Hidden features of the malaria vector mosquito, *Anopheles stephensi*, revealed by a high-quality reference genome

Mahul Chakraborty[1*], Arunachalam Ramaiah[1,2,3*], Adriana Adolfi[4], Paige Halas[4], Bhagyashree Kaduskar[2,3], Luna Thanh Ngo[1], Suvratha Jayaprasad[5], Kiran Paul[5], Saurabh Whadgar[5], Subhashini Srinivasan[3,5], Suresh Subramani[3,6,7], Ethan Bier[2,7], Anthony A. James[4,7,8], J.J. Emerson[1,9,#]

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA
[2]Section of Cell and Developmental Biology, University of California, San Diego, La Jolla, CA 92093-0335, USA
[3]Tata Institute for Genetics and Society, Center at inStem, Bangalore, KA 560065, India
[4]Department of Microbiology & Molecular Genetics, University of California, Irvine, CA 92697, USA
[5]Institute of Bioinformatics and Applied Biotechnology, Bangalore, KA 560100, India
[6]Section of Molecular Biology, University of California, San Diego, La Jolla, CA 92093-0322, USA
[7]Tata Institute for Genetics and Society, University of California, San Diego, La Jolla, CA 92093-0335, USA
[8]Department of Molecular Biology & Biochemistry, University of California, Irvine, CA 92697, USA
[9]Center for Complex Biological Systems, University of California, Irvine, CA 92697, USA

*These authors contributed equally to this work

[#]Correspondence to: J.J. Emerson, jje@uci.edu

ABSTRACT

Anopheles stephensi is a vector of urban malaria in Asia and has recently invaded eastern Africa. Genomic tools for studying its vectorial capacity and engineering genetic interventions depend on the accuracy, completeness, and contiguity of genome assemblies. A new reference-grade genome assembly reveals novel species-specific transposable element families and insertions in functional genetic elements. We rectify missing or incomplete genes, observe gene family expansions and correct ambiguous annotations, uncovering new candidate genes for insecticide resistance, immunity, reproduction, and blood metabolism. We identify 2.4 Mb of the Y-chromosome and seven new male-linked gene candidates. The Y-linked heterochromatin landscape reveals the role of long-terminal repeat retrotransposons in the evolution of this chromosome. A novel Y-linked putative transcription factor is expressed constitutively through male development and adulthood. This reference assembly will be foundational for comparative genomics, as well as for genetic and genomic studies of malaria control in this increasingly important vector species.

MAIN TEXT

Anopheles stephensi is the primary vector of urban malaria in the Indian subcontinent and the Middle East and an emerging malaria vector in Africa (Sharma 1999; Faulde, Rueda, and Khaireh 2014; Seyfarth et al. 2019). Genetic strategies (e.g. CRISPR gene drive) that suppress or modify vector populations are powerful means to curb malaria transmission (James 2005; Gantz et al. 2015; Kyrou et al. 2018; Champer et al. 2020). Success of these strategies depends on availability of accurate and complete genomic target sequences and variants segregating within them (James 2005; Unckless, Clark, and Messer 2017). However, functionally-important genetic elements and variants within them often consist of repetitive sequences that are either mis-assembled or completely missed in draft-quality genome assemblies (Chakraborty et al. 2018; Alkan, Sajjadian, and Eichler 2011; Treangen and Salzberg 2011). Despite being a pioneering model for transgenics and CRISPR gene drive in malaria vectors (Catteruccia et al. 2000; Gantz et al. 2015), the community studying *An. stephensi* still relies on a draft genome assembly that does not achieve the completeness and contiguity of reference-grade genomes (Jiang et al. 2014; Waterhouse et al. 2020). This limitation obscures genes and repetitive genetic elements that are potentially relevant for understanding parasite transmission or for managing vector populations (Matthews et al. 2018). We therefore generated a high-quality reference genome for a laboratory strain of this mosquito sampled from the Indian subcontinent (Nirmala et al. 2006) (supplementary Fig. S1) using deep coverage long reads plus Hi-C scaffolding and then annotated it by full length mRNA sequencing (Iso-Seq). These resources facilitate characterization of regions of the genome less accessible to previous efforts, including gene families associated with insecticide resistance, targets for gene-drive interventions, and recalcitrant regions of the genome rich in repeats, including the Y chromosome.

Anopheles stephensi has three major gene-rich chromosomes (X, 2, 3) and a gene poor, heterochromatic Y chromosome (Fig. 1A) (Sharakhova et al. 2011). In the

Table 1 - Summary of genome assembly and annotation statistics.

Genomic features	**Value**
Total length (bp)	250,632,892
Contig number	566
Contig N50 (bp)	38,117,870
Scaffold Number	560^
Scaffold N50 (bp)	88,747,609*
L50	2
GC content (%)	44.91%
Maximum scaffold length	93,706,523
Minimum scaffold length	1727
N's per 100 Kb	3.60
Alternative haplotypes, bp (# scaffolds)	7,171,998 (66)
Unclassified contigs, bp (# scaffolds)	35,861,427 (458)
Putative Y-chromosome, bp (# scaffolds)	2,431,719 (33)
3 major chromosomes X, 2 & 3, bp (# scaffolds)	205,167,748 (3)
Predicted Genes	14,966
Predicted Transcripts	16,559
5' UTR	9,791
3' UTR	9,290
tRNAs	503

^Except three major chromosomes, we kept others as contigs; *Scaffold N50 is the length of chr3.

new reference assembly, the major chromosomes are represented by just three sequences (scaffold N50 = 88.7 Mb; contig N50 = 38 Mb; N50 = 50% of the genome is

contained within sequence of this length or longer), making this assembly comparable to the *Drosophila melanogaster* reference assembly (Hoskins et al. 2015), widely

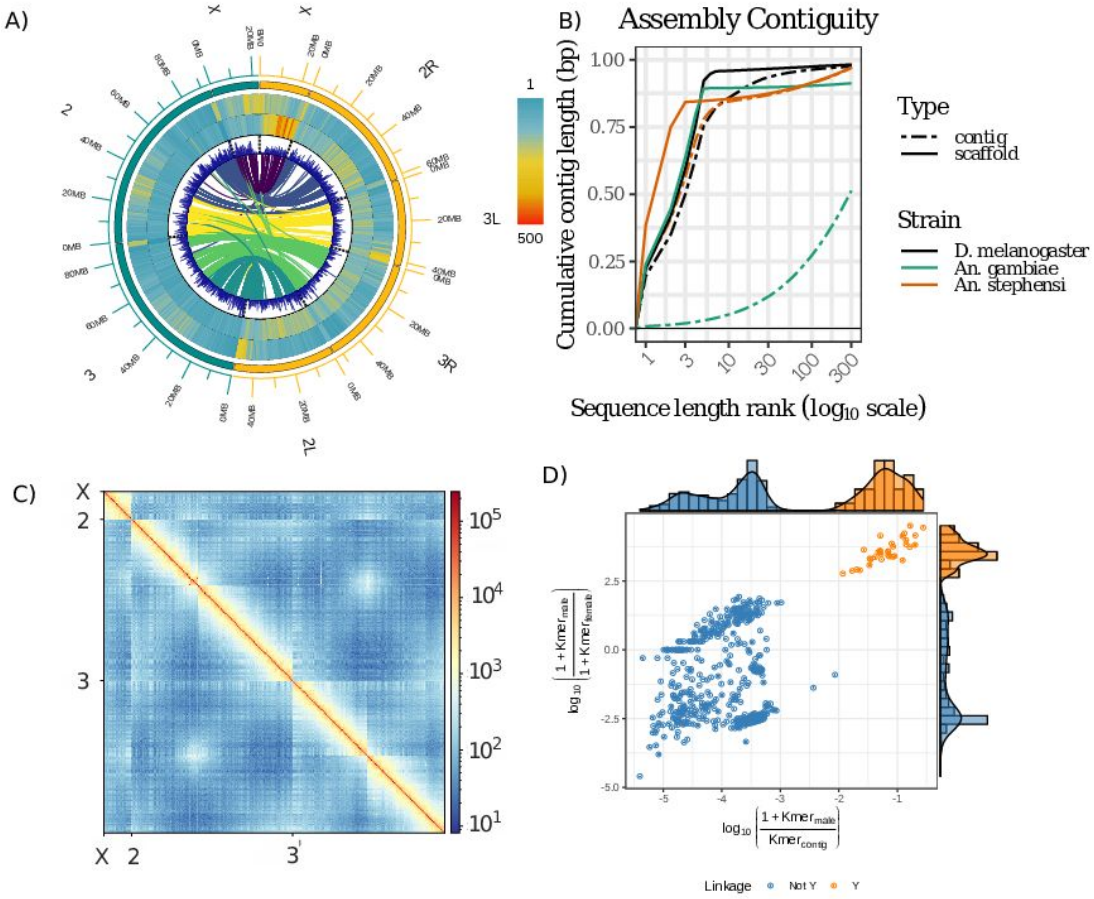

Figure 1. *Anopheles stephensi* genome assembly. A) Distribution of repeats, gene content, and synteny between *An. stephensi* (left, green) and *An. gambiae* (right, yellow) genomes. Each successive track from outside to inside represents TE density, satellite density (supplementary Fig. S2), and gene density across the chromosome arms in 500 kb windows. The innermost track describes the syntenic relationship between *An. stephensi* and *An. gambiae* chromosome arms. B) Comparison of assembly contiguities between *An. stephensi*, *An. gambiae*, and *D. melanogaster* reference assemblies. C) Hi-C contact map of the *An. stephensi* scaffolds. Density of Hi-C contacts are highest at the diagonals, suggesting consistency between assembly and the Hi-C map. D) Identification of putative Y contigs using the density of male-specific k-mers on the x-axis and the ratio of male and female k-mers on the y-axis.

