# Exon-Enriched Libraries Reveal Large Genic Differences Between *Aedes aegypti* from Senegal, West Africa, and Populations Outside Africa

Laura B. Dickson,*,[1,2] Corey L. Campbell,*,[1] Punita Juneja,[†] Francis M. Jiggins,[†] Massamba Sylla,*
and William C. Black IV,*,[3]
*Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, Colorado 80523 and
[†]Department of Genetics, University of Cambridge, CB24 6BG, United Kingdom

**ABSTRACT** *Aedes aegypti* is one of the most studied mosquito species, and the principal vector of several arboviruses pathogenic to humans. Recently failure to oviposit, low fecundity, and poor egg-to-adult survival were observed when *Ae. aegypti* from Senegal (*SenAae*) West Africa were crossed with *Ae. aegypti* (*Aaa*) from outside of Africa, and in *SenAae* intercrosses. Fluorescent *in situ* hybridization analyses indicated rearrangements on chromosome 1, and pericentric inversions on chromosomes 2 and 3. Herein, high throughput sequencing (HTS) of exon-enriched libraries was used to compare chromosome-wide genetic diversity among *Aaa* collections from rural Thailand and Mexico, a sylvatic collection from southeastern Senegal (PK10), and an urban collection from western Senegal (Kaolack). Sex-specific polymorphisms were analyzed in Thailand and PK10 to assess genetic differences between sexes. Expected heterozygosity was greatest in *SenAae*. $F_{ST}$ distributions of 15,735 genes among all six pairwise comparisons of the four collections indicated that Mexican and Thailand collections are genetically similar, while $F_{ST}$ distributions between PK10 and Kaolack were distinct. All four comparisons of *SenAae* with *Aaa* indicated extreme differentiation. $F_{ST}$ was uniform between sexes across all chromosomes in Thailand, but were different, especially on the sex autosome 1, in PK10. These patterns correlate with the reproductive isolation noted earlier. We hypothesize that cryptic *Ae. aegypti* taxa may exist in West Africa, and the large genic differences between *Aaa* and *SenAae* detected in the present study have accumulated over a long period following the evolution of chromosome rearrangements in allopatric populations that subsequently cause reproductive isolation when these populations became sympatric.

The mosquito, *Aedes aegypti* (L), is the principal vector in tropical and subtropical regions world-wide of dengue (DENV1–4)(Guzman and

Harris 2015), Yellow Fever (YF) (Beasley *et al.* 2015) and Zika (Li *et al.* 2012) flaviviruses, and of the Chikungunya alphavirus (Lo Presti *et al.* 2016). The ecology and population biology of the species is well understood (Mattingly 1957), and the ease of colonization and laboratory manipulation have helped it to become the most studied of all mosquito species. There are two recognized subspecies of *Ae. aegypti*: *Ae. aegypti* (*Aaa*) is a domestic form that has spread circumglobally between 35° north and south latitudes, typically in close association with humans, while *Ae. aegypti formosus* (*Aaf*) is a sylvan mosquito that occurs primarily in forest habitats. Most studies of these subspecies have occurred in East Africa, where the subspecies are sympatric. McClelland (1974) formally defined *Aaf* populations as having a dark black cuticle, no pale scales on the first abdominal tergite, breeding in natural containers such as tree holes, and feeding primarily on wild animals. In contrast, *Aaa* is a more domestic subspecies that has spread out of Africa. *Aaa* breeds in artificial containers provided by humans, will breed indoors, and has a

preference for feeding on human blood. It is described as being a lighter colored mosquito, with a range of pale scaling on the abdominal tergites. Presumably, *Aaa* arose from ancestral *Aaf*, and eventually became adapted to human habitats. It then spread from Africa to many different tropical and subtropical regions of the world through human commerce and exploration (Tabachnick 1991).

Detailed mark-release-recapture studies in Kenya (Trpis and Hausermann 1975) demonstrated that immature mosquitoes collected from sylvan, peridomestic, or domestic breeding containers showed an overwhelming preference for their respective habitat as adults. Laboratory experiments crossing *Aaa* and *Aaf* from Kenya showed no evidence of assortative mating between the two subspecies (Moore 1979), and found no decrease in fecundity in hybrids, nor any morphological defects. In rural Kenya, *Aaf* collected from the forest had very different allozyme frequencies when compared to *Aaa* collected indoors, although some populations used in the study were laboratory colonies several generations from the field (Scott and McClelland 1975). Another study in Kenya showed that *Aaa* populations within villages are panmictic, but genetically differentiated from *Aaa* in other villages, even those <2 km away. This suggested that gene flow between villages is infrequent, and *Aaf* living in the intervening forest do not serve as a "genetic bridge," so that *Aaf* and *Aaa* represent separate breeding populations (Tabachnick and Powell 1978). Further investigation demonstrated that *Aaf* collected from the forest or disturbed areas around villages (*e.g.*, coconut groves) were genetically distinct from *Aaa* collections made indoors (Tabachnick *et al.* 1979).

In contrast, *Ae. aegypti* in West Africa has received relatively little attention. While scale patterns correlate with genetic differences in East Africa (Tabachnick *et al.* 1979), this is not the case in West Africa (Huber *et al.* 2008; Moore *et al.* 2013; Paupy *et al.* 2008, 2010; Sylla *et al.* 2009), where cuticle color is predominantly black, with a continuum of scaling patterns. Collections from West Africa with varied scale patterns have been sampled in both the dry and rainy seasons from different vegetation zones, and from domestic *vs.* sylvan habitats. Regardless of collection site, most variation was associated with geographic distance, vegetative zones, ecological habitats, and season, rather than marker type or scale pattern (Brown *et al.* 2014, 2011; Huber *et al.* 2008; Paupy *et al.* 2008; Sylla *et al.* 2009). Accordingly, we designated *Ae. aegypti* in West Africa as *SenAae* because cuticle color is predominantly black (therefore not *Aaa*), and mosquitoes collected in both sylvatic and urban sites exhibit a continuum of scaling patterns (therefore not *Aaf*).

In an attempt to clarify genetic relationships among, and within, West African *Ae. aegypti*, we recently reported crossing experiments of *Aaa* $P_1$ laboratory strains with Senegal *Ae. aegypti* (*SenAae*) (Dickson *et al.* 2016), wherein we were unable to build intercross families due to poor $F_1$ oviposition and low $F_2$ egg-to-adult survival. Egg hatch was predicted based upon Haldane's rule (Haldane 1922) to be low when a female mates with a hybrid $F_1$ male, but average when a male mates with a hybrid $F_1$ female. Instead, most females mated to hybrid males laid no eggs, and egg-pupal survival was reduced when males mated with hybrid females. Linkage and fluorescent *in situ* hybridization (FISH) analyses identified a likely rearrangement on chromosome 1, and identified two rDNA cistrons on that chromosome, and pericentric inversions on chromosomes 2 and 3 (Dickson *et al.* 2016). The latter results suggest that the reproductive isolation observed within *SenAae* may be associated with chromosome rearrangements. These observations differ markedly from the 1979 study of hybridization and mating behavior between *Aaa* and *Aaf* in East Africa, which found no evidence of hybrid breakdown or of assortative mating (Moore 1979). The goal of the current study was to use high throughput sequencing (HTS) of

exon-enriched gDNA libraries to assess the degree of genetic differentiation between two interfertile *SenAae* (called PK10 and Kaolack), and two interfertile *Aaa* (Thailand and Mexico) collections. Dual replicate pools of equal numbers of males and females were used for these comparisons. Expected heterozygosity ($H_{exp}$) in each collection was compared to give an indication of the relative levels of within-collection genetic diversity. These analyses were extended by assessing sex-specific polymorphisms for the Thailand (*Aaa*) and PK10 (*SenAae*) collections. Standardized variance in SNP frequencies ($F_{ST}$) was compared within and between *Aaa* and *SenAae*, and between females and males in the Thailand (*Aaa*) and PK10 (*SenAae*) collections. SNP frequencies were significantly different among sexes in PK10, with the largest differences occurring on sex chromosome 1, but were homogeneous across all chromosomes in the Thailand collection.

## MATERIALS AND METHODS

### Mosquito collections

Senegalese *Ae. aegypti* populations were collected as larvae near domestic sites in urban and village environments, as well as from sylvatic habitats throughout Senegal (Sylla *et al.* 2009) (Supplemental Material, Figure S1). In February 2010, *Ae. aegypti* were collected in Kaolack (14°9′22.35′′N, 16°4′32.31′′W) in West Central Senegal, and in the PK10 forest (12°36′43.00′′N, 12°14′46.80′′W) in southeastern Senegal in September 2011. A complete description of mosquitoes found in PK10 has been published (Sylla *et al.* 2013). The PK10 colony was established from larvae taken from treeholes at the forest–savannah margin (Sylla *et al.* 2013). Both 2009 and 2011 PK10 collections consisted of adults without scales on the first abdominal tergite (McClelland 1974). The Merida, Mexico (Vergel) (20°57′16.82′′N, 89°35′20.24′′W) collection of *Aaa* was collected in 2007. The Pai Lom, Mae Sot Thailand (16°42′46.54′′N, 98°34′28.74′′E) 2002 collection of *Aaa* consisted of DNA individually isolated from adults (Bosio *et al.* 2005). Sex-specific polymorphisms were analyzed in PK10 (*SenAae*) and Thailand (*Aaa*), but not in Mexico (*Aaa*) or Kaolack (*SenAae*) collections because mosquitoes were not identified to sex prior to DNA isolation.

Larvae were removed mostly from artificial containers (*e.g.*, tires, water storage containers, and discarded trash), reared to adults, blood fed on the senior author's arm, and collected eggs were returned to Fort Collins where they were established as laboratory colonies. In each collection, these were raised to adults, transferred to 1-pint cages, anesthetized with Triethylamine (FlyNap Carolina Biological Supply Company, Burlington, NC), and classified according to the presence/absence of scales on the first abdominal tergite (McClelland 1974; Sylla *et al.* 2009, 2013). At least 50 adults of each sex were individually identified and stored in Purell Advanced Hand Sanitizer for eventual extraction of DNA (Black and DuTeau 1997). All colonized adults were maintained at 28°, 70–80% relative humidity, and for a 12:12 hr photoperiod.

