Uncovering transposable element variants and their potential adaptive impact in urban populations of the 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

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. To assess the role of TEs in rapid urban adaptation, the first step is to accurately annotate TE insertions in the genomes of natural populations collected in urban settings.

Results

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 in these genomes and in an additional high quality *An. coluzzii* genome available and identified 64 previously undescribed TE 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

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 identified several TE

insertions that could potentially impact both 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 (8). This extraordinary adaptative capacity makes this malaria vector a threat for malaria control.

69 Thus, elucidating the natural genetic variants underlying the ecological and the physiological
70 responses to fluctuating environments in this species is key for its control.
71
72 At the molecular level, a variety of genetic mechanisms have been related back to the myriad of
73 adaptation processes present in this mosquito. The most prominent and historically studied
74 examples are chromosomal inversions (9, 10). *An. coluzzii* exhibits a large number of
75 polymorphic chromosomal rearrangements (11, 12). Many of these inversions have been
76 associated to environmental adaptation through environmental clines and/or correlation with
77 specific climatic variables (10, 13), such as the inversion 2La associated with aridity tolerance
78 capacity in adults (14, 15). Other types of rearrangements, such as gene duplications, have been
79 involved in insecticide resistance. For example, the acetylcholinesterase (*Ace-1*) gene has been
80 duplicated, maintaining at least a sensitive and a resistance copy, in order to counteract the
81 fitness cost of the resistant phenotype (16-18). Moreover, a recent genome-wide analysis showed
82 that genes containing copy number variants were enriched for insecticide functions (19). Other
83 examples of gene selection due to anthropogenic activities have been found in genes related with
84 detoxification or immunity, particularly in new colonized urban settings (3, 20-22). These
85 adaptive processes have been repeated across West and Central African populations, reducing
86 the efficacity of vector control measures (6). However, while several of the candidate genes
87 responsible for the adaptive capacity of *An. coluzzii* have been identified, our knowledge of the
88 genetic variants underlying differences in these genes lags behind. In particular, very little is
89 known about natural variation in transposable element (TE) insertions in *An. coluzzii*.
90

91 TEs are key players in multiple adaptive processes across species, due to their capacity to
92 generate a wide variety of mutations and to contribute to rapid responses to environmental
93 changes (23, 24). TEs can disperse across the genome regulatory sequences such as promoters,
94 enhancers, insulators, and repressive elements thus affecting nearby gene expression (25).
95 Additionally, they can also act as substrates for ectopic recombination leading to structural
96 mutations such as chromosomal rearrangements (26-28). However, TEs are often ignored when
97 analyzing functional variants in genomes. This is because due to their repetitive nature, TE
98 insertions are difficult to annotate and reads derived from TEs are often discarded in genome-
99 wide analyses (29). Long-read sequencing techniques are needed to get a comprehensive view of
100 TE variation in genomes, as these technique allow the annotation of TE insertions in the genome
101 rather than inferring their position (30, 31).
102
103 Although TE insertions have been annotated genome-wide in several anopheline species
104 including *An. coluzzii*, most studies to date have characterized the TE repertoire in a single
105 genome for each species (32-40). To capture the full extent of TE natural variation and the
106 potential consequences of TE insertions, it is necessary to evaluate multiple genomes in order to
107 comprehensively assess diversity within a species (41-43). This becomes especially relevant
108 when attempting to identify recent TE insertions and their effect in the genome structure and
109 genome function, given that they might be restricted to local populations. So far, our knowledge
110 of *An. coluzzii* genome variation due to TE insertions is limited to a few well-characterized
111 families that have been found to vary across genomes (44-48).
112

113 In this work, we sequenced, using long-read technologies, and assembled the genomes of *An.*
114 *coluzzii* larvae collected in six natural breeding sites in two major cities in Central Africa:
115 Douala (Cameroon) and Libreville (Gabon). We performed a *de novo* TE annotation of the six
116 newly assembled genomes, and we also annotated the previously available *An. coluzzii* genome
117 from Yaoundé (Cameroon) (49). We identified 64 new anopheline TE families and showed that
118 the availability of multiple genomes substantially improves the discovery of TE variants. We
119 further analyzed individual TE insertions that could be acting as enhancers and promoters and
120 that are located nearby genes with functions relevant for the vectorial capacity of the mosquitoes.
121
122
123 **RESULTS**
124 **Six new whole-genome assemblies of *An. coluzzii* from two major cities in Central Africa**
125 To explore the TE diversity in *An. coluzzii*, we used long-read sequencing and performed whole
126 genome assemblies and scaffolding of larvae collected from six natural urban breeding sites:
127 three from Douala, Cameroon, and three from Libreville, Gabon, Central Africa (Figure 1A,
128 Additional File 1: Table S1A; see Material and Methods). Additionally, we performed a
129 reference-guided scaffolding of the available *An. coluzzii* reference genome *AcolN1* using the
130 chromosome level assembly *AgamP4* of *An. gambiae* (49). Although the genomes analyzed
131 varied in sequencing coverage and read length, these differences only had an effect on contig
132 N50 but did not affect other assembly and scaffolding metrics (Additional File 1: Table S1B).
133 Overall the number of scaffolds varied from 5 to 107, with the median being 20, however, the
134 scaffolds' N50 was similar across the seven genomes (Table 1).

135 We assessed the genomes completeness using BUSCO with the dipteran set of genes (50). We
136 obtained percentages of complete genes ranging from 94.2% to 96.6% except for the *DLA155B*
137 sample which had a lower completeness value (89.5%; Table 1). These completeness values for
138 most (5/6) of the samples were similar to those from the *AcolN1* genome assembly which
139 contained 98.9% complete genes (Table 1; Additional file 1: Table S1).
140 Overall the analyzed genomes are comparable in terms of scaffold contiguity and completeness
141 (Table 1 and Additional File 1: Table S1).
142
143 **Table 1. Genome assemblies and scaffolds' statistics for the *An. coluzzii* genomes analyzed**
144 **in this work.**

Genome	Long reads coverage	Illumina coverage	Assembly size (Mb)	Number of contigs	Number of scaffolds	Scaffolds N50 (kb)	Complete BUSCO genes	Genes transferred	TE families identified
DLA112	55X	59X	252	3,917	107	54,591	96.6%	13,469	244
DLA155B	28X	19X	236	2,081	24	52,031	89.5%	13,303	243
DLA146	28X	42X	247	2,036	14	54,960	95.1%	13,314	193
LBV88	31X	41X	245	2,576	19	54,450	94.5%	13,328	280
LBV136	34X	130X	236	2,911	28	52,053	95.2%	13,307	172
LBV11[a]	89X	61X	246	2,608	20	53,712	94.2%	13,393	294
AcolN1[b]	~270X	-	251	205	5	53,057	98.9%	13,487	283

145 Three genomes were collected in Douala (DLA) and three in Libreville (LBV). [a] *LBV11* was
146 sequenced using PacBio technologies, while the other five genomes were sequenced using
147 Oxford Nanopore Technologies. [b] Genome statistics for *AcolN1*, the high quality *de novo*
148 genome assembly reported by Kingan *et al.*, (49) are also included.
149

150 **64 new anopheline TE families discovered in *An. coluzzii***

151 To identify the TE families present in each of the genomes, we used the *TEdenovo* pipeline from
152 the REPET package (51). After several rounds of manual curation, we identified between 172
153 and 294 TE families for each genome (Table 1; Additional file 1: Table S2). Remarkably, while
154 using a single reference would have only allowed the identification of a median of 244 TE
155 families, clustering the TE libraries from an increasing number of genomes allowed the
156 identification of a total of 435 well supported TE families (Figure 1B; see Material and
157 Methods). Interestingly, 64 of these families (32 DNA, 9 LINEs and 23 LTRs) are described here
158 for the first time. The majority of the new families (43/64) had partial matches to other known
159 TEs, thus allowing us to classify them at the superfamily level (Additional file 1: Table S3). The
160 use of multiple references was especially relevant for identifying these previously undescribed
161 families given that using a single genome would have only allowed to identify a median of 37
162 (25-48) novel TE families (Figure 1B).