considered a gold standard for metazoan genome assembly (Ashburner and Bergman 2005) (Fig. 1B, Table 1, supplementary Fig. S3-4, supplementary Table S1, supplemental text). The new *An. stephensi* assembly recovers 98.9% of 1,066 complete

single copy arthropod orthologs (i.e. Benchmarking Universal Single Copy Orthologs or BUSCOs (Waterhouse et al. 2017)). The reference assembly of the *D. melanogaster* genome captures 99.0% of BUSCOs, indicating that their completeness is comparable (Table 1, supplementary Fig. S5, supplementary Table S2). The *An. stephensi* assembly is more contiguous and complete than the latest version (AgamP4) of the African malaria vector *An. gambiae* genome (BUSCO 95.3%, contig N50 = 85 kb, scaffold N50 = 49 Mb), arguably the best characterized genome among malaria vectors (Fig. 1B, supplementary Figs. S3,S5)(Neafsey et al. 2015; Matthews et al. 2018). Furthermore, the concordance of short reads mapped to the assembly suggests that assembly errors are rare (short read consensus QV = 49.2, or ~1 discrepancy per 83 kb), which is further supported by uniformly mapping long reads and a high resolution Hi-C contact map (Fig. 1C, supplementary Fig. S6, Table 1). We also assembled 33 putative Y contigs totaling 2.4 Mb, representing the most extensive Y chromosome sequence yet recovered in any *Anopheles* species (Fig. 1D) (Neafsey et al. 2015; Jiang et al. 2014). Finally, to assist disease interventions using endosymbionts (Gonzalez-Ceron et al. 2003; Bando et al. 2013; Wang et al. 2017; Bai et al. 2019), we assembled *de novo* the first complete genome of the facultative endosymbiont *Serratia marcescens* from *Anopheles* using sequences identified in the *An. stephensi* long read data (Chen, Blom, and Walker 2017) (supplementary Fig. S7).

As naturally-occurring driving genetic elements, transposable elements (TEs) are invaluable for synthetic drives (Ribeiro and Kidwell 1994; Rasgon and Gould 2005; Marshall and Akbari 2016; Macias et al. 2017) and transgenic tools (Jasinskiene et al. 1998; Catteruccia et al. 2000; Arensburger et al. 2005). Although TEs comprise 11% (22.5 Mb) of the assembled *An. stephensi* genome, most LTR and non-LTR retrotransposons we identified were not present in the existing draft genome assembly (Fig. 2A, supplementary Table S3). These include species-specific and evolutionarily recent retrotransposons, which highlight the dynamic landscape of new TEs in this species and provide a resource for modeling the spread of synthetic drive elements (Fig. 2B). Although TEs are typically depleted within exons of protein-coding genes

(Chakraborty et al. 2020), we observed 1,368 TE sequences in transcripts of 8,381 genes, 68% (939/1,368) of which were not found in the earlier assembly (Jiang et al.

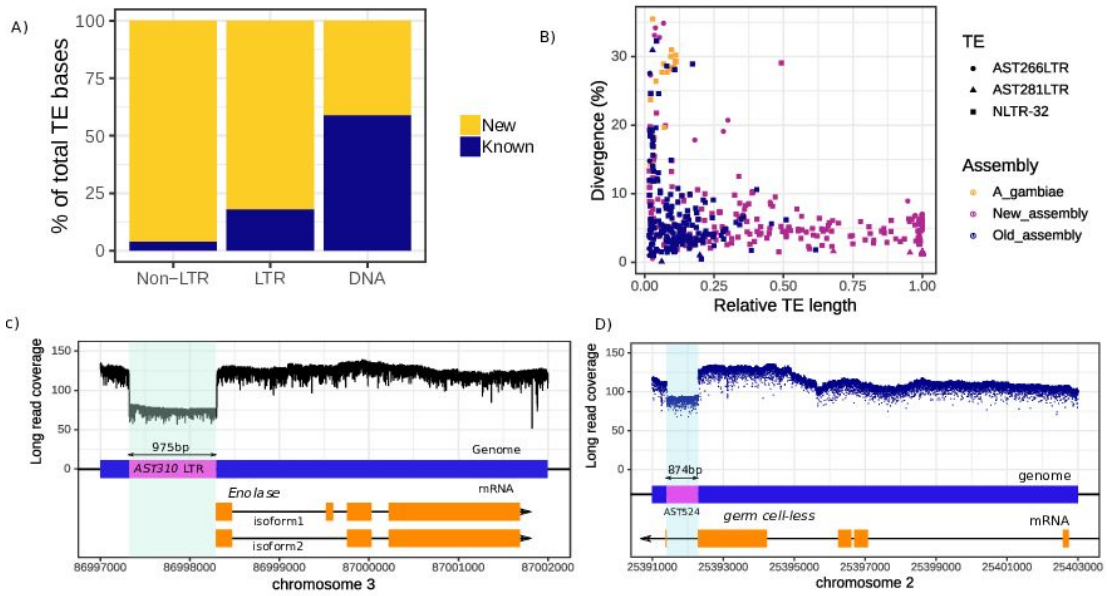

Figure 2. TEs and their role in genetic variation in *An. stephensi.* A) Proportion of TEs (counted in bp) that were uncovered by the new reference assembly of *An. stephensi*. Most LTR and non-LTR TEs are identified for the first time. B) Similar to many other LTR and non-LTR retrotransposososns, LTR elements AST266LTR and AST281LTR and non-LTR element NLTR-32 do not have any closely-related counterpart in *An. gambiae*. As shown here, only small parts of these TE sequences were known. C) Insertion of a polymorphic LTR fragment immediately upstream of the highly-conserved gene *Enolase*.The insertion creates a gap between the promoters and the transcription start site in half of the alleles and may disrupt transcription of the gene. D) A polymorphic DNA element AST524 located in the 3' UTR of *gcl* creates a null *gcl* allele.

2014) (supplementary Table S4). Due to a low level of residual heterozygosity in the sequenced strain (supplementary Fig. S1; see Methods), we discovered several TEs segregating in heterozygous genomic regions (supplementary Table S5-6, supplementary Fig. S8), some of which likely have functional consequences. For example, a 975 bp LTR fragment inserted immediately at the 5' end of the *Enolase* gene (*Eno*) may interrupt its transcription (Fig. 2C). In *D. melanogaster*, null mutants of *Eno* show severe fitness and phenotypic defects that range from flightlessness to lethality (Volkenhoff et al. 2015; Buszczak et al. 2007), whereas reduction in *Eno* expression protects *Drosophila* from cadmium and lead toxicity (Zhou et al. 2017).

Because *Eno* is highly conserved between *An. stephensi* and *D. melanogaster* (88% of 441 amino acids are identical between the two), we anticipate that this structural variant (SV) allele of

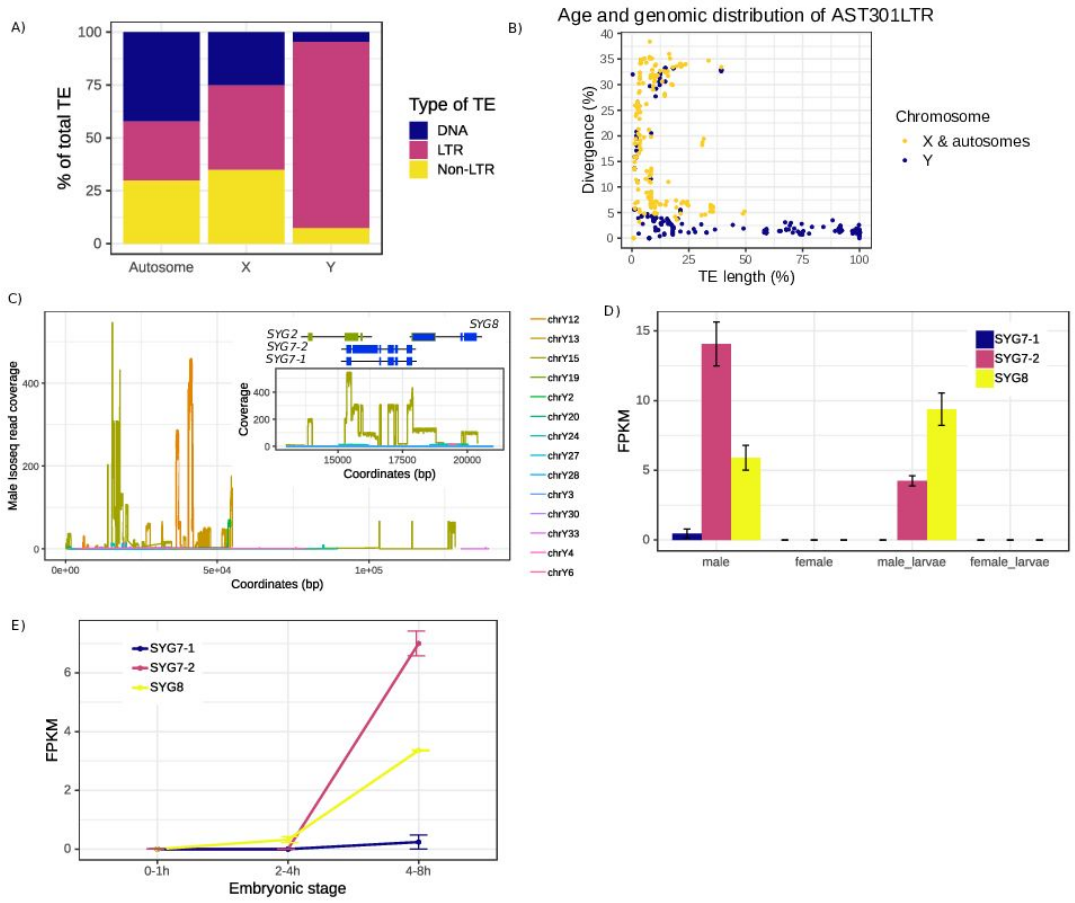

Figure 3. Structural and functional elements of the Y chromosome. A) Comparison of TE compositions in the autosomal, X and Y chromosomal sequences. Not only are most of Y sequences repeats (supplementary Fig. S9), but the majority of Y TEs are LTR elements. B) LTR retrotransposon AST301 is present in intact copies only in Y contigs. Its counterparts in the autosomes and X chromosome are fragmented and more diverged than the AST301 sequences found on the Y sequences. C) Distribution of uniquely-mapping Iso-seq read coverage across the 14/33 Y sequences that show non-zero coverage. Colors represent individual Y contigs. A 7 kb region zoomed in (upper-right corner) to highlight the Iso-seq evidence for three Y-linked genes on Y15. D) Transcript abundance of *SYG7* and *SYG8* in male and female adults and larvae. As shown here, neither gene is expressed in females. E) Expression of *SYG7* and *SYG8* in early embryos, where both begin to be expressed after 4 hours.