### Sequencing

HTS of exon-enriched libraries (Juneja *et al.* 2015) was used to quantify the extent of genomic differences within, and among, the four *Ae. aegypti* collections analyzed in this study, and between sexes within PK10 (*SenAae*) and Thailand (*Aaa*) collections. Twelve libraries were constructed: (1) PK10 males replicate 1 (*n* =12 mosquitoes), (2) PK10 males replicate 2 (*n* = 12), (3) PK10 females replicate 1 (*n* = 12), (4) PK10 females replicate 2 (*n* = 12), (5) Thailand males replicate 1 (*n* = 12), (6) Thailand males replicate 2 (*n* = 12), (7) Thailand females replicate 1, (8) Thailand females replicate 2 (*n* = 12), (9)

| Collection | Monomorphic | Number of Nucleotides/Indels | | | | | Total | % Polymorphic |
|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | | |
| PK10 | 30,229,087 | 2,580,103 | 148,547 | 9276 | 379 | 12 | 32,967,403 | 8.31% |
| Female | 28,164,602 | 1,949,160 | 83,760 | 3976 | 131 | 6 | 30,201,634 | 6.74% |
| Male | 28,324,103 | 1,988,143 | 90,424 | 4624 | 174 | 5 | 30,407,471 | 6.85% |
| Kaolack | 30,259,280 | 2,250,995 | 121,064 | 7291 | 305 | 9 | 32,638,942 | 7.29% |
| Merida | 28,592,543 | 1,459,996 | 42,017 | 1418 | 48 | 1 | 30,096,023 | 5.00% |
| Thailand | 30,224,052 | 1,754,936 | 54,368 | 2035 | 72 | 2 | 32,035,463 | 5.65% |
| Female | 26,580,423 | 1,119,509 | 25,775 | 816 | 30 | 1 | 27,726,553 | 4.13% |
| Male | 27,166,304 | 1,222,100 | 30,373 | 979 | 35 | 0 | 28,419,790 | 4.41% |

For example, 32,967,403 sites were analyzed in PK10. Of these, 30,229,087 of these were monomorphic, while 2,580,103 sites had two alternative nucleotides/indels, and 148,547 had three alternative nucleotides/indels. The percentage polymorphic is 100×[1 − (monomorphic/total)].

Kaolack equal numbers of both sexes replicate 1 ($n = 14$), (10) Kaolack equal numbers of both sexes replicate 2 ($n = 14$), (11) Merida, Mexico equal numbers of both sexes replicate 1 ($n = 22$), and (12) Merida, Mexico equal numbers of both sexes replicate 2 ($n = 22$). Thus, each of the four collections was represented by 28, 44, or 48 individuals.

Prior to pooling, DNA in individual mosquitoes was quantified using Pico Green (Life Technologies, Thermo Fisher Scientific Inc.), and equal amounts of DNA per mosquito were pooled. A Covaris S2 sonicator (Covaris Ltd, Brighton, U.K.) sheared pooled DNA to an average size of 500 bp. Sonication conditions were: duty cycle 10%, intensity 5.0, cycles per burst 200, duration 40 sec, mode frequency sweeping, displayed power 23 W, and Temperature 5.5°–6°. Each TruSeq DNA LT (v.2) library was prepared using 1 μg of sheared genomic DNA following the manufacturer's recommendations. Equimolar quantities of prepared libraries were pooled and enriched for coding sequences by exome capture using custom SeqCap EZ Developer probes (Nimblegen) (Juneja *et al.* 2015). Overlapping probes covering the protein coding sequence (not including UTRs) in the AaegL3.3 gene annotations (https://www.vectorbase.org/organisms/aedes-aegypti/liverpool-lvp/AaegL3.3) were produced by Nimblegen based on exonic coordinates specified by the Jiggins laboratory. In total, 26.7 Mb of the genome (2%) was targeted for enrichment. Exome capture coordinates are available at https://www.jiggins.gen.cam.ac.uk/data/Aaegypti_exome.bed. Enrichment followed the Nimblegen SeqCap EZ protocol. Briefly, pooled TruSeq libraries were hybridized to the probes for 64 hr, unbound DNA was washed away, and the targeted DNA was eluted and amplified. These were then sequenced on two lanes of a HiSeq2000 (Illumina) for paired-end 2 × 100 nt sequencing. TruSeq library preparation, exome capture, and sequencing were performed by the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (Oxford, UK), and produced reads with quality scores >30.

**Alignments and population genetics pipeline**

A custom concatenated reference file "All3U" was built based on the Vectorbase AaegL3.3 gene build, and SNPs were called with respect to this build. The reference was used for the detailed analysis of 18,840 transcripts among 15,735 genes. A FORTRAN program used the coordinates of the features of each gene appearing in *Aedes-aegypti* Liverpool_ BASEFEATURES_Aaegl3.3.GTF to extract: (1) 600 bp of the *cis* region 5′ to each annotated transcript (5′ nontranscribed = 5′NTR), (2) all 5′UTRs, (3) exons, (4) introns, (5) 3′UTRs, and (6) a 600 bp *cis* region in the 3′ direction (3′ nontranscribed = 3′NTR). Each SNP was given an ordered SNPID number according to its position in the gene,

the location of the gene in the supercontig, and where the supercontig mapped physically (Timoshevskiy *et al.* 2014). There were 1948 mapped genes containing 41,119,004 SNPs on chromosome 1, 3429 genes containing 70,643,307 SNPs on chromosome 2, and 2382 genes with 44,883,168 SNPs on chromosome 3. There were 7862 unmapped genes containing 144,703,188 SNPs. In total there were 15,735 genes containing 301,348,667 SNPs.

Fastq files from each library were aligned to the All3U reference file using GSNAP (v. 2013-10-28), allowing 10% divergence from the reference (Wu and Nacu 2010). SAMtools (Li *et al.* 2009) converted GSNAP outputs to *.mpileup files. The "readcounts" command in Varscan2 (v2.3.5) (Koboldt *et al.* 2012) used the *.mpileup file to report SNPs, with default values of a minimum of 15× coverage in each library, and a minimum base quality of 30 (http://varscan.sourceforge.net/using-varscan.html). A series of FORTRAN programs then read the "readcount" files to produce a flat file containing the SNPID number, the Aaegl3.3 reference nucleotide, its coverage at each sequenced position, and the coverages of insertions or deletions at that position.

The program "Trim" removed all SNPs that had <15 counts, and capped all SNP coverage at 2000 to avoid regions of repetitive DNA. The program "2 × 2" read two files, one from each replicate at a collection site, and produced a single "2 × 2.out" file that contained SNPs that occurred in both replicate datasets for each collection (Mexico, Thailand male, Thailand female, PK10 male, PK10 female, and Kaolack). A program "2 × 2 × 2" then read two 2 × 2.out files to produce a list of SNPs common to a pair of collections. A program "Tables" then read the two "2 × 2.out" files and the list to produce a contingency table for each SNP common to both collections. The contingency file contained: (1) the SNPID coordinate followed by four lines containing (2) the name of the library containing that SNP, and (3) the coverage for each nucleotide at that SNP. To ensure high confidence SNP calls, replicate SNP frequencies were compared using a heterogeneity $\chi^2$ test with *nid*-1 degrees of freedom, where *nid* is the number of nucleotides and indels segregating at a SNP. Any SNP in which nucleotide frequencies were significantly different between either replicate were discarded. A program read the contingency table file to produce a file that contains for each SNP: (1) the SNPID, (2) the VectorBase Gene ID (*e.g.*, AAEL001234), (3) coverages of all nucleotide/indels in the first collection, (4) coverages of all nucleotide/indels in the second collection, (5) numbers of alternate nucleotides at that SNP, (6) a list of polymorphic nucleotides/indels, (7) the Aaegl3.3 reference nucleotide, (8) the alternate nucleotide (second most common nucleotide, also referred to as the minor allele), (9) mutation type (*i.e.*, "transition," "transversion," "insertion," or "deletion"), (10) whether the mutation
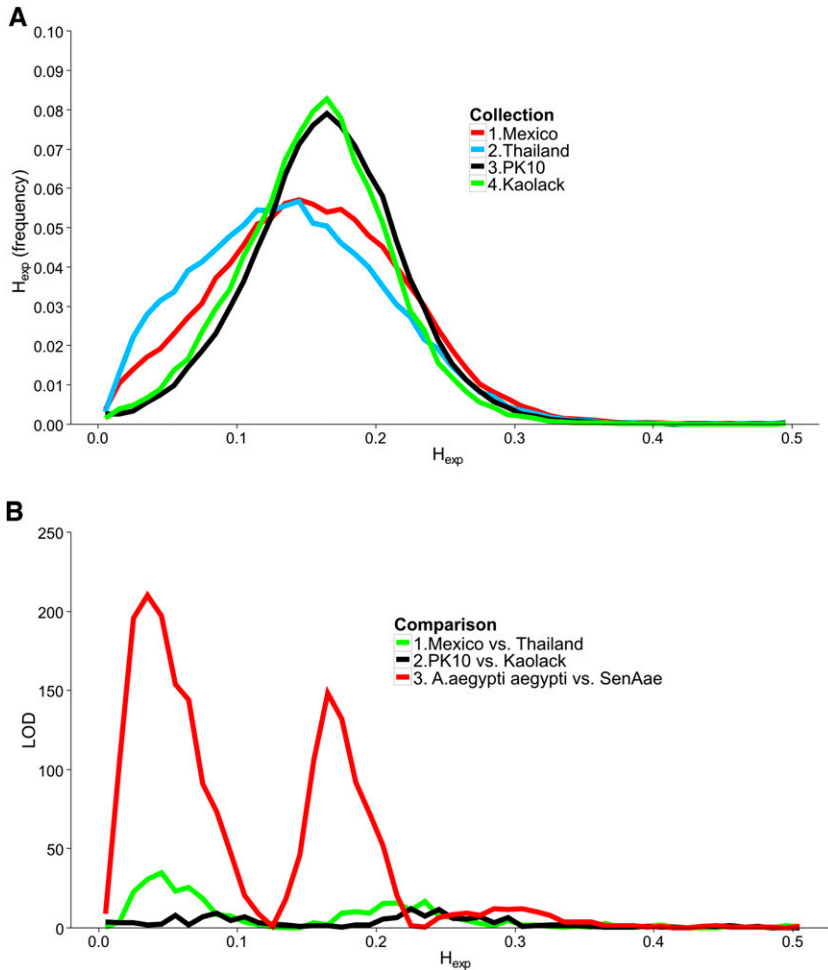
**Figure 1** Expected heterozygosity ($H_{exp}$) distribution for each collection. (A) *Aaa* Mexico (red line), *Aaa* Thailand (bright blue), *SenAae* PK-10 (black), and *SenAae* Kaolack (green). (B) $H_{exp}$ distributions were compared with one another using a heterogeneity $\chi^2$ test of the number of genes in each 0.01 bin along the abscissa in the histogram. *Aaa vs. SenAae* comparisons (red line) have larger LOD differences in $H_{exp}$ scores than *Aaa vs. Aaa* (green line) or *SenAae vs. SenAae* (black line).