163 To further characterize these novel families, we estimated the average number of insertions in
164 the seven *An. coluzzii* genomes, and their distribution and abundance in other species from the
165 *Anopheles* genus (Figure 2; Additional file 3: Figure S2; Additional file 1: Table S3). To do this,
166 we first annotated individual TE insertions in the seven *An. coluzzii* genomes using the *TEannot*
167 pipeline from the REPET package (52). To ensure that our annotation was as complete as
168 possible, in addition to the 435 families identified using REPET *de novo* approach, we also
169 added to our library 85 TE families from the *TEfam* database that we found to also be present in
170 the *An. coluzzii* genomes reported in this work (see material and methods). Most probably, these
171 85 families were either not identified by REPET or were discarded during manual curation
172 (Additional file 1: Table S4; see Material and Methods) (53). The final total of 520 families were

173 classified into 23 superfamilies and then further grouped into four orders (DNA, LINE, LTR and
174 SINE; Figure 1C).
175
176 Copies from all 64 new families were found in all seven *An. coluzzii* genomes, further suggesting
177 that these are *bona fide* families. Although the majority of families contain full-length copies in
178 at least one of the seven genomes analyzed, truncated copies were the most abundant (Figure 2B;
179 Additional file 1: Table S3). We identified a median of 72 insertions (ranging from 16 to 1445)
180 per family and genome (Figure 2B; Additional file 3: Figure S2B). Two out of the four TRIM
181 elements identified (*Acol_LTR_Ele 4* and *Acol_LTR_Ele 6*) are among the most abundant new
182 families, with more than 150 insertions (Figure 2B). TRIM elements are non-autonomous
183 retrotransposons flanked by LTRs and lacking coding capacity (Figure 2A). These elements have
184 not been previously described in anopheline genomes and are still underexplored in insect
185 genomes in general (54-56). However, they might be important players in insect genome
186 evolution: in plants there are some examples of TRIM elements showing the capacity to
187 restructure genomes by acting as target sites for retrotransposon insertions, alter host gene
188 structure, and transduce host genes (57, 58). While we found TRIM elements in *An. coluzzii*
189 genomes to be underrepresented in gene bodies (χ^2 test *p-value* > 0.05), they were
190 overrepresented in nested insertions (χ^2 test *p-value* < 0.05).
191
192 We also assessed the phylogenetic distribution of the 64 new TE families in 15 species of the
193 *Anopheles* genus, including the eight members of the *An. gambiae* complex, two more distantly
194 related mosquitoes species (*Culex quinquefasciatus*, *Aedes aegypti*) and *Drosophila*
195 *melanogaster* (Additional file 1: Table S3) (36, 59, 60). We found that the new families were

196 unevenly distributed among the members of the *Anopheles* genus (Figure 2C and Additional file
197 3: Figure S2C). Ten families were exclusively found in members of the Pyretophorus series,
198 suggesting that these elements emerged after the split of this series from the Cellia subgenus.
199 Moreover, 13 families were also found in at least one of the other three non-anopheline species
200 (Additional file 3: Figure S2C). The distribution of these 13 families was patchy, with some of
201 them present only in distantly related species while others were present in members of the
202 *Anopheles* genus or in members of the Pyretophorus series. These suggests that some of these
203 families might have been acquired through horizontal transfer events (Additional file 1: Table
204 S3) (40).
205
206 **TE superfamily abundance varies across the seven genomes analyzed**
207 The percentage of the genome represented by TEs across the seven genomes varied between
208 16.94% and 20.21% (Table 2). Note that differences across genomes in assembly and scaffolding
209 statistics did not explained the differences in TE content or superfamily abundance (Additional
210 File 1: Table S1B).
211 We found a positive correlation between TE content and genome size as has been previously
212 described in *Anopheles* and other species (Pearson's r = 0.90, significance = 0.007; Additional
213 file 4: Figure S3) (40, 61). As expected due to heterochromatin being a TE rich region and thus
214 challenging to assemble (62), most of the differences in TE content across genomes were found
215 in the heterochromatin compartment (Table 2; χ^2 test for variance, *p-value* = 3.57e-3).
216
217 To assess whether differences in TE content at the family and superfamily level existed among
218 the seven genomes, we focused on the TE copy number in euchromatic regions as we are mostly

219 interested in the potential functional impact of TE insertions. We found significant differences at
220 the order and superfamily levels (χ^2 test *p-value* = 1.07e-21 and *p-value* = 1.69e-14,
221 respectively). The largest differences were found in the LTR order: LTRs were more abundant in
222 the *DLA112* and *LBV88* genomes and less abundant in *AcolN1* (Figure 3A). At the superfamily
223 level, we found that the largest differences were in the *Gypsy* superfamily, which belongs to the
224 LTR order. We also observed an enrichment of the *RTE* superfamily in *LBV11*, of the *CR1* and
225 *Bel-Pao* superfamilies in *DLA112*, and a depletion of the *CR1* superfamily in *AcolN1* (Figure 3).
226 Therefore, most of the differences in TE content between the evaluated genomes appear to be in
227 retrotransposon families.
228
229 **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

230 TE copy number, TE content in megabases and percentage of the genome represented by TEs.
231 Values are given for the whole genome and for the euchromatin and heterochromatin
232 compartments separately.
233
234 **TEs are nonrandomly distributed throughout the genome**
235 As expected, we found that the percentage of TEs in euchromatin, 11.73%-13.40%, is much
236 lower than the percentage of TEs in heterochromatin, 65.76%-74.77% (χ^2 test *p-value* < 0.05;
237 Table 2 and Figure 4A). None of the TE families identified were exclusive to either the
238 euchromatin or heterochromatin. However, 45 families were enriched in the euchromatin (χ^2 test,
239 *p-value* < 0.01) including 12 out of the 32 *mTA* MITE families (Additional file 1: Table S5). This
240 is in line with what has been previously reported in *Ae. aegypti* (63). We also observed that the
241 TE distribution was uneven between the chromosomes, and as expected, the X chromosome had
242 a larger fraction of its euchromatin spanned by TEs (Figure 4B) (64).
243
244 Finally, we also determine the distribution of TE insertions regarding genes. We divided the
245 genome in five regions: 1 kb upstream, exon, intron, 1 kb downstream and intergenic (65). More
246 than half of the genes (7239) in *An. coluzzii* had TEs either in their body or 1 kb upstream or
247 downstream. Many of these genes (3888/7239) had insertions in all seven genomes, while 1065
248 genes have an insertion only in one genome. We found that the number of insertions in
249 intergenic regions was higher than expected by chance while the number of insertions in exons
250 was lower (χ^2 test *p-value* < 0.001; Figure 4C; Additional file 1: Table S6). The upstream and
251 downstream regions behaved differently: the downstream region had a smaller amount of TEs
252 than expected by chance and the upstream region was neither enriched nor depleted for TE

253 insertions (χ^2 test *p-value* = 0; Additional file 1: Table S6). This is possibly linked with the
254 chromatin state of these regions given that downstream regions are more commonly in a closed
255 chromatin state (65).
256
257 Focusing on the TE orders, we observed that LTR elements were more abundant on intergenic
258 regions while SINEs were more abundant on introns, and DNA elements were more abundant in
259 introns and in the upstream region (χ^2 test *p-value* < 2.03e-3; Figure 4C; Additional file 1: Table
260 S7A). MITEs, which are non-autonomous DNA elements have been reported to be more
261 abundant in the introns and flanking regions of genes (66). We observed the same behavior for
262 *mTA* and *m3bp* MITEs, which are more abundant in upstream regions and introns, and *m8bp*
263 MITEs which are more abundant in introns (Additional file 1: Table S7B).
264
265 Overall, TEs are not randomly distributed in the genome, as they are more abundant in
266 heterochromatic than in euchromatic regions, more abundant in the X chromosome than in
267 autosomes, and more abundant in intergenic regions than in gene bodies or gene flanking
268 regions.
269
270 **MITE insertions are present in several inversion breakpoints**
271 TEs have been suggested to be involved in chromosome rearrangements within the *An. gambiae*
272 complex. Indeed, TEs have been found in close proximity to the breakpoints of the 2La in *An.*
273 *gambiae* and *An. melas*, and to the breakpoints of the 2Rb inversions in *An. gambiae* and *An.*
274 *coluzzii* (67, 68). We thus explored the TE content in the breakpoints of the 2La and 2Rb, and
275 three other common polymorphic inversions in *An. coluzzii*: 2Rc, 2Rd, and 2Ru (69). We were

276 able to identify the two breakpoints regions in at least one genome for all the insertions except
277 for the inversion *2Ru* for which we could only identify the distal one (Figure 5). The analysis of
278 these breakpoint regions suggested that analyzed genomes have the standard conformation for all
279 five inversions (see Material and Methods; Additional file 1: Table S8). We identified several
280 TEs nearby the proximal and the distal breakpoints of 2La and 2Rb, in agreement with previous
281 studies (Figure 5) (26, 67, 68). For the standard 2La proximal breakpoint, Sharakhov et al. (67)
282 identified several DNA transposons and a SINE insertion. We also identified a cluster of MITE
283 insertions, which are DNA transposons; however, we additionally identified an *Outcast* (LINE)
284 element (Figure 5). Regarding the standard 2La distal breakpoint, we observed two MITEs
285 similar to one of the insertions in the proximal breakpoint, which was in agreement with the
286 findings by Sharakhov et al. (67) (Figure 5). We also observed similar behavior in the 2Rb
287 breakpoints, such as the one described by Lobo et al., (68): tandem repeats flanking the inversion
288 in the standard and inverted forms, and TEs in the internal sequences of both breakpoints (Figure
289 5). For the 2Rd inversion, we identified MITEs near both breakpoints. Finally, we have also
290 described here for the first time, TE insertions that are present in the distal breakpoint of
291 inversion 2Ru but not near the estimated proximal breakpoint; although in the latter case we
292 were able to identify reads spanning the breakpoints in the seven genomes (70).
293