Eno might be deleterious, although it also could confer some degree of resistance to heavy metal toxins. Another TE, a 874 bp DNA element, is inserted into the 3' UTR of

the gene, *germ cell-less* (*gcl*), the *Drosophila* ortholog of which determines germ cell development (Robertson et al. 1999)(Fig. 2D). All full-length Iso-Seq reads from this gene are from the non-insertion allele, suggesting that the insertion allele is a *gcl* null.

Targeting Y-linked sequences can be the basis for suppressing vector populations (Prowse et al. 2019). In contrast to the autosomes and the X chromosome, 72% of the Y sequences (1.7 Mb) comprise LTR elements (Fig. 3A, supplementary Fig. S9). While most full length LTR elements in the Y chromosome also are present in the other chromosomes, a 3.7 kb retrotransposon, AST301, is represented by 46 highly-similar (<2.5% divergent) full-length copies only in the Y sequences. Its matching sequences elsewhere in the genome are small and evolutionary distant, consistent with AST301 being primarily active in the Y chromosome (Fig. 3B). The proliferation of AST301 may be a consequence of the Y chromosome's lack of recombination, which can lead to irreversible acquisition of deleterious mutations in a process called Muller's ratchet (H. J. Muller 1964). Despite the high repeat content, we uncovered seven Y-linked genes that are supported by multiple uniquely mapping Iso-Seq reads (Fig. 3C, supplementary Table S7, supplementary Fig. S10-16). We also recovered the three previously-identified Y-linked genes and filled sequence gaps in sYG1 (Hall et al. 2013). Two of the newly-discovered Y-linked genes (*Syg7* and *Syg8*) sit in a cluster of three overlapping Y-linked genes, all of which show strong expression in male larvae and adults but no expression in larval or adult females (Fig. 3D). These genes show low or absent expression in the early (0-2 hours) embryos but are expressed in the later stages (>4 hours) (Fig. 3E). Translation of open reading frames from *Syg7* transcripts shows the presence of a myb/SANT-like domain in Adf-1 (MADF) domain in the encoded protein (supplementary Fig. S17). Due to the DNA binding function of the MADF in certain *D. melanogaster* transcription factors (Bhaskar and Courey 2002; Shukla et al. 2014), we hypothesize that *Syg7* is a male-specific transcription factor.

An alternative to suppression schemes aimed at reducing mosquito numbers is modification of mosquito populations to prevent them from serving as parasite vectors. Promoters induced in females by blood feeding can be repurposed to express effector

molecules that impede malaria parasite transmission (Kokoza et al. 2010; Isaacs et al. 2011; Shane et al. 2018). Following a blood meal, hundreds of genes are induced, many of which stay upregulated for days after the blood meal (Fig. 4A, supplementary

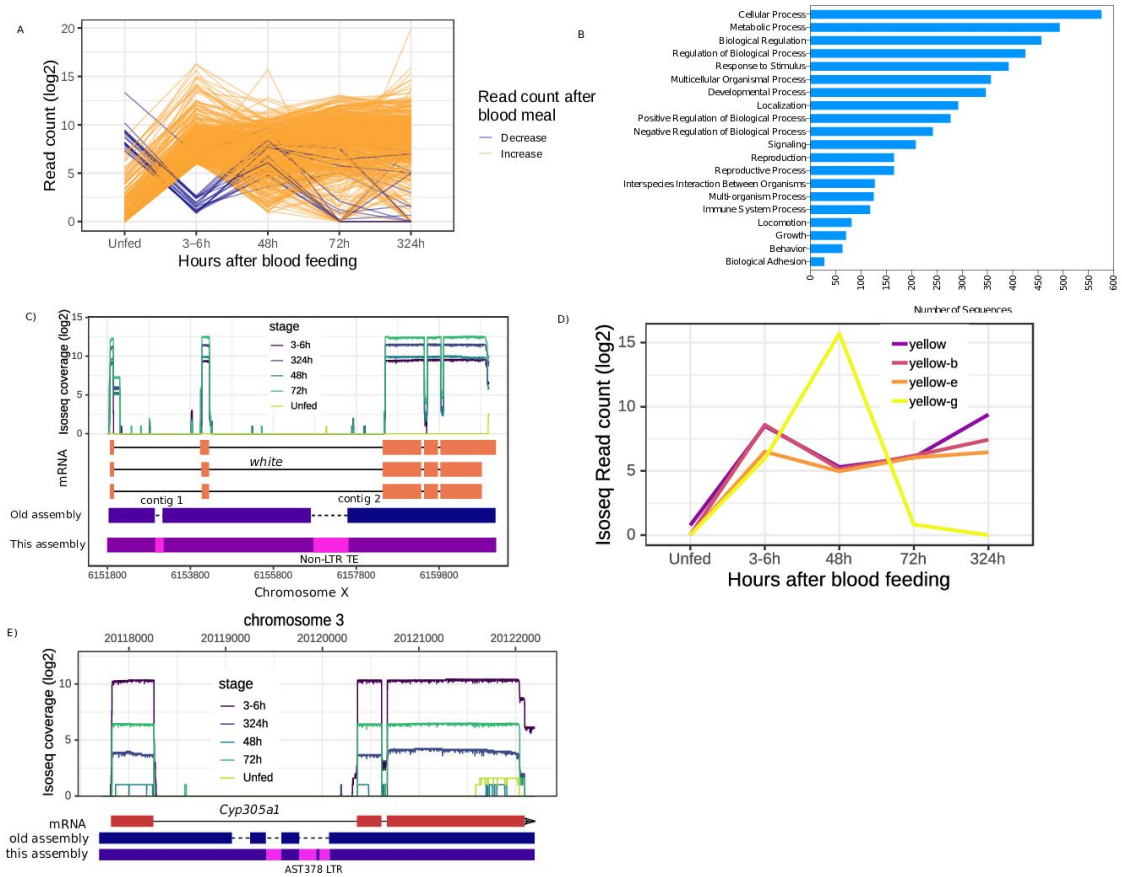

Figure 4. Gene expression changes in adult female mosquitoes after a blood meal. A) Transcript abundance of genes that are in the top 1% (>~64 fold) of the PBM transcript abundance changes. As evident here, more genes show upregulation than downregulation, although expression changes of some genes may not be due to the blood meal. B) GO analysis of the genes from A. Consistent with the role of the blood meal in mosquito biology, the differentially-expressed genes are involved in reproduction, immunity, and biological regulation (i.e. endocrine system). C) Despite being a common genetic marker, sequence of the PBM upregulated *white* gene was fragmented in the draft assembly of *An. stephensi*. D) Transcript abundance of four *yellow* genes (*yellow*, *yellow-b*, *yellow-e*, *yellow-g*) before and after blood meal. All genes show a similar transcript profile until 6 hours PBM after which *yellow-g* transcripts become more abundant. E) A *Cyp450* similar to *D. melanogaster Cyp305a1* shows PBM upregulation. However, parts of its introns were absent in the draft assembly.

Table S8). When ranked by expression-fold changes post blood meal (PBM), the top 1% (>64 fold change) include genes involved in digestion and metabolism, immunity,

endocrine pathways, and reproduction (Fig. 4B). However, sequences of 32% (235/724) of these genes are either fragmented or missing repetitive sequences in the draft *An. stephensi* assembly (Jiang et al. 2014)(supplementary Table S9). For example, despite being a key genetic marker in a wide range of insects and showing upregulation following a blood meal, the eye pigmentation gene *white* was fragmented in the draft assembly due to the presence of an 853 bp intronic non-LTR retrotransposon (Fig. 4C). Similarly, despite the active roles of blood-inducible and constitutively-expressed protease genes encoding trypsins and chymotrypsins in digestion and metabolism of blood (H. M. Muller et al. 1995; Jahan et al. 1999), not all of their complete sequences in *An. stephensi* were known before this study (supplementary Table S9). Interestingly, a Yellow protein gene (*yellow-g*) that was shown previously to be essential for female reproduction in *An. gambiae* and was used as a target in a CRISPR gene drive (Hammond et al. 2016) was found to be upregulated PBM. Yet, neither *yellow-g* nor the three other members of the Yellow protein gene family (*yellow-b*, *yellow-e*, *yellow*) that showed PBM elevated transcript levels were previously annotated (Fig. 4D) (Marinotti et al. 2005). Although transcript levels of all four genes increase in tandem until 6 hours PBM, the *yellow-g* transcript level continues to climb until 48 hours PBM and then reverts to the pre-BM level (Fig. 4D). In contrast, the other three *yellow* genes maintained a similar transcript level even after 13 days (Fig. 4D). The PBM upregulation pattern of the four yellow genes in *An. stephensi* is consistent with –roles in female reproduction– although *yellow-b*, *yellow-e*, and *yellow* are probably required longer than *yellow-g*. Interestingly, a Cytochrome P450 monooxygenase (Cyp450) gene, which bears similarity to *D. melanogaster Cyp305a1*, also was upregulated (Fig. 4E). *Cyp305a1* acts as an epoxidase in the Juvenile hormone biosynthesis pathway and helps maintain intestinal progenitor cells in *D. melanogaster* (Rahman et al. 2017). PBM upregulation of *Cyp305a1* suggests that it may have a potential role in cellular homeostasis in the midgut after a blood meal (Taracena et al. 2018). The *cis*-regulatory elements of the blood-meal-inducible genes and the antimicrobial peptide genes we identified can be leveraged to explore new effector molecule candidates to block

malaria parasite transmission (Abraham et al. 2005; Kokoza et al. 2010) (supplementary text, supplementary Fig. S18, supplementary Table S10).