is a synonymous or a replacement substitution for SNPs in codons, and (11) whether the SNP resides in a first, second, or third codon position, or occurs within a 5′UTR, 5′NTR, 3′UTR, 3′NTR, or an intron. Transfer RNAs were annotated using the anticodon tool at http://lowelab. ucsc.edu/tRNAscan-SE/ (Schattner *et al.* 2005).

A program was written to calculate for each SNP, a between-collection component ($a_s$), a within-collection component ($b_s$), and $F_{ST}$ calculated from $a_s$ and $b_s$ following Fumagalli *et al.* (2013), where:

$$a_s = \frac{4n_i\left(\hat{p}_{(i,s)} - \hat{p}_s\right)^2 + 4n_j\left(\hat{p}_{(j,s)} - \hat{p}_s\right)^2 - b_s}{2\left[2n_in_j/(n_i + n_j)\right]}$$

and

$$b_s = \frac{n_i\alpha_{(i,s)} + n_j\alpha_{(j,s)}}{n_i + n_j - 1}$$

where

$$\alpha_{(i,s)} = 2\hat{p}_{(i,s)}\left(1 - \hat{p}_{(i,s)}\right) \text{ and } \alpha_{(j,s)} = 2\hat{p}_{(j,s)}\left(1 - \hat{p}_{(j,s)}\right)$$

$\alpha_{(i,s)}$ is also known as the expected heterozygosity ($H_{exp}$). $\hat{p}_{(i,s)}$ is the coverage of a nucleotide at SNP($s$) divided by the total coverage of $s$ in collection ($i$). $n_i$ and $n_j$ are the number of mosquitoes sampled in collections $i$ and $j$, and $\hat{p}_s$ is the coverage of a nucleotide at $s$ in both

$i$ and $j$ collections, divided by the total coverage of $s$ in both $i$ and $j$ collections. The estimate of $F_{ST}$ for $s$ is:

$$F_{ST}(s) = \frac{a_s}{a_s + b_s}$$

and, for an entire gene ($g$) with $m$ SNPs, is:

$$F_{ST}(g) = \frac{\sum_{s=1}^{m} a_s}{\sum_{s=1}^{m}(a_s + b_s)}$$

Graphing and statistical metrics were carried out in R-Bioconductor (Huber *et al.* 2015). Venn diagrams (Figure S2 and Figure S3) were prepared using the webtool Venny (Oliveros 2010).

### Data availability

All fastq files are available through the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under accession numbers PRJNA258086 and SRP061709.

## RESULTS

### Genetic diversity within Ae. aegypti collections

Numbers of sequences in each of the six libraries appear in Table 1. To estimate the amount of genetic diversity within each collection, the percentage of polymorphic SNPs were calculated. The greatest

| Comparison | Number of Genes | $H_{exp}$ (1) > $H_{exp}$(2) | $H_{exp}$ (1) < $H_{exp}$ (2) | % $H_{exp}$ (1) > $H_{exp}$ (2) (%) | Average Difference | Probability in Pairwise *t*-Test |
|---|---|---|---|---|---|---|
| (1) Mexico *vs.* (2) Thailand | 15,735 | 10,628 | 5107 | 68 | 0.0018 | ≤2.2E-16 |
| (1) PK10 *vs.* (2) Kaolack | 15,735 | 15,211 | 524 | 97 | 0.0082 | ≤2.2E-16 |
| (1) PK10 *vs.* (2) Mexico | 15,735 | 13,650 | 2085 | 87 | 0.0054 | ≤2.2E-16 |
| (1) PK10 *vs.* (2) Thailand | 15,735 | 14,340 | 1395 | 91 | 0.0072 | ≤2.2E-16 |
| (1) Kaolack *vs.* (2) Mexico | 15,735 | 4509 | 11,226 | 29 | 0.0029 | ≤2.2E-16 |
| (1) Kaolack *vs.* (2) Thailand | 15,735 | 6841 | 8894 | 43 | 0.0010 | ≤2.2E-16 |
| (1) PK10 female *vs.* (2) male | 15,722 | 8499 | 7223 | 54 | 0.0008 | ≤2.2E-16 |
| (1) Thailand female *vs.* (2) male | 15,481 | 6910 | 8571 | 45 | −0.0018 | ≤2.2E-16 |

For example, in comparing the Mexican and Thailand collections, 15,735 genes occurred in common between the two collections. In 10,628 of these genes, $H_{exp}$ was greatest in the Mexican collection, while $H_{exp}$ was greatest in 5107 genes in Thailand. $H_{exp}$ was greatest in the Mexican collection in 68% of comparisons, and the average difference in $H_{exp}$ between Mexican and Thailand collections was 0.0018, which was significantly greater than zero.

percentage of polymorphism was observed in the PK10 (8.3%) and Kaolack (7.3%) collections. Figure 1A shows the distribution of $H_{exp}$ in all four collections. The distribution of $H_{exp}$ was shifted to the right in PK10 and Kaolack collections, whereas the $H_{exp}$ distributions for the two *Aaa* collections were skewed toward lower values. Figure 1B shows the LOD scores associated with three heterogeneity $\chi^2$ analyses. Minor differences in $H_{exp}$ distributions exist between Mexico and Thailand (green line), or between Kaolack and PK10 (black line). However, the differences in $H_{exp}$ distributions between *Aaa* and *SenAae* were large and highly significant (red line).

Table 2 shows a gene-wise comparison of $H_{exp}$ in all six pairwise comparisons of collections, and between sexes, in PK10 and Thailand. $H_{exp}$ in Mexico was greater that $H_{exp}$ in Thailand at 67% of SNPs. $H_{exp}$ in PK10 was greater than $H_{exp}$ in the three other collections in 86–96% of SNPs. Interestingly, $H_{exp}$ in Kaolack was less than $H_{exp}$ in Thailand or Mexico.

Next, $H_{exp}$ values were ranked from smallest to largest, and genes with the lowest 1% of average $H_{exp}$, and highest 1% of average $H_{exp}$ were identified. Of a total of 490 genes in the lower 1% tail of the $H_{exp}$ distributions, 376 occurred in only one collection, 80 occurred in two collections, 29 occurred in three collections, and five genes had low $H_{exp}$ in all collections (Figure S2 and Table S1). AAEL008073-RA encodes a Hyaluronan/mRNA-binding protein that is known in humans to bind the mRNA of type-1 plasminogen activator inhibitor (PAI-1), and is thought to be involved in regulation of mRNA stability (Heaton *et al.* 2001). The second gene, AAEL014847-RA, encodes innexin 2 (AAEL014847, *inx2*), and is involved in postembryonic development in drosophilids (Holcroft *et al.* 2013). The third is the 28S rDNA gene, while AAEL007163-RA and AAEL014638-RA encode hypothetical proteins.

In the upper 1% (Table S2) (*n* = 326), 160 genes occurred in only one collection, 63 occurred in two collections, 58 occurred in three collections, while 44 genes were common in all collections (Figure S3). Half (22) coded for noncoding RNAs (ncRNAs), 21 of which were tRNAs, and one was the U2 spliceosomal RNA gene (AAEL017645). Three encode ankyrins—a family of adaptor proteins that help attach integral membrane proteins to the cytoskeleton; they have binding sites for the β-subunit of spectrin, and 12 families of integral membrane proteins. Ankyrins appear to be required to maintain the integrity of plasma membranes, and to anchor specific ion channels, ion exchangers, and ion transporters in the plasma membrane. Two encode mitochondrial membrane ATP synthase ($F_1F_0$ ATP synthase or Complex V) that produce ATP from ADP in the presence of a proton gradient across the mitochondrial membrane, and two encode lipid storage droplets surface binding protein 2(lsd2), which is essential for embryogenesis, and required for normal deposition of neutral lipids in the oocyte.

## Genetic differences among Ae. aegypti collections

To assess genetic diversity between pairs of collections, $F_{ST}$ values were calculated for each of the six pairwise population comparisons, and the distribution of the $F_{ST}$ values were plotted with a bin width of 0.01 using "hist(FST),breaks=seq(−0.05, 1, 0.01)" in R 3.1.0 (R Development Core Team 2014) (Figure 2). Each of these distributions are plotted individually in Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10, and Figure S11. The red line in Figure 2A is the distribution of pairwise $F_{ST}$ values for the genes that occurred in common between Thailand and Mexico (*Aaa* collections), while the dark blue line is the distribution of pairwise $F_{ST}$ for genes shared between PK10 and Kaolack (*SenAae* collections). The remaining curves are all comparisons between *Aaa* and *SenAae* (*i.e.*, Kaolack *vs.* Thailand, Kaolack *vs.* Mexico, PK10 *vs.* Thailand, and PK10 *vs.* Mexico). Table 3 lists the numbers of genes and SNPs in each pairwise comparison. Also listed are the mean, SD, median, and modal $F_{ST}$ values, and the lower 5% and upper 95% $F_{ST}$ cutoff values.

