A near-chromosome level genome assembly of *Anopheles stephensi*

Afiya Razia Chida[1*], Samathmika Ravi[1*], Suvratha Jayaprasad[1*], Kiran Paul[1], Jaysmita Saha[1], Chinjusha Suresh[1], Saurabh Whadgar[1], Naveen Kumar[2], Raksha Rao K[1], Chaitali Ghosh[2], Bibha Choudhary[1], Suresh Subramani[2], Subhashini Srinivasan[12!]

[1]Institute of Bioinformatics and Applied Biotechnology, Biotech Park, Electronic City Phase I, Bengaluru 560100, India
[2]Tata Institute for Genetics and Society Center at inStem – GKVK campus, Bellary Road, Bangalore 560065, India

*These three authors equally contributed to the work.
[!]ssubha@ibab.ac.in (Corresponding Author)

Key Words: homology-based assembly, simulated mate-pair, comparative genomics, genome browser, olfactory receptors, cytochromeP450, SNPs, gene expression profile, developmental transcriptome

Abstract

Mosquito-borne diseases remain a major healthcare risk to growing economies like India. A chromosome-level quality reference genome of *Anopheles stephensi*, the vector causing malaria, is one of the prerequisites for successful management of malaria in India using state-of-the-art genetic technologies, such as gene editing or gene drive. Here, we describe a method to improve the publicly accessible assembly of the genomes of both Indian and Pakistani (SDA-500) strains of *An. stephensi* using a homology-based iterative approach. We have improved the assembly of one strain by using simulated mate-pair reads of incrementally increasing insert sizes from the other strain iteratively. The L50 for the Indian variety was improved from 37 to 9 after the iterations. From the scaffolds adding up to 214 Mb after the iterative approach, 180 Mb were placed onto chromosomes resulting in an assembly with an L50 of 2. The percent Ns (unspecified nucleotides A, C, G or T) for Indian strain remains low at 6.9 percent and based on the gene annotation/synteny, the genome is roughly more than 85% complete. There are 21,378 predicted proteins from IndV3s, of which 12,148 are validated using transcriptome and/or orthology to other proteomes. There are 54 olfactory receptors (ORs) compared to 42 in *An. funestus* and 79 in *An. gambiae*, displaying wider variations in number of ORs across *Anopheles* genera. A limited whole genome sequencing of individuals to 30X coverage from India reveals a genetic diversity of 1 mutation every 68 bases, which is 10 times more than what is reported for *An. stephensi*. A comparative analysis of the genomes of three *Anopheles* species suggests switching of chromosomal arms while displaying limited inter-chromosomal gene flow.

Introduction

There are ~450 recorded species in the genus *Anopheles* (*An.*) with roughly 100 being vectors of malaria in various endemic regions in India, Africa and elsewhere. Among the malaria vectors, only 30-40 transmit parasites of the genus *Plasmodium*. Efforts to obtain reference genome sequences for various vectors that spread malaria by transmitting *Plasmodium* are under way. Chromosome level assemblies of *An. gambiae* and *An. funestus*, both malaria vectors from Africa, have been reported[1,2]. Draft genomes of an *An. stephensi* strain from India[3] and sixteen other *Anopheles* genera[4] have also been reported.

An. stephensi is an urban species of mosquito that is responsible for causing roughly 12% malaria in India. The major genome resource so far available for *An. stephensi* is the draft genome reported in 2014 [3]. In this report, a number of sequencing technologies including Pacbio, 454, Illumina and BAC-end sequencing were used to generate the draft assembly. The 454 reads were sequenced from multiple mate-pair libraries and include 12.2X coverage from single-end reads, 2.2X coverage from 3 kilobase (kb) paired-end reads, 3.4X coverage from 8 kb paired-end reads, and 1.7X coverage from 20 kb paired-end reads. The majority of 454 reads was in the range of 194 to 395 base-pairs (bp) in length. Illumina short reads with a coverage of 86.4X, read length of 101 bp

paired end reads and an average insert size of approximately 200 bp were sequenced. Also, ten cells of PacBio reads using RS1 sequencing for male genomic DNA produced 5.2X coverage with a median read length of 1,295 bp. A hybrid assembly combining 454 and Illumina data was further improved by filling the gaps with error corrected PacBio reads and scaffolding using BAC-ends with an insert size of ~140kb. The resulting assembly contained 23,371 scaffolds spanning 221 Mb, including 11.8 Mb (5.3%) of gaps filled with Ns (unspecified nucleotides). The L50 of the assembly was 37. The N50 scaffold size was 1.59 Mb and the longest scaffold was 5.9 Mb. Considering that the assembly included various sequencing technologies and strategies, the assembly has the least bias from the use of particular technology platforms.

More recently, a draft genome assembly of sixteen diverse species of *Anopheles* mosquitoes was generated, including a strain of *An. stephensi* from Pakistan (SDA-500)[4]. The contigs for the Pakistani strain from this report were produced using short read sequences from Illumina sequencer with libraries ranging from small to medium and large insert sizes with a coverage of about 100X and read length of 101 bp. From this, the scaffold level assembly for SDA-500 strain of *An. stephensi* was reported by assembling reads with ALLPATHS-LG (P)[4]. The study reported an N50 of 1.6 Mb and the L50 of 85.

Yet another valuable resource for *An. stephensi* is the low-resolution physical map of the chromosomes[5].This study reported a physical map consisting of 422 DNA markers hybridized to 379 chromosomal sites of the *An. stephensi* polytene chromosomes providing a resolution of 0.6 Mb. Of these 241 are cDNA markers for which both location and sequences are accessible[3].

Reference-guided improvement of draft genome assembly of individuals from the same species has been used previously. Mate-pair libraries from one *Arabidopsis thaliana* strain were shared across many strains to build super-scaffolds for all individuals[6].Also, assisted assembly of closely related species significantly improved the contiguity of low coverage mammalian assemblies[7]. For example, the draft genomes of four species including bush baby, African elephant, rabbit and guinea pig from the "Mammal24 - 2X" project were built using both human and canine references[7].

Synteny between species has also been exploited to improve draft genomes. Recently, a chromosome level genome of *Lates calcarifer* was assembled from a draft genome using long read sequencing, transcriptome data, optical/genetic mapping and synteny to two closely related seabasses[8]. In yet another report, 16 out of 60 chromosomes of the Tibetan antelope were reconstructed from draft assemblies using its homology to cattle[9]. In fact, using independent mapping data and conserved synteny between the cattle and human genomes, 91% of the cattle genome was placed onto 30 chromosomes[10]. In a review article, synteny has been used to filter, organize and process local similarities between genome sequences of related organisms to build a coherent global chromosomal context[11]. Similarly, the malarial strain, *Plasmodium falciparum*

HB3, was improved using the reference of *P. falciparum* 3D7 combined with an assisted assembly approach that significantly improved the contiguity of the former[7].

Here, we have created a near-chromosome level assembly of an *An. stephensi* Indian strain maintained at Walter Reed (IndV3s) using draft assemblies of two strains downloaded from VectorBase.

Results

Homology-based assembly and pseudomolecule generation

We utilized contigs/scaffolds from draft genomes of two different strains of *An. stephensi* to iteratively improve the draft assemblies of both obtained from public repositories using complementary information from one to the other. Table 1 describes the assembly metrics through the iterative improvement process. The initial assembly of Indian strain (IndV1) downloaded from VectorBase had a L50 of 37 and that for SDA-500 (PakV1) assembly was 85. In the first iteration, the L50 of IndV1 was reduced to 19 (IndV2) with the help of simulated mate-pairs of varying insert sizes from PakV1. During the second iteration, the L50 of PakV1 was reduced to 22 (PakV2) using the simulated mate-pairs of varying insert sizes from IndV2 of the assembly of Indian strain obtained after first iteration. The subsequent iteration resulted in assemblies with L50 of the Indian strain dropping to 9 from 37 and that for the Pakistani strain to 12 from 85. Further iteration did not show significant improvement suggesting saturation of complementing information in the initial draft assemblies of the two strains.

In the absence of HiC and/or optical mapping data, the published physical mapping data was used to order the scaffolds and stitch them into pseudomolecules. It is important to note here that the scaffolds were long enough for super scaffolding using the 230 out of the 241 cDNA markers[5] for which the sequences were available[3]. Out of 230 markers 209 (90%) could be mapped with confidence to scaffolds. Figure 1a shows the overlap between sequences from the physical mapping information for *An. stephensi* and the scaffolds generated from the third iteration of homology-based assembly for both the Indian (IndV3) and Pakistani strains (PakV3) of *An. stephensi*. The almost null overlap of physical maps placed on scaffolds, as shown in Figure 1A, helped in uniquely assigning chromosomes to scaffolds. However, physical markers from multiple arms hit a few scaffolds challenging the assignment of chromosomes for some long scaffolds (see Table 6 under methods). As shown in Figure 1b, the scaffolds have been assigned orientation within the chromosomes, such as “f” (forward) and “r” 12364795(reverse), based on the linearity of physical maps. The equation for stitching scaffolds is elaborated along with the scaffold IDs (the numbers following ‘f’ and ‘r’). The karyogram in Figure 1c shows the final order and orientation of the physical markers on each chromosome after building the pseudomolecules. The

total bases incorporated into pseudomolecules after stitching constituted more than 85% of the estimated genome size excluding Y chromosome.

Strain/ Iterations	**Iteration 1 (IndV1/PakV1)**	**Iteration 2 (IndV2/PakV2)**	**Iteration 3 (IndV3/PakV3)**	**Pseudomolecules (IndV3s/PakV3s)**
Longest scaffold (in Mb)	5.9 /3.4	9.3 /9.8	24.3 /17.1	58 / 62.5
L50	37/85	19 /22	9 /12	2 / 2
N50 (Mb)	1.7/0.8	3.5 /3.2	6.5 /6.9	43 / 51.4
Assembly size (Mb)	210 /228	213 / 230	213 /244	180 / 193.5
Percentage Ns	5.62 /12.95	6.9 /14.8	6.91/19.71	6.44 / 16.93

Table 1: Assembly statistics for Indian and Pakistani strains after every iteration.