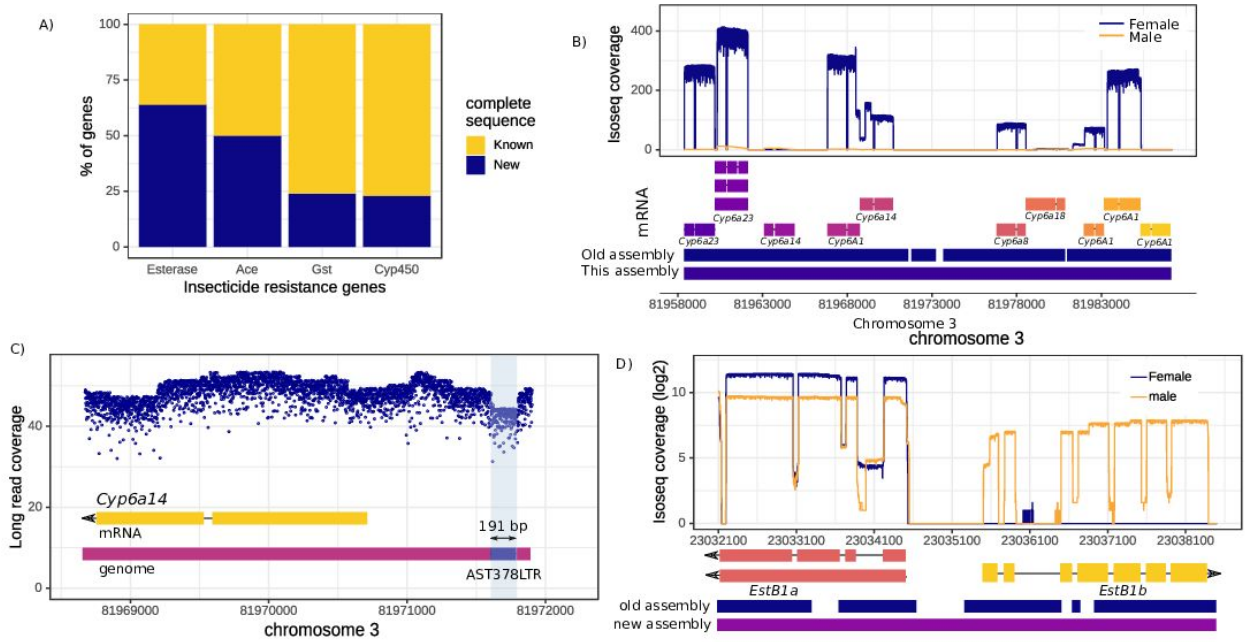

Figure 5. Putative insecticide resistance genes in *An. stephensi*. A) Proportion of various candidate insecticide-resistance genes that are either fragmented or missing repetitive regions in the draft *An. stephensi* assembly. B) An array of tandemly-located *Cyp450* genes that include *Cyp6a14*, a candidate gene for DDT resistance in *D. melanogaster* (Pedra et al. 2004). In the earlier assembly, this cluster was broken into three sequences, undermining investigation of the functional effects of these genes. Most genes show female-biased expression. C) A polymorphic AST378 LTR fragment insertion is segregating inside the *Cyp450* array shown in B, suggesting presence of more than one SV allele in this genomic region. D) Tandemly located *Esterase B1* genes show different sex-biased expression patterns. *EstB1a* do not show any strong bias towards either sex, whereas *EstB1b* shows male-biased expression. *EstB1* amplification causes organophosphate resistance in *Culex.* The *EstB1a* and *EstB1b* sequences were broken into five pieces in the earlier *An. stephensi* assembly.

We identified 94 CYP450, 29 *Gst*, 2 Acetylcholine esterases (*ace-1* and *ace-2*), and 14 *Est* genes, providing a comprehensive resource for discovery and delineation of the molecular basis of insecticide resistance (supplementary Table S11). Sequences of several of these genes were either fragmented or missing repetitive sequences in the draft assembly (Fig. 5A) (Jiang et al. 2014). For example, we recovered the complete *kdr* sequence and corrected sequence ambiguities in *ace-2* (supplementary Fig. S19-20). We also resolved tandem arrays of insecticide resistance genes, as evidenced

by a 28 kb region consisting of Cyp450s similar to *D. melanogaster Cyp6a14*, *Cyp6a23*, *Cyp6a8*, *Cyp6a18* and *Musca domestica Cyp6A1* (Fig. 5B). In *D. melanogaster*, *Cyp6a14* is a candidate gene for DDT resistance (Pedra et al. 2004), suggesting a similar function for its *An. stephensi* counterpart. One *Cyp6a14* in the array has a polymorphic 191 bp LTR TE fragment inserted 1 kb to the 5' end of the transcription start site, implying the presence of more than one SV allele in this complex region (Fig. 5C). We also resolved previously fragmented tandem copies of *Esterase B1* (*Est-B1a* and *Est-B1b*) counterparts, which have been shown previously in *Culex quinquefasciatus* to provide resistance to organophosphates (Mouchès et al. 1986) (Fig. 5D). Interestingly, the Cyp450s in the array and *Est-B1b* show opposite sex-biased expression, suggesting that the molecular basis of insecticide resistance may differ between sexes (Fig. 5B,5D)(Schmidt et al. 2010).

CRISPR and gene drive-based strategies promise to transform the management of disease vectors and pest populations (Gantz and Bier 2016). However, safety and effectiveness of these approaches rely on an accurate description of the functional and fitness effects of the genomic sequences and their variants (James 2005; Carballar-Lejarazú and James 2017). Draft assemblies are poorly suited for this purpose because they miss repetitive sequences or genes that are central to the vector's biology and evolution. Incomplete information about the correct copies or sequence of a gene may mislead conclusions about functional significance of the gene or the target sequence (Chakraborty et al. 2018) and may lead to mistargeting or misuse in a gene drive. The *An. stephensi* reference assembly solves these problems, revealing previously invisible or uncharacterized structural and functional genomic elements that shape various aspects of the vector biology of *An. stephensi*. Additionally, functionally-important SVs are segregating even in this inbred lab strock, indicating a significant role of structural genetic variation in phenotypic variation in this species. Finally, recent advances in technology have sparked enthusiasm for sequencing all eukaryotes in the tree of life (Lewin et al. 2018). The assembly we report here is timely, as it constitutes the first malaria vector to reach the exacting reference standards called

for by these ambitious proposals, and will stand alongside established references like those for human and fruit flies (Figure S3). This new assembly of *An. stephensi* provides a comprehensive and accurate map of genomic functional elements and will serve as a foundation for the new age of active genetics in *An. stephensi*.

METHODS

Mosquitoes

Anopheles stephensi mosquitoes of a strain from Indian subcontinent (gift of M. Jacobs-Lorena, Johns Hopkins University) were maintained in insectary conditions (27°C and 77% humidity) with a photoperiod of 12 h light:12h dark including 30 minutes of dawn and dusk at the University of California, Irvine (UCI). Larvae were reared in distilled water and fed ground TetraMin® fish food mixed with yeast powder. Adults had unlimited access to sucrose solutions (10% wt/vol) and females were provided with blood meals consisting of defibrinated calf blood (Colorado Serum Co., Denver) through the Hemotek® membrane feeding system.

Genome sequencing

Genomic DNA from adult mosquitoes was extracted using the Qiagen Blood & Cell Culture DNA Midi Kit following the previously-described protocol (Chakraborty et al. 2016). The genomic DNA was sheared with 10 plunges of size 21 blunt needles, followed by 10 plunges of size 24 blunt end needles. We generated our PacBio reads using 31 SMRTcells on the RSII platform (P6-C4 chemistry) at the UC San Diego Genomics Core and 2 SMRTcells on Sequel I platform at Nucleome (Hyderabad, India). From the same genomic DNA, we also generated 3.7 GB of 300 bp paired-end reads at the UC San Diego genomics core and 32.37 GB of 150 bp paired-end Illumina reads from Nucleome. To identify the Y-linked contigs, we generated 27.8 GB and 28.5 GB 100 bp paired-end Illumina reads from male and female genomic DNA, respectively, at UCI GHTF.

RNA extraction and sequencing

Total RNA was extracted from a total of six samples prepared from pooled individuals: 5-7-day-old, sugar-fed males, 5-7- day-old, sugar-fed females, and blood-fed females 3-6 h, 24 h, 48 h, and 72 h after feeding. The male pool consisted of 15 individuals, while female pools comprised 10 individuals each. All samples were isolated from the same mosquito cage. To do so, sugar-fed male and female samples were collected, and a blood meal offered for 1 h. Unfed females were removed from the cage and blood-fed females retrieved at each of the indicated time points. At the time of collection, samples were immersed in 500 μL of RNAlater RNA Stabilization Reagent (Qiagen) and stored at 4°C. Total RNA was extracted using the RNeasy Mini Kit (Qiagen) following manufacturer's instructions for the Purification of Total RNA from Animal Tissues. Extracted samples were treated with DNA-free Kit (Ambion) to remove traces of genomic DNA. Finally, samples were cleaned using the RNA Clean & Concentrator Kit (Zymo Research). mRNA selection, cDNA synthesis and Iso-Seq library prep was performed at UCI GHTF following the manufacturer's (Pacific Biosciences) protocol. For each of the six samples, one SMRTcell of Iso-seq reads were generated on the Sequel I platform.