Four trends are evident. First, $F_{ST}$ values between Thailand and Mexico are positively skewed, indicating that the majority of genes are similar between the two *Aaa* collections, despite the fact that the collection sites are ∼14,950 km apart (shortest arc distance). Second, pairwise $F_{ST}$ values between the Kaolack and PK10 collections, which are ∼450 km apart, are shifted to the right, indicating that genetic differences between the two *SenAae* collections were greater than between the two *Aaa* collections. Third, $F_{ST}$ values distributions between Senegal and the Thailand and Mexico collections are strongly shifted to the right, indicating large $F_{ST}$ differences between *SenAae* and *Aaa*. LOD scores of a heterogeneity $\chi^2$ test are consistent with these observations (Figure 2B). Fourth, in examining Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10, and Figure S11, it is clear that the distribution of $F_{ST}$ values between any two collections is not normally distributed, but rather is highly skewed. This is especially evident when comparing sexes in a collection (Figure S4 and Figure S5). Furthermore, the $F_{ST}$ distribution shifts to the right as collections diverge genetically. For example, in comparing Mexico and Thailand collections (Figure S6), the mean $F_{ST}$ (0.165 ± 0.130) is larger than the median (0.130), which in turn is larger than the mode (0.103). Furthermore, the curve underlying the lower 5% $F_{ST}$ values, and the curve underlying the upper 95% $F_{ST}$ values are asymmetrical in shape and distribution (Figure S6). Only when collections become extremely divergent (*e.g.*, Figure S8, Figure S9, Figure S10, and Figure S11) do the mean, median, and mode begin to converge. All of these inequalities arise because the genome-wide distribution of $F_{ST}$ values is not normal. Therefore, centrality measures fail to capture the genome-wide variation in $F_{ST}$, and summaries of genetic relationships based upon point estimates (mean, median, and mode) are misleading, and also miss the
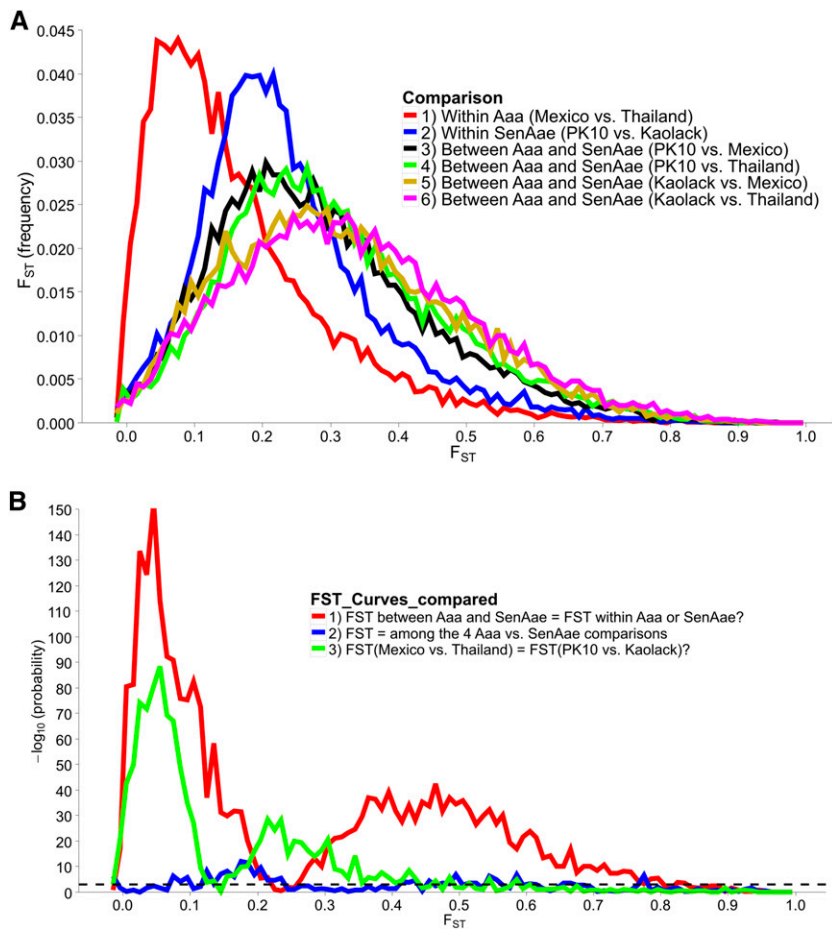
**Figure 2** (A) Distributions of pairwise $F_{ST}$ values among all six comparisons of the four collection sites. $F_{ST}$ values between *Aaa* collections (red line) are skewed toward smaller values, while $F_{ST}$ values between *SenAae* collections (blue line) were generally larger, and shifted to the right. All four $F_{ST}$ distributions between *Aaa* and *SenAae* [*i.e.*, PK10 *vs.* Mexico (black), PK10 *vs.* Thailand (green), Kaolack *vs.* Mexico (tan), and Kaolack *vs.* Thailand (magenta)] had long right-hand tails, indicating larger $F_{ST}$ values, and fewer small $F_{ST}$ values. (B) $F_{ST}$ distributions were compared with one another using a heterogeneity $\chi^2$ test of the number of genes in each 0.01 bin along the abscissa in the histogram. The red line is a comparison of $F_{ST}$ between *Aaa* [red in (A)] and between *SenAae* [blue in (A)] with the four $F_{ST}$ distributions curves [black, green, tan and magenta in (A)] between *Aaa* and *SenAae*. The left-hand portion of the red curve indicates a large and significant excess of small $F_{ST}$ values within the *Aaa* collection, and within *SenAae* collections as compared with $F_{ST}$ values between *Aaa* and *SenAae*. The right-hand region of the red curve indicates a large and significant deficiency of large $F_{ST}$ values within *Aaa* collections, and within *SenAae* collections. The blue line in (B) is a comparison of the four *Aaa* vs. *SenAae* $F_{ST}$ distributions (*i.e.*, PK10 *vs.* Mexico, PK10 *vs.* Thailand, Kaolack *vs.* Mexico, and Kaolack *vs.* Thailand). LOD values are never large, indicating that the four curves are similar. The green line is a comparison of the $F_{ST}$ distributions within *Aaa* [red in (A)], and within *SenAae* [red in (A)]. The left-hand portion of the green curve indicates a large and significant excess of small $F_{ST}$ values within *Aaa* collection as compared within *SenAae* $F_{ST}$ values. The right hand region of the green curve indicates a large and significant deficiency of large $F_{ST}$ values within *Aaa* collections.

point of a population genomics analysis, which is to examine and compare patterns of variation across the entire genome.

## Genetic differences between sexes in Ae. aegypti collections

Our previous observations of chromosomal rearrangements on chromosome 1, along with the presence of the sex-determining locus (SDL) on chromosome 1 (Dickson *et al.* 2016), lead us to explore possible sex-specific chromosomal differences between dual replicate pools of males and females in the deep sequencing data in PK10 and Thailand. Roughly the same percentage of SNPS were polymorphic in the sexes in PK10 (6.74% in females and 6.85% in males) and Thailand (4.13% in females and 4.41% in males) (Table 1). $H_{exp}$ in PK10 females was greater than $H_{exp}$ in males at 54% of genes. $H_{exp}$ in Thailand females was less than $H_{exp}$ in males in 45% of genes.

In addition to the $F_{ST}$ sampling distributions between sexes (Figure S4 and Figure S5), $F_{ST}$ values between sexes in both the PK10 and Thailand collections were mapped for each gene across the three chromosomes (Figure 3). Visual comparison of $F_{ST}$ values on chromosome 1 in the PK10 and Thailand strains indicates that sex-specific differences were generally larger on chromosome 1 of the PK10 strain. Comparison of $F_{ST}$ values on chromosome 1 between PK10 and Thailand were highly significant (pairwise-$t$ = 26.9, d.f. = 1934, $p$-value = $1.15 \times 10^{-135}$), and PK10 $F_{ST}$ values were greater than

Thailand $F_{ST}$ values in 1387 of the 1935 genes compared (71.6%). These differences were distributed uniformly along chromosome 1, and did not appear to pile up in one or a few regions, as would be expected if the differences were associated with a single sex-determining locus. However, $\sim$50% of the *Ae. aegypti* genome remains unannotated, and recent work on mapping with the same markers used in this study revealed many cases where genes were misassembled (Juneja *et al.* 2014). Thus, it is possible that large $F_{ST}$ values associated with markers on chromosomes 2 and 3 could actually map to chromosome 1. On chromosome 2, $F_{ST}$ values were also significantly greater in PK10 ($t$ = 12.6, d.f. = 3391, $p$-value = $1.31 \times 10^{-35}$), and PK10 $F_{ST}$ values were greater than Thailand $F_{ST}$ values in 2110 of 3392 genes compared (62.2%). In contrast, on chromosome 3, $F_{ST}$ values were actually significantly lower between PK10 sexes ($t$ = $-6.50$, d.f. = 2354, $p$-value = $9.64 \times 10^{-11}$), and PK10 $F_{ST}$ values were greater than Thailand $F_{ST}$ values in 1050 of 2355 genes compared (44.6%).

