bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Impact of transposable elements on the genome of the urban malaria vector *Anopheles coluzzii*

Carlos Vargas-Chavez[1], Neil Michel Longo Pendy[2,3], Sandrine E. Nsango[4], Laura Aguilera[1], Diego Ayala[2,5]*, Josefa González[1]*

[1]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona, Spain.

[2]CIRMF, Franceville, Gabon.

[3] Ecole doctorale en infectiologie tropicale (EDR), Franceville, Gabon.

[4] Faculté de Médecine et des Sciences Pharmaceutiques, Université de Douala,
BP 2701, Douala, Cameroun

[5]MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France

Emails: carlos.vargas@ibe.upf-csic.es, longo2michel@gmail.com, nsango2013@yahoo.fr, mlaura.aguilera@gmail.com, diego.ayala@ird.fr, josefa.gonzalez@csic.es

*Corresponding authors

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Background

Anopheles coluzzii is one of the primary vectors of human malaria in sub-Saharan Africa. Recently, it has colonized the main cities of Central Africa threatening vector control programs. The adaptation of *An. coluzzii* to urban environments is partly due to an increased tolerance to organic pollution and insecticides. While some of the molecular mechanisms for ecological adaptation, including chromosome rearrangements and introgressions, are known, the role of transposable elements (TEs) in the adaptive processes of this species has not been studied yet.

Results

To better understand the role of TEs in rapid urban adaptation, we sequenced using long-reads six *An. coluzzii* genomes from natural breeding sites in two major Central Africa cities. We *de novo* annotated the complete set of TEs and identified 64 previously undescribed families. TEs were non-randomly distributed throughout the genome with significant differences in the number of insertions of several superfamilies across the studied genomes. We identified seven putatively active families with insertions near genes with functions related to vectorial capacity. Moreover, we identified several TE insertions providing promoter and transcription factor binding sites to insecticide resistance and immune-related genes.

Conclusions

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The analysis of multiple genomes sequenced using long-read technologies allowed us to generate the most comprehensive TE annotations in this species to date. We found that TEs have an impact in both the genome architecture and the regulation of functionally relevant genes in *An. coluzzii*. These results provide a basis for future studies of the impact of TEs on the biology of *An. coluzzii*.

KEYWORDS

Long-read sequencing, Insecticide resistance, Innate immunity, Comparative genomics, Chromosome inversions

BACKGROUND

The deadly success of the malaria mosquito *Anopheles coluzzii* is rooted in its extraordinary ecological plasticity, inhabiting virtually every habitat in West and Central Africa where it spreads the human malaria parasite (1, 2). Noteworthy, the larvae of *An. coluzzii* exploit more disturbed and anthropogenic sites than its sister species *An. gambiae*. *An. coluzzii* exhibits a higher tolerance to salinity and organic pollution, and as a consequence, is the predominant species in coastal and urban areas (2-4). However, this mosquito not only has a greater resilience to ion-rich aquatic environments, but it has also become resistant to DDT and pyrethroid insecticides used for vector control (5). Actually, insecticide resistant populations of this malaria mosquito are present across its geographical range, driving *An. coluzzii* evolution across the continent (6, 7). The adaptive flexibility of this mosquito has been also highlighted by its rapid competence to expand its range of peak biting times in order to avoid insecticide treated bed-nets

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(8). This extraordinary adaptative capacity makes this malaria vector a threat for malaria control. Thus, elucidating the natural genetic variants underlying the ecological and the physiological responses to fluctuating environments in this species is key for its control.

At the molecular level, a variety of genetic mechanisms have been related back to the myriad of adaptation processes present in this mosquito. The most prominent and historically studied examples are chromosomal inversions (9, 10). *An. coluzzii* exhibits a large number of polymorphic chromosomal rearrangements (11, 12). Many of these inversions have been associated to environmental adaptation through environmental clines and/or correlation with specific climatic variables (10, 13), such as the inversion 2La associated with aridity tolerance capacity in adults (14, 15). Other types of rearrangements, such as gene duplications, have been involved in insecticide resistance. For example, the acetylcholinesterase (*Ace-1*) gene has been duplicated, maintaining at least a sensitive and a resistance copy, in order to counteract the fitness cost of the resistant phenotype (16-18). Moreover, a recent genome-wide analysis showed that genes containing copy number variants were enriched for insecticide functions (19). Other examples of gene selection due to anthropogenic activities have been found in genes related with detoxification or immunity, particularly in new colonized urban settings (3, 20-22). These adaptive processes have been repeated across West and Central African populations, reducing the efficacy of vector control measures (6). However, while several of the candidate genes responsible for the adaptive capacity of *An. coluzzii* have been identified, our knowledge of the genetic variants underlying differences in these genes lags behind. In particular, very little is known about natural variation in transposable element (TE) insertions in *An. coluzzii*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TEs are key players in multiple adaptive processes across a large variety of species, due to their capacity to generate a wide variety of mutations and their contribution to a rapid responses to environmental change (23, 24). TEs can disperse across the genome regulatory sequences such as promoters, enhancers, insulators, and repressive elements thus affecting nearby gene expression (25). Additionally, they can also act as substrates for ectopic recombination leading to structural mutations such as chromosomal rearrangements (26-28). However, TEs are often ignored when analyzing functional variants in genomes. This is because due to their repetitive nature, TE insertions are difficult to annotate and reads derived from TEs are often discarded in genome-wide analyses (Goerner-Potvin and Bourque 2018). Long-read sequencing techniques are needed to get a comprehensive view of TE variation in genomes, as these technique allow the annotation of TE insertions in the genome rather than inferring their position (29, 30).

Although TE insertions have been annotated genome-wide in several anopheline species including *An. coluzzii*, multiple studies to date have characterized the TE repertoire in a single genome for each species (31-39). To capture the full extent of TE natural variation and the potential consequences of TE insertions it is necessary to evaluate multiple genomes in order to comprehensively assess diversity within a species (40-42). This becomes especially relevant when attempting to identify recent TE insertions and their effect in the genome structure and genome function, given that they might be restricted to local populations. So far, our knowledge of *An. coluzzii* genome variation due to TE insertions is limited to a few well-characterized families that have been found to vary across genomes (43-47).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In this work, we sequenced using long-read technologies and assembled the genomes of natural *An. coluzzii* larvae collected in six natural breeding sites in two major cities in Central Africa: Douala (Cameroon) and Libreville (Gabon). We performed a *de novo* TE annotation of the six newly assembled genomes, and we also annotated the previously available *An. coluzzii* genome from Yaoundé (Cameroon) (48). We identified 64 new anopheline TE families and showed that the availability of multiple genomes substantially improves the discovery of TE variants. We further analyzed individual TE insertions that could be acting as enhancers and promoters and that are located nearby genes with functions relevant for the vectorial capacity of the mosquitoes.

RESULTS

Six new whole-genome assemblies of *An. coluzzii* from two major cities in Central Africa

To explore the TE diversity in *An. coluzzii*, we used long-read sequencing and performed whole genome assemblies and scaffolding of larvae collected from six natural urban breeding sites: three from Douala, Cameroon, and three from Libreville, Gabon, Central Africa (Figure 1A, see Material and Methods). Additionally, we performed a reference-guided scaffolding of the available *An. coluzzii* reference genome *AcolN1* using the chromosome level assembly *AgamP4* of *An. gambiae* (48). While the number of scaffolds varied from 5 to 107, with the median being 20, the scaffolds' N50 was similar across the seven genomes (Table 1).

We assessed the genomes completeness using BUSCO with the dipteran set of genes (49). We obtained percentages of complete genes ranging from 94.2% to 96.6% except for the *DLA155B* sample which had a lower completeness value (89.5%; Table 1). These completeness values for

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

most (5/6) of the samples were similar to those from the *AcolN1* genome assembly which contained 98.9% complete genes (Table 1; Additional file 1: Table S1).

Table 1. Genome assemblies and scaffolds' statistics for the *An. coluzzii* genomes we analyzed..

Genome	Long reads coverage	Illumina coverage	Assembly size (Mb)	Number of contigs	Number of scaffolds	N50 of scaffolds (kb)	Complete BUSCO genes (%)	TE families identified
DLA112	55X	59X	252	3917	107	54591	96.6	244
DLA155B	28X	19X	236	2081	24	52031	89.5	243
DLA146	28X	42X	247	2036	14	54960	95.1	193
LBV88	31X	41X	245	2576	19	54450	94.5	280
LBV136	34X	130X	236	2911	28	52053	95.2	172
LBV11[a]	89X	61X	246	2608	20	53712	94.2	294
AcolN1[b]	~270X	-	251	205	5	53057	98.9	283

Three genomes were collected in Douala (DLA) and three in Libreville (LBV). [a] *LBV11* was sequenced using PacBio technologies, while the other five genomes were sequenced using Oxford Nanopore Technologies. [b] Genome statistics for *AcolN1*, the high quality *de novo* genome assembly reported by Kingan *et al*., (48) are also included.

64 new anopheline TE families discovered in *An. coluzzii*

To identify the TE families present in each of the genomes, we used the *TEdenovo* pipeline from the REPET package (50). After several rounds of manual curation, we identified between 172

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and 294 TE families for each genome (Table 1; Additional file 1: Table S2). Remarkably, while using a single reference would have only allowed the identification of a median of 244 TE families, clustering the TE libraries from an increasing number of genomes allowed the identification of a total of 435 well supported TE families (Figure 1B; see Material and Methods). Interestingly, 64 of these families (32 DNA, 9 LINEs and 23 LTRs) are described here for the first time. The majority of the new families (43/64) had partial matches to other known TEs, thus allowing us to classify them at the superfamily level (Additional file 1: Table S3). The use of multiple references was especially relevant for identifying these previously undescribed families given that using a single genome would have only allowed to identify a median of 37 (25-48) novel TE families (Figure 1B).

To further characterize these novel families, we estimated the average number of insertions in the seven *An. coluzzii* genomes, and their distribution and abundance in other species from the *Anopheles* genus (Figure 2; Additional file 2: Figure S1; Additional file 1: Table S3). To do this, we first annotated individual TE insertions in the seven *An. coluzzii* genomes using the *TEannot* pipeline from the REPET package (51). To ensure that our annotation was as complete as possible, besides the 435 families previously identified using REPET, we also included in our library 85 TE families from other mosquito species that we found to be present in *An. coluzzii* (Additional file 1: Table S4; see Material and Methods) (52). The final total of 520 families were classified into 23 superfamilies and then further grouped into four orders (DNA, LINE, LTR and SINE; Figure 1C).

Copies from all 64 new families were found in all seven *An. coluzzii* genomes, further suggesting that these are *bona fide* families. Although the majority of families contain full-length copies in

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

at least one of the seven genomes analyzed, truncated copies were the most abundant (Figure 2B; Additional file 1: Table S3). We identified a median of 72 insertions (ranging from 16 to 1,445) per family and genome (Figure 2B; Additional file 2: Figure S1B). Two out of the four TRIM elements identified (*Acol_LTR_Ele 4* and *Acol_LTR_Ele 6*) are among the most abundant new families, with more than 150 insertions (Figure 2B). TRIM elements are non-autonomous retrotransposons flanked by LTRs and lacking coding capacity (Figure 2A). These elements have not been previously described in anopheline genomes and are still underexplored in insect genomes in general (53-55). However, they might be important players in insect genome evolution: in plants there are some examples of TRIM elements showing the capacity to restructure genomes by acting as target sites for retrotransposon insertions, alter host gene structure, and transduce host genes (56, 57).

We also assessed the phylogenetic distribution of the 64 new TE families in 15 species of the *Anopheles* genus, including the eight members of the *An. gambiae* complex, two more distantly related mosquitoes species (*Culex quinquefasciatus*, *Aedes aegypti*) and *Drosophila melanogaster* (Additional file 1: Table S3) (35, 58, 59). We found that the new families were unevenly distributed among the members of the *Anopheles* genus (Figure 2C and Additional file 2: Figure S1C). Ten families were exclusively found in members of the Pyretophorus series, suggesting that these elements emerged after the split of this series from the Cellia subgenus. Moreover, 13 families were also found in at least one of the other three non-anopheline species (Additional file 2: Figure S1C). The distribution of these 13 families was patchy, with some of them present only in distantly related species while others were present in members of the *Anopheles* genus or in members of the Pyretophorus series. These suggests that some of these

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

families might have been acquired through horizontal transfer events (Additional file 1: Table S3) (39).