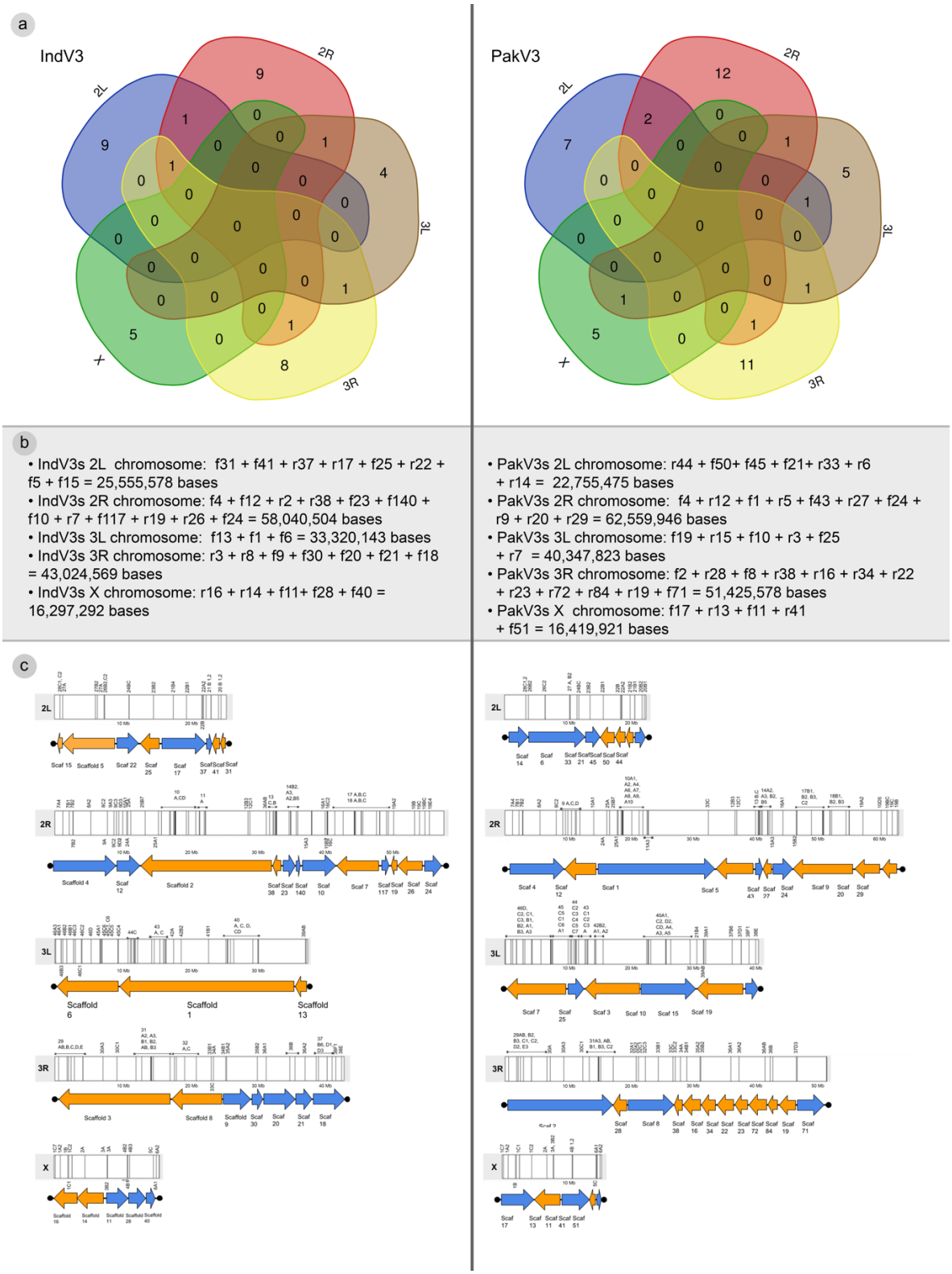

Figure 1: a) Intersection of scaffolds from IndV3 and PakV3 strains generated by homology-based assembly with marker data from physical maps for *An. stephensi*. b) Ordering of scaffolds into five

chromosome arms based on the linearity of physical map markers using position and orientation for IndV3 and PakV3. c) Karyogram showing stitched near-complete chromosomes, IndV3s and PakV3s with DNA marker location on scaffolds from IndV3 and PakV3.

Validation using synteny

The assemblies of IndV3s and PakV3s are validated using synteny to each other and synteny to other completed *Anopheles* genomes. In Figure 2, we show the synteny of IndV3s against PakV3s (Figure 2a), *An. gambiae* (Figure 2b) *and An. funestus* (Figure 2c). The synteny of the X-chromosome and the 3L arm of IndV3s shows the least variation with PakV3s assembly, suggesting that the overall assembly quality is good. The not-so-perfect synteny between the two strains of *An. stephensi* in 2R, 2L and 3R may reveal real differences between the genomes of the two strains and/or potential mis-assembly. Synteny of chromosomal arms 2L and 3L between IndV3s and PakV3s with insertions in IndV3s suggests incomplete assembly of PakV3s as expected because of lack of long reads in the original SDA-500 assembly (Figure 2a). The synteny between chromosomes of IndV3s against *An. gambiae* and *An. funestus* suggests a near complete assembly of IndV3s without major gaps (Figure 2b and 2c). Figure 2 also shows potential intra-species (Figure 2d) and inter-species (Figure 2e and 2f) inversion events.

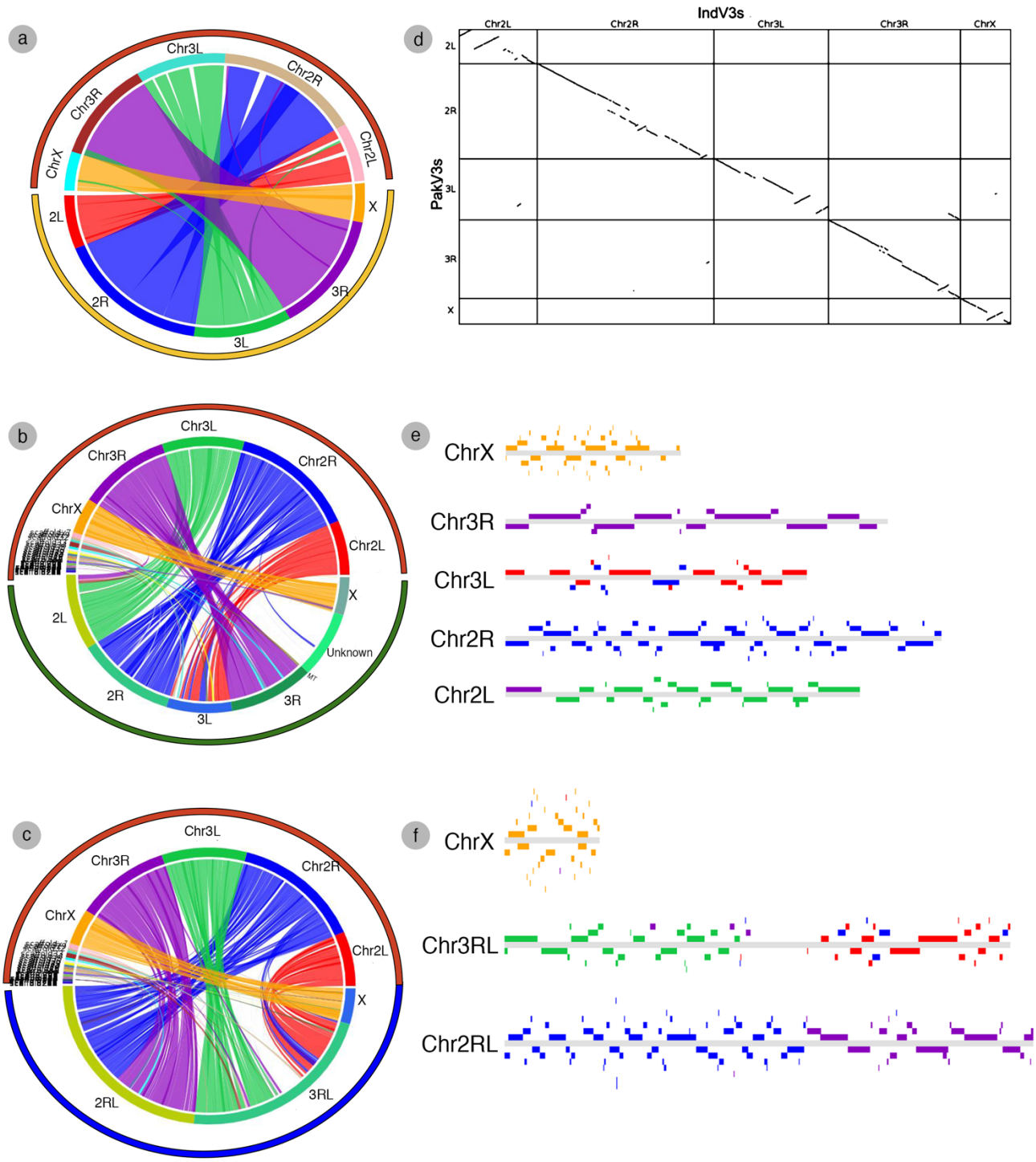

Figure 2: a) Circos plot showing synteny between IndV3s (brown arc) against PakV3s (yellow arc). b) Synteny of IndV3s (brown arc) against *An. gambiae* (green arc) and c) Synteny of IndV3s against *An. funestus* (blue arc). d) represents the dot plot of IndV3s on PakV3s. e) and f) represent block chromosomal diagrams of IndV3s on every chromosome of *An. gambiae* and *An. funestus,* respectively.