Genome assembly

We used 42.4 GB or 180× of long reads (assuming haploid genome size G = 235 Mb) to generate two draft assemblies of *An. stephensi* using Canu v1.7 (Koren et al. 2017) and Falcon (Chin et al. 2013). Falcon was used to assemble the heterozygous regions (supplementary Fig. S1) that were recalcitrant to Canu. We filled the gaps in the canu assembly using the Falcon primary contigs following the two-steps merging approach with Quickmerge v0.3, where the Canu assembly was used as the reference assembly in the first merging step (Chakraborty et al. 2016; Solares et al. 2018). The resulting assembly was processed with finisherSC to remove the redundant contigs and fto fill the

further gaps with raw reads (Lam et al. 2015). This PacBio assembly (613 contigs, contig N50 = 38.1 Mb, 257.1 Mb in total) was polished twice with arrow (smrtanalysis v5.2.1) and twice with Pilon v1.22 using ~400X (80 Gb) 150 bp PE Illumina reads (Walker et al. 2014).

Identification of polymorphic mutations

To identify the variants segregating in the sequenced strain, we aligned the alternate haplotype contigs (a_ctg.fa) identified by Falcon to the scaffolded assembly. Then we called the indels using SVMU. An indel was marked as a TE based on its overlap with the Repeatmasker annotated TEs. To estimate heterozygosity, we mapped the Illumina reads to the chromosome scaffolds using bowtie2 and converted the alignments to a sorted bam file using SAMtools (v1.9). A VCF file containing the SNPs and small indels were generated using freebayes (v1.3.2-40-gcce27fc) and pairwise nucleotide diversity (pi) was calculated over 25 kb windows using vcftools (vcftools --window-pi 25000; v0.1.14v0.1.14).

Microbial sequence decontamination

Microbial contigs in assembly were identified using Kraken v2.0.7-beta (Wood and Salzberg 2014), which assigned taxonomic labels to the 613 contigs (supplementary Fig. S21). Kraken mapped k-mers (31-35 nt default) from the 613 contig sequences against the databases from the six domain sets: bacteria, archaea, viral, UniVec_Core, fungi, and protozoa from NCBI and the genome sets including representative reference mosquito from VectorBase v2019-02 and *Drosophila* genomes (n=24; supplementary Table S12). The databases map k-mers to the lowest common ancestor (LCA) of all genomes known to contain a given k-mer. Kraken label for each contig was further classified as either *Anopheles*, contaminating (non-*Anopheles*), or unclassified (no hit in the database) (supplementary Fig. S21). To prevent false positives in the results, low-complexity sequences in the assembly were masked with dustmasker (blast v2.8.1)

(Morgulis et al. 2008) prior to running Kraken. The mitochondrial genome of *An. stephensi* was identified by aligning the existing mitogenome (GenBank No. KT899888) against the contigs using nucmer in MUMmer v4.0.0b (Marçais et al. 2018).

Scaffolding

To *de novo* scaffold the microbial decontaminated 566 contigs, we collected HiC data from adult male and female mosquitoes. We flash-froze the adult mosquitoes and sent them to Arima Genomics (San Diego) to generate a HiC library using the Arima kit. This library was sequenced on a single flow cell of an Illumina HiSeq 2500 instrument, generating 326 GB of Illumina 150 bp paired-end reads. We mapped the HiC reads to the *An. stephensi* contigs using Juicer (Durand et al. 2016) and used the resulting contact map to scaffold the contigs using 3D-DNA (Dudchenko et al. 2017). The order and orientation of the three chromosomes were examined by nucmer in MUMmer alignment of 20 gene/probe physical map data (chrX, 5 probes; chr2, 7; chr3, 8) generated from FISH on polytene chromosomes (supplementary Fig. S22; supplementary Table S13) (Jiang et al. 2014) against Hi-C chromosome assemblies.

QV estimation and assembly statistics

To estimate the error rate in our final assembly, we mapped the paired-end Illumina reads to the assembly using bwa mem (bwa v0.7.17-5). The alignments were converted to bam format and then sorted using SAMtools (v.1.8-11). We called the variants using freeBayes v0.9.21 (Garrison and Marth 2012) (command: freebayes -C 2 -0 -O -q 20 -z 0.10 -E 0 -X -u -p 2 -F 0.75) and followed the approach of (Koren et al. 2017) to calculate QV (10*-log10(2981/250,632,892)). Briefly, we counted the number of bases comprising homozygous variants in the assembly (2981) and then divided it by the total mapped bases that had a coverage of at least three (250,632,892). We used QUAST v5.0.3 to obtain assembly statistics (Mikheenko, Saveliev, and Gurevich 2016). Out of 250 Mb, 205 Mb (82%) were scaffolded into the three chromosome-length scaffolds that correspond to the three *An. stephensi* chromosomes (chrX, 22.7 Mb; chr2, 93.7 Mb;

chr3, 88.7 Mb). Of the remaining sequences, we used Hi-C data to identify the 7 Mb (2.8%) of mis-joined sections of alternate haplotypes homologous to the chromosome scaffolds, while the remaining 35 Mb (14%) could not be assigned to a chromosome arm. Based on analysis of the duplicated BUSCOs, an additional 28 unplaced contigs were marked as candidates for alternate haplotypes (supplementary Table S14; supplemental Text). The final Hi-C map was visualized using HiCExplorer v3.4.2 (Ramírez et al. 2018).

Tandem repeats and transposable elements

Tandem repeats in the chromosomes of *An. stephensi* were annotated using Tandem Repeat Finder (Benson 1999). The number and copy number of micro-, mini- and macro-satellites spanning in each 100 kb non-overlapping window of the three chromosomes were identified (supplementary Fig. S23). The satellite classification was made as described in (Jiang et al. 2014). In brief, tandem repeats were classified as micro- (1-6 bases), mini- (7-99 bases) and macro- (>=100 bases) satellites (supplementary Fig. S24). Mini- and macro-satellites were considered only if they had a copy number of more than 2. All these three simple repeats were considered only if they had at least 80% sequence identity, and set some cutoff (>=2 copy number; >=80% identity) to screen high confidence repeats, then the overall abundance was calculated (supplementary Fig. S25).

We created a custom TE library using the EDTA pipeline (Ou et al. 2019) and Repeatmodeler (http://www.repeatmasker.org/RepeatModeler/) to annotate the TEs. LTR retrotransposons and DNA elements were identified *de novo* using EDTA, but because EDTA does not identify Non-LTR elements, Repeatmodeler was used to identify these. The two libraries were combined and the final library was used with Repeatmasker to annotate the genome-wide TEs.

Annotation using Iso-Seq

In total, six samples (5 females; 1 male) of *An. stephensi* mosquitoes were used for Iso-Seq sequencing (see RNA extraction and sequencing). Raw PacBio long-molecule sequencing data was processed using the SMRT analysis v7.0.0 Iso-Seq3 pipeline (Gordon et al. 2015) (supplemental Text). Briefly, CCS was used to generate the full-length (FL) reads for which all 5'-end primer, polyA tail and 3'-end primer have been sequenced and then Lima was used to identify and remove the 5'- and 3'-end cDNA primers from the FL reads. The resulting bam files were processed with Iso-Seq3 to refine and cluster the reads, which were polished with Arrow. This *de novo* pipeline outputs FASTQ files containing two sets of error-corrected, full length isoforms: i) the high-quality set contains isoforms supported by at least two FL reads with an accuracy of at least 99% and ii) while the low-quality set contains isoforms with an accuracy <99% that occurred due to insufficient coverage or rare transcripts. The high quality isoforms were collapsed with Cupcake and were used in Talon for annotation (Wyman et al. 2019). We combined the high quality isoforms with other lines of evidence using MAKER2 to create a final annotation (see below) (supplementary Table S15).

MAKER Annotation

The final annotation of the genome was performed using MAKER2 (Holt and Yandell 2011), which combines empirical evidence and *ab initio* gene prediction to produce final annotations (supplemental Text). We used MAKER2 for three cycles of gene predictions. First, the Iso-Seq data were used as evidence for training MAKER2 for gene predictions. We also used transcriptome and peptide sequence data from *An. gambiae* (PEST4.12) and *An. funestus* (FUMOZ 3.1) as alternative evidence to support the predicted gene models. Prior to gene annotation, repeats were masked using RepeatMasker included in MAKER2. Mapping of EST and protein evidence to the genome by MAKER2 using BLASTn and BLASTx, respectively, yielded 12,324 genes

transcribing 14,888 mRNAs. The output of first round gene models were used for the second round, where MAKER2 ran SNAP and AUGUSTUS for *ab initio* gene predictions. Next, another round of SNAP and AUGUSTUS predictions were performed to synthesize the final annotations that produced 14,966 genes, transcribing 16,559 mRNAs. In total, we identified 56,388 exons, 9,791 5'-end UTRs, 9290 3'end UTRs and 503 tRNAs (Table 1; supplementary Table S15). We also predicted *ab initio* an additional 14,192 mRNAs/proteins but due to weak support they were not considered. The final MAKER annotation was assessed using recommended AED (annotation edit distance) and Pfam statistical metrics. For the AED, 92% of the annotated genes with <0.5 AED was observed, which is more than the recommended score (>90%), indicating that the annotated genes conform to the gold standard. The Pfam content score (65.4%) also was above the recommended range 55-65%, indicating that the *An. stephensi* proteome was well annotated.

The gene models were functionally annotated in MAKER2 through a homology BLAST search to UniProt-Sprot database, while the domains in the annotated proteins were assigned from the InterProScan database (Supplemental Text). Comparison of gene model annotations of our assembly with the *An. stephensi* draft assembly using OrthoFinder v2.3.7 (Emms and Kelly 2019) showed 3653 (22.1%) transcripts were identified as being unique to our assembly (supplementary Table S16) (i.e. assembly-specific), in which 1009 were supported by Iso-Seq reads (supplementary Table S16). Gene ontology of the differentially expressed genes after the blood meal was analyzed using BLAST2GO (Conesa and Götz 2008).