The *Gypsy* superfamily has the largest copy number differences across genomes

The percentage of the genome represented by TEs across the seven genomes varied between 16.94% and 20.21% (Table 2). We found a positive correlation between TE content and genome size as has been previously described in *Anopheles* and other species (Pearson's r = 0.90, significance = .007; Additional file 3: Figure S2) (39, 60). As expected due to heterochromatin being a TE rich region and thus challenging to assemble (61), most of the differences in TE content across genomes were found in the heterochromatin compartment (Table 2; χ^2 test for variance, p-value = 3.57e-3).

To assess whether differences in TE content at the family and superfamily level existed among the seven genomes, we focused on the TE copy number in euchromatic regions. We found significant differences at the order and superfamily levels (χ^2 p-value = 1.07e-21 and p-value = 1.69e-14, respectively). The largest differences were found in the LTR order: LTRs were more abundant in the *DLA112* and *LBV88* genomes and less abundant in *AcolN1* (Figure 3A). At the superfamily level, we found that the largest differences were in the *Gypsy* superfamily, which belongs to the LTR order. We also observed an enrichment of the *RTE* superfamily in *LBV11*, of the *CR1* and *Bel-Pao* superfamilies in *DLA112*, and a depletion of the *CR1* superfamily in *AcolN1* (Figure 3). Therefore, most of the differences in TE content between the evaluated genomes appear to be in retrotransposon families.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. TE content in the seven genomes analyzed.

Genome	Whole genome			Euchromatin			Heterochromatin		
	TE copy number	Mb	Genome %	Copy number	Mb	Region %	Copy number	Mb	Region %
DLA112	72901	48.00	19.02	49853	28.18	12.67	22930	19.74	70.34
DLA155B	62999	40.08	16.94	45592	25.39	11.86	17371	14.66	65.76
DLA146	68658	45.42	18.40	47874	27.22	12.36	20682	18.15	68.35
LBV88	68593	45.81	18.70	48922	28.06	12.81	19582	17.68	68.74
LBV136	64343	40.79	17.26	45792	24.97	11.73	18406	15.73	67.59
LBV11	71803	47.59	19.58	50187	28.95	13.40	21564	18.60	70.22
AcolN1	75745	50.81	20.21	48537	26.10	11.95	27205	24.70	74.77

TE copy number, TE content in megabases and percentage of the genome represented by TEs. Values are given for the whole genome and for the euchromatin and heterochromatin compartments separately.

TEs are nonrandomly distributed throughout the genome

As expected, we found that the percentage of TEs in euchromatin, 11.73%-13.40%, is much lower than the percentage of TEs in heterochromatin, 65.76%-74.77%, (Table 2 and Figure 4A). None of the TE families identified were exclusive to either the euchromatin or heterochromatin. However, 45 families were enriched in the euchromatin (χ^2 test, p-value < 0.01) including 12 out of the 32 *mTA* MITE families (Additional file 1: Table S5). This is in line with what has been previously reported in *Ae. aegypti* (62). We also observed that the TE distribution was uneven

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between the chromosomes, and as expected, the X chromosome had a larger fraction of its euchromatin spanned by TEs (Figure 4B) (63).

Finally, we also determine the distribution of TE insertions regarding genes. We divided the genome in five regions: 1 kb upstream, exon, intron, 1 kb downstream and intergenic (64). More than half of the genes (7,239) in *An. coluzzii* had TEs either in their body or 1 kb upstream or downstream. Many of these genes (3,888/7,239) had insertions in all seven genomes, while 1,065 genes have an insertion only in one genome. We found that the number of insertions in intergenic regions was higher than expected by chance while the number of insertions in exons was lower (χ^2 p-value < 0.001; Figure 4C; Additional file 1: Table S6). The upstream and downstream regions behaved differently: the downstream region had a smaller amount of TEs than expected by chance and the upstream region was neither enriched nor depleted for TE insertions (p-value = 0; Additional file 1: Table S6). This is possibly linked with the chromatin state of these regions given that downstream regions are more commonly in a closed chromatin state (64).

Focusing on the TE orders, we observed that LTR elements were more abundant on intergenic regions while SINEs were more abundant on introns, and DNA elements were more abundant in introns and in the upstream region (χ^2 p-value < 2.03e-3; Figure 4C; Additional file 1: Table S7A). MITEs, which are non-autonomous DNA elements have been reported to be more abundant in the introns and flanking regions of genes (65). We observed the same behavior for *mTA* and *m3bp* MITEs, which are more abundant in upstream regions and introns, and *m8bp* MITEs which are more abundant in introns (Additional file 1: Table S7B).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Overall, TEs are not randomly distributed in the genome, as they are more abundant in heterochromatic than in euchromatic regions, more abundant in the X chromosome than in autosomes, and more abundant in intergenic regions than in gene bodies or gene flanking regions.

MITE insertions are present in several inversion breakpoints

TEs have been suggested to be involved in chromosome rearrangements within the *An. gambiae* complex. Indeed, TEs have been found in close proximity to the breakpoints of the 2La in *An. gambiae* and *An. melas*, and to the breakpoints of the 2Rb inversions in *An. gambiae* and *An. coluzzii* (66, 67). We thus explored the TE content in the breakpoints of the 2La and 2Rb, and three other common polymorphic inversions in *An. coluzzii*: 2Rc, 2Rd, and 2Ru (68). The analysis of the breakpoint regions suggested that our genomes have the standard conformation for all five inversions (see Material and Methods; Additional file 1: Table S8). We identified several TEs nearby the proximal and the distal breakpoints of 2La and 2Rb, in agreement with previous studies (Figure 5) (26, 66, 67). For the standard 2La proximal breakpoint, Sharakhov et al. (66) identified several DNA transposons and a SINE insertion. We also identified a cluster of MITE insertions, which are DNA transposons; however, we additionally identified an *Outcast* (LINE) element (Figure 5). Regarding the standard 2La distal breakpoint, we observed two MITEs similar to one of the insertions in the proximal breakpoint, which was in agreement with the findings by Sharakhov et al. (66) (Figure 5). We also observed similar behavior in the 2Rb breakpoints, such as the one described by Lobo et al., (67): tandem repeats flanking the inversion in the standard and inverted forms, and TEs in the internal sequences of both breakpoints (Figure

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5). For the 2Rd inversion, we identified MITEs near both breakpoints. Finally, we have also described here for the first time, TE insertions that are present in the distal breakpoint of inversion 2Ru but not near the estimated proximal breakpoint; although in the latter case we were able to identify reads spanning the breakpoints in the seven genomes (69).

TE insertions from active families might affect the regulation of functionally relevant genes

To identify potentially active TE families, we first estimated their relative age by analyzing the TE landscapes (70, 71). We observed an "L" shape landscape in all genomes which is indicative of a recent TE burst (Additional file 4: Figure S3) (72). This "L" shape landscape, dominated by retrotransposons, had previously been described for the sister species *An. gambiae* (71, 73), where numerous Gypsy LTR Retrotransposons (up to 75%) might currently be active (74, 75). We further investigated the families in the peak of the landscape and we identified eight families with more than two identical full-length fragments and with more than half of their copies identical to the consensus (Additional file 1: Table S9). Additionally, we assessed the potential ability of our candidates to actively transpose by identifying their intact open read frames (ORFs), LTRs (in the case of LTR retrotransposons), and target site duplications (TSDs), and determined that seven of these families are potentially fully capable of transposing, and thus confirming that these families might be responsible for the recent retrotransposon burst in *An. coluzzii* (Additional file 1: Table S9).

To assess the potential functional consequences of the TE insertions from these seven putatively active families, we focused on insertions that occurred in introns, exons, and 1 kb upstream or downstream of a gene. We identified 80 genes with insertions from these families, with five

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes containing up to two insertions in the same gene region Additional file 1: Table Sand since these are all recent insertions, one plausible explanation for those found at high frequencies is that they are subject to positive selection (76) (Additional file 1: Table S10; Additional file 5: Figure S4). We found that 8 insertions were present in all seven genomes analyzed, 24 were present in two or more genomes while 53 were present in a single genome (Additional file 1: Table S11). We focused on the genes containing insertions in two or more genomes to look for functional enrichment. However, we found no significant GO enrichment terms using PANTHER (77). No significant GO enrichment was either found when considering all genes with nearby insertions.

To further investigate the potential role of TE insertions from active families on the function of nearby genes, we looked for functional information on all the genes, and focused on seven of them that have functions related to vectorial capacity: insecticide resistance, immunity, and biting ability (Table 3). We checked whether the TE insertions nearby these genes contained binding sites for transcription factors or promoter motifs (Additional file 1: Table S12; Additional file 1: Table S13). We focused on identifying binding sites for three transcription factors that are known to be involved in response to xenobiotics (cap'n'collar: *cnc*) and in immune response and development (dorsal: *dl* and signal transducer and activator of transcription: *STAT*) given the availability of matrix profiles from *D. melanogaster* (78, 79). We identified binding sites for either *dl*, *STAT* or both in three insertions; interestingly the *Acol_gypsy_Ele18* and the *Acol_copia_Ele8* insertions have more than three binding sites for the same transcription factors, suggesting that they might be functional sites (Table 3) (80). Additionally, the genes that contained these TEs insertions also contained binding sites for these

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

same transcription factor, which suggests that these factors already played a prior role in their regulation. We also identified a putative promoter sequence in the *Acol_copia_Ele24* insertion found upstream of the CLIPA1 protease encoded by AGAP011794 which could also lead to changes in the regulation of this gene (Additional file 1: Table S14).

Table 3. TE insertions from putatively active families.

TE family	**Insert size (bp)**	**TE Freq.**	**Gene**	**Function**	**Possible phenotype [Reference]**	**TFBS**
Acol_copia_Ele24	2233	5/5	AGAP012452	Concanavalin A-like lectin/glucanase	Insecticide resistance [1]	-
Acol_copia_Ele8	3230 (200)*	2/4	AGAP012466	cuticular protein RR-2 family 146	Development, insecticide resistance [2, 3]	*dl* (3) and *STAT* (5)
Acol_copia_Ele24	167	4/4				-
Acol_gypsy_Ele65	185	3/3	AGAP010620	Peptidase S1, PA clan	Immunity, digestion [4, 5]	-
Acol_gypsy_Ele18	4858	1/6	AGAP029191	Defective proboscis extension response	"Bendy" proboscis [6]	*dl* (3-7) and *STAT* (1-6)
Acol_copia_Ele24	168	1/6	AGAP011794	CLIPA1 protein	Digestion, immunity or development [7]	-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acol_gypsy_Ele18	235	1/7	AGAP002633	Gustatory receptor 53	Vectorial capacity [8]	-
Acol_gypsy_Ele65	141	1/3	AGAP028069	Peptidase S1, PA clan	Immunity, digestion [5, 6]	*dl* (1)

TE Freq. specifies the number of genomes where the TE insertion was found and the number of genomes where the gene was correctly transferred. References in the Phenotype column are as follows: 1 (81), 2 (82), 3 (83), 4 (84), 5 (85), 6 (86), 7 (87), 8 (88). The number in parenthesis in the transcription factor binding site (TFBS) column refers to the number (or range) of TFBS found in the TE. *The insertion size in parenthesis refers to an insertion found in one of the genomes corresponding to a solo-LTR insertion.