294 **Six of the seven potential active families are LTR insertions**

295 To identify potentially active TE families, we first estimated their relative age by analyzing the
296 TE landscapes (71, 72). We observed an "L" shape landscape in all genomes which is indicative
297 of a recent TE burst (Additional file 5: Figure S4) (73). This "L" shape landscape, dominated by
298 retrotransposons, had previously been described for the sister species *An. gambiae* (72, 74),

299 where numerous Gypsy LTR Retrotransposons (up to 75%) might currently be active (75, 76).
300 We further investigated the families in the peak of the landscape and we identified eight families
301 with more than two identical full-length fragments and with more than half of their copies
302 identical to the consensus (Additional file 1: Table S9A). Additionally, we assessed the potential
303 ability of our candidates to actively transpose by identifying their intact open read frames
304 (ORFs), LTRs (in the case of LTR retrotransposons), and target site duplications (TSDs), and
305 determined that seven of these families are potentially fully capable of transposing. For the LTR
306 families, we further evaluated the identity between LTRs of the same copy. Mean identities
307 ranged from 97.38% to 99.44% across the six LTR families, with 17.64% to 40,35% of the
308 copies having identical LTRs (Additional File 1: Table S9B). These results further suggest that
309 these families might be responsible for the recent retrotransposon burst in *An. coluzzii*
310 (Additional file 1: Table S9A). Note that LTR elements accounted for most of the differences in
311 TE content across genomes (Figure 3).
312
313 As a first step towards assessing the potential functional consequences of the TE insertions from
314 these seven putatively active families, we focused on insertions that occurred in introns, exons,
315 and 1 kb upstream or downstream of a gene. We identified 66 genes with insertions from these
316 families, with four genes containing up to two insertions in the same gene region (Additional File
317 6: Figure S5; Additional File 1: Table S10 and Table S11). Six of the genes have functions
318 related to vectorial capacity: insecticide resistance, immunity, and biting ability (Table 3). We
319 checked whether the TE insertions nearby these genes contained binding sites for transcription
320 factors or promoter motifs (Additional file 1: Table S12; Additional file 1: Table S13). We
321 focused on identifying binding sites for three transcription factors that are known to be involved

322 in response to xenobiotics (cap'n'collar: *cnc*) and in immune response and development (dorsal:
323 *dl* and signal transducer and activator of transcription: *STAT*) given the availability of matrix
324 profiles from *D. melanogaster* (77, 78). We identified binding sites for either *dl*, *STAT* or both in
325 three insertions; interestingly the *Acol_gypsy_Ele18* and the *Acol_copia_Ele8* insertions have
326 more than three binding sites for the same transcription factors, suggesting that they might be
327 functional sites (Table 3) (79). Additionally, the genes that contained these TE insertions also
328 contained binding sites for the same transcription factors, which suggests that these transcription
329 factors already played a prior role in their regulation. We also identified a putative promoter
330 sequence in the *Acol_copia_Ele24* insertion found upstream of the CLIPA1 protease encoded by
331 AGAP011794 which could also potentially lead to changes in the regulation of this gene
332 (Additional file 1: Table S14).
333
334 **Table 3. Seven TE insertions from putatively active families are located nearby genes**
335 **related with vectorial capacity.**

TE family	**Insert size (bp)**	**TE Freq.**	**Gene**	**Function**	**Possible phenotype [Reference]**	**TFBS**
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]	-
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 structure was correctly transferred (genes where some exons were missing were not taken into consideration in this analysis). References in the Phenotype column are as follows: 1 (80), 2 (81), 3 (82), 4 (83), 5 (84), 6 (85), 7 (86), 8 (87). 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 (88-92). 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

349 acetylcholinesterase *ace-1* gene (93-95). We first investigated whether the seven genomes
350 analyzed in this work contained these resistance alleles. We found the *kdr-west* mutation in the
351 six genomes from Douala and Libreville but not in *AcolN1* genome (49). None of the other
352 mutations were identified, however a previously undescribed nonsynonymous substitution
353 (L1688M) in the fourth domain of *para* was identified in the aforementioned six genomes.
354 Whether this replacement also increases insecticide resistance is yet to be assessed.
355
356 TEs have been hypothesized to play a relevant role specifically in response to insecticides (96-
357 98), and a few individual insertions affecting insecticide tolerance in anopheline mosquitoes
358 have already been described (99). Thus, we searched for TE insertions in the neighborhood of
359 insecticide-related genes that could potentially lead to differences in their regulation. We focused
360 on eight well-known insecticide resistance genes: *Ace1*, *Cyp6p3*, *Gstd1-6*, *para*, *rdl, Cyp6m2*,
361 *Cyp6z1* and *Gste2*. We also considered genes that have been shown to be differentially expressed
362 in *An. gambiae* when exposed to insecticides (Additional file 1: Table S14; Additional file 7:
363 Figure S6) (3, 100-102). We found that 21 out of the 43 genes analyzed contained at least one TE
364 insertion. We also observed that *para* had the largest number of TE insertions (48 in average per
365 genome, mainly in its introns) from this set of genes. This is an exception, given that the average
366 number of insertions per gene is 2.95 for members of this set which falls within the expected
367 number of insertions per gene in all the genome (t-test, *p-values* > 0.2).
368
369 Only one of the insertions, a solo LTR element of *Acol_Pao_Bel_Ele43* from the *Pao-Bel*
370 superfamily and present in all the genomes analyzed, was located in the 3' UTR of *GSTE2*.
371 Interestingly, an upstream insertion possibly affecting the expression level of this gene has

372 previously been identified in *An. funestus* (99). To determine if TEs could influence the
373 regulation of insecticide-resistance genes, we focused on polymorphic (present in two or more
374 genomes) and fixed (present in all seven genomes analyzed) insertions located in introns or 1 kb
375 upstream or downstream of the gene. We searched for *cnc* binding sites, and for those insertions
376 located in gene upstream regions we also looked for promoter motifs (Additional file 1: Table
377 S12; Additional file 1: Table S13). We identified 15 insertions in 10 genes containing either *cnc*
378 binding sites or promoter sequences. One insertion located in *CYP4C28* and two insertions in
379 *para* contained binding sites for *cnc*, although the genes did not contain binding sites for this
380 transcription factor. Additionally, we identified 12 insertions containing promoter motifs and
381 located nearby nine genes (Figure 6). In some cases, such as the *Acol_m2bp_Ele10* MITE
382 insertion in *ABCA4* or the *tSINE* insertion in *GSTMS2*, while the same TE insertion was found in
383 six and seven genomes respectively, the promoter motifs were found only in four and one
384 genome respectively (Figure 6; Additional file 1: Table S13). We analyzed the consensus
385 sequence of these two families and we found that while the *Acol_m2bp_Ele10* had the promoter
386 motif, the *tSINE* did not, suggesting that some of the *Acol_m2bp_Ele10* elements lost the
387 promoter motifs while the *tSINE* copies acquired them.
388

389 **Immune response genes could also be potentially affected by TEs**

390 Mosquitoes breeding in urban and polluted aquatic environments overexpress immune-related
391 genes suggesting that immune response is relevant for urban adaptation (103). To assess the
392 potential role of TEs in immune response, we searched for TE insertions in genes putatively
393 involved in immunity according to ImmunoDB (104) (Additional file 1: Table S15). We
394 identified 466 TE insertions in 148 out of the 281 genes analyzed. The number of insertions in

395 each gene varied greatly going from 58 genes with a single insertion to AGAP000940, a gene
396 coding for a C-type lectin and spanning 107.2 kb, with 48 insertions. The frequency of these
397 insertions was also variable with 186 (41.1%) of the insertions being fixed, 202 (44.6%)
398 polymorphic and 65 (14.3%) unique. We further explored polymorphic and fixed insertions and
399 identified binding sites for *dl* and *STAT* and promoter motifs. We found that 19 TEs contained
400 bindings sites for *dl*, 21 TEs contained binding sites for *STAT* and 12 TEs contained binding sites
401 both for *dl* and *STAT* (Additional file 1: Table S15). Additionally, we identified 81 insertions, in
402 the upstream region of 56 genes, which carried putative promoter sequences.
403
404 We identified TE insertions in three different antimicrobial peptides (AMPs). AMPs form the
405 first line of host defense against infection and are a key component of the innate immune system,
406 however none had transcription factor binding sites (TFBS) for *dl* or *STAT*. It is important to
407 keep in mind that there are other TF that participate in the regulation of AMPs and that both *dl*
408 and *STAT* are also involved in other biological processes (105). Interestingly we also identified
409 TEs with TFBS for *dl* in the vicinity of *STAT1* and *STAT2* which might lead to novel regulatory
410 mechanisms of the JAK/STAT signaling pathway. Furthermore, 11 of the 156 genes containing
411 TE insertions are differentially expressed in response to a *Plasmodium* invasion. These genes
412 participate in several pathways of the immune response including the small regulatory RNA
413 pathway, pathogen recognition, the nitric oxide response and ookinete melanization (78, 106-
414 108). Four of the TEs affecting these genes added TFBS and promoter sequences, thus
415 suggesting that these TE insertions can potentially influence the response to this pathogen (109)
416 (Table 4).
417

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

419

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)
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	-