Figure 4 suggests greater differences in SNP frequencies between sexes on chromosome 1 within PK10 than among genes on all three chromosomes within Thailand. To test this further, the distribution of $F_{ST}$ values between sexes was plotted for all three chromosomes (Figure 4). The distribution of $F_{ST}$ values are very similar among all three chromosomes in the Thailand collection (Figure 4A). All three distributions were compared against one another using a 3 × 2 heterogeneity $\chi^2$ test of the number of genes in each 0.01 bin. None of the values

| Collections Compared | No. Genes | No. SNPs | Mean | Median | Mode | SD | Low 5% | High 95% |
|---|---|---|---|---|---|---|---|---|
| Thai sexes | 15,345 | 1,109,671 | 0.015 | 0.007 | −0.003 | 0.032 | −0.013 | 0.068 |
| PK10 sexes | 15,442 | 1,882,229 | 0.030 | 0.014 | 0.000 | 0.047 | −0.009 | 0.129 |
| Mexico/Thailand | 15,435 | 1,165,626 | 0.165 | 0.130 | 0.103 | 0.133 | 0.017 | 0.429 |
| PK10/Kaolack | 15,473 | 1,721,304 | 0.242 | 0.218 | 0.172 | 0.133 | 0.067 | 0.502 |
| PK10/Mexico | 15,453 | 1,266,724 | 0.289 | 0.267 | 0.185 | 0.155 | 0.074 | 0.582 |
| PK10/Thailand | 15,467 | 1,507,065 | 0.309 | 0.285 | 0.240 | 0.158 | 0.085 | 0.610 |
| Kaolack/Mexico | 15,452 | 1,665,639 | 0.316 | 0.297 | 0.298 | 0.168 | 0.073 | 0.616 |
| Kaolack/Thailand | 15,478 | 1,665,753 | 0.334 | 0.320 | 0.277 | 0.173 | 0.075 | 0.650 |

crossed the LOD = 3.0 threshold. In contrast, the distribution of $F_{ST}$ values are different between sexes on all three chromosomes in PK10 (Figure 4B). Specifically, the $F_{ST}$ distribution for chromosome 1 is strongly shifted to the right, while, for chromosome 2, the distribution is slightly shifted to the right, and for chromosome 3, the distribution is positively skewed. $F_{ST}$ from 0 to 0.2 cross the LOD = 3.0 threshold. Thus, while markers on all three chromosomes are at similar frequencies in the two sexes in the Thailand *Aaa* populations, they are at strikingly different frequencies between PK10 sexes.

Figure 5 repeats the analyses on Figure 3 and Figure 4, but compares $F_{ST}$ values according to the gene region in which the SNP occurred including codon positions 1, 2, and 3, introns, 5′UTR, and 600 bp on the 5′NTR, 3′UTR, and 600 bp on the 3′NTR. When comparing sexes in Thailand, $F_{ST}$ values were greatest among SNPs in the 5′ and 3′ UTRs, and least among SNPs in the second codon position. Values were virtually equivalent on all three chromosomes, and including unmapped regions. $F_{ST}$ values were also compared among the four types of mutations (transition, transversion, insertion, and deletion) giving rise to the SNP. $F_{ST}$ values were equivalent between transitions and transversions, slightly lower for insertions, and lowest for deletions. Again, $F_{ST}$ values were virtually equivalent on all three chromosomes and in unmapped regions. $F_{ST}$ values were finally compared among SNPs occurring in codons that encoded a silent or replacement substitution. While $F_{ST}$ values among silent SNPs were higher than among replacement SNPs, $F_{ST}$ values were virtually equivalent among all three chromosomes and including unmapped regions.

In contrast, when comparing sexes in PK10, $F_{ST}$ values between sexes were greatest among SNPs in the third codon position and in the 5′ and 3′ UTRs, and least among SNPs in the second codon position and introns. Values differed greatly among chromosomes and unmapped regions. The largest $F_{ST}$ values occurred on chromosome 1 among all eight gene regions. The second largest $F_{ST}$ values occurred on chromosome 2 among all eight gene regions, while the smallest $F_{ST}$ values occurred on chromosome 3. The unmapped markers represent an average of all three chromosomes. Similar trends are evident among the transition, transversion, insertion, and deletion SNPs, and among SNPs encoding silent and replacement substitutions.

## DISCUSSION

In keeping with earlier studies, the percentage of polymorphic SNPs and $H_{exp}$ was greatest in the African *Ae. aegypti* collections. This is consistent with the idea that *Ae. aegypti* arose as a species somewhere in the African continent (Tabachnick 1991, 1979). A total of 490 genes occurred in the lower 1% tail of the four $H_{exp}$ distributions, and most (77% = 376/490) occurred in only one collection, while only five genes (1%) had low $H_{exp}$ in all collections (Figure S2). In contrast, 324 genes occurred in the upper 1% tail of the four $H_{exp}$ distributions, and a smaller percentage (49% = 160/324) occurred in only one collection,

while 44 genes (13.8%) were common in all four collections (Figure S3). Many of the genes in Table S1, have hundreds or thousands of SNPs. This probably reflects a pattern of neutral evolution, wherein a forward mutation rate generates new SNPs independently among genes and among SNPs within genes. In the context of this project it would be assumed that a mutation (SNP) arose only once in single collections (infinite alleles model), and drift would cause the majority of these to become extinct in a few generations, depending upon the effective population sizes. In contrast, the genes that occur in common among the four collections are probably maintained at a low $H_{exp}$ by purifying selection. For example, 24 of the low $H_{exp}$ genes encode rRNA genes that would be expected to be under strong purifying selection, and homogeneous due to the molecular drive operating in the rDNA cistron (Table S2).

By comparison, the 324 genes that had a high $H_{exp}$ (Table S3) do not follow a pattern consistent with neutral evolution, and, instead, are either not constrained by purifying selection, or are subject to some form of diversifying selection. Curiously 35.4% (115/325) of these encode either tRNAS(86), tRNA pseudogenes(19), or, while annotated as tRNAs in VectorBase, 10 could not be confirmed with tRNAscan-SE/ (Schattner *et al.* 2005). Forty-four high $H_{exp}$ genes (14%) occurred in common among all collections (Figure S2). Half of these 44 genes coded for noncoding RNAs (ncRNAs), 21 of which were tRNAs, and one was the U2 spliceosomal RNA gene (AAEL017645). This suggests a common lack of purifying selection, or diversifying selection, acting upon these genes in all collections.

*Ae. aegypti* has 962 annotated nuclear-encoded transfer RNA (tRNA) genes arranged in 61 genomic clusters, with 111 predicted as pseudogenes (Behura and Severson 2011; Lawson *et al.* 2007; Schattner *et al.* 2005). Members of the tRNA functional class were present in both the high $H_{exp}$ (116 tRNAs) and low $H_{exp}$ (22 tRNAs) (Table S1 and Table S2). The majority of low $H_{exp}$ occurred in *SenAae* (Table S2), while the highest $H_{exp}$ occurred in all collections. Of these, 62% had anticodons predicted to have preferred codon usage (Behura and Severson 2011). Curiously, 32% (7/22) of the low $H_{exp}$, and 19% (19/113) of high $H_{exp}$ tRNAs, were predicted to be pseudogenes (Schattner *et al.* 2005).