TE insertions could influence the regulation of genes involved in insecticide resistance

The usage of pyrethroids, carbamates, and DDT as vector control mechanisms has led to the rapid dispersion of insecticide resistance alleles in natural populations (89-93). Among the best characterized resistance point mutations are L1014F (*kdr-west*), L1014S (*kdr-east*), and N1575Y in the voltage gated sodium channel *para* (also known as *vgsc*), and G119S in the acetylcholinesterase *ace-1* gene (94-96). We first investigated whether the seven genomes analyzed in this work contained these resistance alleles. We found the *kdr-west* mutation in the six genomes from Douala and Libreville but not in *AcolN1* genome (48). None of the other mutations were identified, however a previously undescribed nonsynonymous substitution (L1688M) in the fourth domain of *para* was identified in the aforementioned six genomes. Whether this replacement also increases insecticide resistance is yet to be assessed.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TEs have been hypothesized to play a relevant role specifically in response to insecticides (97-99), and a few individual insertions affecting insecticide tolerance in anopheline mosquitoes have already been described (100). Thus, we searched for TE insertions in the neighborhood of insecticide-related genes that could lead to differences in their regulation. We focused on well-known insecticide resistance genes as well as considering genes that have been shown to be differentially expressed in *An. gambiae* when exposed to insecticides (Additional file 1: Table S14; Additional file 6: Figure S5) (3, 101-103). We found that 23 out of the 43 genes analyzed contained at least one TE insertion. We also observed that *para* had the largest number of TE insertions (48 in average per genome, mainly in its introns) from this set of genes. This is an exception, given that the average number of insertions per gene is 2.95 for members of this set which falls within the expected number of insertions per gene in all the genome (t-test, p-values > 0.2).

Only one of the insertions, a solo LTR element of *Acol_Pao_Bel_Ele43* from the *Pao-Bel* superfamily and present in all the genomes analyzed, was located in the 3' UTR of *GSTE2*. Interestingly, an upstream insertion possibly affecting the expression level of this gene has previously been identified in *An. funestus* (100). To determine if TEs could influence the regulation of insecticide-resistance genes, we focused on polymorphic (present in two or more genomes) and fixed (present in all seven genomes analyzed) insertions located in introns or 1 kb upstream of the gene. We searched for *cnc* binding sites, and for those insertions located in gene upstream regions we also looked for promoter motifs (Additional file 1: Table S12; Additional file 1: Table S13). We identified 15 insertions in 10 genes containing either *cnc* binding sites or promoter sequences. One insertion located in *CYP4C28* and two insertions in *para* contained

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

binding sites for *cnc*, although the genes did not contain binding sites for this transcription factor. Additionally, we identified 12 insertions containing promoter motifs and located nearby nine genes (Figure 6). In some cases, such as the *Acol_m2bp_Ele10* MITE insertion in *ABCA4* or the *tSINE* insertion in *GSTMS2*, while the same TE insertion was found in six and seven genomes respectively, the promoter motifs were found only in four and one genome respectively (Figure 6; Additional file 1: Table S13). We analyzed the consensus sequence of these two families and we found that while the *Acol_m2bp_Ele10* had the promoter motif, the *tSINE* did not, suggesting that some of the *Acol_m2bp_Ele10* elements lost the promoter motifs while the *tSINE* copies acquired them.

Immune response genes could also be affected by TEs

Mosquitoes breeding in urban and polluted aquatic environments overexpress immune-related genes suggesting that immune response is relevant for urban adaptation (104). To assess the potential role of TEs in immune response, we searched for TE insertions in genes putatively involved in immunity according to ImmunoDB (105)(Additional file 1: Table S15). We identified 466 TE insertions in 156 out of the 281 genes analyzed. The number of insertions in each gene varied greatly going from 60 genes with a single insertion to AGAP000940, a gene coding for a C-type lectin and spanning 107.2 kb, with 48 insertions. The frequency of these insertions was also variable with 184 (39.5%) of the insertions being fixed, 208 (44.6%) polymorphic and 74 (15.9%) unique. We further explored polymorphic and fixed insertions and identified binding sites for *dl* and *STAT* and promoter motifs. We found that 20 TEs contained bindings sites for *dl*, 23 TEs contained binding sites for *STAT* and 12 TEs contained binding sites

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

both for *dl* and *STAT* (Additional file 1: Table S15). Additionally, we identified 82 insertions, in the upstream region of 58 genes, which carried putative promoter sequences.

We identified TE insertions in three different antimicrobial peptides (AMPs). AMPs form the first line of host defense against infection and are a key component of the innate immune system, however none had transcription factor binding sites (TFBS) for *dl* or *STAT*. It is important to keep in mind that there are other TF that participate in the regulation of AMPs and that both *dl* and *STAT* are also involved in other biological processes (106). Interestingly we also identified TEs with TFBS for *dl* in the vicinity of STAT1 and STAT2 which might lead to novel regulatory mechanisms of the JAK/STAT signaling pathway. Furthermore, 11 of the 156 genes containing TE insertions are differentially expressed in response to a *Plasmodium* invasion. These genes participate in several pathways of the immune response including the small regulatory RNA pathway, pathogen recognition, the nitric oxide response and ookinete melanization (79, 107-109). Four of the TEs affecting these genes added TFBS and promoter sequences, thus suggesting that these TE insertions can presumably influence the response to this pathogen (110) (Table 4).

Table 4. TE insertions in *Plasmodium* responsive genes from the immune system.

Gene ID	**Gene symbol**	**Function**	**# of TE insertions**	**Family**	**Frequency**	**Promoter**	**TFBS**
AGAP002625	CTL9	CTLs	1	-	-	No	-
AGAP003663	RM62B	SRRPs	2	*Acol_mTA_Ele11*	7/7	No	*dl* (1), *STAT* (1)

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AGAP004845	SCRB8	SCRs	4	*Acol_otherMITEs_Eles16*	7/7	No	*STAT* (1)
				Acol_Pao_Bel_Ele35	7/7	Yes	*STAT* (1)
AGAP005203	PGRPLC1	PGRPs	1	-	-	No	-
AGAP008844	GALE1	GALEs	1	*Acol_ m3bp_Ele11*	7/7	Yes	-
AGAP009033	HPX2	PRDXs	1	-	-	No	-
AGAP009887	R2D2	SRRPs	1	-	-	No	-
AGAP011204	AUB	SRRPs	3	-	-	No	-
AGAP011717	AGO1	SRRPs	16	*Acol_ mTA_Ele31*	6/7	No	*dl* (1)
AGAP011780	CLIPA4	CLIPs	1	-	-	No	-
AGAP011792	CLIPA7	CLIPs	1	-	-	No	-

Family and frequency are only shown for TEs with TFBS or promoter sequences. In the Function column the following abbreviations are used: C-Type Lectins (CTLs), Small Regulatory RNA Pathway Members (SRRPs), Scavenger Receptors (SCRs), Peptidoglycan Recognition Proteins (PGRPs), Galactoside-Binding Lectins (GALEs), Peroxidases (PRDXs), CLIP-Domain Serine Proteases (CLIPs).

DISCUSSION

In this study, we *de novo* annotated transposable element (TE) insertions in seven genomes of *An. coluzzii*, six of them newly sequenced here. A comprehensive genome-wide TE annotation was possible because we used long-read technologies to perform the genome sequencing and assembly. Long-reads allow identifying TE insertions with high confidence given that the entire TE insertion sequence can be spanned by a single read (29, 30). While the genome-wide TE

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

repertoire has been studied in other anopheline species, particularly in *An. gambiae*, to our knowledge there are no other studies that have explored TE variation in multiple genomes from a single species (31, 32, 35, 39, 71, 111). We observed that increasing the number of available genomes analyzed allowed us to increase the number of identified TE families from a median of 244 (172-294) to 435 (Figure 1B). Moreover, having the full sequences of seven genomes also allowed us to discover 64 new TE families, including four TRIM families previously undescribed in anopheline genomes that are likely to be important players in genome evolution (56, 57). The wide range of families identified across genomes was not directly related to the quality of the genome assembly taking into consideration the more generally used quality parameters such as read length, number of contigs, and contig N50 (112). This suggests that there are possibly other characteristics of each genome that affect the identification of high quality TE families, such as biases in the location of the TE insertions given that TE families are challenging to identify in regions with low complexity or with numerous nested TEs. Nonetheless, the identification of TE families is dependent on the methodology used to perform TE annotations, therefore different annotation strategies could lead to the discovery of still undescribed families (59).

The availability of several genome assemblies also allowed us to determine that the majority of the intraspecies differences in the TE content were in heterochromatic regions, most likely due to differences in the quality of the genome assembly. Nevertheless, there were also significant differences in the TE content in euchromatic regions, reflecting true intraspecific variability as has been previously observed in several organisms including Drosophila (76, 113), mammals (114, 115), maize (116) and Arabidopsis (117). TE insertions were not randomly distributed

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

throughout the genome and instead were consistently enriched in intergenic regions, most likely due to purifying selection, as suggested in the wild grass *Brachypodium distachyon* (118). In Drosophila, TE enrichment in intergenic regions was also observed in addition to enrichment in the intronic region, which we did not observe in *An. coluzzii* (119). We also analyzed the TE content in the breakpoints of five common polymorphic inversions, three of them analyzed here for the first time. We found TE insertions in all but one of the inversion breakpoints, with MITE elements the most common TE family, as already described in the 2Rd' inversion in *An. arabiensis* (26) (Figure 5).

The choice to use samples from urban environments allowed us to take a first look into the role of TEs in rapid adaptation to novel habitats (120). We focused on insertions from recently active families located near genes that are relevant for the vectorial capacity of *An. coluzzii*. Because adaptation can also happen from standing variation, in the case of insecticide resistance genes, which have been shown to be shaped by TE insertions in several organisms, and immune-related genes, we analyzed all insertions independently of their age (100, 121, 122). While the role of nonsynonymous substitutions and copy number variation in resistance to insecticides commonly used in urban environments has been studied, the potential role of TEs has not yet been comprehensively assessed in *An. coluzzii* or any other anopheline species (19, 103, 123-125). In the genomes we assessed, we identified several insertions that were polymorphic or fixed nearby functionally relevant genes (Table 3, Table 4 and Figure 6). Some of the identified candidate insertions contained binding sites for transcription factors related to the function of nearby genes and promoter regions. Besides adding regulatory regions, TEs can also affect the regulation of nearby genes by affecting gene splicing and generating long non-coding RNAs among many

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

other molecular mechanisms (25, 126-129). Thus, it is possible that the candidate TE insertions identified that lack binding sites and promoters could be affecting nearby genes through other molecular mechanisms. Our results are a first approximation to the potential role of TEs in *An. coluzzii* adaptation to the challenging environment that urban ecosystems entail. Establishing a direct link between the TEs and the traits involved in urban adaptation will require sampling a larger number of individuals and characterizing the phenotypes associated with the insertions.

CONCLUSIONS

The long-read sequencing of seven *An. coluzzii* genomes from urban environments allowed us to capture to a larger extent the diversity of TE families and TE insertions and to assess their impact in the genome architecture and genome function in this species. While there was an enrichment of TE insertions in intergenic regions, we found several insertions located in the 1 kb flanking regions or inside genes relevant for the vectorial capacity of this species. Furthermore, we found that some of these TE insertions are adding regulatory regions and as such they could influence the regulation of these genes. The genomic resources and the results that we present in this work provide a basis for future studies of the impact of TEs in the biology of *An. coluzzii.* This will allow increasing our knowledge on a species which besides being interesting from an evolutionary perspective, given its high levels of genetic diversity and the strong anthropogenic pressures it faces, is of great importance to human health. A better understanding of the biology of *An. coluzzii* and its ability to rapidly adapt to urban environments will further facilitate the development of novel strategies to combat malaria. Better management strategies can be

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

implemented if we understand and are able to predict changes in the frequency of genetic variants relevant for the vectorial capacity of this species.

MATERIALS AND METHODS

Sample collection and DNA isolation

We sampled *An. coluzzii* larvae in two cities of Central Africa: Libreville, Gabon, in January 2016 and Douala, Cameroon, in April 2018. A systematic inspection of potential breeding sites was conducted to determine the presence of *Anopheles* larvae. We manually separated the anopheline from the culicine larvae based on morphological recognition and positioning of their bodies on or under the water surface (Robert, 2017). We collected immature 3rd and 4th stage larvae of *Anopheles* from water bodies using the standard dipping method (Service, 1993). We collected 25 larvae from each site and stored them in 1.5 ml of absolute ethanol. After each daily sampling session, the samples were stored at -20 °C.