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

425

426

427 **DISCUSSION**

428 In this study, we *de novo* annotated transposable element (TE) insertions in seven genomes of
429 *An. coluzzii,* six of them sequenced here. A comprehensive genome-wide TE annotation was
430 possible because we used long-read technologies to perform the genome sequencing and
431 assembly. Long-reads allow identifying TE insertions with high confidence given that the entire
432 TE insertion sequence can be spanned by a single read (30, 31). While the genome-wide TE
433 repertoire has been studied in other anopheline species, particularly in *An. gambiae*, to our
434 knowledge there are no other studies that have explored TE variation in multiple genomes from a
435 single species (32, 33, 36, 40, 72, 110). We observed that increasing the number of available
436 genomes analyzed allowed us to increase the number of identified TE families from a median of
437 244 (172-294) to 435 (Figure 1B). Moreover, having the full sequences of seven genomes also
438 allowed us to discover 64 new TE families, including four TRIM families previously
439 undescribed in anopheline genomes. This might be relevant as TRIM elements have been shown
440 to be important players in genome evolution in other species (57, 58). The wide range of families
441 identified across genomes was not directly related to the quality of the genome assembly taking
442 into consideration the more generally used quality parameters such as read length, number of
443 contigs, and contig N50 (111). This suggests that there are possibly other characteristics of each
444 genome that affect the identification of high-quality TE families, such as biases in the location of
445 the TE insertions given that TE families are challenging to identify in regions with low
446 complexity or with numerous nested TEs. Nonetheless, the identification of TE families is
447 dependent on the methodology used to perform TE annotations, therefore different annotation
448 strategies could lead to the discovery of still undescribed families (60).
449

450 The availability of several genome assemblies also allowed us to determine that the majority of
451 the intraspecies differences in the TE content were in heterochromatic regions, most likely due to
452 differences in the quality of the genome assembly. Nevertheless, there were also significant
453 differences in the TE content in euchromatic regions, reflecting true intraspecific variability as
454 has been previously observed in several organisms including Drosophila (112, 113), mammals
455 (114, 115), maize (116) and Arabidopsis (117). TE insertions were not randomly distributed
456 throughout the genome and instead were consistently enriched in intergenic regions, most likely
457 due to purifying selection, as suggested in the wild grass *Brachypodium distachyon* (118). In
458 Drosophila, TE enrichment in intergenic regions was also observed in addition to enrichment in
459 the intronic region, which we did not observe in *An. coluzzii* (119). We also analyzed the TE
460 content in the breakpoints of five common polymorphic inversions, three of them analyzed here
461 for the first time. We found TE insertions in all but one of the inversion breakpoints, with MITE
462 elements being the most common TE family, as already described in the 2Rd' inversion in *An.*
463 *arabiensis* (26) (Figure 5).
464
465 As a first step towards identifying the potential role of TEs in rapid adaptation to novel habitats
466 (120), we focused on insertions from recently active families located near genes that are relevant
467 for the vectorial capacity of *An. coluzzii* (Table 3). Because adaptation can also happen from
468 standing variation, in the case of insecticide resistance genes, which have been shown to be
469 shaped by TE insertions in several organisms, and immune-related genes, we analyzed all
470 insertions independently of their age (Figure 6 and Table 4) (99, 121, 122). While the role of
471 nonsynonymous substitutions and copy number variation in resistance to insecticides commonly
472 used in urban environments has been studied, the potential role of TEs has not yet been

473 comprehensively assessed in *An. coluzzii* or any other anopheline species (19, 102, 123-125). In
474 the genomes we assessed, we identified several insertions that were polymorphic or fixed nearby
475 functionally relevant genes (Table 3, Table 4 and Figure 6). Some of the identified candidate
476 insertions contained binding sites for transcription factors related to the function of the nearby
477 genes and promoter regions. Besides adding regulatory regions, TEs can also affect the
478 regulation of nearby genes by affecting gene splicing and generating long non-coding RNAs
479 among many other molecular mechanisms (25, 126-129). Thus, it is possible that the candidate
480 TE insertions identified, which lack binding sites and promoters, could be affecting nearby genes
481 through other molecular mechanisms. Our results are a first approximation to the potential role
482 of TEs in *An. coluzzii* adaptation to the challenging environment that urban ecosystems entail.
483 Establishing a direct link between the TEs and the traits involved in urban adaptation will require
484 sampling a larger number of individuals and characterizing the phenotypes associated with the
485 insertions.
486
487
488 **CONCLUSIONS**
489 The long-read sequencing of seven *An. coluzzii* genomes from urban environments allowed us to
490 capture to a larger extent the diversity of TE families and TE insertions and to assess their
491 potential impact in the genome architecture and genome function in this species. While there was
492 an enrichment of TE insertions in intergenic regions, we found several insertions located in the 1
493 kb flanking regions or inside genes relevant for the vectorial capacity of this species.
494 Furthermore, we found that some of these TE insertions are adding regulatory regions suggesting
495 that they could influence the regulation of these genes. Further studies are needed to confirm the

496 potential functional effect of these insertions. The genomic resources and the results that we
497 present in this work provide a basis for future studies of the impact of TEs in the biology of *An.*
498 *coluzzii.* This will allow increasing our knowledge on a species which besides being interesting
499 from an evolutionary perspective, given its high levels of genetic diversity and the strong
500 anthropogenic pressures it faces, is of great importance to human health. A better understanding
501 of the biology of *An. coluzzii* and its ability to rapidly adapt to urban environments will further
502 facilitate the development of novel strategies to combat malaria. Better management strategies
503 can be implemented if we understand and are able to predict changes in the frequency of genetic
504 variants relevant for the vectorial capacity of this species.
505
506
507 **MATERIALS AND METHODS**
508 **Sample collection and DNA isolation**
509 We sampled *An. coluzzii* larvae in two cities of Central Africa: Libreville, Gabon, in January
510 2016 and Douala, Cameroon, in April 2018 (Additional File 1: Table S1). A systematic
511 inspection of potential breeding sites was conducted to determine the presence of *Anopheles*
512 larvae. We manually separated the anopheline from the culicine larvae based on morphological
513 recognition and positioning of their bodies on or under the water surface (Robert, 2017). We
514 collected immature 3rd and 4th stage larvae of *Anopheles* from water bodies using the standard
515 dipping method (Service, 1993). Larvae were stored in 1.5 ml of absolute ethanol. After each
516 daily sampling session, the samples were stored at -20 °C.
517

518 All the samples were PCR tested to differentiate *An. coluzzii* larvae from *An. gambiae* larvae
519 before library preparation, using primers SINE200_F (TCGCCTTAGACCTTGCGTTA) and
520 SINE200_R (CGCTTCAAGAATTCGAGATAC) (46). These primers target a single copy
521 *SINE200* transposable element insertion that is fixed in *An. coluzzii* and absent in *An. gambiae.*
522 For PacBio sequencing, DNA from a single *An. coluzzii* larva from the *LBV11* site was extracted
523 using the MagAttract HMW DNA extraction kit (Qiagen) following manufacturer's instructions.
524 Briefly, the larva was air-dried and lysed in 240 µl of buffer ATL (proteinase K added) shaking
525 overnight at 56 ºC. Next, the DNA was isolated using the MagAttract magnetic beads and eluted
526 twice in 50 µl of buffer AE. The DNA concentration was measured using a Qubit fluorometer.
527 For Nanopore sequencing, DNA from six larvae from each of the five breeding sites was
528 extracted either with the QiaAMP UCP DNA kit (Qiagen) or MagAttract HMW DNA extraction
529 kit (Qiagen). We performed individual larvae extractions as our objective was to use the
530 minimum number of larvae possible to avoid the presence of excess polymorphisms that could
531 affect the genome assembly. For the QiaAMP UCP DNA kit, we followed the manufacturer's
532 instructions. Each larva was air-dried and lysed in 200 µl of buffer AUT (proteinase K added)
533 shaking overnight at 56 ºC, then DNA was isolated using a QIAamp UCP MinElute column and
534 eluted twice in 25 µl of buffer AUE. For the MagAttract HMW DNA extraction kit, we followed
535 manufacturer's instructions but using lower buffer amounts to increase DNA concentration.
536 Briefly, each larva was lysed in 120 µl of buffer ATL (proteinase K added) shaking overnight at
537 56 ºC, then DNA was isolated using the MagAttract magnetic beads and eluted twice in 25 µl of
538 buffer AE. The DNA concentration was measured using a Qubit fluorometer. Both elutions of
539 the same sample were mixed before library preparation. For Illumina sequencing, DNA from one