Validation and quantification with RNAseq and Iso-Seq

To quantify transcript abundance using Iso-seq reads, raw reads were mapped to the genome assembly using minimap2 (Li 2018) and the gene-specific transcript abundance was measured using bedtools, requiring that each Iso-seq read overlaps at least 75% of gene length annotated with TALON (bedtools coverage -mean -F 0.75 -a

talon.gff -b minimap.bam) (Quinlan and Hall 2010). To take variation due to sequencing yield per SMRTcell into account while calculating transcript abundance, Iso-seq coverage of each gene was divided by a normalization factor, which was calculated by dividing the total read counts for each sample by the total read counts from the unfed female sample.

To obtain transcript levels of Y-linked genes from embryos, larvae, and adults, we used publicly available RNA-seq data (supplementary Table S17). RNA-seq reads were mapped to the genome using HISAT2 and the per-base read coverage was calculated from the sorted bam files using samtools depth. Additionally, the bam files were processed with stringtie to generate sample-specific transcript annotation in GTF format (Pertea et al. 2016). Sample specific GTF files were merged with stringtie to generate the final GTF. To obtain the gene model and transcript isoforms of *kdr*, stringtie annotated transcripts that covered the entire predicted ORF based on homology with *D. melanogaster para* were used.

Identification of hidden elements

To identify incomplete or ambiguous sequences in the earlier draft assembly (Jiang et al. 2014), the contigs in the previous version of *An. stephensi* assembly were aligned to the new assembly using nucmer (Marçais et al. 2018) and then alignments due to repeats were filtered using delta-filter to generate 1-to-1 mapping (delta-filter -r -q) between the two assemblies. The resulting delta file was converted into tab-separated alignment format using show-cords utility in MUMmer. Breaks in alignment between the two assemblies were considered as discrepancies between the assemblies and occurrence of such breaks within a genome feature (i.e. genes or repeats; bedtools intersect -f 1.0 -a feature.bed -b alignment.bed) was assumed to be due to fragmentation or ambiguities in the feature sequence in the draft assembly. To rule out assembly errors in the new assembly as the cause of discrepancy between the two assemblies, 20 features disagreeing between the assemblies were randomly selected

from each category and manually inspected in IGV. At least 3 long reads spanning an entire feature was used as evidence for correct assembly of the features.

Identification of Y contigs

To identify the putative Y contigs, male and female specific k-mers were identified from male and female paired-end Illumina reads using Jellyfish (Marçais and Kingsford 2011)(supplementary Fig. S26). Density of male and female k-mers for each contig was calculated and the contigs showing more than two-fold higher density of male specific k-mers were designated as putative Y contigs. Interestingly, the *Serratia* genome we assembled also showed similar male k-mer enrichment as the Y contigs.

Experimental validation of Y-linked contigs

The k-mer based approach employed to identify male-specific kmers that occur at a rate 20-fold higher than female-specific kmers in the *An. stephensi* scaffolds (supplementary Fig. S27). In order to verify putative Y-linked sequences, ten 2-3 days-old male or female *An. stephensi* mosquitoes per replicate were used for the experiment. Genomic DNA was extracted from each sample using DNeasy Blood and Tissue Kit (Cat # 69504). Gene specific primers (Y15 forward(F) - ATT TTA GTT ATT TAG AGG CTT CGA, Y15 reverse(R) - GCG TAT GAT AGA AAC CGC AT; Y22 F - ATG CCA AAA AAA CGG TTG CG, Y22 R - CTA GCT CTT GTA AAG AGT CAC CTT; Y28 F - ATG CTA CAA AAC AGT GCC TT, Y28 R - TTA GGT CAG ATA TAG ACA CAG ACA CA) were designed based on the genome sequence to amplify ≥500 bp products using Polymerase Chain Reaction (PCR) reaction. The amplification was done using Q5 high fidelity 2X Master Mix (Cat # M0492). Amplicons were resolved in agarose gels and male versus female amplification was compared. The PCR products were gel eluted and Sanger sequenced (supplementary Fig. S27) (Genewiz) with forward PCR primer. The identity of the sequencing was confirmed by aligning the amplicon sequences against the *An. stephensi* genome assembly using BLAST.

Identification of putative immune gene family

Studying the patterns of evolution in innate immune genes facilitate understanding the evolutionary dynamics of *An. stephensi* and pathogens they harbor. A total of 1649 manually curated immune proteins of *An. gambiae* (Agam 385), *Ae. aegypti* (Aaeg 422), *Cu. quinquefasciatus* (Cpip 495) and *D. melanogaster* (Dmel 347) in ImmuneDB (supplementary Table S10) (Waterhouse et al. 2007) were used as databases to search for the putative immune-related proteins in MAKER2-annotated protein sequences of the *An. stephensi* assembly using sequence alignment and phylogenetic orthology inference based method in OrthoFinder v2. The number of single copy orthogroup/orthologous proteins (one-to-one) and co-orthologous and paralogous proteins in *An. stephensi* were identified (one-to-many; many-to-one; many-to-many).

Data availability

The raw PacBio, illumina and Hi-C sequencing data and *An. stephensi* genome assembly were deposited in the NCBI BioProject database (Accession number PRJNA629843). The annotations and other genomic features can be accessed at http://3.93.125.130/tigs/anstephdb/.

Code availability

All codes used in the study, including those used to make figures are available at https://github.com/mahulchak/stephensi_genome.

Acknowledgements

MC and JJE were supported by NIH grants K99GM129411 and R01GM123303-1, respectively. EB was supported by NIH grant R01 GM117321 and Paul G. Allen Frontiers Group Distinguished Investigators Award. AR and BK were supported by Tata Institute for

Genetics and Society(TIGS)-India. This work was supported in part by the TIGS-UCSD and TIGS-India. AAJ is a Donald Bren Professor at the University of California, Irvine. We thank Judith Coleman for help with mosquito collection and Yi Liao for helpful discussions.

Author contributions

MC, AAJ, and JJE conceived the experimental approach; AA, PH and BK performed laboratory research; MC, AR, SJ, KP, and SW performed bioinformatics analysis; SSr coordinated all activities at IBAB and contributed to the annotation pipeline. MC and AR wrote the manuscript draft; EB, SSu, AAJ and JJE edited the draft. SSu coordinated the project activities between TIGS-India and TIGS-UC San Diego, and was involved in planning of the sequencing strategies. All authors contributed to the finalized version of the manuscript.

Competing interests

EB has equity interest in two companies: Synbal Inc. and Agragene, Inc. These companies that may potentially benefit from the research results. E.B. also serves on the Synbal Inc.'s Board of Directors and Scientific Advisory Board, and on Agragene Inc.'s Scientific Advisory Board. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. All other authors declare no conflict of interest.

References

Abraham, E. G., M. Donnelly-Doman, H. Fujioka, A. Ghosh, L. Moreira, and M. Jacobs-Lorena. 2005. “Driving Midgut-Specific Expression and Secretion of a Foreign Protein in Transgenic Mosquitoes with AgAper1 Regulatory Elements.” *Insect Molecular Biology* 14 (3): 271–79.

Alkan, Can, Saba Sajjadian, and Evan E. Eichler. 2011. “Limitations of next-Generation Genome Sequence Assembly.” *Nature Methods* 8 (1): 61–65.

Arensburger, Peter, Yu-Jung Kim, Jamison Orsetti, Channa Aluvihare, David A. O’Brochta, and Peter W. Atkinson. 2005. “An Active Transposable Element, Herves, from the African Malaria Mosquito Anopheles Gambiae.” *Genetics* 169 (2): 697–708.

Ashburner, Michael, and Casey M. Bergman. 2005. “Drosophila Melanogaster: A Case Study of a Model Genomic Sequence and Its Consequences.” *Genome Research* 15 (12): 1661–67.

Bai, Liang, Lili Wang, Joel Vega-Rodríguez, Guandong Wang, and Sibao Wang. 2019. “A Gut Symbiotic Bacterium Serratia Marcescens Renders Mosquito Resistance to Plasmodium Infection Through Activation of Mosquito Immune Responses.” *Frontiers in Microbiology* 10 (July): 1580.

Bando, Hironori, Kiyoshi Okado, Wamdaogo M. Guelbeogo, Athanase Badolo, Hiroka Aonuma, Bryce Nelson, Shinya Fukumoto, Xuenan Xuan, N ’fale Sagnon, and Hirotaka Kanuka. 2013. “Intra-Specific Diversity of Serratia Marcescens in Anopheles Mosquito Midgut Defines Plasmodium Transmission Capacity.” *Scientific Reports* 3: 1641.

Benson, G. 1999. “Tandem Repeats Finder: A Program to Analyze DNA Sequences.” *Nucleic Acids Research* 27 (2): 573–80.

Bhaskar, Vinay, and Albert J. Courey. 2002. “The MADF-BESS Domain Factor Dip3 Potentiates Synergistic Activation by Dorsal and Twist.” *Gene* 299 (1-2): 173–84.

Buszczak, Michael, Shelley Paterno, Daniel Lighthouse, Julia Bachman, Jamie Planck, Stephenie Owen, Andrew D. Skora, et al. 2007. “The Carnegie Protein Trap Library: A Versatile Tool for Drosophila Developmental Studies.” *Genetics* 175 (3): 1505–31.

Carballar-Lejarazú, Rebeca, and Anthony A. James. 2017. “Population Modification of Anopheline Species to Control Malaria Transmission.” *Pathogens and Global Health* 111 (8): 424–35.