All the samples were PCR tested to differentiate *An. coluzzii* larvae from *An. gambiae* larvae before library preparation, using primers SINE200_F (TCGCCTTAGACCTTGCGTTA) and SINE200_R (CGCTTCAAGAATTCGAGATAC) (45). For PacBio sequencing, DNA from a single *An. coluzzii* larva from the *LBV11* site was extracted using the MagAttract HMW DNA extraction kit (Qiagen) following manufacturer’s instructions. Briefly, the larva was air-dried and lysed in 240 µl of buffer ATL (proteinase K added) shaking overnight at 56 ºC. Next, the DNA was isolated using the MagAttract magnetic beads and eluted twice in 50 µl of buffer AE. The DNA concentration was measured using a Qubit fluorometer. For Nanopore sequencing, DNA

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from six larvae from each of the five breeding sites was extracted either with the QiaAMP UCP DNA kit (Qiagen) or MagAttract HMW DNA extraction kit (Qiagen). For the QiaAMP UCP DNA kit, we followed the manufacturer's instructions. Each larva was air-dried and lysed in 200 µl of buffer AUT (proteinase K added) shaking overnight at 56 ºC, then DNA was isolated using a QIAamp UCP MinElute column and eluted twice in 25 µl of buffer AUE. For the MagAttract HMW DNA extraction kit, we followed manufacturer's instructions but using lower buffer amounts to increase DNA concentration. Briefly, each larva was lysed in 120 µl of buffer ATL (proteinase K added) shaking overnight at 56 ºC, then DNA was isolated using the MagAttract magnetic beads and eluted twice in 25 µl of buffer AE. The DNA concentration was measured using a Qubit fluorometer. Both elutions of the same sample were mixed before library preparation. For Illumina sequencing, DNA from one larva from each of the six different breeding sites was extracted following the same extraction protocol as for Nanopore sequencing.

Library preparation and sequencing

Quality control of the DNA sample for PacBio sequencing (Qubit, NanoDrop and Fragment analyzer) was performed at the Center for Genomic Research facility of the University of Liverpool prior to library preparation. The library was prepared by shearing DNA to obtain fragments of approximately 30 kb and sequenced on 2 SMRT cells using Sequel SMRT cell, 3.0 chemistry. Nanopore libraries were constructed using the Native Barcoding Expansion 1-12 (PCR-free) and the Ligation Sequencing Kit following manufacturer's instructions. A minimum of 400 ng of DNA from each larva was used to start with the library workflow. For each breeding site, six larvae were barcoded, and equal amounts of each barcoded sample were pooled prior to sequencing. The samples from the same breeding site were ran in a single R9.4 flow cell

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in a 48-hour run, except for sample *DLA112* which was run in two flow cells. The DNA concentration was assessed during the whole procedure to ensure enough DNA was available for sequencing.

The quality control of the samples, library preparation and Illumina sequencing was performed at the Center for Genomic Research facility of the University of Liverpool. Low input libraries were prepared with the NEBNext Ultra II FS DNA library kit (300 bp inserts) on the Mosquito platform, using a 1/10 reduced volume protocol. Paired-end sequencing was performed on the Illumina Novaseq platform using S2 chemistry (2x150 bp).

Genome Assemblies

The PacBio sequenced genome was assembled using *Canu* version 1.8 (130) with an estimated genome size of 250Mb and parameters: '*stopOnLowCoverage*=5, *corMinCoverage*=0, *correctedErrorRate*=0.105, *CorMhapFilterThreshold*=0.0000000002, *corMhapOptions="--threshold 0.80 --num-hashes 512 --num-min-matches 3 --ordered-sketch-size 1000 --ordered-kmer-size 14 --min-olap-length 2000 --repeat-idf-scale 50*" *mhapMemory*=60g, *mhapBlockSize*=500, *ovlMerDistinct*=0.975'. Next, we identified and removed allelic variants using *purge_haplotigs* version 1.0.4 (131) with the "*-l 15 -m 100 -h 195*" parameters. The Nanopore genomes were assembled using *Canu* version 1.8 using the same parameters as previously described, except for *correctedErrorRate* which was set to 0.16, followed by a round of polishing using *racon* version 1.3.3 (132), followed by *nanopolish* version 0.11.1 (133) and *pilon* version 1.23-0 (134) with the fix parameter set on '*bases*'. *Pilon* requires high coverage short-read data to perform the polishing and these data came from the aforementioned single

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

larvae sequenced from each of the sites. Finally, *blobtools* version 1.1.1 (135) was used to remove contamination from all six genome assemblies taking into consideration fragment sizes, their taxonomic assignation and the coverage using the Illumina reads.

As a proxy of the completeness, the BUSCO values for the six newly assembled genomes plus the *AcolN1* genome were obtained using BUSCO version 3.0.2 (49) with the *diptera_odb9* set as reference. Finally, the contigs for all seven assemblies were ordered and merged with *RaGOO* v1.1 (136) using the chromosome level *An. gambiae* AgamP4 assembly.

Gene annotation transfer

The *gff* for the genome annotation for AgamP4 was transferred into the newly assembled genomes using *Liftoff* (137) with default parameters. The annotation was manually inspected using *UGENE* version 35 (138) and whenever needed the annotation was accordingly corrected. 96% of the AgamP4 genes were correctly transferred.

Construction of the curated TE library and *de novo* TE annotation

We ran the *TEdenovo* pipeline (50) independently on each of the seven genomes with default parameters. The obtained consensus in each genome were further filtered by discarding those generated with only one sequence, with less than one full-length fragment mapping to the genome, or with less than three full-length copies (Additional file 1: Table S2). The remaining consensuses were manually curated to remove redundant sequences and artifacts by manual inspection of coverage plots generated using the *plotCoverage* tool from REPET and visualization of the structural features on the genome browser IGV version 2.4.19 (139).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To ensure that we identified as much of the TE diversity as possible, the *TEfam* (tefam.biochem.vt.edu) database, which contains the TE libraries for several species of mosquitoes, was used to annotate the seven genomes using *RepeatMasker* version open-4.0.9 (Smit et al. 2015). Families with more than three matches longer than 90% in any genome were selected and their hit with the highest identity from each genome was extracted. These sequences were added to the REPET library and all the consensuses were clustered using CD-HIT version 4.8.1 (140) with the -*c* and -*s* parameters set to 0.8. 85 clusters contain sequences only identified by TEfam. The sequences belonging to the same cluster were used to perform a multiple sequence alignment and the consensuses were obtained.

The consensuses were classified using PASTEC (141) with default parameters. Next their bidirectional best-hits were calculated using BLAST (142) against the TEfam (tefam.biochem.vt.edu), AnoTExcel (143) and Repbase (144) databases. When more than 80% of a consensus matched to a feature from the databases with an identity higher than 80%, the classification was transferred to the consensus. While not an order *per se*, MITEs were grouped together for subsequent analysis. Additionally, we classified the families based on the conservation of features characteristic of their orders into putative autonomous, putative autonomous lacking terminal inverted repeats (TIRs) or long terminal repeats (LTRs), putative non-autonomous, such as MITEs and TRIMs, and degenerated (Additional file 1: Table S4)(Fonseca et al 2019). These classified consensuses were used to re-annotate the assembled genomes with the *TEannot* pipeline using default parameters and we discarded copies whose length overlapped >80% with satellite annotations (51).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transfer of TE annotations to the *AcolN1* reference genome

We transferred the TE annotations from the six genomes we sequenced to the *AcolN1* genome. First, we selected only TEs mapping to genes (including 1 kb upstream and 1 kb downstream) in each of the six genomes and built a *gff* file including two 1 kb long "anchors" adjacent to each TE. We transferred these features considering each anchor and the TE as exons using the Liftoff tool with the *-exclude_partial -overlap 1 -s 0.8* parameters (137). We discarded transfers where the transferred TE was shorter than 10 bp or any of the anchors was shorter than 500 bp. Discarded transfers and TEs not transferred were used for a second round where a new *gff* was created with two 1 kb long anchors but this time located 500 bp away from each end of the TE. and the previously described transfer process was performed. A third round of transfer was performed this time with anchors located 1 kb away from the TE insertion. The TE positions and family of each transferred TE were conserved. Finally, for all non-transferred TEs we generated a new *gff* file with only the two anchors and no TE and transferred these features using the same methodology. In these cases, the distance between both anchors was conserved as the transferred TE coordinates and the TE family was conserved.

We discarded TEs that were not transferred to genes (plus 1 kb upstream and downstream) in the *AcolN1* genome. Using GenomicRanges we identified overlaps between TEs in the *AcolN1* genome and transferred TEs. We allowed a distance of up to 10 bp between matches and when a TE from the same family was found in the same position we considered the TE as present. Finally, for TEs that were transferred using only the anchors we identified overlaps between the six genomes to calculate the frequencies of these non-reference TE insertions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identification of newly described families in other species

We analyzed all 10 available fully sequenced species from the Pyretophorus series, which belongs to the Cellia subgenus. We also included an additional five *Anopheles* species, three from each of the other series from the Cellia subgenus and two from the other subgenera with available fully sequenced species. As outgroups we included the genomes of *Cx. quinquefasciatus*, *Ae. aegypti* and *D. melanogaster*. *RepeatMasker* version open-4.0.9 (Smit et al. 2015) was run with default parameters using the 64 newly described families as the library on the following genomes: *An. albimanus* (AalbS2), *An. atroparvus* (AatrE3), *An. farauti* (AfarF2), *An. funestus* (AfunF3), *An. stephensi* (AsteS1), *An. epiroticus* (AepiE1), *An. christyi* (AchrA1), *An. merus* (AmerM2), *An. gambiae* (AgamP4), *An. coluzzii* (AcolN1), *An. melas* (AmelC2), *An. arabiensis* (AaraD1), *An. quadriannulatus* (AquaS1), *An. bwambae* (Abwa2) and *An. fontenillei* (ASM881789v1), *Cx. quinquefasciatus* (CulPip1.0), *Ae. aegypti* (AaegL5.0) and *D. melanogaster* (ISO1 release 6).

Identification of heterochromatin

The coordinates for the pericentric heterochromatin, compact intercalary heterochromatin, and diffuse intercalary heterochromatin in *An. gambiae* AgamP3 were obtained from a previous work (61). The *An. gambiae* AgamP3 genome assembly was mapped against the seven *An. coluzzii* genome assemblies using *progressiveMauve* (145) and the corresponding coordinates on each of the assemblies were retrieved. To identify families enriched in either euchromatin or heterochromatin a χ^2 test of independence was performed.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transfer of known inversion breakpoints

The coordinates for inversions 2La, 2Rb, 2Rc and 2Rd were obtained from Corbett-Detig et al., (68) and for 2Ru from (69). 50 kb regions flanking each side of the insertion were obtained and mapped using *minimap2* (146) against the scaffolded genome assemblies to transfer the breakpoints. To validate the breakpoint, we determined if long reads spanned the breakpoint using the genome browser IGV version 2.4.19 (139).

Detection of putatively active TE families

To identify potentially active TE families, we identified families with more than two identical full-length fragment copies in at least six of the seven annotated genomes. We determined the fraction of identical copies of these families by identifying all their insertions in the genome and calculating the sequence identity of all their bases against the consensus by performing a nucleotide BLAST. Given that the polishing of the genomes using Illumina reads could have modified the sequence of the insertions thus affecting the age estimation, we used dnaPipeTE (147) to estimate the relative age of the TE families using the raw Illumina reads for the six genomes that we sequenced. We compared the TE landscape obtained using dnaPipeTE with that obtained using the BLAST procedure, using a Kolmogorov-Smirnov test corrected for multiple testing using the Benjamini–Hochberg procedure (Additional file 1: Table S16). Given that we observed few significant differences, we continued using the landscape data obtained using the BLAST procedure. We identified the families where the majority of the bases of their insertions were on the peak of identical sequences in the TE landscape (>50% of the bases with >99% base identity) in more than five of the seven genomes we analyzed. Finally, we assessed the ability to

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

actively transpose of strong candidates by identifying their intact ORFs, LTRs (in the case of LTR retrotransposons) and target site duplication (TSD).

Classification of TEs by their genomic location

To determine the location of TEs we used the *findOverlaps* function from the *GenomicAlignments* R package (148) using default parameters. Both the TE and the gene annotation were converted to *GenomicRanges* objects ignoring strand information in the case of TEs.