The scaffolds classified as unknown in *An. gambiae* genome, which may represent Y-chromosome and/or microbial genomes, is not syntenic with unplaced scaffolds of IndV3s suggesting the possibility of unique microbial compositions in *An. stephensi* compared to *An. gambiae*. However, several unplaced scaffolds from IndV3s are syntenic to *An. gambiae* and *An. funestus* chromosomes, suggesting that they are part of *An. stephensi* chromosomes but could not be placed on chromosomes in IndV3s assembly because of the low-resolution physical map (Figure 2b, 2c). The synteny of IndV3s with the rather high-resolution assembly of *An. funestus* with chromosomal arms assembling into a single scaffold reveals that the centromeric region between the two R and L arms of chromosome 3 is much larger than that of chromosome 2 (Figure 2c). Interestingly, in *An. funestus* the R arm of chromosome 3 maps to 3L arms of *An. stephensi* and 2L of *An. gambiae* respectively. Also, the 2L arm of *An. funestus* is syntenic with 3R arms of both *An. stephensi* and *An. gambiae*. Thus, chromosomal arm switching (twisting) must be common among *Anopheles* genus. However, high synteny between arms show limited gene flow between chromosomal arms.

Physical markers from the end of both 3R and 3L arms hit scaffold18 of length 3.8 Mb of IndV3 and scaffold19 of length 4.2 Mb of PakV3 and hence had to be further resolved using numbers/linearity of markers and level of homology to marker DNAs. After resolving the conflict between 3R and 3L, the scaffolds were assigned to the end of chromosomal arm 3R. Upon creating synteny between scaffold18 with chromosomes 3 and chromosome 2 of *An. funestus* (Figure S1), it is clear that scaffold18 in IndV3 and scaffold19 of PakV3 joins 3R and 3L arms of chromosome 3 of *An. stephensi*.

Completeness and annotation

In this section, we evaluated the overall completeness of the two assemblies IndV3s and PakV3s using diverse approaches. These include i) percentage of physical markers hitting the scaffolds, ii) total predicted proteins, iii) a set of elite proteins from BUSCO, iv) synteny to other *Anopheles* genomes summarized in Table 2. Additionally, we assessed the gene structures for a set of predicted genes from IndV3s to validate the quality of assembly with respect to gene loci (Table 3).

The number of physical markers mapped to each assembly, 87% to IndV3s and 84% to PakV3s, provides a macroscopic measure of the completeness of each chromosome. For example, 15/18 markers (83%) from X chromosome mapped to both IndV3s assembly and 84/88 (95%) of 2R, 37/40 (92%) from 3R, 52/56 (92%) from 3L and 21/28 (75%) markers from 2L were mapped to IndV3s assembly. Furthermore, a total of 180 Mb and 194 Mb of scaffolds from IndV3 (Figure 1b) could be placed on the chromosomes using physical markers to obtain IndV3s and PakV3s assemblies (Table 2).

Gene prediction using Augustus with *Aedes aegypti* as a model, the species closest to *An. stephensi* offered by the tool, resulted in the prediction of 21,378 and 20,083 proteins for IndV3s and PakV3s, respectively. These are comparable to the 20,318 proteins predicted in-house from the high-resolution chromosome-level assembly of *An. funestus* using exactly the same method. The proteins were validated using both publicly available transcriptome data from *An. stephensi* and orthology to the proteomes of other species. For transcriptome-based validation midgut transcriptome and Indian peptides database from VectorBase was used and for orthology, predicted genes from other *Anopheles* genus, Swiss-Prot and TrEMBL was used. Table 2 shows the number of predicted proteins in IndV3s and PakV3s, respectively. Figure 3a and 3b depict the intersection of predicted proteins validated by each database. Of the 21,378 genes predicted for IndV3s, 12,148 are validated using one and/or all the above databases. For PakV3s, 11,303 of the 20,083 proteins predicted by Augustus were validated using these databases. An annotated GTF file is made available via the genome browser site listed under Data Availability section.

Parameter	**IndV3s assembly**	**PakV3s assembly**
Completion based on physical map	87% 21/28 (75%) on 2L, 84/88 (95%) on 2R, 52/56 (92%) on 3L, 37/40 (92%) on 3R, 15/18 (83%) of X	84% 21/28 (75%) on 2L, 74/88 (84%) on 2R, 50/56 (89%) on 3L, 36/40 (90%) on 3R, 15/18 (83%) on X
Predicted proteins	21,378 (20,318 in *An. funestus*)	20,083
Validated proteins	12,148	11,303
BUSCO (out of 3285 *Diptera* specific)	C:88.1% [S:87.9%, D:0.2%] F:1.7%, M:10.2%	C:82.2% [S:79.3%, D:2.9%] F:1.2%, M:16.6%
Total ATGC content based on synteny to *Anopheles* genomes	168/180 million bases (93%)	160/193 million bases (84%)

Table 2: Completeness of assemblies IndV3s and PakV3s using various approaches including physical markers, predicted proteins from Augustus, validated proteins from public databases and ATGC content from synteny analysis. (BUSCO notation, C: Complete BUSCOs, S: Complete and single copy BUSCOs, D: complete and duplicated BUSCOs, F: Fragmented BUSCOs, M: Missing BUSCOs).

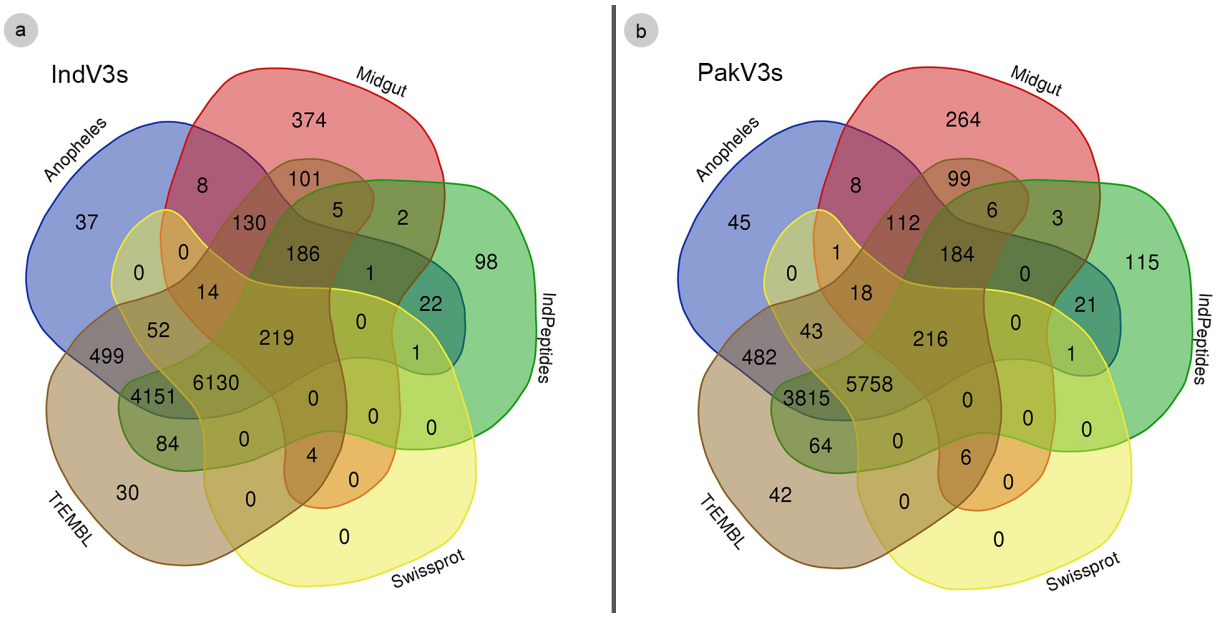

Figure 3: Venn diagrams depicting the intersection of proteins validated by 18 other *Anopheles* proteomes, female midgut transcriptome, Indian peptides gene set, Swiss-Prot and UniProt TrEMBL databases for IndV3s (3a) and PakV3s (3b), respectively.

The completeness of the proteome in IndV3s and PakV3s assemblies are also assessed with BUSCO, which utilizes evolutionarily informed expectations of gene content. Out of the total of 3285 complete single-copy elite marker genes for *Diptera* used by BUSCO, we found 2894 genes (88%) in IndV3s and 2701 (82%) in PakV3s assemblies.

Genes of interest to vector biologists

To assess the completeness of assembly across gene loci, we assessed gene structures of selected predicted genes across IndV3s. The selected genes are of importance based on functional implications in the development of antimalarial strategies (detailed in methods section Table 7) These include genes that are implicated in insecticide resistance, parasitic infection and other DNA break repair mechanisms as listed in Table 3. Based on the well-studied orthologs, we found that the gene structure for all the genes of interest (see Table 3) is predicted at full length. The links listed for Augustus ID in Table 3 can be used to view the gene structure and download sequences. For example, all 20 exons are predicted intact for the *KDR* gene with exon structures and sizes similar to orthologs from other species.

Gene	Function/Role	Chr:Start-End	Strand	Augustus ID
SPO11	involved in double stranded DNA break in early stages of meiotic recombination[12]	Chr2R:12179314-12181513	-	g3704
Cardinal	enzyme involved in ommochrome (eye-pigment) pathway[13]	Chr2R:23870838-23873635	-	g5128
ACE1	target of organophosphate insecticides[14]	Chr2R:31272014-31272014	-	g6054
CYP6P9a/b	responsible for pyrethroid resistance[15]	Chr2R:37541130-37542720	-	g6842
CYP6P4	responsible for pyrethroid resistance[15]	Chr2R:37543818-37551831	-	g6843
RAD51	DNA break repair protein[16]	Chr2R:38454130-38455210	-	g6908
RAB5	endocytic protein implicated in cellular entry of Dengue and West Nile viruses[17]	Chr3R:258801-260265	+	g12024
CYP6M10	responsible for pyrethroid resistance[18]	Chr3R:6197773-6199346	+	g12863
KDR	target for pyrethroid insecticides[19]	Chr3R:41729809-41760265	+	g16696
Kynurenine hydroxylase	1. precursor in the biosynthesis of eye-pigment[20] 2. xanthurenic acid, side product of kynurenine pathway, helps *P. falciparum* reproduce in mosquito midgut[21]	Chr3L:9580284-9583072	-	g18313
FREP1	helps *P. falciparum* invade mosquito midgut[22]	Chr3L:11747343-11804938	+	g18569
MRE11	double strand break repair nuclease[16]	Scaffold 1023: 49-2089	+	g24450

Table 3: Genes of interest from a functional perspective in IndV3s.

Developmental transcriptome analysis

Developmental transcriptome data for various stages from embryonic to adult stages were obtained from public sources (SRP013839, Table 4). The coverage of Illumina data for each stage varied from 10X to 35X and is computed based on the assumption that 20% of the genome is transcribed.