540 larva from each of the six different breeding sites was extracted following the same extraction
541 protocol as for Nanopore sequencing.
542
543 **Library preparation and sequencing**
544 Quality control of the DNA sample for PacBio sequencing (Qubit, NanoDrop and Fragment
545 analyzer) was performed at the Center for Genomic Research facility of the University of
546 Liverpool prior to library preparation. The library was prepared by shearing DNA to obtain
547 fragments of approximately 30 kb and sequenced on 2 SMRT cells using Sequel SMRT cell, 3.0
548 chemistry. Nanopore libraries were constructed using the Native Barcoding Expansion 1-12
549 (PCR-free) and the Ligation Sequencing Kit following manufacturer's instructions. A minimum
550 of 400 ng of DNA from each larva was used to start with the library workflow. For each
551 breeding site, six larvae were barcoded, and equal amounts of each barcoded sample were pooled
552 prior to sequencing. The samples from the same breeding site were ran in a single R9.4 flow cell
553 in a 48-hour run, except for sample *DLA112* which was run in two flow cells. The DNA
554 concentration was assessed during the whole procedure to ensure enough DNA was available for
555 sequencing.
556
557 The quality control of the samples, library preparation and Illumina sequencing was performed at
558 the Center for Genomic Research facility of the University of Liverpool. Low input libraries
559 were prepared with the NEBNext Ultra II FS DNA library kit (300 bp inserts) on the Mosquito
560 platform, using a 1/10 reduced volume protocol. Paired-end sequencing was performed on the
561 Illumina Novaseq platform using S2 chemistry (2x150 bp).
562

563 **Genome Assemblies**

564 The PacBio sequenced genome was assembled using *Canu* version 1.8 (130) with an estimated
565 genome size of 250Mb and parameters: '*stopOnLowCoverage*=5, *corMinCoverage*=0,
566 *correctedErrorRate*=0.105, *CorMhapFilterThreshold*=0.0000000002, *corMhapOptions="--*
567 *threshold 0.80 --num-hashes 512 --num-min-matches 3 --ordered-sketch-size 1000 --ordered-*
568 *kmer-size 14 --min-olap-length 2000 --repeat-idf-scale 50*" *mhapMemory*=60g,
569 *mhapBlockSize*=500, *ovlMerDistinct*=0.975'. The parameter stopOnLowCoverage was set to 5 to
570 prevent fragmentation given that some of our samples had medium coverage. corMinCoverage
571 was set to 0 to conserve the full length of the reads during the correction stage.
572 correctedErrorRate was set to 0.105 following the recommendations in Canu's manual for low
573 coverage genome assemblies. All of the remaining parameters were set to reduce disk space and
574 run time following the recommendations for repetitive genomes.Next, we identified and removed
575 allelic variants using *purge_haplotigs* version 1.0.4 (131) with the "*-l 15 -m 100 -h 195*"
576 parameters.
577 The Nanopore genomes were assembled using *Canu* version 1.8 using the same parameters as
578 previously described, except for *correctedErrorRate* which was set to 0.16, followed by a round
579 of polishing using *racon* version 1.3.3 (132), followed by *nanopolish* version 0.11.1 (133) and
580 *pilon* version 1.23-0 (134) with the fix parameter set on '*bases*'. *Pilon* requires high coverage
581 short-read data to perform the polishing and these data came from the aforementioned single
582 larvae sequenced from each of the sites. Finally, *blobtools* version 1.1.1 (135) was used to
583 remove contamination from all six genome assemblies taking into consideration fragment sizes,
584 their taxonomic assignation and the coverage using the Illumina reads.
585

586 As a proxy of the completeness, the BUSCO values for the six newly assembled genomes plus
587 the *AcolN1* genome were obtained using BUSCO version 3.0.2 (50) with the *diptera_odb9* set as
588 reference. Finally, the contigs for all seven assemblies were ordered and merged with *RaGOO*
589 v1.1 (136) using the chromosome level *An. gambiae* AgamP4 assembly.
590
591 **Gene annotation transfer**
592 The *gff* for the genome annotation for AgamP4 was transferred into the newly assembled
593 genomes using *Liftoff* (137) with default parameters. The annotation was manually inspected
594 using *UGENE* version 35 (138) and whenever needed the annotation was accordingly corrected.
595 96% of the AgamP4 genes were correctly transferred.
596
597 **Construction of the curated TE library and *de novo* TE annotation**
598 We ran the *TEdenovo* pipeline (51) independently on each of the seven genomes with default
599 parameters. The obtained consensus in each genome were further filtered by discarding those
600 generated with (i) only one sequence; (ii) with less than one full-length fragment mapping to the
601 genome; (iii) with less than three full-length copies; and (iv) shorter than 100bp (Additional file
602 1: Table S2). The remaining consensuses were manually curated to remove redundant sequences
603 and artifacts by manual inspection of coverage plots generated using the *plotCoverage* tool from
604 REPET and visualization of the structural features on the genome browser IGV version 2.4.19
605 (139).
606
607 To ensure that we identified as much of the TE diversity as possible, the *TEfam*
608 (tefam.biochem.vt.edu) database, which contains the TE libraries for several species of

609 mosquitoes, was used to annotate the seven genomes using *RepeatMasker* version open-4.0.9
610 (Smit et al. 2015). Families with more than three matches longer than 90% in any genome were
611 selected and their hit with the highest identity from each genome was extracted. These sequences
612 were added to the REPET library and all the consensuses were clustered using CD-HIT version
613 4.8.1 (140) with the *-c* and *-s* parameters set to 0.8. These filters ensured that all TEs with an
614 identity greater than 80% throughout more than 80% of their sequence were grouped in the same
615 family. 85 clusters contained sequences only identified by *TEfam*. The sequences belonging to
616 the same cluster were used to perform a multiple sequence alignment and the consensuses were
617 obtained.
618
619 The consensuses were classified using PASTEC (141) with default parameters. Next their
620 bidirectional best-hits were calculated using BLAST (142) against the *TEfam*
621 (tefam.biochem.vt.edu), AnoTExcel (143) and Repbase (144) databases. When more than 80%
622 of a consensus matched to a feature from the databases with an identity higher than 80%, the
623 classification was transferred to the consensus. While not an order *per se*, MITEs were grouped
624 together for subsequent analysis. Additionally, we classified the families based on the
625 conservation of features characteristic of their orders into putative autonomous, putative
626 autonomous lacking terminal inverted repeats (TIRs) or long terminal repeats (LTRs), putative
627 non-autonomous, such as MITEs and TRIMs, and degenerated (Additional file 1: Table S4)
628 (Fonseca et al 2019). These classified consensuses were used to re-annotate the assembled
629 genomes with the *TEannot* pipeline using default parameters and we discarded copies whose
630 length overlapped >80% with satellite annotations (52).
631

632 **Transfer of TE annotations to the *AcolN1* reference genome**

633 We transferred the euchromatic TE annotations from the six genomes we sequenced to the
634 *AcolN1* genome. First, we built a gff file composed by the coordinates of two 500 bp long
635 "anchors" adjacent to each TE. We transferred these features considering each pair of anchors as
636 exons from a single gene using the *Liftoff* tool with the -exclude_partial -overlap 1 -s 0.8
637 parameters (137). We conserved only transfers where both anchors were transferred to the
638 AcolN1 genome. When both anchors were separated by less than 10 bp we considered the TE to
639 be absent. This allowed us to identify most TE insertions, both present and absent in the AcolN1
640 genome. Next, following the same strategy we transferred these regions from the AcolN1
641 genome to the other six genomes. To summarize this information, we built a matrix containing
642 the status for all of the TE insertions transferred to the AcolN1 genome in every genome. When
643 the anchors were found more than 10 bp away the TE was considered to be present and this was
644 represented with a 1, when the distance was less than 10 bp the TE was considered to be absent
645 and it was represented with a -1. When any of the anchors was not transferred the TE was
646 considered to be not transferred and represented with a 0.

647 Overall, we were able to transfer to the AcolN1 genome 71.93% to 75.29% of the TEs present in
648 each of the six genomes leading to a total of 67,548 TEs transferred. We then attempt to
649 transferred back these 67,548 from the AcolN1 reference to the remaining six genomes and we
650 were able to confidently transfer 53,893 (79.78%) of these regions to at least three of the
651 genomes with 32,185 (47.65%) transferred to all six genomes. We checked whether the TEs that
652 failed to be transferred were enriched for nested TEs and we found that this was the case: while
653 68.58% and 51.18% of the TEs located more and less than 500 bp away from other TEs,

654 respectively, were transferred, only 38.26% of the TEs overlapping other TEs were transferred to
655 the six genomes.
656
657 **Identification of newly described families in other species**
658 We analyzed all 10 available fully sequenced species from the Pyretophorus series, which
659 belongs to the Cellia subgenus. We also included an additional five *Anopheles* species, three
660 from each of the other series from the Cellia subgenus and two from the other subgenera with
661 available fully sequenced species. As outgroups we included the genomes of *Cx.*
662 *quinquefasciatus*, *Ae. aegypti* and *D. melanogaster*. *RepeatMasker* version open-4.0.9 (Smit et
663 al. 2015) was run with default parameters using the 64 newly described families as the library on
664 the following genomes: *An. albimanus* (AalbS2), *An. atroparvus* (AatrE3), *An. farauti* (AfarF2),
665 *An. funestus* (AfunF3), *An. stephensi* (AsteS1), *An. epiroticus* (AepiE1), *An. christyi* (AchrA1),
666 *An. merus* (AmerM2), *An. gambiae* (AgamP4), *An. coluzzii* (AcolN1), *An. melas* (AmelC2), *An.*
667 *arabiensis* (AaraD1), *An. quadriannulatus* (AquaS1), *An. bwambae* (Abwa2) and *An. fontenillei*
668 (ASM881789v1), *Cx. quinquefasciatus* (CulPip1.0), *Ae. aegypti* (AaegL5.0) and *D.*
669 *melanogaster* (ISO1 release 6).
670
671 **Identification of heterochromatin**
672 The coordinates for the pericentric heterochromatin, compact intercalary heterochromatin, and
673 diffuse intercalary heterochromatin in *An. gambiae* AgamP3 were obtained from a previous work
674 (62). The *An. gambiae* AgamP3 genome assembly was mapped against the seven *An. coluzzii*
675 genome assemblies using *progressiveMauve* (145) and the corresponding coordinates on each of