Insecticide resistance genes

A list with a total of 43 relevant insecticide resistance genes was generated taking several works into consideration (3, 101-103) (Additional file 1: Table S14). To determine the position of the L to M nonsynonymous substitution that we observed in AGAP004707 (*para*) we used the position from the CAM12801.1 reference sequence.

Immune-related genes

The full list of 414 immune-related genes from *An. gambiae* was downloaded from ImmunoDB (105). We conserved the 281 most reliable genes filtering by the STATUS field and conserving only those with A or B scores.

TFBS and promoter identification

The matrices for *dl* (MA0022.1), *cnc::maf-S* (MA0530.1) and *Stat92E* (MA0532.1) were downloaded from JASPAR (http://jaspar.genereg.net/) (149). The sequences for the TEs of

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interest were obtained using *getSeq* from the *Biostrings* R package. The TFBS in the sequences were identified using the web version of FIMO (150) from the MEME SUITE (151) with default parameters. The ElemeNT online tool was used to identify promoter motifs (152).

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the genome sequencing data obtained in this work, as well as the genome assembly are available in NCBI SRA and NCBI Genbank respectively, under the BioProject accession number PRJNA676011.

Competing interests

The authors declare that they have no competing interests.

Funding

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This study was supported by the Ministry of Economy, Industry and Competitiveness of Spain (BFU2017-82937-P) to JG. DA was supported by an ANR grant (ANR-18-CE35-0002-01 – WILDING). NMLP was funded by AUF and CIRMF scholarships.

Authors' contributions

DA and JG conceived and designed the experiments. NMLP, SEN and LA performed the data generation. CVC, DA and JG performed the data analysis. CVC and JG wrote and revised the manuscript with input from all authors. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the González Lab for comments on the manuscript. We thank the Ecology of Vectorial Systems team at the CIMRF (Franceville, Gabon) for their support in field collections. We thank Jean Pierre Agbor and Serge Donfanck for their commitment in larvae collections in Douala (Cameroon).

Authors' information

Twitter handles: @VargasChavezC (Carlos Vargas-Chavez); @d_ayalag (Diego Ayala); @ GonzalezLab_BCN (Josefa González).

REFERENCES

1. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 2015;347(6217).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Tene Fossog B, Ayala D, Acevedo P, Kengne P, Ngomo Abeso Mebuy I, Makanga B, et al. Habitat segregation and ecological character displacement in cryptic African malaria mosquitoes. Evolutionary Applications. 2015;8(4):326-45.

3. Fossog Tene B, Poupardin R, Costantini C, Awono-Ambene P, Wondji CS, Ranson H, et al. Resistance to DDT in an Urban Setting: Common Mechanisms Implicated in Both M and S Forms of Anopheles gambiae in the City of Yaoundé Cameroon. PLOS ONE. 2013;8(4):e61408.

4. Kengne P, Charmantier G, Blondeau-Bidet E, Costantini C, Ayala D. Tolerance of disease-vector mosquitoes to brackish water and their osmoregulatory ability. Ecosphere. 2019;10(10):e02783.

5. Vontas J, Grigoraki L, Morgan J, Tsakireli D, Fuseini G, Segura L, et al. Rapid selection of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities. Proceedings of the National Academy of Sciences. 2018;115(18):4619.

6. Fouet C, Kamdem C, Gamez S, White BJ. Extensive genetic diversity among populations of the malaria mosquito Anopheles moucheti revealed by population genomics. Infection, Genetics and Evolution. 2017;48:27-33.

7. Wiebe A, Longbottom J, Gleave K, Shearer FM, Sinka ME, Massey NC, et al. Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. Malaria journal. 2017;16(1):85-.

8. Perugini E, Guelbeogo WM, Calzetta M, Manzi S, Virgillito C, Caputo B, et al. Behavioural plasticity of Anopheles coluzzii and Anopheles arabiensis undermines LLIN community protective effect in a Sudanese-savannah village in Burkina Faso. Parasites & vectors. 2020;13(1):277-.

9. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. A Polytene Chromosome Analysis of the <em>Anopheles gambiae</em> Species Complex. Science. 2002;298(5597):1415.

10. Ayala D, Acevedo P, Pombi M, Dia I, Boccolini D, Costantini C, et al. Chromosome inversions and ecological plasticity in the main African malaria mosquitoes. Evolution. 2017;71(3):686-701.

11. Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, Bassole IHN, et al. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. BMC Ecology. 2009;9(1):16.

12. Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, Ose K, et al. Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. BMC Ecology. 2009;9(1):17.

13. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Transactions of The Royal Society of Tropical Medicine and Hygiene. 1979;73(5):483-97.

14. Fouet C, Gray E, Besansky NJ, Costantini C. Adaptation to Aridity in the Malaria Mosquito Anopheles gambiae: Chromosomal Inversion Polymorphism and Body Size Influence Resistance to Desiccation. PLOS ONE. 2012;7(4):e34841.

15. Ayala D, Zhang S, Chateau M, Fouet C, Morlais I, Costantini C, et al. Association mapping desiccation resistance within chromosomal inversions in the African malaria vector Anopheles gambiae. Molecular Ecology. 2019;28(6):1333-42.

16. Labbé P, Berthomieu A, Berticat C, Alout H, Raymond M, Lenormand T, et al. Independent Duplications of the Acetylcholinesterase Gene Conferring Insecticide Resistance in the Mosquito Culex pipiens. Molecular Biology and Evolution. 2007;24(4):1056-67.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Assogba BS, Djogbénou LS, Milesi P, Berthomieu A, Perez J, Ayala D, et al. An ace-1 gene duplication resorbs the fitness cost associated with resistance in Anopheles gambiae, the main malaria mosquito. Scientific Reports. 2015;5(1):14529.

18. Weetman D, Djogbenou LS, Lucas E. Copy number variation (CNV) and insecticide resistance in mosquitoes: evolving knowledge or an evolving problem? Current opinion in insect science. 2018;27:82-8.

19. Lucas ER, Miles A, Harding NJ, Clarkson CS, Lawniczak MKN, Kwiatkowski DP, et al. Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes. Genome Research. 2019;29(8):1250-61.

20. Mitri C, Markianos K, Guelbeogo WM, Bischoff E, Gneme A, Eiglmeier K, et al. The kdr-bearing haplotype and susceptibility to Plasmodium falciparum in Anopheles gambiae: genetic correlation and functional testing. Malaria Journal. 2015;14(1):391.

21. Kamdem C, Fouet C, Gamez S, White BJ. Pollutants and Insecticides Drive Local Adaptation in African Malaria Mosquitoes. Mol Biol Evol. 2017;34(5):1261-75.

22. King SA, Onayifeke B, Akorli J, Sibomana I, Chabi J, Manful-Gwira T, et al. The Role of Detoxification Enzymes in the Adaptation of the Major Malaria Vector Anopheles gambiae (Giles; Diptera: Culicidae) to Polluted Water. Journal of Medical Entomology. 2017;54(6):1674-83.

23. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Molecular Ecology. 2013;22(6):1503-17.

24. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Molecular Ecology. 2019;28(6):1537-49.

25. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. Nature Reviews Genetics. 2017;18(2):71-86.

26. Mathiopoulos KD, Della Torre A, Predazzi V, Petrarca V, Coluzzi M. Cloning of inversion breakpoints in the Anopheles gambiae complex traces a transposable element at the inversion junction. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(21):12444-9.

27. Gray YH. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. 2000;16(10):461-8.

28. Reis M, Vieira CP, Lata R, Posnien N, Vieira J. Origin and Consequences of Chromosomal Inversions in the virilis Group of Drosophila. Genome Biology and Evolution. 2018;10(12):3152-66.

29. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nature Reviews Genetics. 2020;21(10):597-614.

30. Shahid S, Slotkin RK. The current revolution in transposable element biology enabled by long reads. Current Opinion in Plant Biology. 2020;54:49-56.

31. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*. Science. 2002;298(5591):129-49.

32. Marinotti O, Cerqueira GC, De Almeida LGP, Ferro MIT, Da Silva Loreto EL, Zaha A, et al. The Genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids Research. 2013;41(15):7387-400.

33. Jiang X, Peery A, Hall AB, Sharma A, Chen XG, Waterhouse RM, et al. Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. Genome biology. 2014;15(9):459-.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

34. Zhou D, Zhang D, Ding G, Shi L, Hou Q, Ye Y, et al. Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. BMC Genomics. 2014;15(1).

35. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science. 2015;347(6217).

36. Lau YL, Lee WC, Chen J, Zhong Z, Jian J, Amir A, et al. Draft genomes of Anopheles cracens and Anopheles maculatus: Comparison of simian malaria and human malaria vectors in peninsular Malaysia. PLoS ONE. 2016;11(6):1-24.

37. Chakraborty M, Ramaiah A, Adolfi A, Halas P, Kaduskar B, Ngo LT, et al. Hidden features of the malaria vector mosquito, <em>Anopheles stephensi</em>, revealed by a high-quality reference genome. bioRxiv. 2020:2020.05.24.113019.

38. Compton A, Liang J, Chen C, Lukyanchikova V, Qi Y, Potters M, et al. The beginning of the end: a chromosomal assembly of the New World malaria mosquito ends with a novel telomere. bioRxiv. 2020:2020.04.17.047084.

39. de Melo ES, Wallau GdL. Transposable elements are constantly exchanged by horizontal transfer reshaping mosquito genomes. bioRxiv. 2020:2020.06.23.166744.

40. Yang X, Lee WP, Ye K, Lee C. One reference genome is not enough. Genome Biology. 2019;20(1):19-21.

41. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nature Plants. 2020;6(8):914-20.

42. Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, et al. Discovery and population genomics of structural variation in a songbird genus. Nature Communications. 2020;11(1):3403.

43. Quesneville H, Nouaud D, Anxolabéhère D. P elements and MITE relatives in the whole genome sequence of Anopheles gambiae. BMC Genomics. 2006;7.

44. Boulesteix M, Simard F, Antonio-Nkondjio C, Awono-Ambene HP, Fontenille D, Biémont C. Insertion polymorphism of transposable elements and population structure of Anopheles gambiae M and S molecular forms in Cameroon. Molecular Ecology. 2007;16(2):441-52.

45. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, Della Torre A. Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. Malaria Journal. 2008;7:1-10.

46. Esnault C, Boulesteix M, Duchemin JB, Koffi AA, Chandre F, Dabiré R, et al. High genetic differentiation between the M and S molecular forms of Anopheles gambiae in Africa. PloS one. 2008;3(4):e1968-e.

47. Salgueiro P, Moreno M, Simard F, O'Brochta D, Pinto J. New Insights into the Population Structure of Anopheles gambiae s.s. in the Gulf of Guinea Islands Revealed by Herves Transposable Elements. PLoS ONE. 2013;8(4).

48. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, et al. A high-quality de novo genome assembly from a single mosquito using pacbio sequencing. Genes. 2019;10(1).

49. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

50. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element Diversification in De Novo Annotation Approaches. PLOS ONE. 2011;6(1):e16526.

51. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS computational biology. 2005;1(2):166-75.

52. Platt RN, 2nd, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome biology and evolution. 2016;8(2):403-10.

53. Marsano RM, Leronni D, D'Addabbo P, Viggiano L, Tarasco E, Caizzi R. Mosquitoes LTR retrotransposons: a deeper view into the genomic sequence of Culex quinquefasciatus. PloS one. 2012;7(2):e30770-e.

54. Zhou Y, Cahan SH. A Novel Family of Terminal-Repeat Retrotransposon in Miniature (TRIM) in the Genome of the Red Harvester Ant, Pogonomyrmex barbatus. PLoS ONE. 2012;7(12).

55. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics. 2014;15(1):86.

56. Witte C-P, Le QH, Bureau T, Kumar A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proceedings of the National Academy of Sciences. 2001;98(24):13778.

57. Gao D, Li Y, Kim KD, Abernathy B, Jackson SA. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome Biology. 2016;17(1):7.