S. No	Samples	Accession ID	Number of reads	Read length (bp)	Sequenced bases (Mb)
1 (EM4-8)	Embryo 4 to 8 hours	SRR514863	8335955	39	323
2 (EM8-12)	Embryo 8 to 12 hours	SRR515304	14390100	38	532
3 (larva)	Larva	SRR515305	14725066	41	574
4 (pupa)	Pupa	SRR515306	12192512	41	500
5 (AdFe)	Adult Female	SRR515307	14263361	41	586
6 (AdMa)	Adult Male	SRR515308	16487004	41	676
7 (PBM)	Ovary post blood meal	SRR515309	16466546	41	676
8 (PEM)	Ovary post emergence	SRR515310	14422518	39	567
9 (FeCa)	Female carcass	SRR515315	11175963	41	457
10 (EM0-1)	Embryo 0 to 1 hour	SRR515316	5524113	41	225
11 (EM2-4))	Embryo 2 to 4 hours	SRR515341	5210129	41	213

Table 4: Transcriptome data summarizing the various developmental stages including embryonic stages 0-1 hour, 2-4 hours, 4-8 hours and 8-12 hours, larva, pupa, adult female, adult male, PBM (ovary Post Blood Meal), PEM (ovary post emergence), FeCa (female carcass).

It is interesting to note that all the genes of interest listed in Table 3 show stage specific expression (Figure 4a). For example, *CYP6M10a, CYP6P4* and *CYP6P9a/b* are highly expressed in the larva stage compared to all other stages (Figure 4a3, 4a4, 4a5). On the other hand, the *KH* gene shows low expression in larva and adult males (Figure 4a8). Genes *ACE1*, *FREP1, KDR*, and *SPO11* show very low overall expression as can be seen by the range in the Y-axis (Figure 4a1, 4a6, 4a7 and 4a12). However, *SPO11* expression is higher in embryonic early stage (Figure 4a12) and in post blood meal ovary. *ACE1* and *KDR* show detectable peaks in adult male (Figure 4a1, 4a7) and larva. Among other interesting findings is that *MRE11* and *RAD51*, both implicated in DNA repair, are upregulated in post blood meal ovary and the profiles of these genes correlate well across stages (Figure 4a9,4a11). The genes *cardinal* and *KH* show negative correlation except in adult female, with *KH* upregulated in PBM and PEM stages and downregulated in adult male (Figure 4a2, 4a8). The gene *Cardinal* is upregulated in adult male and is downregulated in PBM and PEM stages (Figure 4a2).

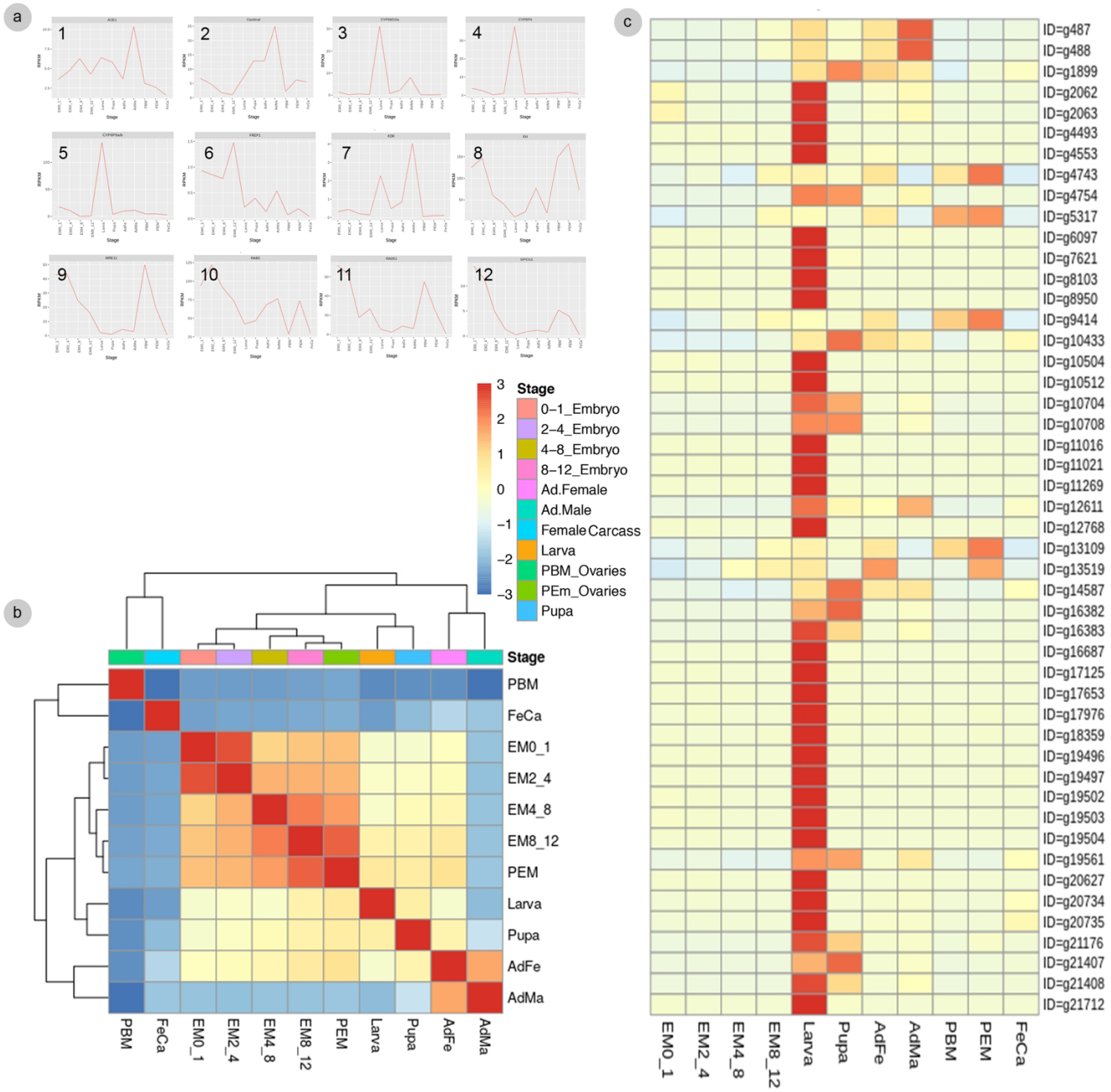

Figure 4: a) Expression profiles of genes of interest listed in Table 3, b) Correlation of genome-wide gene expression between different developmental stages, c) Heatmap of gene expression sorted by larva.

Genome-wide gene expression patterns show that adult females correlate with pupa and larva stages while adult male, PBM and FeCa show no correlation to early developmental stages (Figure 4b). As expected, there is good correlation between various embryonic stages (Figure 4b). Also, the overall gene expression profile of adult male and adult female correlates well without any

correlation to female carcass (FeCa). Genome-wide gene expression profiles also provide opportunities to discover stage-specific novel target genes. For example, Figure 4c shows genome-wide gene expression profile sorted descending order of gene expression (RPKM) in larva, suggesting stage-specific expression of several genes in each stage. Gene expression data sorted by other stages are provided in a link listed under Data Availability.

Evolution of Olfactory receptors (ORs)

Olfactory receptor genes (ORs) are of interest from an evolutionary point of view in vectors because of their role in host choice and therefore disease transmission. The number of ORs in different vectors vary widely. *Ae. aegypti* has 110 ORs[23] whereas *An. gambiae* encodes for only 79 ORs [24]. The total number of OR proteins predicted from IndV3s that are orthologous to *An. gambiae* ORs is 54. Using a similar approach, we found 42 unique ORs from the recently reported high-quality genome of *An. funestus*[2]. The protein sequences of the 54 ORs from *An. stephensi* and 42 from *An. funestus* (Text S1) are used to create a phylogenetic tree (Figure 5), which shows the orthology of the 54 *An. stephensi* ORs with respect to 42 from *An. funestus*. The corresponding orthologs of *An. gambiae* ORs are in purple at the branch points. Figure S2 shows a phylogenetic tree including orthologs of ORs from all three *Anopheles* species. Table S1 shows the BLAST-based orthology used in decorating the tree in Figure 5 with *An. gambiae* accessions, which is consistent with the ClustalW based tree presented in Figure S2.

The lower number of total ORs predicted from both *An. stephensi* and *An. funestus* is clearly at odds with the level of completion of both the genomes. Furthermore, in all three species some orthologous ORs appear in tandem in the genome, as reported previously[3]. For example, the 3 ORs from *An. stephensi* with accession IDs of i) g13438 with orthology to OR-47 of *An. gambiae* and 8961 of *An. funestus*, ii) g13438_2 and iii) g13438_3 are in tandem in the genome reported here. Orthologs for g13438_2 and g13438_3 from IndV3s are missing in *An. funestus* but expanded into six ORs in *An. gambiae* with accessions of 13, 15, 16, 17, 55 (see green dot in Figure 5). On the other hand, the orthologs of ORs 46 and 47 from *An. gambiae* are missing in *An. stephensi* but are present in tandem in *An. funestus* (cyan dot in Figure 5) with accessions Afun_g15671 and Afun_15671_1. The orthologs of *An. stephensi* with accessions of g15209_2, g15210 and g15212 (orange dot in Figure 5) are missing in *An. funestus* but are collinear orthologs of 6 ORs in *An. gambiae* with accessions ranging from 69 to 74 (see highlighted in orange in Figure. S2). There are orthologs for 5 *An. gambiae* ORs with accession IDs of 12, 19, 20, 21, 50 missing in both *An. stephensi* and *An. funestus*, which cluster into a group (see highlighted in green in Figure. S2) and is also reported missing in *D. melanogaster*[24]. While this could result from potential incomplete genomes of *An. stephensi*, considering that the *An. funestus* reference genome is of very high-resolution, it is unlikely that the orthologs to the same genes would be missing in the two independently reported near completed genomes and hence can be considered unique to *An. gambiae*.

