne Disease Pilot Grant Program and DARPA .
333

334 **References**

335 1. Gloria-Soria A, Brown JE, Kramer V, Hardstone Yoshimizu M, Powell JR. Origin of the dengue fever
336 mosquito, *Aedes aegypti*, in California. PLoS Negl Trop Dis. 2014;8(7):e3029. doi:
337 10.1371/journal.pntd.0003029. PubMed PMID: 25077804; PubMed Central PMCID: PMC4117443.
338 2. Cornel AJ, Holeman J, Nieman CC, Lee Y, Smith C, Amorino M, et al. Surveillance, insecticide
339 resistance and control of an invasive *Aedes aegypti* (Diptera: Culicidae) population in California.
340 F1000Res. 2016;5:194. doi: 10.12688/f1000research.8107.3. PubMed PMID: 27158450; PubMed
341 Central PMCID: PMC4857756.
342 3. California Surveilliance Gateway Maps [Internet]. California Vectorborne Disease Surveillance
343 System. 2007 [cited 2016]. Available from: http://maps.calsurv.org/invasive.
344 4. Honorio NA, Silva Wda C, Leite PJ, Goncalves JM, Lounibos LP, Lourenco-de-Oliveira R. Dispersal of
345 *Aedes aegypti* and *Aedes albopictus* (Diptera: Culicidae) in an urban endemic dengue area in the
346 State of Rio de Janeiro, Brazil. Mem Inst Oswaldo Cruz. 2003;98(2):191-8. PubMed PMID: 12764433.
347 5. Harrington LC, Scott TW, Lerdthusnee K, Coleman RC, Costero A, Clark GG, et al. Dispersal of the
348 dengue vector Aedes aegypti within and between rural communities. Am J Trop Med Hyg.
349 2005;72(2):209-20. PubMed PMID: 15741559.
350 6. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the
351 *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92-
352 5. doi: 10.1126/science.aal3327. PubMed PMID: 28336562.
353 7. Nieman CC, Yamasaki Y, Collier TC, Lee Y. A DNA extraction protocol for improved DNA yield from
354 individual mosquitoes. F1000Res. 2015;4:1314. Epub Nov 20, 2015. doi:
355 10.12688/f1000research.7413.1.
356 8. Yamasaki YK, Nieman CC, Chang AN, Collier TC, Main BJ, Lee Y. Improved tools for genomic DNA
357 library construction of small insects. F1000Res [Internet]. 2016; 5:[211 p.]. Available from:
358 http://f1000research.com/posters/5-211.
359 9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
360 Bioinformatics. 2014. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404.
361 10. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Cornell
362 University Library; 2013. p. arXiv:1303.3997v2.
363 11. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint
364 [Internet]. 2012. Available from: http://arxiv.org/abs/1207.3907.

365 12. Crawford JE, Lazzaro BP. Assessing the accuracy and power of population genetic inference from
366 low-pass next-generation sequencing data. Frontiers in genetics. 2012;3:66. doi:
367 10.3389/fgene.2012.00066. PubMed PMID: 22536207; PubMed Central PMCID: PMCPMC3334522.
368 13. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989;5:164-6.
369 14. Miles A, Harding N. scikit-allel - Explore and analyse genetic variation. 0.21.2 ed.
370 https://github.com/cggh/scikit-allel: GitHub; 2016.
371 15. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of Aedes aegypti, a
372 major arbovirus vector. Science. 2007;316(5832):1718-23. doi: 10.1126/science.1138878. PubMed
373 PMID: 17510324; PubMed Central PMCID: PMC2868357.
374 16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and
375 VCFtools. Bioinformatics. 2011;27(15):2156-8. doi: 10.1093/bioinformatics/btr330. PubMed PMID:
376 21653522; PubMed Central PMCID: PMC3137218.
377 17. Behura SK, Lobo NF, Haas B, deBruyn B, Lovin DD, Shumway MF, et al. Complete sequences of
378 mitochondria genomes of Aedes aegypti and Culex quinquefasciatus and comparative analysis of
379 mitochondrial DNA fragments inserted in the nuclear genomes. Insect Biochem Mol Biol.
380 2011;41(10):770-7. doi: 10.1016/j.ibmb.2011.05.006. PubMed PMID: 21640823; PubMed Central
381 PMCID: PMCPMC3162086.
382 18. Hohna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, et al. RevBayes: Bayesian
383 Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.
384 Systematic biology. 2016;65(4):726-36. doi: 10.1093/sysbio/syw021. PubMed PMID: 27235697;
385 PubMed Central PMCID: PMCPMC4911942.
386 19. Rambaut A, Drummond AJ. Tracer 2014 Available from: http://beast.bio.ed.ac.uk/Tracer.
387 20. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al. Mosquito
388 genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science.
389 2015;347(6217):1258522. doi: 10.1126/science.1258522. PubMed PMID: 25554792.
390 21. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data.
391 Genetics. 1992;132(2):583-9. PubMed PMID: 1427045; PubMed Central PMCID: PMCPMC1205159.
392 22. Main BJ, Lee Y, Collier TC, Norris LC, Brisco K, Fofana A, et al. Complex genome evolution in
393 *Anopheles coluzzii* associated with increased insecticide usage in Mali. Mol Ecol. 2015;24(20):5145-
394 57. doi: 10.1111/mec.13382. PubMed PMID: 26359110; PubMed Central PMCID: PMC4615556.