58. Barrón MG, Paupy C, Rahola N, Akone-Ella O, Ngangue MF, Wilson-Bahun TA, et al. A new species in the major malaria vector complex sheds light on reticulated species evolution. Scientific Reports. 2019;9(1):1-13.

59. Vargas-Chavez C, González J. Transposable elements in Anopheles species: refining annotation strategies towards population-level analysis. In: Dupuis ORaJ, editor. Population Genomics: Insects: Springer; 2021.

60. Sessegolo C, Burlet N, Haudry A. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biology Letters. 2016;12(8):0-3.

61. Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, Smith CD, et al. Genome mapping and characterization of the Anopheles gambiae heterochromatin. BMC Genomics. 2010;11(1):459.

62. Tu Z. Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, Aedes aegypti. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(14):7475-80.

63. Xia A, Sharakhova MV, Leman SC, Tu Z, Bailey JA, Smith CD, et al. Genome landscape and evolutionary plasticity of chromosomes in malaria mosquitoes. PLoS ONE. 2010;5(5).

64. Ruiz JL, Ranford-Cartwright LC, Gómez-Díaz E. The regulatory genome of the malaria vector <em>Anopheles gambiae</em>: integrating chromatin accessibility and gene expression. bioRxiv. 2020:2020.06.22.164228.

65. Tu Z. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(4):1699-704.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

66. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, Santolamazza F, et al. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(16):6258-62.

67. Lobo NF, Sangaré DM, Regier AA, Reidenbach KR, Bretz DA, Sharakhova MV, et al. Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. Malaria journal. 2010;9:293-.

68. Corbett-Detig RB, Said I, Calzetta M, Genetti M, McBroome J, Maurer NW, et al. Fine-Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the <em>Anopheles gambiae</em> Species Complex Using Proximity-Ligation Sequencing. Genetics. 2019;213(4):1495.

69. White BJ, Cheng C, Sangaré D, Lobo NF, Collins FH, Besansky NJ. The population genomics of trans-specific inversion polymorphisms in Anopheles gambiae. Genetics. 2009;183(1):275-88.

70. Smit A, Hubley R, Green P. *RepeatMasker Open-4.0* 2013-2015 [Available from: http://www.repeatmasker.org.

71. Diesel JF, Ortiz MF, Marinotti O, Vasconcelos ATR, Loreto ELS. A re-annotation of the Anopheles darlingi mobilome. Genetics and Molecular Biology. 2019;42(1):125-31.

72. Fonseca PM, Moura RD, Wallau GL, Loreto ELS. The mobilome of Drosophila incompta, a flower-breeding species: comparison of transposable element landscapes among generalist and specialist flies. Chromosome Research. 2019;27(3):203-19.

73. Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evolutionary Biology. 2019;19(1):11-.

74. Tubío JMC, Naveira H, Costas J. Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of Anopheles gambiae. Molecular Biology and Evolution. 2005;22(1):29-39.

75. Tubío JMC, Tojo M, Bassaganyas L, Escaramis G, Sharakhov IV, Sharakhova MV, et al. Evolutionary Dynamics of the Ty3/Gypsy LTR Retrotransposons in the Genome of Anopheles gambiae. PLOS ONE. 2011;6(1):e16328.

76. Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. PLOS Genetics. 2019;15(2):e1007900.

77. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic acids research. 2019;47(D1):D419-D26.

78. Ingham VA, Pignatelli P, Moore JD, Wagstaff S, Ranson H. The transcription factor Maf-S regulates metabolic resistance to insecticides in the malaria vector Anopheles gambiae. BMC genomics. 2017;18(1):669-.

79. Osta MA, Christophides GK, Vlachou D, Kafatos FC. Innate immunity in the malaria vector <em>Anopheles gambiae</em>: comparative and functional genomics. Journal of Experimental Biology. 2004;207(15):2551.

80. Xie D, Chen C-C, Ptaszek LM, Xiao S, Cao X, Fang F, et al. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome research. 2010;20(6):804-15.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

81. Bonizzoni M, Afrane Y, Dunn WA, Atieli FK, Zhou G, Zhong D, et al. Comparative Transcriptome Analyses of Deltamethrin-Resistant and -Susceptible Anopheles gambiae Mosquitoes from Kenya by RNA-Seq. PLOS ONE. 2012;7(9):e44607.

82. Vannini L, Willis JH. Localization of RR-1 and RR-2 cuticular proteins within the cuticle of Anopheles gambiae. Arthropod structure & development. 2017;46(1):13-29.

83. Balabanidou V, Kefi M, Aivaliotis M, Koidou V, Girotti JR, Mijailovsky SJ, et al. Mosquitoes cloak their legs to resist insecticides. Proceedings Biological sciences. 2019;286(1907):20191091-.

84. Sriwichai P, Rongsiryam Y, Jariyapan N, Sattabongkot J, Apiwathnasorn C, Nacapunchai D, et al. Cloning of a Trypsin-like Serine Protease and expression Patterns during *Plasmodium falciparum* invasion in the mosquito, *Anopheles dirus* (Peyton and Harrison). Archives of Insect Biochemistry and Physiology. 2012;80(3):151-65.

85. Dias-Lopes G, Borges-Veloso A, Saboia-Vahia L, Domont GB, Britto C, Cuervo P, et al. Expression of active trypsin-like serine peptidases in the midgut of sugar-feeding female Anopheles aquasalis. Parasites & vectors. 2015;8:296-.

86. Hughes GL, Ren X, Ramirez JL, Sakamoto JM, Bailey JA, Jedlicka AE, et al. Wolbachia Infections in Anopheles gambiae Cells: Transcriptomic Characterization of a Novel Host-Symbiont Interaction. PLOS Pathogens. 2011;7(2):e1001296.

87. Cao X, Gulati M, Jiang H. Serine protease-related proteins in the malaria mosquito, Anopheles gambiae. Insect biochemistry and molecular biology. 2017;88:48-62.

88. Kent LB, Walden KKO, Robertson HM. The Gr Family of Candidate Gustatory and Olfactory Receptors in the Yellow-Fever Mosquito Aedes aegypti. Chemical Senses. 2008;33(1):79-93.

89. Dabiré RK, Namountougou M, Diabaté A, Soma DD, Bado J, Toé HK, et al. Distribution and frequency of kdr mutations within Anopheles gambiae s.l. populations and first report of the ace.1 G119S mutation in Anopheles arabiensis from Burkina Faso (West Africa). PloS one. 2014;9(7):e101484-e.

90. Silva APB, Santos JMM, Martins AJ. Mutations in the voltage-gated sodium channel gene of anophelines and their association with resistance to pyrethroids - a review. Parasites & vectors. 2014;7:450-.

91. Cheung J, Mahmood A, Kalathur R, Liu L, Carlier PR. Structure of the G119S Mutant Acetylcholinesterase of the Malaria Vector Anopheles gambiae Reveals Basis of Insecticide Resistance. Structure (London, England : 1993). 2018;26(1):130-6.e2.

92. Elanga-Ndille E, Nouage L, Ndo C, Binyang A, Assatse T, Nguiffo-Nguete D, et al. The G119S Acetylcholinesterase (Ace-1) Target Site Mutation Confers Carbamate Resistance in the Major Malaria Vector Anopheles gambiae from Cameroon: A Challenge for the Coming IRS Implementation. Genes. 2019;10(10):790.

93. Fadel AN, Ibrahim SS, Tchouakui M, Terence E, Wondji MJ, Tchoupo M, et al. A combination of metabolic resistance and high frequency of the 1014F kdr mutation is driving pyrethroid resistance in Anopheles coluzzii population from Guinea savanna of Cameroon. Parasites & Vectors. 2019;12(1):263.

94. Santolamazza F, Calzetta M, Etang J, Barrese E, Dia I, Caccone A, et al. Distribution of knock-down resistance mutations in Anopheles gambiae molecular forms in west and west-central Africa. Malaria Journal. 2008;7(1):74.

95. Jones CM, Liyanapathirana M, Agossa FR, Weetman D, Ranson H, Donnelly MJ, et al. Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sodium channel of Anopheles gambiae. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(17):6614-9.

96. Essandoh J, Yawson AE, 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. Malaria Journal. 2013;12(1):404.

97. Wilson TG. Transposable Elements as Initiators of Insecticide Resistance. Journal of Economic Entomology. 1993;86(3):645-51.

98. ffrench-Constant R, Daborn P, Feyereisen R. Resistance and the jumping gene. BioEssays. 2006;28(1):6-8.

99. Rostant WG, Wedell N, Hosken DJ. Chapter 2 - Transposable Elements and Insecticide Resistance. In: Goodwin SF, Friedmann T, Dunlap JC, editors. Advances in Genetics. 78: Academic Press; 2012. p. 169-201.

100. Weedall GD, Riveron JM, Hearn J, Irving H, Kamdem C, Fouet C, et al. An Africa-wide genomic evolution of insecticide resistance in the malaria vector Anopheles funestus involves selective sweeps, copy number variations, gene conversion and transposons. PLOS Genetics. 2020;16(6):e1008822.

101. Main BJ, Everitt A, Cornel AJ, Hormozdiari F, Lanzaro GC. Genetic variation associated with increased insecticide resistance in the malaria mosquito, Anopheles coluzzii. Parasites & vectors. 2018;11(1):225-.

102. Adolfi A, Poulton B, Anthousi A, Macilwee S, Ranson H, Lycett GJ. Functional genetic validation of key genes conferring insecticide resistance in the major African malaria vector, <em>Anopheles gambiae</em>. Proceedings of the National Academy of Sciences. 2019;116(51):25764.

103. Bamou R, Sonhafouo-Chiana N, Mavridis K, Tchuinkam T, Wondji CS, Vontas J, et al. Status of Insecticide Resistance and Its Mechanisms in Anopheles gambiae and Anopheles coluzzii Populations from Forest Settings in South Cameroon. Genes. 2019;10(10):741.

104. Cassone BJ, Kamdem C, Cheng C, Tan JC, Hahn MW, Costantini C, et al. Gene expression divergence between malaria vector sibling species Anopheles gambiae and An. coluzzii from rural and urban Yaoundé Cameroon. Molecular Ecology. 2014;23(9):2242-59.

105. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, et al. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science (New York, NY). 2007;316(5832):1738-43.

106. Clayton AM, Dong Y, Dimopoulos G. The ***Anopheles*** Innate Immune System in the Defense against Malaria Infection. Journal of Innate Immunity. 2014;6(2):169-81.

107. Volz J, Müller H-M, Zdanowicz A, Kafatos FC, Osta MA. A genetic module regulates the melanization response of Anopheles to Plasmodium. Cellular Microbiology. 2006;8(9):1392-405.

108. Oliveira GdA, Lieberman J, Barillas-Mury C. Epithelial nitration by a peroxidase/NOX5 system mediates mosquito antiplasmodial immunity. Science (New York, NY). 2012;335(6070):856-9.

109. Dennison NJ, BenMarzouk-Hidalgo OJ, Dimopoulos G. MicroRNA-regulation of Anopheles gambiae immunity to Plasmodium falciparum infection and midgut microbiota. Developmental & Comparative Immunology. 2015;49(1):170-8.

110. Ruiz JL, Yerbanga RS, Lefèvre T, Ouedraogo JB, Corces VG, Gómez-Díaz E. Chromatin changes in Anopheles gambiae induced by Plasmodium falciparum infection. Epigenetics & Chromatin. 2019;12(1):5.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

111. Fernández-Medina RD, Ribeiro JMC, Carareto CMA, Velasque L, Struchiner CJ. Losing identity: structural diversity of transposable elements belonging to different classes in the genome of Anopheles gambiae. BMC genomics. 2012;13.

112. Ou S, Liu J, Chougule KM, Fungtammasan A, Seetharam AS, Stein JC, et al. Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nature Communications. 2020;11(1):2288.

113. Kofler R, Nolte V, Schlötterer C. Tempo and Mode of Transposable Element Activity in Drosophila. PLOS Genetics. 2015;11(7):e1005406.

114. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. Mobile DNA. 2015;6:21-.

115. Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. Nature Communications. 2020;11(1):1796.

116. Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, et al. European maize genomes highlight intraspecies variation in repeat and gene content. Nature Genetics. 2020;52(9):950-7.

117. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. eLife. 2016;5:e15716.

118. Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass Brachypodium distachyon. New Phytologist. 2020;227(6):1736-48.

119. Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. Molecular biology and evolution. 2013;30(10):2311-27.

120. Johnson MTJ, Munshi-South J. Evolution of life in urban environments. Science. 2017;358(6363):eaam8327.

121. Mateo L, Ullastres A, González J. A Transposable Element Insertion Confers Xenobiotic Resistance in Drosophila. PLOS Genetics. 2014;10(8):e1004560.

122. Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. Transposable elements contribute to the genomic response to insecticides in Drosophila melanogaster. Philosophical Transactions of the Royal Society B: Biological Sciences. 2020;375(1795):20190341.

123. Kamgang B, Tchapga W, Ngoagouni C, Sangbakembi-Ngounou C, Wondji M, Riveron JM, et al. Exploring insecticide resistance mechanisms in three major malaria vectors from Bangui in Central African Republic. Pathogens and global health. 2018;112(7):349-59.

124. Grau-Bové X, Tomlinson S, O’Reilly AO, Harding NJ, Miles A, Kwiatkowski D, et al. Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and Interkaryotypic Introgression. Molecular Biology and Evolution. 2020;37(10):2900-17.

125. The Anopheles gambiae Genomes Consortium. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. Genome Research. 2020;30(10):1533-46.

126. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome research. 2014;24(12):1963-76.

127. Jiang J-C, Upton KR. Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines. Mobile DNA. 2019;10:16-.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

128. Villanueva-Cañas JL, Horvath V, Aguilera L, González J. Diverse families of transposable elements affect the transcriptional regulation of stress-response genes in Drosophila melanogaster. Nucleic Acids Research. 2019;47(13):6842-57.

129. Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philosophical Transactions of the Royal Society B: Biological Sciences. 2020;375(1795):20190347.

130. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research. 2017;27(5):722-36.

131. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.

132. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome research. 2017;27(5):737-46.

133. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods. 2015;12(8):733-5.

134. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014;9(11):e112963.

135. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000Research. 2017;6(1287).

136. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology. 2019;20(1):224.

137. Shumate A, Salzberg SL. Liftoff: an accurate gene annotation mapping tool. bioRxiv. 2020:2020.06.24.169680.

138. Okonechnikov K, Golosova O, Fursov M, the Ut. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166-7.

139. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-6.

140. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England). 2012;28(23):3150-2.

141. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An Automatic Transposable Element Classification Tool. PLOS ONE. 2014;9(5):e91929.

142. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10:421-.

143. Fernández-Medina RD, Struchiner CJ, Ribeiro JMC. Novel transposable elements from Anopheles gambiae. BMC Genomics. 2011;12(1):260-.

144. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6(1):4-9.

145. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE. 2010;5(6):e11147.

146. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.

147. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome biology and evolution. 2015;7(4):1192-205.

148. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology. 2013;9(8):e1003118.

149. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research. 2004;32(suppl_1):D91-D4.

150. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017-8.

151. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-8.

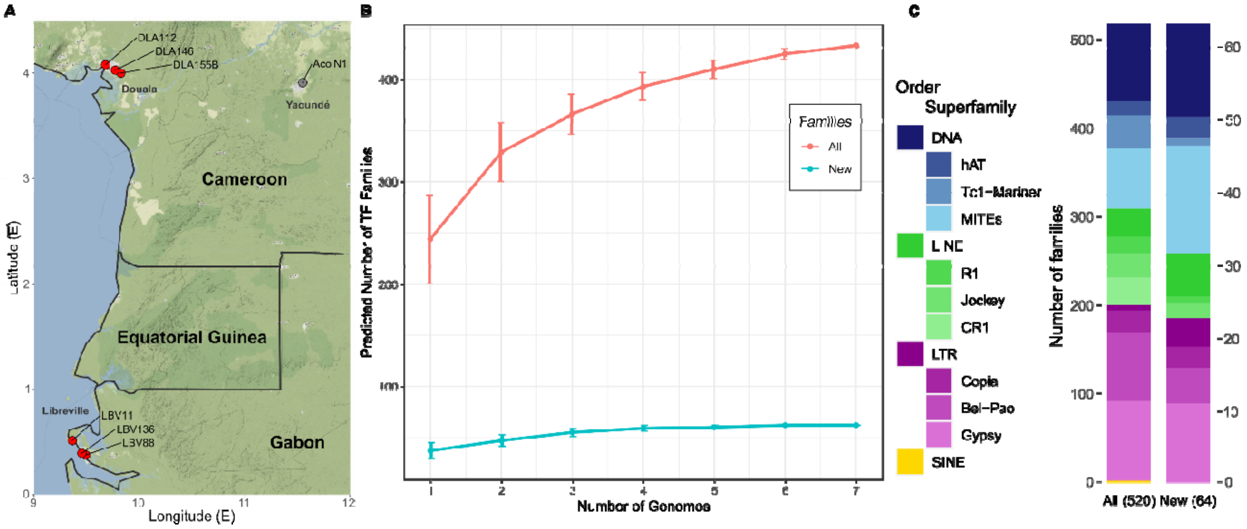
152. Sloutskin A, Danino YM, Orenstein Y, Zehavi Y, Doniger T, Shamir R, et al. ElemeNT: a computational tool for detecting core promoter elements. Transcription. 2015;6(3):41-50.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURES

Figure 1. Transposable elements in *An. coluzzii*.

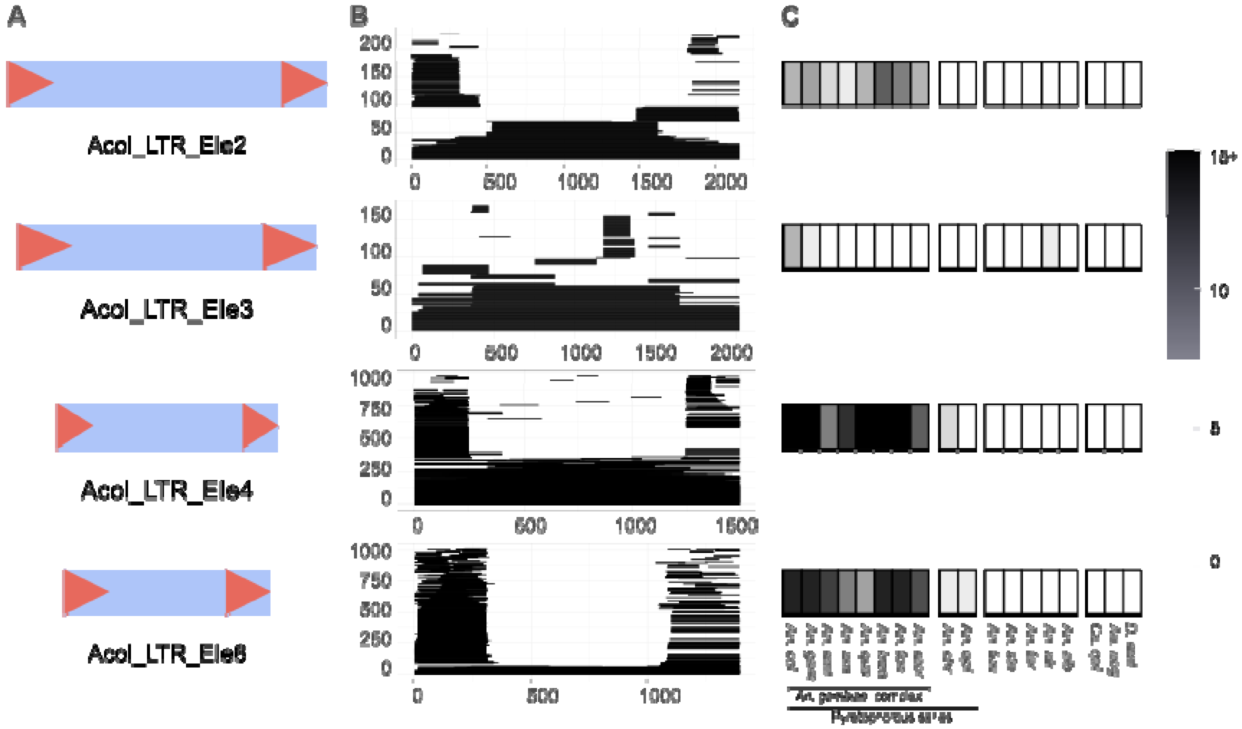
A) Geographic location of the six breeding sites analyzed (in red) and of the place of origin of the Ngousso colony (in grey) which was used to generate the *AcolN1* genome. B) Number of TE families identified when using a single genome or when using all possible combinations of more than one genome. The red line shows the total number of TE families and the blue line shows the number of newly described families. C) Classification of all TE families and newly described families in *An. coluzzii*. The three most abundant superfamilies from each order are shown.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. Structure, abundance and phylogenetic distribution of novel TE families.

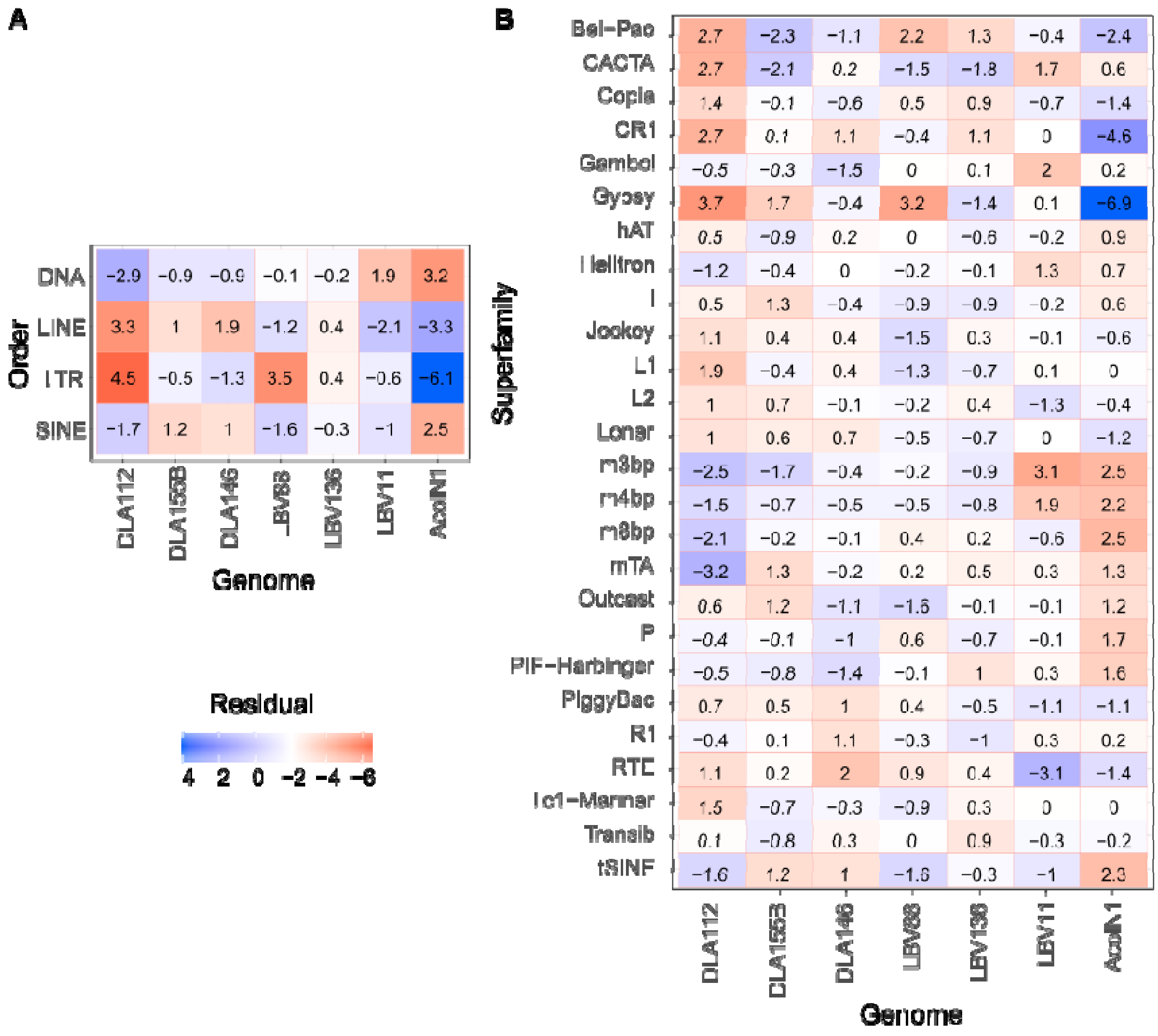
The four newly identified TRIMs families are shown, for the remaining 60 novel families see Additional file 2: Figure S1. A) The structure of each new family is displayed: the light blue box represents the full extension of the TE and the red arrows represent LTRs. B) All insertions for each TE family are shown as a coverage plot where each line represents a copy in a genome. C) Phylogenetic distribution of the TE family insertions in 15 members of the *Anopheles* genus, *Culex quinquefasciatus, Ae. Aegypti* and *D. melanogaster*. The number of insertions with more than 80% identity and spanning at least 80% of the consensus, in each species is shown using a black and white gradient. Species with no insertions are shown in white while species with 15 or more insertions are shown in black.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. Differences in TE content between the seven *An. coluzzii* genomes.