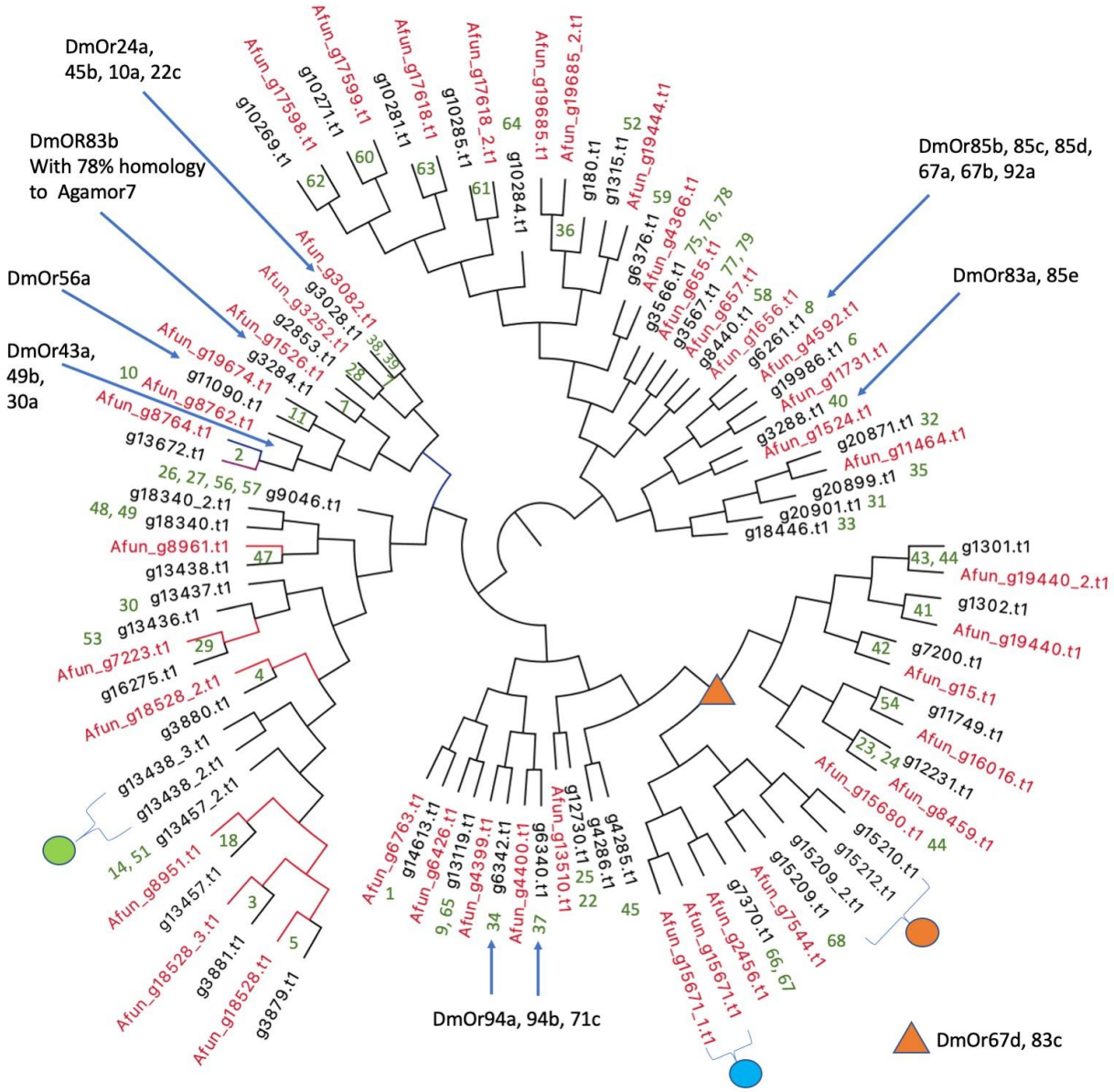

Figure 5: Phylogenetic relationships of the 54 ORs of *An. stephensi* (black) predicted from IndV3s with the 42 ORs predicted for *An. funestus* (red). Embedded numbers in green are the accession IDs of *An. gambiae* ORs. Green dot represents ORs missing in *An. funestus* and expanded into 5 ORs in *An. gambiae* including accessions 13, 15, 16, 17, 55. Cyan dot represents ORs missing in *An. stephensi*, duplicated in tandem in *An. funestus* and duplicated in *An. gambiae* with accessions 43-44. Orange dot is missing in *An. funestus* and expanded into 6 ORs in *An. gambiae* with accession IDs of 69-74. Blue arrows and orange triangle show ORs of *D. melanogaster* based on their orthology to *An. gambiae* reported elsewhere[24].

Genetic diversity among individuals using whole genome sequencing

We sequenced at 30X coverage, 6 lab female individuals from one urban city and 1 male individual from another city to estimate the diversity among individuals compared to IndV3s reference. The number of variants in these individuals with respect to IndV3s reference is roughly 2.6 million suggesting an overall SNP density of 1 every 68 bases of the genome compared to 1 in 1,000 for the human genome. This is ten times higher than the number of SNPs (319,751) reported for *An. stephensi* using diverse individuals from a lab strain, which may be because the individuals were taken from the same colony that was used in creating the reference[3]. The X chromosome has the lowest SNP density with 1 in 85, which is in tune with what is reported for *An. gambiae* and *An. stephensi* [3,25]. In Table 5, we show SNP density for X chromosome from both female and male individuals to make sure the lower SNP density in X is not resulting from lower coverage of X from hemizygous male. In females the SNP density is at 1 in 85 bases and is only slightly higher than the SNP density of 1 in 114 bases in male.

Sample	**Total number of SNPs**	**Number of SNPs in ChrX**	**Number of SNPs in Chr2L**	**Number of SNPs in Chr2R**	**Number of SNPs in Chr3L**	**Number of SNPs in Chr3R**
MQ1 (female)	2,583,297	172,357	389,985	882,228	545,816	592,911
MQ2 (female)	2,587,876	204,980	394,843	796,302	606,930	584,821
MQ3 (female)	2,661,744	201,570	398,831	845,061	595,744	620,538
MQ4 (female)	2,743,999	212,733	413,882	817,748	649,369	650,267
MQ5 (female)	2,776,507	180,361	432,634	872,075	673,586	617,851
MQ6 (female)	2,569,980	221,061	420,930	698,455	635,528	594,006
Mos3 (male)	2,618,085	142,454	400,325	868,274	585,467	621,565
Average	**2,648,784**	**190,788**	**407,347**	**825,735**	**613,206**	**611,708**
No. of Genes	**21,378**	**1,962**	**2,851**	**7,165**	**4,493**	**4,907**
Size of IndV3s (bp)	**180,469,025**	**16,301,284**	**25,562,564**	**58,051,482**	**37,523,138**	**43,030,557**
SNP density (1/bp)	**1/68**	**1/85**	**1/63**	**1/70**	**1/61**	**1/70**

Table 5: SNP statistics of 7 *An. stephensi* individuals from whole genome sequencing.

Discussion

Here, we report a near-chromosome-level assembly of *An. stephensi*, a major malaria causing vector in urban India. Draft assemblies of two strains of *An. stephensi* were obtained from VectorBase and their complementarity was used to improve the assembly of each strain using simulated mate-pairs from the other in an iterative fashion. The L50 for the Indian strain improved from 37 to 9 after iterative improvement and to 2 after building pseudomolecules. The longest scaffolds increased in lengths from 6 Mb to 24 Mb to 58 Mb for IndV3s. Considering that the genome size is roughly 225 million bases with 3 chromosomes, the assembly statistics achieved are good enough to be validated and stitched into chromosomes using physical markers[5]. The total genome size after stitching is 180 Mb and 193 Mb for Indian and Pakistani strains, respectively, covering more than 80% of the estimated genome size. Considering that there is a low level of representation from the heterochromatin region of the genome and that the Y chromosome remains unassembled, placing the scaffolds with lengths adding up to 180 million bases, covering 80% of the estimated genome size, onto chromosomes suggests much higher representation of euchromatin regions of the chromosome assembled here. Furthermore, the lower percentage in the assembled genome is also resulting from the low resolution of the physical mapping data, which may be missing in shorter scaffolds (Figure 2b). The linearity of the final chromosomes was validated using synteny to each other and to the chromosomes of *An. gambiae and An. funestus*. The full synteny between the X chromosomes of IndV3s and PakV3s suggests potentially fewer mis-assemblies. There are no major gaps in the synteny between IndV3s and chromosome-level genomes of *An. gambiae* and *An. funestus* (Figure 2b/2e and 2c/2f) suggesting near-chromosome level assembly of IndV3s. Also, synteny of IndV3s to *An. gambiae* and *An. funestus* showed the least shuffling of genes between the chromosomal arms of the three malaria vectors compared here. However, arm switching appears to be common among species under *Anopheles* genus.

We used multiple approaches to assess the completeness and usefulness of IndV3s and PakV3s assemblies reported here. These include assessing i) the percentage of physical maps located on the chromosomes, ii) the percentage of predicted genes validated via databases, iii) gene structure validation using genes of interest, and iv) synteny with other *Anopheles* genomes. Using Augustus gene prediction software with *Aedes aegypti* as a training model, we were able to predict 21,378 genes in IndV3s and have validated 12,148 using multiple databases including transcriptome data from the same species and proteome data from other species. The number of predicted and validated proteins from IndV3s is in line with the corresponding numbers for *An. funestus* with 20,316 predicted genes, the most closely related to *An. stephensi*[26] for which the genome is recently deciphered. To assess the robustness of the assembly around gene loci, we looked for completeness of a set of functional genes of interest to vector biologists. All selected genes were predicted full length as validated by the gene structures of the orthologs from other well studied species. Almost all the genes showed conservation in the size of orthologous exons within the folding domain.

Table 3 provides direct links to these gene structures on the browser (link under Data Availability section).