Why do aedine mosquitoes have many tRNA isoacceptor gene copies? Moreover, why is a subset of pseudogenes highly conserved? Varied codon usage, which would partially contribute to multiple tRNA copy numbers, has been suggested to be pertinent to specific metabolic pathways (Behura and Severson 2011), and could affect stage-specific expression patterns. Moreover, tRNAs also participate in regulatory activity, as tRNA fragments have been found to regulate sperm maturation in mammals (Sharma *et al.* 2016). Supportive evidence in aedine mosquitoes has also been described; for example, there is evidence for general RNA transfer in seminal fluid (Alfonso-Parra *et al.* 2016). Moreover, tRNA fragments (15–30 nt) have been isolated from small
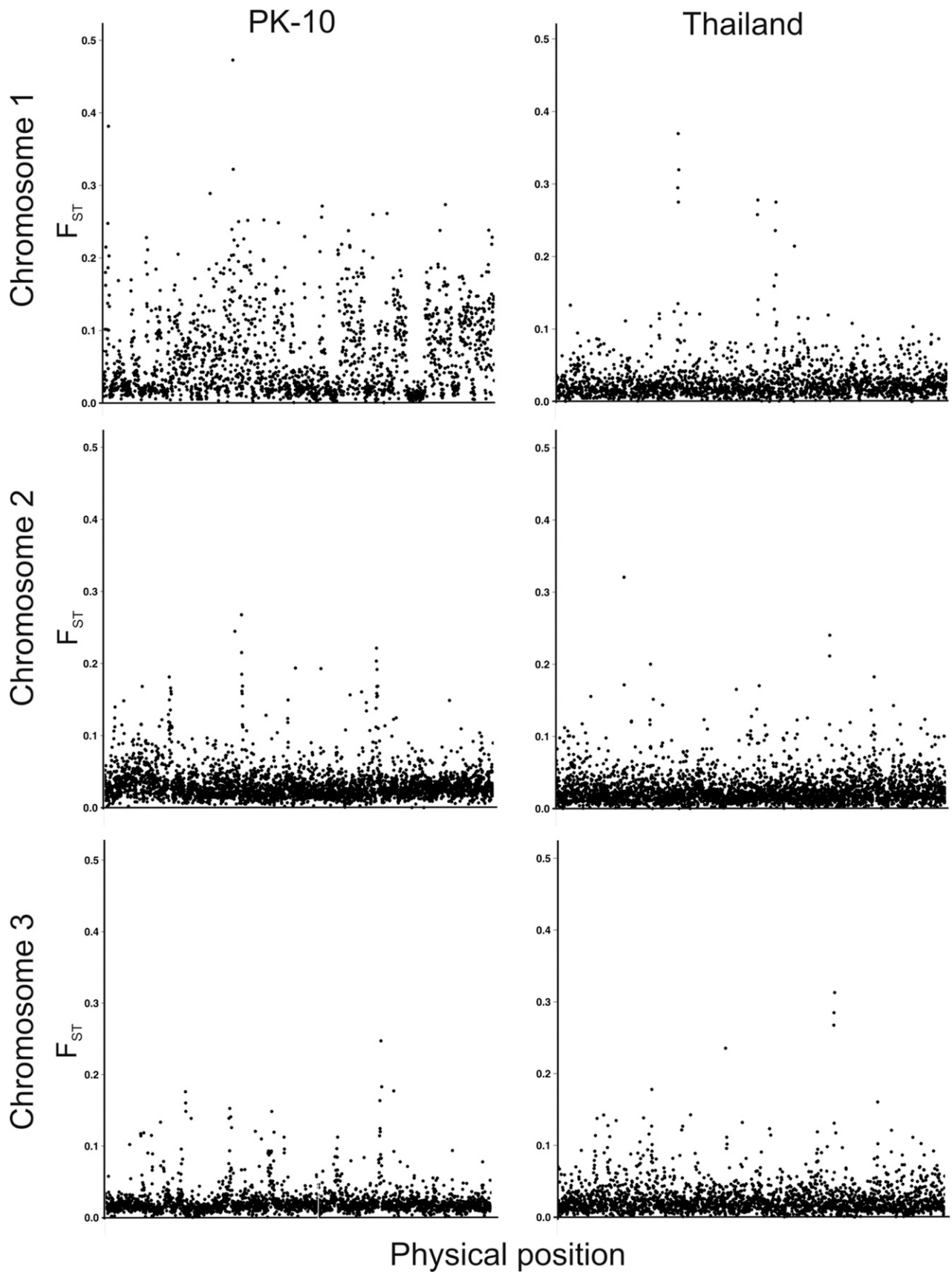
**Figure 3** Comparison of $F_{ST}$ values between sexes for genes that have been mapped across the three chromosomes in the PK10 (*SenAae*) and Thailand (*Aaa*) collections.
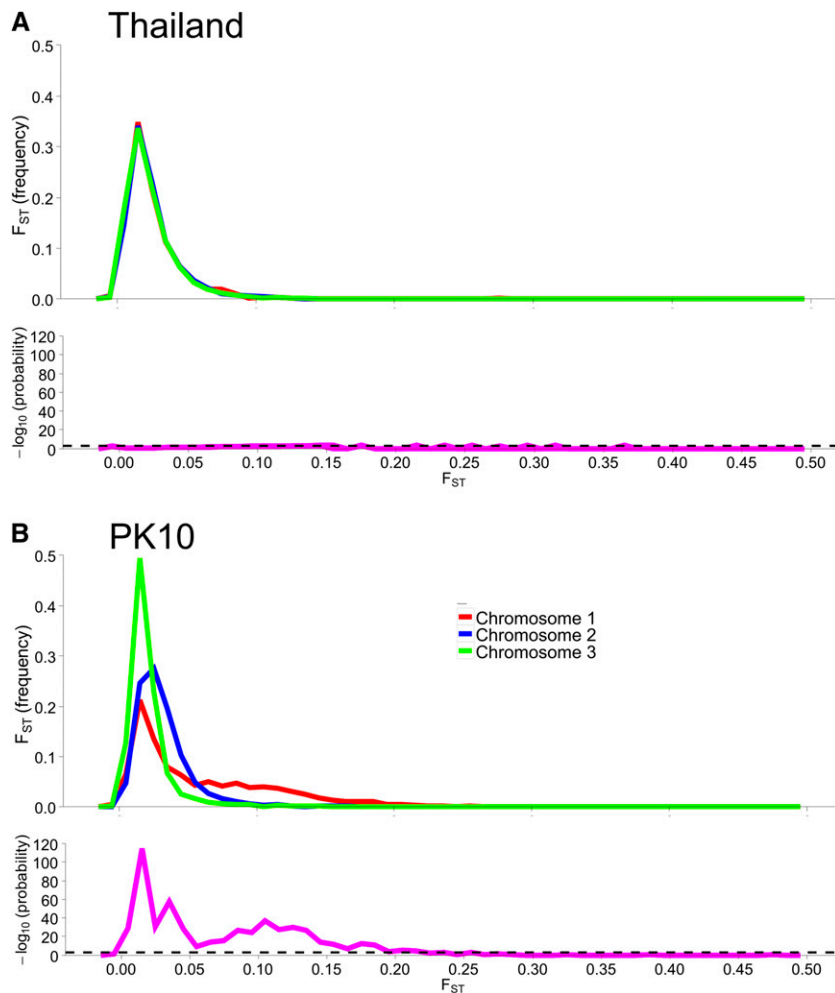
**A** Thailand

**B** PK10

- Chromosome 1
- Chromosome 2
- Chromosome 3

**Figure 4** Distribution of $F_{ST}$ and LOD values between sexes plotted on each chromosome. $F_{ST}$ distributions were compared with one another using a heterogeneity $\chi^2$ test of the number of genes in each 0.01 bin along the abscissa in the histogram. (A) $F_{ST}$ distributions are very similar among all three chromosomes in the Thailand collection. The magenta line indicates the LOD value across the chromosomes, while the black dashed line is the LOD = 3 cutoff. LOD values never exceed the LOD = 3 cutoff. (B) $F_{ST}$ distributions differ among all three chromosomes in PK10 (*SenAae*). LOD values across the chromosomes exceed the LOD = 3 cutoff across all three chromosomes.

RNA (sRNA) sequencing pools (Hess *et al.* 2011). Although the mechanistic details of possible tRNA regulation are yet to be elucidated in mosquitoes, these studies, coupled with the presence of over 100 tRNA genes showing high $H_{exp}$ values, is consistent with the hypothesis that tRNA polymorphisms may arise during speciation.

$F_{ST}$ differences were most pronounced between the two *SenAae* collections, and between *SenAae* and *Aaa*. $F_{ST}$ values were ranked from smallest to largest, and genes with the lowest 1% of $F_{ST}$ are listed in Table S3, and those with the highest 1% of $F_{ST}$ appear in Table S4. Fifty genes occurred within the lowest 1%. Of these, none occurred only once or twice, 29 occurred three times, none occurred four times, 15 occurred in five comparisons, and six occurred in all comparisons, suggesting these are highly conserved across populations. AAEL017730 was predicted to encode a novel miRNA, but was not similar to any known hexapod miRNAs (Griffiths-Jones *et al.* 2008). AAEL006000-RC is predicted to encode cytochrome *c* oxidase polypeptide 7A1—one of the nuclear-coded polypeptide chains of cytochrome *c* oxidase, the terminal oxidase in mitochondrial electron transport. AAEL008026-RA encodes an MADF (myb/SANT-like domain in Adf-1), where the SANT domain (acronym for Swi3, Ada2, N-Cor, TFIIB) is a highly conserved motif. In *Drosophila*, Adf-1 encodes a transcription factor that binds the promoters of a diverse group of genes (England *et al.* 1992). AAEL016955-RA encodes a tRNA-Asp, while AAEL006284-RA and AAEL006943-RA encode hypothetical proteins.

Similarly, the upper 1% of $F_{ST}$ values were interrogated for each of the six pairwise comparisons (Table S4). Of 48 genes, only CYP6P12

(AAEL014891) occurred in all comparisons. The remaining genes all appeared three times. CYP6P12 is a major insecticide detoxification gene (Saavedra-Rodriguez *et al.* 2014). Several other genes known to be involved in insecticide resistance occurred in the upper 1% of $F_{ST}$ values. AAEL012491-RA encodes a different CYP6P12, while four epsilon glutathione-S-transferases (2, 4, 6, and 7) appear in this list. This in consistent with an hypothesis that insecticidal pressure [very high in Mexico (Flores *et al.* 2013) and Thailand (Bosio *et al.* 2005), very low or nonexistent in Senegal (Sylla *et al.* 2013)] has promoted directional selection, and caused large differences in the frequencies of SNPs. Reductions in heterozygosity usually accompany directional selection with insecticides. The average $H_{exp}$ among the four epsilon GSTs was PK10 (0.166), Kaolack (0.210), Mexico (0.061), and Thailand (0.023), but $H_{exp}$ among the CYP6P12 was: PK10 (0.130), Kaolack (0.140), Mexico (0.166), and Thailand (0.038).

In the process of completing this study, we learned an important technical aspect of next generation sequencing (NGS) when applied to species (such as *Ae. aegypti*) with large genomes that contain a high percentage of repetitive DNA (Warren and Crampton 1991). Early on in this investigation, we tried to perform the same NGS analyses but without exon enrichment. From 23 to 37 $\times$ 10$^6$ sequences were recovered in each of the 12 libraries; however, codon regions were detected in <5% of all recovered sequences. Reassociation kinetics of the *Ae. aegypti* genome (Warren and Crampton 1991) indicated that 40% of the genome consisted of fold-back, highly repetitive, and middle repetitive DNA. Perusal of VectorBase indicates that much of the
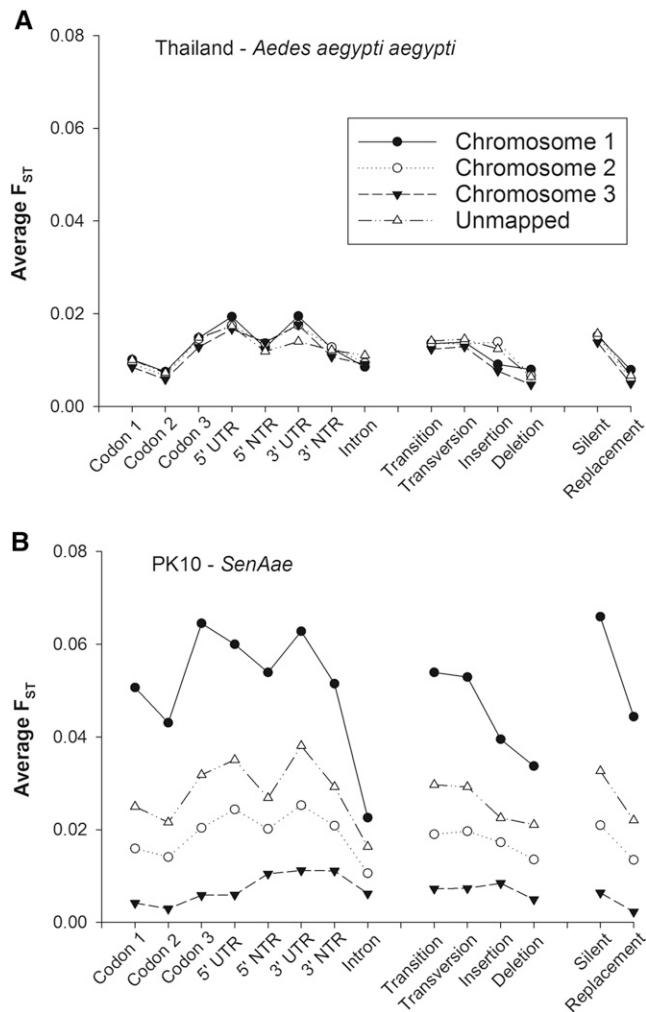
**Figure 5** Average sex-specific $F_{ST}$ values for SNPs in different regions of a gene. Values were plotted by chromosome in (A) Thailand (*Aaa*) females *vs.* males, and (B) PK10 (*SenAae*) females *vs.* males.

remaining "unique" DNA consists of transposable elements. Most importantly, analysis of these unenriched datasets provided no resolution of subspecies (Figure 1 and Figure 2), nor did it identify many differences between sexes (Figure 4 and Figure 5).