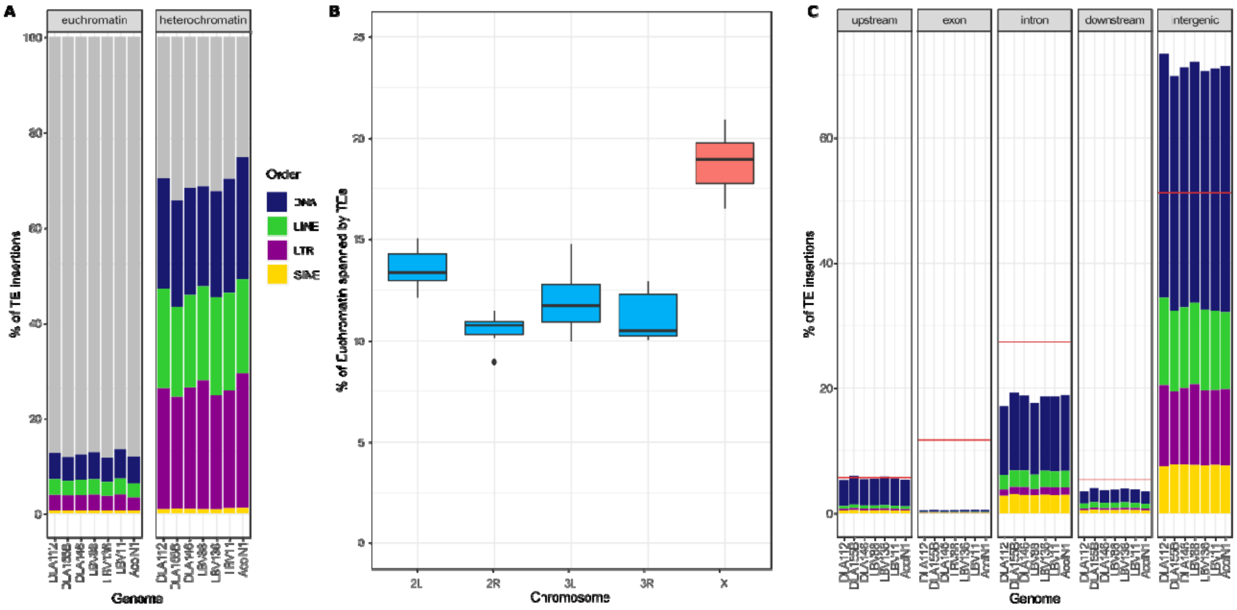
Differences are shown at the (A) order and (B) superfamily levels. χ^2 tests were performed for the number of insertions and the Person's residuals are shown. Note that MITEs are divided into the m3bp, m4bp, m8bp and mTA superfamilies.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. TE insertions distribution throughout the genomes.

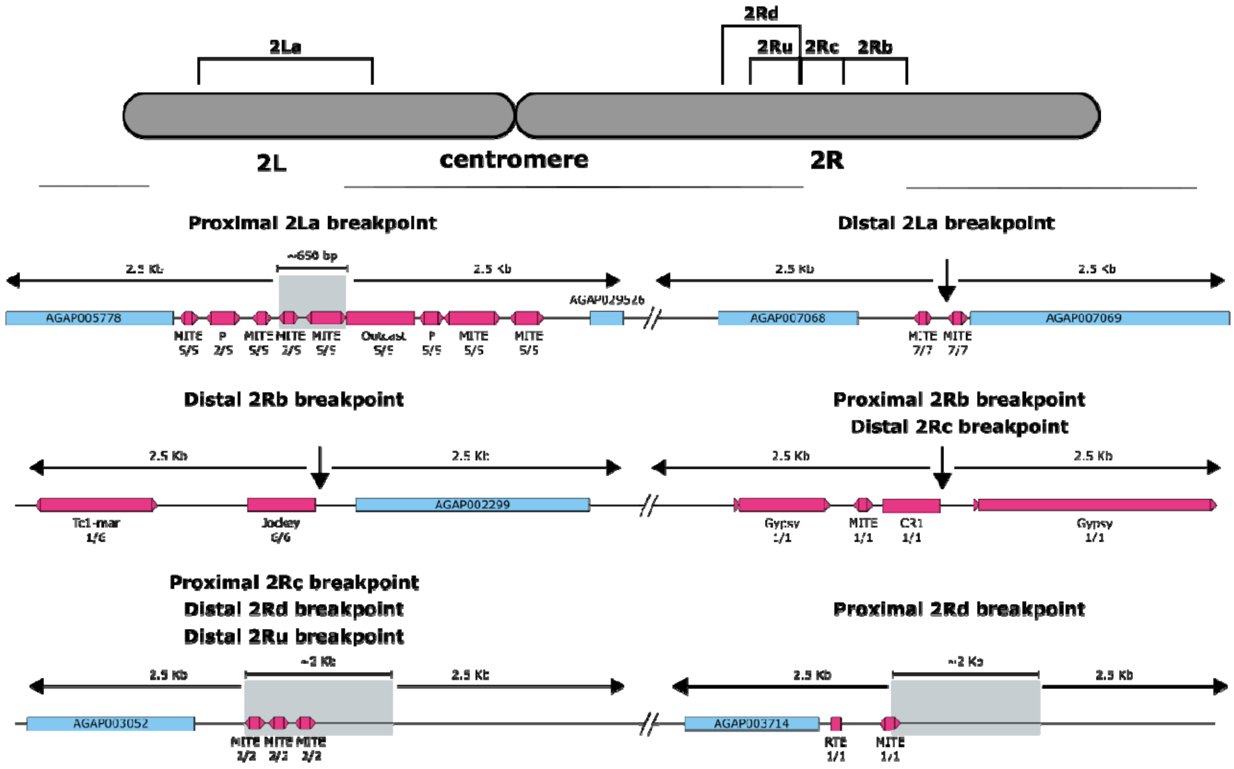
A) Percentage of euchromatin and heterochromatin occupied by TEs in each of the seven analyzed genomes. Each order is shown in a different color. B) Boxplots of the percentage of the euchromatin of each chromosome covered by TEs. Autosomes are shown in blue and the X chromosome in red. C) Percentage of TE insertions in each genome that fall in a specific genomic region. A red line is used to display the expected percentage that should be covered by TEs taking in consideration the size of the genomic region. Each order is shown in a different color as in A).



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

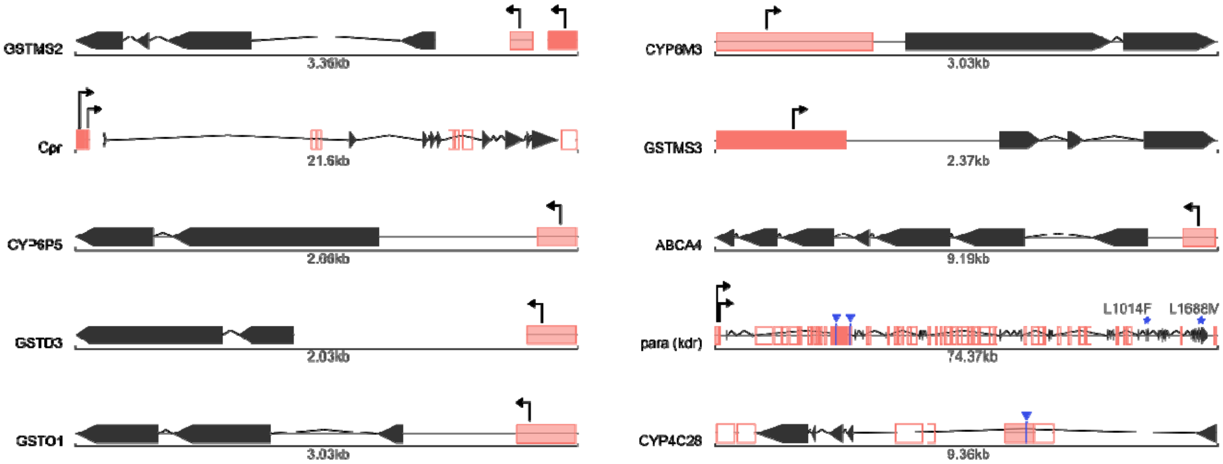
Figure 5. TE insertions near known inversion breakpoints.

Diagram of the chromosome 2 with the analyzed inversions. For each inversion both breakpoints, proximal (closer to the centromere) and distal (farther from the centromere), plus 2.5 kb to each side are shown. When the position of a breakpoint was not identified at the single base pair level, the interval where the breakpoint is predicted to be is shown in a grey box. Genes are shown as blue boxes while TEs are shown as pink boxes. Below each TE, the family of the TE is shown and below the family name the number of genomes where the insertion was found and the number of genomes where the breakpoint region was identified. Note that breakpoints are shared among some of the inversions.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6. TE insertions in the neighborhood of genes involved in insecticide resistance. The gene structure is shown in black with arrows representing the exons. TE insertions are depicted as red boxes. When containing a TFBS for *cnc* or a promoter they are filled in red, otherwise they are empty. The red color is darker on fixed TEs and lighter on polymorphic TEs. Promoters are shown as arrows while *cnc* binding sites are shown in blue. Resistance alleles are shown for *para* (*kdr*).



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ADDITIONAL FILES

File name: Additional file 1

File format: Microsoft Excel Binary File format (xls)

Title of data: Supplementary Tables

Description of data: Supplementary Tables

File name: Additional file 2

File format: Portable document format (pdf)

Title of data: Figure S1. Novel TE families

Description of data: Newly described families. A) The structure of each new family is displayed: the light blue box represents the full extension of the TE and the red arrows represent LTRs. B) All insertions for each TE family are shown as a coverage plot where each line represents a copy in a genome. C) Phylogenetic distribution of the TE family insertions in 15 members of the *Anopheles* genus, *Culex quinquefasciatus, Ae. Aegypti* and *D. melanogaster*. The number of insertions with more than 80% identity and spanning at least 80% of the consensus, in each species is shown using a black and white gradient. Species with no insertions are shown in white while species with 50 or more insertions are shown in black.

File name: Additional file 3

File format: Portable document format (pdf)

Title of data: Figure S2. Number of TE insertions vs genome size

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393231; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Description of data: Comparison of the bases spanned by TEs in each genome with their full genome sizes.

File name: Additional file 4

File format: Portable document format (pdf)

Title of data: Figure S3. TE landscapes

Description of data: TE landscapes for the six genomes sequenced in this work generated using dnaPipeTE

File name: Additional file 5

File format: Portable document format (pdf)

Title of data: Figure S4. Genes with TE insertions from active families

Description of data: Diagrams of TE insertions closer than 1 kb to genes showing the gene structure and the TE insertion

File name: Additional file 6

File format: Portable document format (pdf)

Title of data: Figure S5. Genes associated with insecticide resistance with TE insertions

Description of data: Diagrams of genes associated with insecticide resistance showing the gene structure and the TE insertions closer than 1 kb to gene.