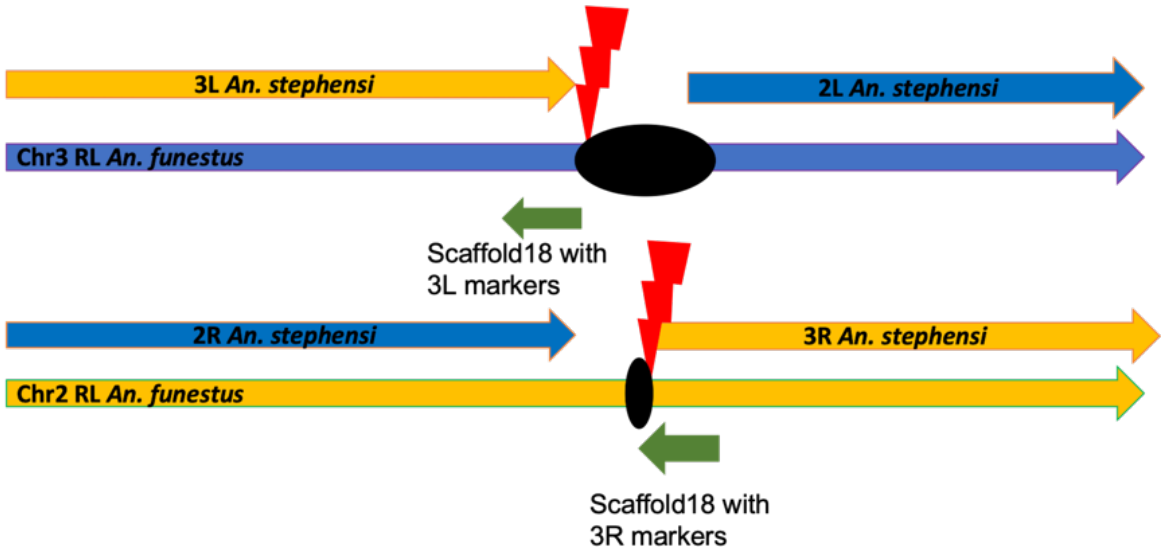

Figure 6: Arms of *An. stephensi* chromosomes syntenic with arms of *An. funestus* chromosome. Lightning rods in red are the potential breakpoints and the ellipses in black are the centromeres.

The end of chromosome 3R of IndV3s near the centromere built using scaffold18 needs special mention. The scaffold18 of IndV3 of length 3.8 Mb and the scaffold19 of PakV3 of length 4.2 Mb could not be uniquely assigned to chromosomal arms based on physical markers. After resolving conflict, the scaffolds were assigned to the end of chromosomal arm 3R. A synteny between scaffold18 of IndV3 with the two completed chromosomes of *An. funestus* (Figure S1) shows that the first half of scaffold18 is syntenic to middle of chromosome 3 and the second half to middle of chromosome 2 of *An. funestus*. Since the 2L arm of *An. funestus* is syntenic to 3R arm of *An. stephensi* and 3R arm of *An. funestus* is syntenic with 3L of *An. stephensi* (see table in Figure S1), scaffold18 of IndV3 and scaffold19 of PakV3 is suspected of spanning the breakpoint/centromere between the two arms of chromosome 3 of *An. stephensi*. Considering that 3R and 2R arms of *A. stephensi* is syntenic to chromosome 2 of *An. funestus* with very small centromere as revealed by lack of gap in the synteny (Figure 2c and 2f), we hypothesize that the centromere for chromosome 3 of *An. stephensi* is short enough for scafold18 of IndV3 and scaffold19 of PakV3 to span across without the need for HiC data (Figure 6). This also explains why markers from both 3L and 3R arms were hitting these scaffolds. Thus, chromosome 3 could have been stitched full length with both arms connecting using scaffold18 using the approach presented here. However, in the assembly submitted to NCBI this scaffold is stitched at the end of arm 3R reaching out to 3L.

The evolution of olfactory receptors is of interest among blood feeding mosquitoes. We have identified 54 ORs from the IndV3s assembly and 42 ORs from the recently reported genome of

An. funestus using the ORs of *An. gambiae* [24]. These numbers are much lower than the 79 ORs predicted for *An. gambiae*. Considering the level of completion of the genomes of *An. stephensi* and *An. funestus* this steep drop in the number of ORs predicted in these two species may suggest active/convergent evolution among OR genes. Also, the multiple OR genes that are found in tandem within the genomes of all three anopheles species with some tandem OR cluster from one species orthologous to tandem OR cluster in the other, suggests multiple tandem duplication and expansion of OR genes in *Anopheles* genera, as reported for salivary genes[3]. The most conserved OR from *An. gambiae* with an accession of GPRor7 (Agamor7) to the OR from *D. melanogaster* with an accession of DmOr83b is also conserved in both *An. stephensi* (g3284) and *An. funestus* (AFun_g1526) suggesting a conserved function within *Diptera*. These are the olfactory coreceptor (orco), which constitutes a heterodimer with other ORs helping the odor molecules to enter into the insect body by forming a ligand-gated ion channel [27]. Also, the branch marked with orange triangle in Figure 5 with two orthologs in *D. melanogaster* has expanded to include 17 orthologs of *An. gambiae* and 10 orthologs for both *An. stephensi* and *An. funestus*, suggesting expansion of this function in anopheles genera.

A chromosome-level genome also provides a great resource to study genome-wide gene expression across tissues and/or developmental stages. Genome-wide gene expression analysis is also an important source to clone full-length genes of interest for a non-model organism. We have taken advantage of the completed genome to profile genome-wide gene expression from early embryonic stages to adult mosquitoes using publicly available data. As shown in Figure 4b there is a good correlation between various embryonic stages including PEM. Also, while there is good correlation between adult male and female, only adult female retain correlation with all early developmental stages. Not surprisingly, there is no correlation between PBM and FeCa, as FeCa is carcasses left after removing post blood meal ovary (PBM). We observe that the selected metabolic genes of interest, implicated in pyrethroid resistant insects, *CYP6M10a*[28], *CYP6P4*[29] and *CYP6P9a/b*[30] are highly expressed in the larva stage compared to all other stages (Figure 4a3, 4a4, 4a5). This is expected because the majority of insecticides are targeted to control larvae. Also, *ACE1* and *KDR* genes, which are implicated in insecticide resistance, is also differentially upregulated in larva with the exception of adult male where it is highly expressed (Figure 4a1 and 4a7). Interestingly, mutations in *KDR* and *ACE1* genes have been shown to cost male fitness [31,32]. The expression profile of *SPO11* gene, with a role in meiosis, is detectably higher in embryonic early stage (Figure 4a12). *Cardinal* and *KH* genes, with the exception in adult female, display inverse correlation across developmental stages. On the other hand, the expression profiles of *MRE11* and *RAD51* genes (Figure 4a9 and 4a11), known to form complexes, are highly correlated with high expression in post blood meal ovary (PBM), perhaps suggesting a role in plasmodium infection.

We believe that this is the first instance to obtain a chromosome level assembly from two draft assemblies of the same species using low-resolution physical marker data. This is also the first

chromosome-level assembly of a malaria vector from India. The assembly reported here offers a genomic context to genes and other genetic elements providing a framework for controlling malaria in India with state-of-the-art technologies such as gene editing and gene drive.

Methods and Materials

Sources of data used in this work

The draft assemblies of both the Indian (IndV1) and Pakistani (PakV1) strains of *An. stephensi* were downloaded from https://www.vectorbase.org/. The sequences are also available on GenBank under the accession ALPR00000000.

The transcriptome data was obtained from public sources (NCBI SRA database: SRP013839), The sequencing coverage of Illumina data for each stage varied from 10X to 35X computed based on 20% of the genome being transcribed. The different developmental stages include embryonic stages 1 hour (EM0-1), 4 hours (EM2-4), 8 hours (EM4-8) and 12 hours (EM8-12, larvae, pupa, adult female (AdFe), adult male (AdMa), PBM (Ovary post blood meal), PEM (Ovary post emergence), FeCa (female carcass).

Individual insects for sequencing was originally maintained at NIMR, Bangalore and were obtained from Dr. Susanta K. Ghosh (see acknowledgement).

Homology-based assembly

The *An. stephensi* genomes of Indian strain (IndV1) and Pakistani strain SDA-500 (PakV1) were downloaded from https://www.vectorbase.org/.
Mate-pairs libraries containing reads carry information about the arrangement/ordering of the chromosomes, which help to build scaffolds from the first pass assembly that is obtained from deep sequencing of paired-end reads. This implies that the sequence information from the mate-pair reads are NOT directly used in the assembly, but only the knowledge of ordering information is inherited.

a) Generation of read libraries

Using the samtools wgsim (version 1.9), mate-pair libraries with varying insert sizes (1 kb, 5 kb, 10 kb, 50 kb, 100 kb, 500 kb, 1 Mb, 2 Mb, 5 Mb) and read length of 50 bp were simulated from PakV1 with flags set for zero base error rate, rate of mutation and incorporation of indels.

b) Confident set of reads for subsequent scaffolding

To assess the quality of the reads generated from each library, they were mapped back against *An. stephensi* IndV1 genome using bowtie2 (version 2.3.5.1)[33]. Flags to suppress discordant and unpaired alignments were set. While processing the alignments, only paired reads mapped in the

right orientation with the correct insert sizes were used for downstream analyses. Care was taken to remove multiple mapped reads from the concordant list of fastq reads.

c) Scaffolding of pre-assembled contigs

Reads from b) were used for super-scaffolding of the IndV1 genome to get the IndV2 assembly using the SSPACE tool (version 3.0)[34]. Now, from the IndV2 assembly, mate-pairs with varying library sizes (1 kb, 5 kb, 10 kb, 50 kb, 100 kb, 500 kb, 1 Mb, 2 Mb, 5 Mb) were simulated again to get the reads concordantly mapping to PakV1 assembly. These reads were used to super-scaffold PakV1 to obtain the PakV2 assembly. This iterative procedure was continued till the L50 and N50 of the assemblies did not improve significantly with further iterations resulting in assemblies IndV3 for the Indian strain and PakV3 for the Pakistani strain (shown in Figure 7).