Catteruccia, F., T. Nolan, T. G. Loukeris, C. Blass, C. Savakis, F. C. Kafatos, and A. Crisanti. 2000. “Stable Germline Transformation of the Malaria Mosquito Anopheles

Stephensi." *Nature* 405 (6789): 959–62.
Chakraborty, Mahul, James G. Baldwin-Brown, Anthony D. Long, and J. J. Emerson. 2016. "Contiguous and Accurate de Novo Assembly of Metazoan Genomes with Modest Long Read Coverage." *Nucleic Acids Research* 44 (19): e147.
Chakraborty, Mahul, Ching-Ho Chang, Danielle E. Khost, Jeffrey Vedanayagam, Jeffrey R. Adrion, Yi Liao, Kristi Montooth, Colin D. Meiklejohn, Amanda M. Larracuente, and J. J. Emerson. 2020. "Evolution of Genome Structure in the Drosophila Simulans Species Complex." *bioRxiv*. https://doi.org/10.1101/2020.02.27.968743.
Chakraborty, Mahul, Nicholas W. VanKuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and J. J. Emerson. 2018. "Hidden Genetic Variation Shapes the Structure of Functional Elements in Drosophila." *Nature Genetics* 50 (1): 20–25.
Champer, Samuel E., Suh Yeon Oh, Chen Liu, Zhaoxin Wen, Andrew G. Clark, Philipp W. Messer, and Jackson Champer. 2020. "Computational and Experimental Performance of CRISPR Homing Gene Drive Strategies with Multiplexed gRNAs." *Science Advances* 6 (10): eaaz0525.
Chen, Shicheng, Jochen Blom, and Edward D. Walker. 2017. "Genomic, Physiologic, and Symbiotic Characterization of Serratia Marcescens Strains Isolated from the Mosquito Anopheles Stephensi." *Frontiers in Microbiology* 8 (August): 1483.
Chin, Chen-Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nature Methods* 10 (6): 563–69.
Conesa, Ana, and Stefan Götz. 2008. "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics." *International Journal of Plant Genomics* 2008: 619832.
Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. "De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science* 356 (6333): 92–95.
Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3 (1): 95–98.
Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.
Enayati, Ahmad Ali, H. Vatandoost, H. Ladonni, Harold Townson, and Janet Hemingway. 2003. "Molecular Evidence for a Kdr-like Pyrethroid Resistance Mechanism in the Malaria Vector Mosquito Anopheles Stephensi." *Medical and Veterinary Entomology* 17 (2): 138–44.

Faulde, Michael K., Leopoldo M. Rueda, and Bouh A. Khaireh. 2014. “First Record of the Asian Malaria Vector Anopheles Stephensi and Its Possible Role in the Resurgence of Malaria in Djibouti, Horn of Africa.” *Acta Tropica* 139 (November): 39–43.

Gantz, Valentino M., and Ethan Bier. 2016. “The Dawn of Active Genetics.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 38 (1): 50–63.

Gantz, Valentino M., Nijole Jasinskiene, Olga Tatarenkova, Aniko Fazekas, Vanessa M. Macias, Ethan Bier, and Anthony A. James. 2015. “Highly Efficient Cas9-Mediated Gene Drive for Population Modification of the Malaria Vector Mosquito Anopheles Stephensi.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (49): E6736–43.

Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1207.3907.

Gonzalez-Ceron, Lilia, Frida Santillan, Mario H. Rodriguez, Domingo Mendez, and Juan E. Hernandez-Avila. 2003. “Bacteria in Midguts of Field-Collected Anopheles Albimanus Block Plasmodium Vivax Sporogonic Development.” *Journal of Medical Entomology* 40 (3): 371–74.

Gordon, Sean P., Elizabeth Tseng, Asaf Salamov, Jiwei Zhang, Xiandong Meng, Zhiying Zhao, Dongwan Kang, et al. 2015. “Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing.” *PloS One* 10 (7): e0132628.

Hall, Andrew Brantley, Yumin Qi, Vladimir Timoshevskiy, Maria V. Sharakhova, Igor V. Sharakhov, and Zhijian Tu. 2013. “Six Novel Y Chromosome Genes in Anopheles Mosquitoes Discovered by Independently Sequencing Males and Females.” *BMC Genomics* 14 (April): 273.

Hammond, Andrew, Roberto Galizi, Kyros Kyrou, Alekos Simoni, Carla Siniscalchi, Dimitris Katsanos, Matthew Gribble, et al. 2016. “A CRISPR-Cas9 Gene Drive System Targeting Female Reproduction in the Malaria Mosquito Vector Anopheles Gambiae.” *Nature Biotechnology* 34 (1): 78–83.

Holt, Carson, and Mark Yandell. 2011. “MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects.” *BMC Bioinformatics* 12 (December): 491.

Hoskins, Roger A., Joseph W. Carlson, Kenneth H. Wan, Soo Park, Ivonne Mendez, Samuel E. Galle, Benjamin W. Booth, et al. 2015. “The Release 6 Reference Sequence of the Drosophila Melanogaster Genome.” *Genome Research* 25 (3): 445–58.

Isaacs, Alison T., Fengwu Li, Nijole Jasinskiene, Xiaoguang Chen, Xavier Nirmala, Osvaldo Marinotti, Joseph M. Vinetz, and Anthony A. James. 2011. “Engineered

Resistance to Plasmodium Falciparum Development in Transgenic Anopheles Stephensi.” *PLoS Pathogens* 7 (4): e1002017.

Jahan, N., P. T. Docherty, P. F. Billingsley, and H. Hurd. 1999. “Blood Digestion in the Mosquito, Anopheles Stephensi: The Effects of Plasmodium Yoelii Nigeriensis on Midgut Enzyme Activities.” *Parasitology* 119 (Pt 6) (December): 535–41.

James, Anthony A. 2005. “Gene Drive Systems in Mosquitoes: Rules of the Road.” *Trends in Parasitology* 21 (2): 64–67.

Jasinskiene, N., C. J. Coates, M. Q. Benedict, A. J. Cornel, C. S. Rafferty, A. A. James, and F. H. Collins. 1998. “Stable Transformation of the Yellow Fever Mosquito, Aedes Aegypti, with the Hermes Element from the Housefly.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (7): 3743–47.

Jiang, Xiaofang, Ashley Peery, A. Brantley Hall, Atashi Sharma, Xiao-Guang Chen, Robert M. Waterhouse, Aleksey Komissarov, et al. 2014. “Genome Analysis of a Major Urban Malaria Vector Mosquito, Anopheles Stephensi.” *Genome Biology* 15 (9): 459.

Kokoza, Vladimir, Abdouelaziz Ahmed, Sang Woon Shin, Nwando Okafor, Zhen Zou, and Alexander S. Raikhel. 2010. “Blocking of Plasmodium Transmission by Cooperative Action of Cecropin A and Defensin A in Transgenic Aedes Aegypti Mosquitoes.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (18): 8111–16.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation.” *Genome Research*, March. https://doi.org/10.1101/gr.215087.116.

Kyrou, Kyros, Andrew M. Hammond, Roberto Galizi, Nace Kranjc, Austin Burt, Andrea K. Beaghton, Tony Nolan, and Andrea Crisanti. 2018. “A CRISPR-Cas9 Gene Drive Targeting Doublesex Causes Complete Population Suppression in Caged Anopheles Gambiae Mosquitoes.” *Nature Biotechnology* 36 (11): 1062–66.

Lam, Ka-Kit, Kurt LaButti, Asif Khalak, and David Tse. 2015. “FinisherSC: A Repeat-Aware Tool for Upgrading de Novo Assembly Using Long Reads.” *Bioinformatics* 31 (19): 3207–9.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Macias, Vanessa M., Alyssa J. Jimenez, Bianca Burini-Kojin, David Pledger, Nijole Jasinskiene, Celine Hien Phong, Karen Chu, et al. 2017. “Nanos-Driven Expression

of piggyBac Transposase Induces Mobilization of a Synthetic Autonomous Transposon in the Malaria Vector Mosquito, Anopheles Stephensi." *Insect Biochemistry and Molecular Biology* 87 (August): 81–89.

Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. 2018. "MUMmer4: A Fast and Versatile Genome Alignment System." *PLoS Computational Biology* 14 (1): e1005944.

Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers." *Bioinformatics* 27 (6): 764–70.

Marinotti, O., Q. K. Nguyen, E. Calvo, A. A. James, and J. M. C. Ribeiro. 2005. "Microarray Analysis of Genes Showing Variable Expression Following a Blood Meal in Anopheles Gambiae." *Insect Molecular Biology* 14 (4): 365–73.

Marshall, John M., and Omar S. Akbari. 2016. "Chapter 9 - Gene Drive Strategies for Population Replacement." In *Genetic Control of Malaria and Dengue*, edited by Zach N. Adelman, 169–200. Boston: Academic Press.

Matthews, Benjamin J., Olga Dudchenko, Sarah B. Kingan, Sergey Koren, Igor Antoshechkin, Jacob E. Crawford, William J. Glassford, et al. 2018. "Improved Reference Genome of Aedes Aegypti Informs Arbovirus Vector Control." *Nature* 563 (7732): 501–7.

Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. "MetaQUAST: Evaluation of Metagenome Assemblies." *Bioinformatics* 32 (7): 1088–90.

Morgulis, Aleksandr, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala, and Alejandro A. Schäffer. 2008. "Database Indexing for Production MegaBLAST Searches." *Bioinformatics* 24 (16): 1757–64.

Mouchès, C., N. Pasteur, J. B. Bergé, O. Hyrien, M. Raymond, B. R. de Saint Vincent, M. de Silvestri, and G. P. Georghiou. 1986. "Amplification of an Esterase Gene Is Responsible for Insecticide Resistance in a California Culex Mosquito." *Science* 233 (4765): 778–80.

Muller, H. J. 1964. "THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE." *Mutation Research* 106 (May): 2–9.

Muller, H. M., F. Catteruccia, J. Vizioli, A. Dellatorre, and A. Crisanti. 1995. "Constitutive and Blood Meal-Induced Trypsin Genes in Anopheles Gambiae." *Experimental Parasitology* 81 (3): 371–85.

Neafsey, Daniel E., Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, et al. 2015. "Mosquito Genomics. Highly Evolvable Malaria Vectors: The Genomes of 16 Anopheles Mosquitoes." *Science* 347 (6217): 1258522.

Nirmala, Xavier, Osvaldo Marinotti, Juan Miguel Sandoval, Sophea Phin, Surendra Gakhar, Nijole Jasinskiene, and Anthony A. James. 2006. "Functional Characterization of the Promoter of the Vitellogenin Gene, AsVg1, of the Malaria

Vector, Anopheles Stephensi.” *Insect Biochemistry and Molecular Biology* 36 (9): 694–700.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 20 (1): 275.