676 the assemblies were retrieved (Additional file 1: Table S1C). To identify families enriched in
677 either euchromatin or heterochromatin a χ^2 test of independence was performed.
678
679 **Transfer of known inversion breakpoints**
680 The coordinates for inversions 2La, 2Rb, 2Rc and 2Rd were obtained from Corbett-Detig et al.
681 (2019) (69) and coordinates for 2Ru were obtained from White et al (2009) (70). 50 kb regions
682 flanking each side of the insertion were obtained and mapped using *minimap2* (146) against the
683 scaffolded genome assemblies to transfer the breakpoints. To validate the breakpoints, we
684 determined if long reads spanned the breakpoints using the genome browser IGV version 2.4.19
685 (139).
686
687 **Detection of putatively active TE families**
688 To identify potentially active TE families, we identified families with more than two identical
689 full-length fragment copies in at least six of the seven annotated genomes. We determined the
690 fraction of identical copies of these families by identifying all their insertions in the genome and
691 calculating the sequence identity of all their bases against the consensus by performing a
692 nucleotide BLAST. Given that long-read have a higher rate of sequencing errors that could affect
693 the age estimation (although we used Illumina reads for polishing), we also used dnaPipeTE
694 (147) to estimate the relative age of the TE families using the raw Illumina reads for the six
695 genomes that we sequenced. We compared the TE landscape obtained using dnaPipeTE with that
696 obtained using the BLAST procedure, using a Kolmogorov-Smirnov test corrected for multiple
697 testing using the Benjamini–Hochberg procedure (Additional file 1: Table S16). Given that we
698 observed few significant differences, we continued using the landscape data obtained using the

699 BLAST procedure. We identified the families where the majority of the bases of their insertions
700 were on the peak of identical sequences in the TE landscape (>50% of the bases with >99% base
701 identity) in more than five of the seven genomes we analyzed. Finally, we assessed the ability to
702 actively transpose of strong candidates by identifying their intact ORFs, LTRs (in the case of
703 LTR retrotransposons) and target site duplication (TSD), and estimated the % identity between
704 the two LTR of each TE copy.
705

706 **Classification of TEs by their genomic location**

707 To determine the location of TEs we used the *findOverlaps* function from the
708 *GenomicAlignments* R package (148) using default parameters. Both the TE and the gene
709 annotation were converted to *GenomicRanges* objects ignoring strand information in the case of
710 TEs.
711

712 **Insecticide resistance genes**

713 A list with a total of 43 relevant insecticide resistance genes was generated taking several works
714 into consideration (3, 100-102) (Additional file 1: Table S14). To determine the position of the L
715 to M nonsynonymous substitution that we observed in AGAP004707 (*para*) we used the position
716 from the CAM12801.1 reference sequence.
717

718 **Immune-related genes**

719 The full list of 414 immune-related genes from *An. gambiae* was downloaded from ImmunoDB
720 (104). We conserved the 281 most reliable genes filtering by the STATUS field and conserving
721 only those with A or B scores (A refers to genes confirmed with high confidence and expert-

722 refined cDNA supplied, and B refers to genes confirmed with high confidence, no refinement
723 required).
724
725 **TFBS and promoter identification**
726 The matrices for *dl* (MA0022.1), *cnc::maf-S* (MA0530.1) and *Stat92E* (MA0532.1) were
727 downloaded from JASPAR (http://jaspar.genereg.net/) (149). The sequences for the TEs of
728 interest were obtained using *getSeq* from the *Biostrings* R package. The TFBS in the sequences
729 were identified using the web version of FIMO (150) from the MEME SUITE (151) with default
730 parameters. The ElemeNT online tool was used to identify promoter motifs (152).
731
732
733 **DECLARATIONS**
734
735 **Ethics approval and consent to participate**
736 Not applicable.
737
738 **Consent for publication**
739 Not applicable.
740
741 **Availability of data and materials**
742 All the genome sequencing data obtained in this work, as well as the genome assemblies are
743 available in NCBI SRA and NCBI Genbank respectively, under the BioProject accession number
744 PRJNA676011. TE annotations in each of the seven genomes are provided as a gff files available

745 at https://digital.csic.es/handle/10261/224416. Transferred annotations across the seven genomes
746 are also available at https://digital.csic.es/handle/10261/224416 as a txt file.
747
748 **Competing interests**
749 The authors declare that they have no competing interests.
750
751 **Funding**
752 This study was supported by the Ministry of Economy, Industry and Competitiveness of Spain
753 (BFU2017-82937-P) to JG. DA was supported by an ANR grant (ANR-18-CE35-0002-01 –
754 WILDING). NMLP was funded by AUF and CIRMF scholarships.
755
756 **Authors' contributions**
757 DA and JG conceived and designed the experiments. NMLP, SEN and LA generated data. CVC,
758 DA and JG performed the data analysis. CVC and JG wrote and revised the manuscript with
759 input from all authors. All authors read and approved the final manuscript.
760
761 **Acknowledgements**
762 We thank members of the González Lab for comments on the manuscript. We thank the Ecology
763 of Vectorial Systems team at the CIMRF (Franceville, Gabon) for their support in field
764 collections. We thank Jean Pierre Agbor and Serge Donfanck for their commitment in larvae
765 collections in Douala (Cameroon).
766
767 **Authors' information**

768 Twitter handles: @VargasChavezC (Carlos Vargas-Chavez); @d_ayalag (Diego Ayala); @
769 GonzalezLab_BCN (Josefa González).
770
771 **REFERENCES**

772 1. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al.
773 Extensive introgression in a malaria vector species complex revealed by phylogenomics.
774 Science. 2015;347(6217).
775 2. Tene Fossog B, Ayala D, Acevedo P, Kengne P, Ngomo Abeso Mebuy I, Makanga B, et
776 al. Habitat segregation and ecological character displacement in cryptic African malaria
777 mosquitoes. Evolutionary Applications. 2015;8(4):326-45.
778 3. Fossog Tene B, Poupardin R, Costantini C, Awono-Ambene P, Wondji CS, Ranson H, et
779 al. Resistance to DDT in an Urban Setting: Common Mechanisms Implicated in Both M and S
780 Forms of Anopheles gambiae in the City of Yaoundé Cameroon. PLOS ONE. 2013;8(4):e61408.
781 4. Kengne P, Charmantier G, Blondeau-Bidet E, Costantini C, Ayala D. Tolerance of
782 disease-vector mosquitoes to brackish water and their osmoregulatory ability. Ecosphere.
783 2019;10(10):e02783.
784 5. Vontas J, Grigoraki L, Morgan J, Tsakireli D, Fuseini G, Segura L, et al. Rapid selection
785 of a pyrethroid metabolic enzyme CYP9K1 by operational malaria control activities. Proceedings
786 of the National Academy of Sciences. 2018;115(18):4619.
787 6. Fouet C, Kamdem C, Gamez S, White BJ. Extensive genetic diversity among populations
788 of the malaria mosquito Anopheles moucheti revealed by population genomics. Infection,
789 Genetics and Evolution. 2017;48:27-33.

790 7. Wiebe A, Longbottom J, Gleave K, Shearer FM, Sinka ME, Massey NC, et al.
791 Geographical distributions of African malaria vector sibling species and evidence for insecticide
792 resistance. Malaria journal. 2017;16(1):85-.
793 8. Perugini E, Guelbeogo WM, Calzetta M, Manzi S, Virgillito C, Caputo B, et al.
794 Behavioural plasticity of Anopheles coluzzii and Anopheles arabiensis undermines LLIN
795 community protective effect in a Sudanese-savannah village in Burkina Faso. Parasites &
796 vectors. 2020;13(1):277-.
797 9. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. A Polytene Chromosome
798 Analysis of the Anopheles gambiae Species Complex. Science.
799 2002;298(5597):1415.
800 10. Ayala D, Acevedo P, Pombi M, Dia I, Boccolini D, Costantini C, et al. Chromosome
801 inversions and ecological plasticity in the main African malaria mosquitoes. Evolution.
802 2017;71(3):686-701.
803 11. Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, Bassole IHN, et al. Living
804 at the edge: biogeographic patterns of habitat segregation conform to speciation by niche
805 expansion in Anopheles gambiae. BMC Ecology. 2009;9(1):16.
806 12. Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, Ose K, et al. Ecological niche
807 partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of
808 speciation. BMC Ecology. 2009;9(1):17.
809 13. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA. Chromosomal differentiation and
810 adaptation to human environments in the Anopheles gambiae complex. Transactions of The
811 Royal Society of Tropical Medicine and Hygiene. 1979;73(5):483-97.