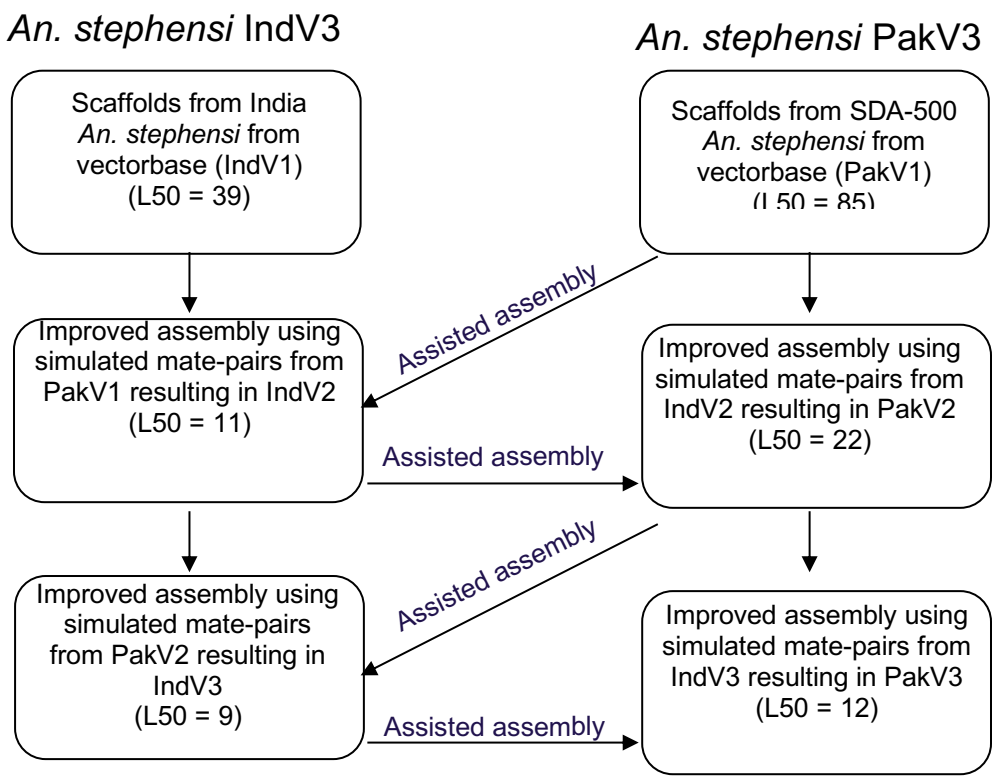

Figure 7: Schematic diagram showing the iterative approach followed to improve the assemblies of both strains of *An. stephensi*.

Pseudomolecule generation using physical marker data

DNA sequences for physical map data for all chromosomal arms were downloaded from the supplementary material of *An. stephensi* draft assembly paper [3]. These were subjected to analysis using BLAST under stringent conditions (blastn -query 3L.fa -db indV3 -outfmt 6 -evalue 1.0e-25 -max_target_seqs 1 -out 3L_indV3.out6) against the assembled genome IndV3. The unique hit of the physical markers onto scaffold IDs was used as a measure of the quality of assembly at this

stage. A Venn diagram was generated to check the assignment of physical markers to scaffolds of the assembly (Figure 1). The scaffold hitting against markers from two chromosomes is resolved using the number and linearity of markers, and in some cases using the level of homology at the DNA level. Chromosomes were put together after orienting the scaffolds using the order of the physical markers. 1000Ns were inserted to separate different scaffolds while stitching. Scaffolds in Table 6 are those with doubtful orientation or overlapping markers.

The 5 chromosomal arms for the Indian strain are given below for a total genome size of 180 Mb, covering more than 80% of the estimated genome size.

- IndV3s 2L chromosome: f31 + f41 + r37 + r17 + f25 + r22 + f5 + f15 = 25,555,578 bases
- IndV3s 2R chromosome: f4 + f12 + r2 + r38 + f23 + f140 + f10 + r7 + f117 + r19 + r26 + f24 = 58,040,504 bases
- IndV3s 3L chromosome: f13 + f1 + f6 = 33,320,143 bases
- IndV3s 3R chromosome: r3 + r8 + f9 + f30 + f20 + f21 + f18 = 43,024,569 bases
- IndV3s X chromosome: r16 + r14 + f11+ f28 + f40 = 16,297,292 bases

The 5 chromosomal arms for Pakistani strain are given below for a total genome size of 198.14 Mb, again covering more than 80% of the estimated genome size.

- PakV3s 2L chromosome: r44 + f50+ f45 + f21+ r33 + r6 + r14 = 22,755,475 bases
- PakV3s 2R chromosome: f4 + r12 + f1 + r5 + f43 + r27 + f24 + r9 + r20 + r29 = 62,559,946 bases
- PakV3s 3L chromosome: f19 + r15 + f10 + r3 + f25 + r7 = 40,347,823 bases
- PakV3s 3R chromosome: f2 + r28 + f8 + r38 + r16 + r34 + r22 + r23 + r72 + r84 + r19 + f71 = 51,425,578 bases
- PakV3s X chromosome: f17 + r13 + f11 + r41 + f51 = 16,419,921 bases

Physical map conflict resolution

The scaffolds that were mapped to markers from multiple chromosomes were assigned to unique chromosomes based on the number of mapped markers, percent identity, e-value and alignment lengths. For example, Table 6 shows scaffolds of IndV3s (ScaffoldID) assigned to markers from multiple chromosome arms (ChromosomeID) and that were resolved as shown in column Assignment.

Assembly	ChromosomeID	ScaffoldID	Size (bp)	Assignment
IndV3	3L or 3R	Scaffold18	3,883,490	3R
IndV3	2L or 2R	Scaffold12	4,818,062	2R
IndV3	2L or 2R	Scaffold2	18,136,279	2R
PakV3	3L or X	Scaffold3	10,618,294	3L

PakV3	3L or 3R	Scaffold19	4,266,331	3L and 3R
PakV3	2R or 3R	Scaffold8	8,712,747	3R
PakV3	2L or 3L	Scaffold10	8,043,466	3L
PakV3	2L or 2R	Scaffold12	5,110,278	2R
PakV3	2L or 2R	Scaffold1	15,821,620	2R

Table 6: Scaffolds of IndV3 and PakV3 which show the presence of physical markers from both the chromosomal arms present in column 'ChromsomeID' required resolution and have been designated a chromosome in column 'Assignment'.

Gene Annotation

AUGUSTUS (version 3.2.3)[35], a eukaryotic gene prediction tool, was used to find protein-coding genes in the IndV3s genome. AUGUSTUS uses *ab initio* gene prediction and reconciles the predicted gene structures with orthology to the proteome from a model organism. The model organism closest to *An. stephensi* of that was made available by AUGUSTUS for use in gene prediction is *Aedes aegypti.* AUGUSTUS provides a gff3 file delineating the exon-intron boundaries for each predicted gene along with its protein sequence. This gff file is uploaded to the browser, the link to which can be found under the Data Availability section. The same gff3 file is used to compute gene expression profiles of all these genes across developmental stages.

Identification of genes of interest

We used orthologs of the genes of interest from public databases by searching for the gene sequences from the closest available species. For the majority of genes, sequences were available in *An. gambiae*. The table below shows the accession and organism used in the identification of these genes from *An. stephensi*.

Gene of Interest	**Source**	**Accession ID**	**Species**
SPO11	UniProt	A0A182YKK2	*Anopheles stephensi*
Cardinal	UniProt	Q9VCW2	*Drosophila melanogaster*
ACE1	UniProt	Q869C3	*Anopheles gambiae*
CYP6P9a/b	UniProt	B5AII6	*Anopheles funestus*
CYP6P4	UniProt	Q7QCY8	*Anopheles gambiae*
RAD51	UniProt	F5HKE1	*Anopheles gambiae*

MRE11	UniProt	Q7QID8	*Anopheles gambiae*
KH	UniProt	Q7Q6A7	*Anopheles gambiae*
RAB5	UniProt	A1YSB1	*Anopheles gambiae*
CYP6M10	UniProt	Q16WQ8	*Aedes aegypti*
KDR	UniProt	Q0N3S3	*Anopheles funestus (fragment)*
FREP1	Niu et al 2017 [36]	-----	*Anopheles stephensi*

Table 7: Online source of data for genes of interest identified in *An. stephensi.*

Finding orthologs of ORs

To find the ORs from *An. stephensi* and *An. funestus*, fasta sequences of 79 OR genes reported for *An. gambiae* is used [24]. Here, protein blast (blastp) v2.7.1 was used for identifying orthologous genes from the predicted proteomes of *An. stephensi* and *An. funestus* using the 79 OR protein sequences from *An. gambiae* as query. The top hits are given in Table S2. The amino acid sequences of ORs from *An. stephensi* and *An. funestus* were extracted from the respective proteomes and tandem ORs were manually split and provided in Text S1. Using clustal 2.1 tool, a multiple sequence alignment of ORs from only *An. stephensi* and *An. funestus* were generated, because these two species are more closely related to each other than *An. gambiae* [26]. Using Figtree (version 1.4.4) we created phylogenetic trees shown in both Figure 5 and Figure S2.

Transcriptome analysis

Developmental transcriptome data was downloaded from NCBI SRA (SRP013839). Mapping was done using transcriptome reads on IndV3s using STAR aligner (version 2.6). Read counts mapped per gene were obtained and normalized RPKM was computed providing a normalized expression profile for all predicted genes across the developmental stages. The genes of interest were taken out and used for plotting the line graph as well as in the heatmap. The entire transcriptome analysis was done using R programming. Packages used in the analyses are pheatmap, stats, corrplot, knitr, ggplot2, dplyr, tidyr, reshape2. A description of the functionality of each tool is given below. For further details on the script and version of packages used, we provide a link to markdown transcriptome report (URL under Data Availability section).

R version 3.6.3 (2020-02-29)
Platform: x86_64-pc-linux-gnu (64-bit)
Running: Ubuntu 18.04.4 LTS

Attached base packages:
stats, graphics, grDevices utils, datasets, methods, base

Other attached packages along with the version
edgeR_3.28.1, limma_3.42.2, reshape2_1.4.3, tidyr_1.0.2, dplyr_0.8.4, ggplot2_3.3.0 knitr_1.28, corrplot_0.84, pheatmap_1.0.12

Packages used for Rmarkdown along with the version
Compiler_3.6.3, htmltools_0.4.0, tools_3.6.3, yaml_2.2.1, Rcpp_1.0.4.6, rmarkdown_2.1 knitr_1.28, xfun_0.12, digest_0.6.25, packrat_0.5.0, rlang_0.4.5, evaluate_0.14

edgeR is a Bioconductor software package for examining differential expression of count data. Here we have used inbuilt RPKM function of edgeR, package [37]

Reshape package was used to Convert an object into a molten data frame which was further used to plot using ggplot package

ggplot is used to create visualization for given data such as a heatmap, line chart etc.