Pedra, J. H. F., L. M. McIntyre, M. E. Scharf, and Barry R. Pittendrigh. 2004. “Genome-Wide Transcription Profile of Field- and Laboratory-Selected Dichlorodiphenyltrichloroethane (DDT)-Resistant Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (18): 7034–39.

Pertea, Mihaela, Daehwan Kim, Geo M. Pertea, Jeffrey T. Leek, and Steven L. Salzberg. 2016. “Transcript-Level Expression Analysis of RNA-Seq Experiments with HISAT, StringTie and Ballgown.” *Nature Protocols* 11 (9): 1650–67.

Prowse, Thomas Aa, Fatwa Adikusuma, Phillip Cassey, Paul Thomas, and Joshua V. Ross. 2019. “A Y-Chromosome Shredding Gene Drive for Controlling Pest Vertebrate Populations.” *eLife* 8 (February). https://doi.org/10.7554/eLife.41873.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.

Rahman, M. M., X. Franch-Marro, J. L. Maestro, D. Martin, and A. Casali. 2017. “Local Juvenile Hormone Activity Regulates Gut Homeostasis and Tumor Growth in Adult Drosophila.” *Scientific Reports* 7 (1): 11677.

Ramírez, Fidel, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. 2018. “High-Resolution TADs Reveal DNA Sequences Underlying Genome Organization in Flies.” *Nature Communications* 9 (1): 189.

Rasgon, J. L., and F. Gould. 2005. “Transposable Element Insertion Location Bias and the Dynamics of Gene Drive in Mosquito Populations.” *Insect Molecular Biology* 14 (5): 493–500.

Ribeiro, J. M., and M. G. Kidwell. 1994. “Transposable Elements as Population Drive Mechanisms: Specification of Critical Parameter Values.” *Journal of Medical Entomology* 31 (1): 10–16.

Robertson, S. E., T. C. Dockendorff, J. L. Leatherman, D. L. Faulkner, and T. A. Jongens. 1999. “Germ Cell-Less Is Required Only during the Establishment of the Germ Cell Lineage of Drosophila and Has Activities Which Are Dependent and Independent of Its Localization to the Nuclear Envelope.” *Developmental Biology* 215 (2): 288–97.

Safi, Noor Halim Zahid, Abdul Ali Ahmadi, Sami Nahzat, Supriya Warusavithana, Naimullah Safi, Reza Valadan, Atie Shemshadian, Marzieh Sharifi, Ahmadali Enayati, and Janet Hemingway. 2019. “Status of Insecticide Resistance and Its

Biochemical and Molecular Mechanisms in Anopheles Stephensi (Diptera: Culicidae) from Afghanistan." *Malaria Journal* 18 (1): 249.

Safi, Noor Halim Zahid, Abdul Ali Ahmadi, Sami Nahzat, Seyyed Payman Ziapour, Seyed Hassan Nikookar, Mahmoud Fazeli-Dinan, Ahmadali Enayati, and Janet Hemingway. 2017. "Evidence of Metabolic Mechanisms Playing a Role in Multiple Insecticides Resistance in Anopheles Stephensi Populations from Afghanistan." *Malaria Journal* 16 (1): 100.

Schmidt, Joshua M., Robert T. Good, Belinda Appleton, Jayne Sherrard, Greta C. Raymant, Michael R. Bogwitz, Jon Martin, et al. 2010. "Copy Number Variation and Transposable Elements Feature in Recent, Ongoing Adaptation at the Cyp6g1 Locus." *PLoS Genetics* 6 (6): e1000998.

Seyfarth, Marco, Bouh A. Khaireh, Abdoulilah A. Abdi, Samatar M. Bouh, and Michael K. Faulde. 2019. "Five Years Following First Detection of Anopheles Stephensi (Diptera: Culicidae) in Djibouti, Horn of Africa: Populations Established-Malaria Emerging." *Parasitology Research* 118 (3): 725–32.

Shane, Jackie L., Christina L. Grogan, Caroline Cwalina, and David J. Lampe. 2018. "Blood Meal-Induced Inhibition of Vector-Borne Disease by Transgenic Microbiota." *Nature Communications* 9 (1): 4127.

Sharakhova, Maria V., Ai Xia, Scotland C. Leman, and Igor V. Sharakhov. 2011. "Arm-Specific Dynamics of Chromosome Evolution in Malaria Mosquitoes." *BMC Evolutionary Biology* 11 (April): 91.

Sharma, V. P. 1999. "Current Scenario of Malaria in India." *Parassitologia* 41 (1-3): 349–53.

Shukla, Vallari, Farhat Habib, Apurv Kulkarni, and Girish S. Ratnaparkhi. 2014. "Gene Duplication, Lineage-Specific Expansion, and Subfunctionalization in the MADF-BESS Family Patterns the Drosophila Wing Hinge." *Genetics* 196 (2): 481–96.

Singh, Om P., Cherry L. Dykes, Manila Lather, Om P. Agrawal, and Tridibes Adak. 2011. "Knockdown Resistance (kdr)-like Mutations in the Voltage-Gated Sodium Channel of a Malaria Vector Anopheles Stephensi and PCR Assays for Their Detection." *Malaria Journal* 10 (March): 59.

Solares, Edwin A., Mahul Chakraborty, Danny E. Miller, Shannon Kalsow, Kate Hall, Anoja G. Perera, J. J. Emerson, and R. Scott Hawley. 2018. "Rapid Low-Cost Assembly of the Drosophila Melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing." *G3* , July. https://doi.org/10.1534/g3.118.200162.

Taracena, Mabel L., Vanessa Bottino-Rojas, Octavio A. C. Talyuli, Ana Beatriz Walter-Nuno, José Henrique M. Oliveira, Yesseinia I. Angleró-Rodriguez, Michael B. Wells, George Dimopoulos, Pedro L. Oliveira, and Gabriela O. Paiva-Silva.

2018. “Regulation of Midgut Cell Proliferation Impacts Aedes Aegypti Susceptibility to Dengue Virus.” *PLoS Neglected Tropical Diseases* 12 (5): e0006498.

Treangen, Todd J., and Steven L. Salzberg. 2011. “Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions.” *Nature Reviews. Genetics* 13 (1): 36–46.

Unckless, Robert L., Andrew G. Clark, and Philipp W. Messer. 2017. “Evolution of Resistance Against CRISPR/Cas9 Gene Drive.” *Genetics* 205 (2): 827–41.

Vatandoost, Hassan, and Ahmad Ali Hanafi-Bojd. 2012. “Indication of Pyrethroid Resistance in the Main Malaria Vector, Anopheles Stephensi from Iran.” *Asian Pacific Journal of Tropical Medicine* 5 (9): 722–26.

Volkenhoff, Anne, Astrid Weiler, Matthias Letzel, Martin Stehling, Christian Klämbt, and Stefanie Schirmeier. 2015. “Glial Glycolysis Is Essential for Neuronal Survival in Drosophila.” *Cell Metabolism* 22 (3): 437–47.

Vontas, John, J-P David, Dimitra Nikou, Janet Hemingway, G. K. Christophides, C. Louis, and Hilary Ranson. 2007. “Transcriptional Analysis of Insecticide Resistance in Anopheles Stephensi Using Cross-Species Microarray Hybridization.” *Insect Molecular Biology* 16 (3): 315–24.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *PloS One* 9 (11): e112963.

Wang, Sibao, André L. A. Dos-Santos, Wei Huang, Kun Connie Liu, Mohammad Ali Oshaghi, Ge Wei, Peter Agre, and Marcelo Jacobs-Lorena. 2017. “Driving Mosquito Refractoriness to Plasmodium Falciparum with Engineered Symbiotic Bacteria.” *Science* 357 (6358): 1399–1402.

Waterhouse, Robert M., Sergey Aganezov, Yoann Anselmetti, Jiyoung Lee, Livio Ruzzante, Maarten J. M. F. Reijnders, Romain Feron, et al. 2020. “Evolutionary Superscaffolding and Chromosome Anchoring to Improve Anopheles Genome Assemblies.” *BMC Biology* 18 (1): 1.

Waterhouse, Robert M., Evgenia V. Kriventseva, Stephan Meister, Zhiyong Xi, Kanwal S. Alvarez, Lyric C. Bartholomay, Carolina Barillas-Mury, et al. 2007. “Evolutionary Dynamics of Immune-Related Genes and Pathways in Disease-Vector Mosquitoes.” *Science* 316 (5832): 1738–43.

Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2017. “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.” *Molecular Biology and Evolution*, December. https://doi.org/10.1093/molbev/msx319.

Wood, Derrick E., and Steven L. Salzberg. 2014. “Kraken: Ultrafast Metagenomic

Sequence Classification Using Exact Alignments." *Genome Biology* 15 (3): R46.
Wyman, Dana, Gabriela Balderrama-Gutierrez, Fairlie Reese, Shan Jiang, Sorena Rahmanian, Weihua Zeng, Brian Williams, et al. 2019. "A Technology-Agnostic Long-Read Analysis Pipeline for Transcriptome Discovery and Quantification." *bioRxiv*. https://doi.org/10.1101/672931.
Yared, Solomon, Araya Gebressielasie, Lambodhar Damodaran, Victoria Bonnell, Karen Lopez, Daniel Janies, and Tamar Carter. 2019. "Insecticide Resistance in Anopheles Stephensi in Somali Region, Eastern Ethiopia." *In Review*.
Zhou, Shanshan, Sarah E. Luoma, Genevieve E. St Armour, Esha Thakkar, Trudy F. C. Mackay, and Robert R. H. Anholt. 2017. "A Drosophila Model for Toxicogenomics: Genetic Variation in Susceptibility to Heavy Metal Exposure." *PLoS Genetics* 13 (7): e1006907.