The results of this study suggest that genes from throughout the genome are similar between *Aaa* from two geographically distinct locations. *SenAae* genomes also differed from one another to a large extent in the present study, but, more notably, differ greatly between *Aaa* and *SenAae*. We have yet to determine how *SenAae* are related to East African *Aaf*, but this same pattern was evident in the recent study of worldwide collections of *Ae. aegypti* (*s.l.*) by Evans *et al.* (2015) using a 50,000 SNP-Chip. In particular, their principal components analysis (PCA) of collections from Africa, Asia and the Pacific, and the Americas (Figure 3B in Evans *et al.* 2015) showed that the five Americas collections formed a very tight overlapping cluster distinct from, but on the same side of the PCA, as the four Asia and the Pacific collections, which formed three distinct clusters. In contrast, African collections (including two *SenAae* but neither PK10 nor Kaolack) formed very large clusters. One apparent difference between the PCA and results in the present study is that Thailand and Mexico collections in the PCA appear to be more distinct than in our comparison of $F_{ST}$ distributions. However, our results do not necessarily contradict those of Evans *et al.*

(2015). First, collection sites in Thailand differed between the two studies. Second, referring to Panel B of the PCA in Figure 3 of Evans *et al.* (2015), the Thailand (purple squares) and Tapachula (gold diamonds) collections are extremely close along the PC1 axis that accounts for 13.2% of the variation in SNP frequencies. Thailand and Tapachula only separate along PC2, which accounts for only 4.5% of the variation. Third, PC1+PC2 account for only 17.7% of the overall variation, indicating that 82% of the variation in SNP frequencies is unexplained by the first two PCs. Fourth, PCA is an analysis of individuals, while our analysis of $F_{ST}$ is based on collections of individuals. Evans *et al.* (2015) does not indicate how much variation exists among individuals in a collection (this appears for our data in Figure 1). Could within-collection variation among individuals account for a large part of the 82% of unexplained variance? Finally, Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10, and Figure S11 show that summary of genetic relationships based upon point estimates (mean, median, mode, and variance) is misleading.

These patterns in genic diversity correlate with the reproductive isolation patterns noted in our earlier study (Dickson *et al.* 2016). In that study, attempts to construct $F_1$ intercross families between *Aaa* laboratory strains and *SenAae* failed due to poor $F_1$ oviposition and low $F_2$ egg-to-adult survival. Backcrosses were performed to test for postzygotic isolation patterns consistent with Haldane's rule modified for species, like *Aedes*, that have an autosomal SDL (Presgraves and Orr 1998). However, these crosses were inconclusive with regards to Haldane's rule. Instead, basic cytogenetic analyses and FISH experiments revealed many rearrangements between *SenAae* and *Aaa*. Linkage analysis of the SDL, and the white-eye locus, identified a likely chromosomal rearrangement on chromosome 1. Another rearrangement was close to the centromere on the p arm of chromosome 2. Two overlapping pericentric inversions were found on chromosome 3, or an insertion of a large fragment into the 3q arm. The reproductive incompatibility observed within *SenAae*, and between *SenAae* and *Aaa*, may be generally associated with chromosome rearrangements on all three chromosomes, and specifically caused by pericentric inversions on chromosomes 2 and 3. Chromosomal rearrangements have frequently been associated with speciation in taxa, including mosquitoes (Kitzmiller 1963). Pericentric inversions are only rarely detected within animal species, but can be abundant among species (Coyne and Orr 2004; Orr and Coyne 1989). Acentric fragments and dicentric "bridge" chromosomes arise when recombination occurs in parents that are heterozygous for pericentric inversions. These, in turn, yield aneuploid gametes and inviable zygotes.

We hypothesize that cryptic *Ae. aegypti* taxa exist in West Africa, or that new subspecies of *Ae. aegypti* are arising in West Africa. We propose that the large genic differences between *Aaa* and *SenAae* documented in the present study have accumulated over a long period of time, possibly following the occurrence of chromosome rearrangements in allopatric populations that subsequently caused the taxa to become reproductively isolated when populations became sympatric. It may be that PK10 represents a collection of at least two fully, or partially, reproductively isolated taxa. If so, it is curious that the largest differences between PK10 females and males occur on the sex autosome 1 (Figure 3). Sex determination is governed by a dominant male-determining "M" allele in the sex determination locus on chromosome 1 in *Ae. aegypti*. The SDL gene "Nix" has recently been identified. *Nix* exhibits persistent M linkage and early embryonic expression (Hall *et al.* 2015); however differences between PK10 females and males were not limited to one region of chromosome 1. It has been proposed that sexual selection may drive the more rapid evolution of male-expressed genes (Wu and Davis 1993; Wu *et al.* 1996; Coulthart and

Singh 1988), as well as variation in expression of male-expressed genes (Harrison *et al.* 2015). More rapid male evolution would also explain Haldane's rule because hybrid male sterility would occur before hybrid female sterility. Experiments in *Drosophila* suggest that faster male evolution causes Haldane's rule for sterility (Hollocher and Wu 1996; Harrison *et al.* 2015; True *et al.* 1996). Unfortunately there is no information to suggest that male-expressed genes occur primarily on chromosome 1 or within the region containing *Nix* (Hall *et al.* 2015).

Herein, we have demonstrated large amounts of genic differentiation between *SenAae* and *Aaa* populations outside Africa, and earlier have provided evidence for chromosome rearrangements (Dickson *et al.* 2016). These observations are important because there is currently a great deal of interest in manipulating or eliminating *Ae. aegypti* populations on a global scale (Zhang *et al.* 2015; Rasic *et al.* 2014). The presence of cryptic *Ae. aegypti* species could negatively impact this effort. There is a great need to define the geographic distributions of chromosome rearrangements that are presumably associated with these cryptic *Ae. aegypti* taxa or subspecies.

## LITERATURE CITED

Alfonso-Parra, C., Y. H. Ahmed-Braimah, E. C. Degner, F. W. Avila, S. M. Villarreal *et al.*, 2016   Mating-induced transcriptome changes in the reproductive tract of female *Aedes aegypti*. PLoS Negl. Trop. Dis. 10(2): e0004451.

Beasley, D. W. C., A. J. McAuley, and D. A. Bente, 2015   Yellow fever virus: genetic and phenotypic diversity and implications for detection, prevention and therapy. Antiviral Res. 115: 48–70.

Behura, S. K., and D. W. Severson, 2011   Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in *Aedes aegypti* and *Anopheles gambiae*. Insect Mol. Biol. 20(2): 177–187.

Black, W. C., and N. M. DuTeau, 1997   RAPD-PCR and SSCP analysis for insect population genetic studies, pp. 361–373 in *The Molecular Biology of Insect Disease Vectors: A Methods Manual*, edited by Crampton, J., C. B. Beard, and C. Louis. Chapman and Hall, New York.

Bosio, C. F., L. C. Harrington, J. W. Jones, R. Sithiprasasna, D. E. Norris *et al.*, 2005   Genetic structure of *Aedes aegypti* populations in Thailand using mitochondrial DNA. Am. J. Trop. Med. Hyg. 72(4): 434–442.

Brown, J. E., C. S. McBride, P. Johnson, S. Ritchie, C. Paupy *et al.*, 2011   Worldwide patterns of genetic differentiation imply multiple 'domestications' of *Aedes aegypti*, a major vector of human diseases. Proc. Biol. Sci. 278: 2446–2454.

Brown, J. E., B. R. Evans, W. Zheng, V. Obas, L. Barrera-Martinez *et al.*, 2014   Human impacts have shaped historical and recent evolution in *Aedes Aegypti*, the dengue and yellow fever mosquito. Evolution 68(2): 514–525.

Coulthart, M. B., and R. S. Singh, 1988   High-level of divergence of male-reproductive-tract proteins, between *Drosophila melanogaster* and its sibling species, *Drosophila simulans*. Mol. Biol. Evol. 5(2): 182–191.

Coyne, J.A., and H. A. Orr, 2004   *Speciation*. Sinauer Associates, Sunderland, MA.

Dickson, L. B., M. V. Sharakhova, V. A. Timoshevskiy, K. L. Fleming, A. Caspary *et al.*, 2016   Reproductive incompatibility involving senegalese *Aedes aegypti* (L.) is associated with chromosome rearrangements. PLoS Negl. Trop. Dis. 10(4): e0004626.

England, B. P., A. Admon, and R. Tjian, 1992   Cloning of *Drosophila* transcription factor Adf-1 reveals homology to Myb oncoproteins. Proc. Natl. Acad. Sci. USA 89(2): 683–687.

Evans, B.R., A. Gloria-Soria, L. Hou, C. McBride, M. Bonizzoni *et al.*, 2015   A multipurpose, high-throughput single-nucleotide polymorphism chip for the dengue and yellow fever mosquito, *Aedes aegypti*. G3 (Bethesda) 5(5): 711–718.

Flores, A. E., G. Ponce, B. G. Silva, S. M. Gutierrez, C. Bobadilla *et al.*, 2013   Wide spread cross resistance to pyrethroids in *Aedes aegypti* (Diptera: Culicidae) from Veracruz state Mexico. J. Econ. Entomol. 106(2): 959–969.