395 23. Keightley PD, Ness RW, Halligan DL, Haddrill PR. Estimation of the spontaneous mutation rate per
396 nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics. 2014;196(1):313-20. doi:
397 10.1534/genetics.113.158758. PubMed PMID: 24214343; PubMed Central PMCID: PMC3872194.
398 24. Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, et al. Estimation of the
399 spontaneous mutation rate in *Heliconius melpomene*. Mol Biol Evol. 2015;32(1):239-43. doi:
400 10.1093/molbev/msu302. PubMed PMID: 25371432; PubMed Central PMCID: PMC4271535.
401 25. Moreno M, Marinotti O, Krzywinski J, Tadei WP, James AA, Achee NL, et al. Complete mtDNA
402 genomes of Anopheles darlingi and an approach to anopheline divergence time. Malar J.
403 2010;9:127. doi: 10.1186/1475-2875-9-127. PubMed PMID: 20470395; PubMed Central PMCID:
404 PMCPMC2877063.
405 26. Lynch M. The origins of genome architecture. Sunderland, Mass.: Sinauer Associates; 2007. xvi, 494
406 p. p.
407 27. Havird JC, Sloan DB. The Roles of Mutation, Selection, and Expression in Determining Relative Rates
408 of Evolution in Mitochondrial versus Nuclear Genomes. Mol Biol Evol. 2016;33(12):3042-53. doi:
409 10.1093/molbev/msw185. PubMed PMID: 27563053; PubMed Central PMCID: PMCPMC5100045.
410 28. Molnar RI, Bartelmes G, Dinkelacker I, Witte H, Sommer RJ. Mutation rates and intraspecific
411 divergence of the mitochondrial genome of Pristionchus pacificus. Mol Biol Evol. 2011;28(8):2317-
412 26. doi: 10.1093/molbev/msr057. PubMed PMID: 21368317.
413 29. Diaz-Nieto LM, Chiappero MB, Diaz de Astarloa C, Macia A, Gardenal CN, Beron CM. Genetic
414 Evidence of Expansion by Passive Transport of Aedes (Stegomyia) aegypti in Eastern Argentina. PLoS
415 Negl Trop Dis. 2016;10(9):e0004839. doi: 10.1371/journal.pntd.0004839. PubMed PMID: 27583521;
416 PubMed Central PMCID: PMCPMC5008820.
417 30. Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, Ferguson HM, et al. Diversity, differentiation,
418 and linkage disequilibrium: prospects for association mapping in the malaria vector *Anopheles*
419 *arabiensis*. G3 (Bethesda). 2014;4(1):121-31. doi: 10.1534/g3.113.008326. PubMed PMID:
420 24281424; PubMed Central PMCID: PMC3887528.
421 31. Black Iv WC, Bernhardt SA. Abundant nuclear copies of mitochondrial origin (NUMTs) in the Aedes
422 aegypti genome. Insect Mol Biol. 2009;18(6):705-13. doi: 10.1111/j.1365-2583.2009.00925.x.
423 PubMed PMID: 19849722.
424 32. Hlaing T, Tun-Lin W, Somboon P, Socheat D, Setha T, Min S, et al. Mitochondrial pseudogenes in the
425 nuclear genome of Aedes aegypti mosquitoes: implications for past and future population genetic

426 studies. BMC genetics. 2009;10:11. doi: 10.1186/1471-2156-10-11. PubMed PMID: 19267896;
427 PubMed Central PMCID: PMCPMC2660364.
428 33. Marinotti O, Cerqueira GC, de Almeida LG, Ferro MI, Loreto EL, Zaha A, et al. The genome of
429 *Anopheles darlingi*, the main neotropical malaria vector. Nucleic Acids Res. 2013;41(15):7387-400.
430 doi: 10.1093/nar/gkt484. PubMed PMID: 23761445; PubMed Central PMCID: PMC3753621.
431 34. Goncalves CM, Melo FF, Bezerra JM, Chaves BA, Silva BM, Silva LD, et al. Distinct variation in vector
432 competence among nine field populations of *Aedes aegypti* from a Brazilian dengue-endemic risk
433 city. Parasit Vectors. 2014;7:320. doi: 10.1186/1756-3305-7-320. PubMed PMID: 25015526; PubMed
434 Central PMCID: PMCPMC4230638.
435 35. Giraldo-Calderon GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an
436 updated bioinformatics resource for invertebrate vectors and other organisms related with human
437 diseases. Nucleic Acids Res. 2015;43(Database issue):D707-13. doi: 10.1093/nar/gku1117. PubMed
438 PMID: 25510499; PubMed Central PMCID: PMC4383932.
439 36. Evans BR, Gloria-Soria A, Hou L, McBride C, Bonizzoni M, Zhao H, et al. A Multipurpose High
440 Throughput SNP Chip for the Dengue and Yellow Fever Mosquito, Aedes aegypti. G3 (Bethesda).
441 2015. doi: 10.1534/g3.114.016196. PubMed PMID: 25721127.
442