812 14. Fouet C, Gray E, Besansky NJ, Costantini C. Adaptation to Aridity in the Malaria
813 Mosquito Anopheles gambiae: Chromosomal Inversion Polymorphism and Body Size Influence
814 Resistance to Desiccation. PLOS ONE. 2012;7(4):e34841.
815 15. Ayala D, Zhang S, Chateau M, Fouet C, Morlais I, Costantini C, et al. Association
816 mapping desiccation resistance within chromosomal inversions in the African malaria vector
817 Anopheles gambiae. Molecular Ecology. 2019;28(6):1333-42.
818 16. Labbé P, Berthomieu A, Berticat C, Alout H, Raymond M, Lenormand T, et al.
819 Independent Duplications of the Acetylcholinesterase Gene Conferring Insecticide Resistance in
820 the Mosquito Culex pipiens. Molecular Biology and Evolution. 2007;24(4):1056-67.
821 17. Assogba BS, Djogbénou LS, Milesi P, Berthomieu A, Perez J, Ayala D, et al. An ace-1
822 gene duplication resorbs the fitness cost associated with resistance in Anopheles gambiae, the
823 main malaria mosquito. Scientific Reports. 2015;5(1):14529.
824 18. Weetman D, Djogbenou LS, Lucas E. Copy number variation (CNV) and insecticide
825 resistance in mosquitoes: evolving knowledge or an evolving problem? Current opinion in insect
826 science. 2018;27:82-8.
827 19. Lucas ER, Miles A, Harding NJ, Clarkson CS, Lawniczak MKN, Kwiatkowski DP, et al.
828 Whole-genome sequencing reveals high complexity of copy number variation at insecticide
829 resistance loci in malaria mosquitoes. Genome Research. 2019;29(8):1250-61.
830 20. Mitri C, Markianos K, Guelbeogo WM, Bischoff E, Gneme A, Eiglmeier K, et al. The
831 kdr-bearing haplotype and susceptibility to Plasmodium falciparum in Anopheles gambiae:
832 genetic correlation and functional testing. Malaria Journal. 2015;14(1):391.
833 21. Kamdem C, Fouet C, Gamez S, White BJ. Pollutants and Insecticides Drive Local
834 Adaptation in African Malaria Mosquitoes. Mol Biol Evol. 2017;34(5):1261-75.

835 22. King SA, Onayifeke B, Akorli J, Sibomana I, Chabi J, Manful-Gwira T, et al. The Role
836 of Detoxification Enzymes in the Adaptation of the Major Malaria Vector Anopheles gambiae
837 (Giles; Diptera: Culicidae) to Polluted Water. Journal of Medical Entomology. 2017;54(6):1674-
838 83.
839 23. Casacuberta E, González J. The impact of transposable elements in environmental
840 adaptation. Molecular Ecology. 2013;22(6):1503-17.
841 24. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution.
842 Molecular Ecology. 2019;28(6):1537-49.
843 25. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From
844 conflicts to benefits. Nature Reviews Genetics. 2017;18(2):71-86.
845 26. Mathiopoulos KD, Della Torre A, Predazzi V, Petrarca V, Coluzzi M. Cloning of
846 inversion breakpoints in the Anopheles gambiae complex traces a transposable element at the
847 inversion junction. Proceedings of the National Academy of Sciences of the United States of
848 America. 1998;95(21):12444-9.
849 27. Gray YH. It takes two transposons to tango: transposable-element-mediated chromosomal
850 rearrangements. Trends Genet. 2000;16(10):461-8.
851 28. Reis M, Vieira CP, Lata R, Posnien N, Vieira J. Origin and Consequences of
852 Chromosomal Inversions in the virilis Group of Drosophila. Genome Biology and Evolution.
853 2018;10(12):3152-66.
854 29. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements.
855 Nature Reviews Genetics. 2018;19(11):688-704.
856 30. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its
857 applications. Nature Reviews Genetics. 2020;21(10):597-614.

858 31. Shahid S, Slotkin RK. The current revolution in transposable element biology enabled by
859 long reads. Current Opinion in Plant Biology. 2020;54:49-56.
860 32. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The
861 Genome Sequence of the Malaria Mosquito Anopheles gambiae. Science.
862 2002;298(5591):129-49.
863 33. Marinotti O, Cerqueira GC, De Almeida LGP, Ferro MIT, Da Silva Loreto EL, Zaha A,
864 et al. The Genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids
865 Research. 2013;41(15):7387-400.
866 34. Jiang X, Peery A, Hall AB, Sharma A, Chen XG, Waterhouse RM, et al. Genome
867 analysis of a major urban malaria vector mosquito, Anopheles stephensi. Genome biology.
868 2014;15(9):459-.
869 35. Zhou D, Zhang D, Ding G, Shi L, Hou Q, Ye Y, et al. Genome sequence of Anopheles
870 sinensis provides insight into genetics basis of mosquito competence for malaria parasites. BMC
871 Genomics. 2014;15(1).
872 36. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al.
873 Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science.
874 2015;347(6217).
875 37. Lau YL, Lee WC, Chen J, Zhong Z, Jian J, Amir A, et al. Draft genomes of Anopheles
876 cracens and Anopheles maculatus: Comparison of simian malaria and human malaria vectors in
877 peninsular Malaysia. PLoS ONE. 2016;11(6):1-24.
878 38. Chakraborty M, Ramaiah A, Adolfi A, Halas P, Kaduskar B, Ngo LT, et al. Hidden
879 features of the malaria vector mosquito, Anopheles stephensi, revealed by a high-
880 quality reference genome. bioRxiv. 2020:2020.05.24.113019.

881 39. Compton A, Liang J, Chen C, Lukyanchikova V, Qi Y, Potters M, et al. The beginning of
882 the end: a chromosomal assembly of the New World malaria mosquito ends with a novel
883 telomere. bioRxiv. 2020:2020.04.17.047084.
884 40. de Melo ES, Wallau GdL. Transposable elements are constantly exchanged by horizontal
885 transfer reshaping mosquito genomes. bioRxiv. 2020:2020.06.23.166744.
886 41. Yang X, Lee WP, Ye K, Lee C. One reference genome is not enough. Genome Biology.
887 2019;20(1):19-21.
888 42. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new
889 reference. Nature Plants. 2020;6(8):914-20.
890 43. Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, et al.
891 Discovery and population genomics of structural variation in a songbird genus. Nature
892 Communications. 2020;11(1):3403.
893 44. Quesneville H, Nouaud D, Anxolabéhère D. P elements and MITE relatives in the whole
894 genome sequence of Anopheles gambiae. BMC Genomics. 2006;7.
895 45. Boulesteix M, Simard F, Antonio-Nkondjio C, Awono-Ambene HP, Fontenille D,
896 Biémont C. Insertion polymorphism of transposable elements and population structure of
897 Anopheles gambiae M and S molecular forms in Cameroon. Molecular Ecology.
898 2007;16(2):441-52.
899 46. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, Della Torre A. Insertion
900 polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae
901 molecular forms. Malaria Journal. 2008;7:1-10.

902 47. Esnault C, Boulesteix M, Duchemin JB, Koffi AA, Chandre F, Dabiré R, et al. High
903 genetic differentiation between the M and S molecular forms of Anopheles gambiae in Africa.
904 PloS one. 2008;3(4):e1968-e.
905 48. Salgueiro P, Moreno M, Simard F, O'Brochta D, Pinto J. New Insights into the
906 Population Structure of Anopheles gambiae s.s. in the Gulf of Guinea Islands Revealed by
907 Herves Transposable Elements. PLoS ONE. 2013;8(4).
908 49. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, et al. A high-
909 quality de novo genome assembly from a single mosquito using pacbio sequencing. Genes.
910 2019;10(1).
911 50. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
912 assessing genome assembly and annotation completeness with single-copy orthologs.
913 Bioinformatics. 2015;31(19):3210-2.
914 51. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element
915 Diversification in De Novo Annotation Approaches. PLOS ONE. 2011;6(1):e16526.
916 52. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al.
917 Combined evidence annotation of transposable elements in genome sequences. PLoS
918 computational biology. 2005;1(2):166-75.
919 53. Platt RN, 2nd, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation
920 Is Vital When Analyzing New Genome Assemblies. Genome biology and evolution.
921 2016;8(2):403-10.
922 54. Marsano RM, Leronni D, D'Addabbo P, Viggiano L, Tarasco E, Caizzi R. Mosquitoes
923 LTR retrotransposons: a deeper view into the genomic sequence of Culex quinquefasciatus. PloS
924 one. 2012;7(2):e30770-e.