Pheatmap was used to create heatmap for the transcriptome data. The heatmap given in supplementary material generally top 100 as the choice range of the data.

Tidyr and dplyr are the data wrangling packages used for data handling and to filter the data based on the condition.

Corrplot is used for plotting the correlation plot between the sample based on read counts.

Knitr is used along with the htmltools, rmarkdown to create HTML output for RCode

Isolation of Genomic DNA

Briefly, the insect tissue was homogenized, and the genomic DNA was isolated using Qiagen Genomic-tp (Qiagen). Later, fluorometric quantification of DNA was done using Qubit 2.0 (Invitrogen).

Library preparation and sequencing

Whole Genome DNA libraries with an average insert size of 200 bp were made using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (New England Biolabs, 2016) using the protocol

recommended by the company. Briefly, around 50 ng of DNA was used for library preparation, DNA was sheared using Adaptive Focused Acoustic technology (Covaris, Inc.) to generate fragments of length around 200 bp. The fragments were end repaired, 3′-adenylated, ligated with Illumina adapters, and PCR enriched with Illumina sequencing indexes. The size selection was performed using solid-phase reversible immobilization (SPRI) beads (Agencourt AMPure XP Beads) from Beckman Coulter. The quality and quantity of the libraries were evaluated using Qubit (Invitrogen) and TapeStation (Agilent). The libraries were diluted and pooled with an equimolar concentration of each library. Cluster generation was done using cBot (Illumina) and paired-end sequenced on Illumina HiSeq 2500 platform using TruSeq SBS Kit v3-HS (200 cycle) (Illumina, San Diego, CA) following the manufacturer's recommendations**.**

SNP analysis for genetic diversity

DNA samples for 7 individuals were extracted and sequenced for 30X coverage. Variant calling was performed using an in-house pipeline, which uses fastqc (version 0.11.5), bowtie2 (version 2.3.5.1), samtools (version 1.9), picard-tools (version 2.18.12), samtools (version 1.9) mpileup [38] and bcftools (version 1.9) to extract variants. The resulting vcf files are filtered using bcftools (version 1.9) [39] for variants above a quality of 10 and a minimum depth of 3 reads.

JBrowse and Database

The entire web framework of the website was taken from MEGHAGEN LLC, which includes the pre-build platform with the capability of genome visualization, blast, and further bioinformatic analysis. The website is built using HTML, CSS & javascript. Current website is hosted on amazon cloud running linux EC2 instance provided by Meghagen LLC. It uses inbuilt servers (Apache, Shiny) to establish user side connections.
JBrowse is open source, fast and full-featured genome browser built with JavaScript and HTML5. It is easily embedded into websites or apps but can also be served as a standalone web page[40]. JBrowse can utilize multiple types of data in a variety of common genomic data formats, including genomic feature data in bioperl databases, GFF files, BED files, and quantitative data in wiggle files[41].

Server version: Apache/2.4.29 (Ubuntu)
Server built: 2020-03-13T12:26:16
JBrowse v1.16.8

Data Availability

The improved IndV3s and PakV3s assemblies for India and SDA-500 are available on NCBI BioProject under the project ID PRJNA473393. Other metadata (annotation files) reported here can be found at http://3.93.125.130/indv3s

JBrowse, Download genome data	**Scripts, transcriptome data**
http://3.93.125.130/indv3s Download GTF	https://github.com/SaurabhWhadgar/indv3stranscriptome

Acknowledgments

TIGS India for funding researchers involved in this work and for funds for sequencing.
Government of Karnataka for funding the computational infrastructure at IBAB.
IBAB facility for providing computational infrastructure used in all data analysis reported here.
The authors wish to acknowledge Dr. Bhagyasree Kaduskar for help validating genes of interest, Dr. Sonia Sen for initiating work on olfactory receptors, Dr. Pushpinder Singh Bawa for library preparation and Dr. Susanta K. Ghosh, Senior Scientist (retd) of NIMR, Bangalore for providing the mosquitoes.

References

1. Sharakhova, M. V. *et al.* Update of the Anopheles gambiae PEST genome assembly. *Genome Biol.* **8**, R5 (2007).
2. Ghurye, J. *et al.* A chromosome-scale assembly of the major African malaria vector Anopheles funestus. *Gigascience* **8**, 1-8 (2019).
3. Jiang, X. *et al.* Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol.* **15**, 459-477 (2014).
4. Neafsey, D. E. *et al.* Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* **347**, 1258522 (2015).
5. Sharakhova, M. V. *et al.* A physical map for an Asian malaria mosquito, Anopheles stephensi. *Am. J. Trop. Med. Hyg.* **83**, 1023–1027 (2010).
6. Schneeberger, K. *et al.* Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10249–10254 (2011).
7. Gnerre, S., Lander, E. S., Lindblad-Toh, K. & Jaffe, D. B. Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biol.* **10**, R88 (2009).
8. Vij, S. *et al.* Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS Genet.* **12**, e1005954 (2016).

9. Kim, J. *et al.* Reference-assisted chromosome assembly. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1785–1790 (2013).
10. Zimin, A. V. *et al.* A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol.* **10**, R42 (2009).
11. Batzoglou, S. The many faces of sequence alignment. *Brief. Bioinformatics* **6**, 6–22 (2005).
12. Keeney, S. Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis. *Genome Dyn Stab* **2**, 81–123 (2008).
13. Ferré, J., Silva, F. J., Real, M. D. & Ménsua, J. L. Pigment patterns in mutants affecting the biosynthesis of pteridines and xanthommatin in Drosophila melanogaster. *Biochem. Genet.* **24**, 545–569 (1986).
14. Essandoh, J., Yawson, A. E. & Weetman, D. Acetylcholinesterase (Ace-1) target site mutation 119S is strongly diagnostic of carbamate and organophosphate resistance in Anopheles gambiae s.s. and Anopheles coluzzii across southern Ghana. *Malar. J.* **12**, 404 (2013).
15. Wondji, C. S. *et al.* Two duplicated P450 genes are associated with pyrethroid resistance in Anopheles funestus, a major malaria vector. *Genome Res.* **19**, 452–459 (2009).
16. Overcash, J. M., Aryan, A., Myles, K. M. & Adelman, Z. N. Understanding the DNA damage response in order to achieve desired gene editing outcomes in mosquitoes. *Chromosome Res.* **23**, 31–42 (2015).
17. Krishnan, M. N. *et al.* Rab 5 is required for the cellular entry of dengue and West Nile viruses. *J. Virol.* **81**, 4881–4885 (2007).
18. David, J.-P., Ismail, H. M., Chandor-Proust, A. & Paine, M. J. I. Role of cytochrome P450s in insecticide resistance: impact on the control of mosquito-borne diseases and use of insecticides on Earth. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **368**, 20120429 (2013).
19. Enayati, A. A., Vatandoost, H., Ladonni, H., Townson, H. & Hemingway, J. Molecular evidence for a kdr-like pyrethroid resistance mechanism in the malaria vector mosquito Anopheles stephensi. *Med. Vet. Entomol.* **17**, 138–144 (2003).
20. Borycz, J., Borycz, J. A., Kubów, A., Lloyd, V. & Meinertzhagen, I. A. Drosophila ABC transporter mutants white, brown and scarlet have altered contents and distribution of biogenic amines in the brain. *J. Exp. Biol.* **211**, 3454–3466 (2008).
21. Billker, O. *et al.* Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature* **392**, 289–292 (1998).
22. Zhang, G. *et al.* Anopheles Midgut FREP1 Mediates Plasmodium Invasion. *J. Biol. Chem.* **290**, 16490–16501 (2015).
23. Bohbot, J. *et al.* Molecular characterization of the Aedes aegypti odorant receptor gene family. *Insect Mol. Biol.* **16**, 525–537 (2007).
24. Hill, C. A. *et al.* G protein-coupled receptors in Anopheles gambiae. *Science* **298**, 176–178 (2002).
25. Holt, R. A. *et al.* The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129–149 (2002).
26. Kamali, M. *et al.* Multigene phylogenetics reveals temporal diversification of major African malaria vectors. *PLoS ONE* **9**, e93580 (2014).
27. Wolff, G. H. & Riffell, J. A. Olfaction, experience and neural mechanisms underlying mosquito host preference. *J. Exp. Biol.* **221**, (2018).
28. David, J.-P. *et al.* Comparative analysis of response to selection with three insecticides in the dengue mosquito Aedes aegypti using mRNA sequencing. *BMC Genomics* **15**, 174 (2014).

29. Bamou, R. *et al.* Status of Insecticide Resistance and Its Mechanisms in Anopheles gambiae and Anopheles coluzzii Populations from Forest Settings in South Cameroon. *Genes (Basel)* **10**, (2019).
30. Weedall, G. D. *et al.* A cytochrome P450 allele confers pyrethroid resistance on a major African malaria vector, reducing insecticide-treated bednet efficacy. *Sci Transl Med* **11**, (2019).
31. Brito, L. P. *et al.* Assessing the effects of Aedes aegypti kdr mutations on pyrethroid resistance and its fitness cost. *PLoS ONE* **8**, e60878 (2013).
32. Djogbénou, L., Labbé, P., Chandre, F., Pasteur, N. & Weill, M. Ace-1 duplication in Anopheles gambiae: a challenge for malaria control. *Malar. J.* **8**, 70 (2009).
33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
34. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
35. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics* **65**, e57 (2019).
36. Niu, G. *et al.* The fibrinogen-like domain of FREP1 protein is a broad-spectrum malaria transmission-blocking vaccine antigen. *J. Biol. Chem.* **292**, 11960–11969 (2017).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
40. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
41. Skinner, M. E. & Holmes, I. H. Setting up the JBrowse genome browser. *Curr Protoc Bioinformatics* **Chapter 9**, Unit 9.13 (2010).