Fumagalli, M., F. G. Vieira, T. S. Korneliussen, T. Linderoth, E. Huerta-Sanchez *et al.*, 2013   Quantifying population genetic differentiation from next-generation sequencing data. Genetics 195(3): 979.

Griffiths-Jones, S., H. K. Saini, S. van Dongen, and A. J. Enright, 2008   miRBase: tools for microRNA genomics. Nucleic Acids Res. 36: D154–D158.

Guzman, M. G., and E. Harris, 2015   Dengue. Lancet 385(9966): 453–465.

Haldane, J. B. S., 1922   Sex ratio and unisexual sterility in hybrid animals. J. Genet. 12(2): 101–109.

Hall, A. B., S. Basu, X. F. Jiang, Y. M. Qi, V. A. Timoshevskiy *et al.*, 2015   A male-determining factor in the mosquito *Aedes aegypti*. Science 348 (6240): 1268–1270.

Harrison, P. W., A. E. Wright, F. Zimmer, R. Dean, S. H. Montgomery *et al.*, 2015   Sexual selection drives evolution and rapid turnover of male gene expression. Proc. Natl. Acad. Sci. USA 112(14): 4393–4398.

Heaton, J. H., W. M. Dlakic, M. Dlakic, and T. D. Gelehrter, 2001   Identification and cDNA cloning of a novel RNA-binding protein that interacts with the cyclic nucleotide-responsive sequence in the type-1 plasminogen activator inhibitor mRNA. J. Biol. Chem. 276(5): 3341–3347.

Hess, A. M., A. N. Prasad, A. Ptitsyn, G. D. Ebel, K. E. Olson *et al.*, 2011   Small RNA profiling of Dengue virus–mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. BMC Microbiol. 11: 45.

Holcroft, C. E., W. D. Jackson, W. H. Lin, K. Bassiri, R. A. Baines *et al.*, 2013   Innexins Ogre and Inx2 are required in glial cells for normal postembryonic development of the *Drosophila* central nervous system. J. Cell Sci. 126(17): 3823–3834.

Hollocher, H., and C. I. Wu, 1996   The genetics of reproductive isolation in the *Drosophila simulans* clade: X *vs.* autosomal effects and male *vs.* female effects. Genetics 143(3): 1243–1255.

Huber, K., Y. Ba, I. Dia, C. Mathiot, A. A. Sall *et al.*, 2008   *Aedes aegypti* in Senegal: genetic diversity and genetic structure of domestic and sylvatic populations. Am. J. Trop. Med. Hyg. 79(2): 218–229.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson *et al.*, 2015   Orchestrating high-throughput genomic analysis with bioconductor. Nat. Methods 12(2): 115–121.

Juneja, P., J. Osei-Poku, Y. S. Ho, C. V. Ariani, W. J. Palmer *et al.*, 2014   Assembly of the genome of the disease vector *Aedes aegypti* onto a genetic linkage map allows mapping of genes affecting disease transmission. PLoS Negl. Trop. Dis. 8(1): e2652.

Juneja, P., C. V. Ariani, Y. S. Ho, J. Akorli, W. J. Palmer *et al.*, 2015   Exome and transcriptome sequencing of *Aedes aegypti* identifies a locus that confers resistance to *Brugia malayi* and alters the immune response. PLoS Pathog. 11(3): e1004765.

Kitzmiller, J. B., 1963   Mosquito cytogenetics. A review of the literature, 1953–62. Bull. World Health Organ. 29: 345–355.

Koboldt, D. C., Q. Y. Zhang, D. E. Larson, D. Shen, M. D. McLellan *et al.*, 2012   VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22(3): 568–576.

Lawson, D., P. Arensburger, P. Atkinson, N. J. Besansky, R. V. Bruggner *et al.*, 2007   VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res. 35: D503–D505.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009   The sequence alignment/map format and SAMtools. Bioinformatics 25(16): 2078–2079.

Li, M. I., P. S. J. Wong, L. C. Ng, and C. H. Tan, 2012   Oral susceptibility of Singapore *Aedes* (*stegomyia*) *aegypti* (Linnaeus) to Zika virus. PLoS Negl. Trop. Dis. 6(8): e1792.

Lo Presti, A., E. Cella, S. Angeletti, and M. Ciccozzi, 2016   Molecular epidemiology, evolution and phylogeny of Chikungunya virus: an updating review. Infect. Genet. Evol. 41: 270–278.

Mattingly, P. F., 1957 Genetical aspects of the *Aedes aegypti* problem. I. taxonom: and bionomics. Ann. Trop. Med. Parasitol. 51(4): 392–408.

McClelland, G.A., 1974 A worldwide survey of variation in scale pattern of the abdominal tergum of *Aedes aegypti* (L.) (Diptera : Culicidae). Trans. R. Ent. Soc. Lond. 126: 239–259.

Moore, D. F., 1979 Hybridization and mating-behavior in *Aedes aegypti* (Diptera, Culicidae). J. Med. Entomol. 16(3): 223–226.

Moore, M., M. Sylla, L. Goss, M. W. Burugu, R. Sang *et al.*, 2013 Dual African origins of global *Aedes aegypti* s.l. Populations revealed by mitochondrial DNA. PLoS Negl. Trop. Dis. 7(4): e2175.

Oliveros, J. C., 2010 *VENNY. An Interactive Tool for Comparing Lists with Venn Diagrams.* BioinfoGP, CNB-CSIC, Spain.

Orr, H. A., and J. A. Coyne, 1989 The genetics of postzygotic isolation in the *Drosophila virilis* group. Genetics 121(3): 527–537.

Paupy, C., C. Brengues, B. Kamgang, J. P. Herve, D. Fontenille *et al.*, 2008 Gene flow between domestic and sylvan populations of *Aedes aegypti* (Diptera: Culicidae) in North Cameroon. J. Med. Entomol. 45(3): 391–400.

Paupy, C., C. Brengues, O. Ndiath, C. Toty, J. P. Herve *et al.*, 2010 Morphological and genetic variability within *Aedes aegypti* in Niakhar, Senegal. Infect. Genet. Evol. 10(4): 473–480.

Presgraves, D. C., and H. A. Orr, 1998 Haldane's rule in taxa lacking a hemizygous X. Science 282(5390): 952–954.

R Development Core Team, 2014 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

Rasic, G., N. M. Endersby, C. Williams, and A. A. Hoffmann, 2014 Using *Wolbachia*-based release for suppression of *Aedes* mosquitoes: insights from genetic data and population simulations. Ecol. Appl. 24(5): 1226–1234.

Saavedra-Rodriguez, K., C. Strode, A. E. Flores, S. Garcia-Luna, G. Reyes-Solis *et al.*, 2014 Differential transcription profiles in *Aedes aegypti* detoxification genes after temephos selection. Insect Mol. Biol. 23(2): 199–215.

Schattner, P., A. N. Brooks, and T. M. Lowe, 2005 The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoR-NAs. Nucleic Acids Res. 33: W686–W689.

Scott, J. A., and G. A. H. McClelland, 1975 Electrophoretic differences between sympatric ecotypes. Nature 256(5516): 405–406.

Sharma, U., C. C. Conine, J. M. Shea, A. Boskovic, A. G. Derr *et al.*, 2016 Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science 351(6271): 391–396.

Sylla, M., C. Bosio, L. Urdaneta-Marquez, M. Ndiaye, and W. C. Black, 2009 Gene flow, subspecies composition, and dengue virus-2 susceptibility among *Aedes aegypti* collections in Senegal. PLoS Negl. Trop. Dis. 3(4): e408.

Sylla, M., M. Ndiaye, and W. C. Black, 2013 *Aedes* species in treeholes and fruit husks between dry and wet seasons in southeastern Senegal. J. Vector Ecol. 38(2): 237–244.

Tabachnick, W. J., 1991 Evolutionary genetics and arthropod-borne disease: the yellow fever mosquito. American Entomologist. 37: 14–26.

Tabachnick, W. J., and J. R. Powell, 1978 Genetic structure of East-African domestic populations of *Aedes aegypti*. Nature 272(5653): 535–537.

Tabachnick, W. J., L. E. Munstermann, and J. R. Powell, 1979 Genetic distinctness of sympatric forms of *Aedes aegypti* in East-Africa. Evolution 33(1): 287–295.

Timoshevskiy, V. A., N. A. Kinney, B. S. deBruyn, C. H. Mao, Z. J. Tu *et al.*, 2014 Genomic composition and evolution of *Aedes aegypti* chromosomes revealed by the analysis of physically mapped supercontigs. BMC Biol. 12: 27.

Trpis, M., and W. Hausermann, 1975 Demonstration of differential domesticity of *Aedes aegypti* (L.) (Diptera, Culicidae) in Africa by mark release recapture. Bull. Entomol. Res. 65(2): 199–208.

True, J. R., B. S. Weir, and C. C. Laurie, 1996 A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. Genetics 142(3): 819–837.

Warren, A. M., and J. M. Crampton, 1991 The *Aedes aegypti* genome—complexity and organization. Genet. Res. 58(3): 225–232.

Wu, C. I., and A. W. Davis, 1993 Evolution of postmating reproductive isolation—the composite nature of Haldane's rule and its genetic bases. Am. Nat. 142(2): 187–212.

Wu, C. I., N. A. Johnson, and M. F. Palopoli, 1996 Haldane's rule and its legacy: why are there so many sterile males? Trends Ecol. Evol. 11(7): 281–284.

Wu, T. D., and S. Nacu, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7): 873–881.

Zhang, D. J., X. Y. Zheng, Z. Y. Xi, K. Bourtzis, and J. R. L. Gilles, 2015 Combining the sterile insect technique with the incompatible insect technique: i–impact of *Wolbachia* infection on the fitness of triple- and double-infected strains of *Aedes albopictus*. PLoS One 10(4): e0121126.

*Communicating editor: R. Kulathinal*