925 55. Zhou Y, Cahan SH. A Novel Family of Terminal-Repeat Retrotransposon in Miniature
926 (TRIM) in the Genome of the Red Harvester Ant, Pogonomyrmex barbatus. PLoS ONE.
927 2012;7(12).
928 56. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, et al. Finding the
929 missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics.
930 2014;15(1):86.
931 57. Witte C-P, Le QH, Bureau T, Kumar A. Terminal-repeat retrotransposons in miniature
932 (TRIM) are involved in restructuring plant genomes. Proceedings of the National Academy of
933 Sciences. 2001;98(24):13778.
934 58. Gao D, Li Y, Kim KD, Abernathy B, Jackson SA. Landscape and evolutionary dynamics
935 of terminal repeat retrotransposons in miniature in plant genomes. Genome Biology.
936 2016;17(1):7.
937 59. Barrón MG, Paupy C, Rahola N, Akone-Ella O, Ngangue MF, Wilson-Bahun TA, et al.
938 A new species in the major malaria vector complex sheds light on reticulated species evolution.
939 Scientific Reports. 2019;9(1):1-13.
940 60. Vargas-Chavez C, González J. Transposable elements in Anopheles species: refining
941 annotation strategies towards population-level analysis. In: Dupuis ORaJ, editor. Population
942 Genomics: Insects: Springer; 2021.
943 61. Sessegolo C, Burlet N, Haudry A. Strong phylogenetic inertia on genome size and
944 transposable element content among 26 species of flies. Biology Letters. 2016;12(8):0-3.
945 62. Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, Smith CD, et al.
946 Genome mapping and characterization of the Anopheles gambiae heterochromatin. BMC
947 Genomics. 2010;11(1):459.

948 63. Tu Z. Three novel families of miniature inverted-repeat transposable elements are
949 associated with genes of the yellow fever mosquito, Aedes aegypti. Proceedings of the National
950 Academy of Sciences of the United States of America. 1997;94(14):7475-80.
951 64. Xia A, Sharakhova MV, Leman SC, Tu Z, Bailey JA, Smith CD, et al. Genome landscape
952 and evolutionary plasticity of chromosomes in malaria mosquitoes. PLoS ONE. 2010;5(5).
953 65. Ruiz JL, Ranford-Cartwright LC, Gómez-Díaz E. The regulatory genome of the malaria
954 vector Anopheles gambiae: integrating chromatin accessibility and gene expression.
955 bioRxiv. 2020:2020.06.22.164228.
956 66. Tu Z. Eight novel families of miniature inverted repeat transposable elements in the
957 African malaria mosquito, Anopheles gambiae. Proceedings of the National Academy of
958 Sciences of the United States of America. 2001;98(4):1699-704.
959 67. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, Santolamazza F, et al.
960 Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in
961 the Anopheles gambiae complex. Proceedings of the National Academy of Sciences of the
962 United States of America. 2006;103(16):6258-62.
963 68. Lobo NF, Sangaré DM, Regier AA, Reidenbach KR, Bretz DA, Sharakhova MV, et al.
964 Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. Malaria journal.
965 2010;9:293-.
966 69. Corbett-Detig RB, Said I, Calzetta M, Genetti M, McBroome J, Maurer NW, et al. Fine-
967 Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the
968 Anopheles gambiae Species Complex Using Proximity-Ligation Sequencing.
969 Genetics. 2019;213(4):1495.

970 70. White BJ, Cheng C, Sangaré D, Lobo NF, Collins FH, Besansky NJ. The population
971 genomics of trans-specific inversion polymorphisms in Anopheles gambiae. Genetics.
972 2009;183(1):275-88.
973 71. Smit A, Hubley R, Green P. *RepeatMasker Open-4.0* 2013-2015 [Available from:
974 http://www.repeatmasker.org
975 72. Diesel JF, Ortiz MF, Marinotti O, Vasconcelos ATR, Loreto ELS. A re-annotation of the
976 Anopheles darlingi mobilome. Genetics and Molecular Biology. 2019;42(1):125-31.
977 73. Fonseca PM, Moura RD, Wallau GL, Loreto ELS. The mobilome of Drosophila
978 incompta, a flower-breeding species: comparison of transposable element landscapes among
979 generalist and specialist flies. Chromosome Research. 2019;27(3):203-19.
980 74. Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, et al. Diversity and
981 evolution of the transposable element repertoire in arthropods with particular reference to
982 insects. BMC Evolutionary Biology. 2019;19(1):11-.
983 75. Tubío JMC, Naveira H, Costas J. Structural and evolutionary analyses of the Ty3/gypsy
984 group of LTR retrotransposons in the genome of Anopheles gambiae. Molecular Biology and
985 Evolution. 2005;22(1):29-39.
986 76. Tubío JMC, Tojo M, Bassaganyas L, Escaramis G, Sharakhov IV, Sharakhova MV, et al.
987 Evolutionary Dynamics of the Ty3/Gypsy LTR Retrotransposons in the Genome of Anopheles
988 gambiae. PLOS ONE. 2011;6(1):e16328.
989 77. Ingham VA, Pignatelli P, Moore JD, Wagstaff S, Ranson H. The transcription factor Maf-
990 S regulates metabolic resistance to insecticides in the malaria vector Anopheles gambiae. BMC
991 genomics. 2017;18(1):669-.

992 78. Osta MA, Christophides GK, Vlachou D, Kafatos FC. Innate immunity in the malaria
993 vector Anopheles gambiae: comparative and functional genomics. Journal of
994 Experimental Biology. 2004;207(15):2551.
995 79. Xie D, Chen C-C, Ptaszek LM, Xiao S, Cao X, Fang F, et al. Rewirable gene regulatory
996 networks in the preimplantation embryonic development of three mammalian species. Genome
997 research. 2010;20(6):804-15.
998 80. Bonizzoni M, Afrane Y, Dunn WA, Atieli FK, Zhou G, Zhong D, et al. Comparative
999 Transcriptome Analyses of Deltamethrin-Resistant and -Susceptible Anopheles gambiae
1000 Mosquitoes from Kenya by RNA-Seq. PLOS ONE. 2012;7(9):e44607.
1001 81. Vannini L, Willis JH. Localization of RR-1 and RR-2 cuticular proteins within the cuticle
1002 of Anopheles gambiae. Arthropod structure & development. 2017;46(1):13-29.
1003 82. Balabanidou V, Kefi M, Aivaliotis M, Koidou V, Girotti JR, Mijailovsky SJ, et al.
1004 Mosquitoes cloak their legs to resist insecticides. Proceedings Biological sciences.
1005 2019;286(1907):20191091-.
1006 83. Sriwichai P, Rongsiryam Y, Jariyapan N, Sattabongkot J, Apiwathnasorn C, Nacapunchai
1007 D, et al. Cloning of a Trypsin-like Serine Protease and expression Patterns during *Plasmodium*
1008 *falciparum* invasion in the mosquito, *Anopheles dirus* (Peyton and Harrison). Archives of Insect
1009 Biochemistry and Physiology. 2012;80(3):151-65.
1010 84. Dias-Lopes G, Borges-Veloso A, Saboia-Vahia L, Domont GB, Britto C, Cuervo P, et al.
1011 Expression of active trypsin-like serine peptidases in the midgut of sugar-feeding female
1012 Anopheles aquasalis. Parasites & vectors. 2015;8:296-.

1013 85. Hughes GL, Ren X, Ramirez JL, Sakamoto JM, Bailey JA, Jedlicka AE, et al. Wolbachia
1014 Infections in Anopheles gambiae Cells: Transcriptomic Characterization of a Novel Host-
1015 Symbiont Interaction. PLOS Pathogens. 2011;7(2):e1001296.
1016 86. Cao X, Gulati M, Jiang H. Serine protease-related proteins in the malaria mosquito,
1017 Anopheles gambiae. Insect biochemistry and molecular biology. 2017;88:48-62.
1018 87. Kent LB, Walden KKO, Robertson HM. The Gr Family of Candidate Gustatory and
1019 Olfactory Receptors in the Yellow-Fever Mosquito Aedes aegypti. Chemical Senses.
1020 2008;33(1):79-93.
1021 88. Dabiré RK, Namountougou M, Diabaté A, Soma DD, Bado J, Toé HK, et al. Distribution
1022 and frequency of kdr mutations within Anopheles gambiae s.l. populations and first report of the
1023 ace.1 G119S mutation in Anopheles arabiensis from Burkina Faso (West Africa). PloS one.
1024 2014;9(7):e101484-e.
1025 89. Silva APB, Santos JMM, Martins AJ. Mutations in the voltage-gated sodium channel
1026 gene of anophelines and their association with resistance to pyrethroids - a review. Parasites &
1027 vectors. 2014;7:450-.
1028 90. Cheung J, Mahmood A, Kalathur R, Liu L, Carlier PR. Structure of the G119S Mutant
1029 Acetylcholinesterase of the Malaria Vector Anopheles gambiae Reveals Basis of Insecticide
1030 Resistance. Structure (London, England : 1993). 2018;26(1):130-6.e2.
1031 91. Elanga-Ndille E, Nouage L, Ndo C, Binyang A, Assatse T, Nguiffo-Nguete D, et al. The
1032 G119S Acetylcholinesterase (Ace-1) Target Site Mutation Confers Carbamate Resistance in the
1033 Major Malaria Vector Anopheles gambiae from Cameroon: A Challenge for the Coming IRS
1034 Implementation. Genes. 2019;10(10):790.

1035 92. Fadel AN, Ibrahim SS, Tchouakui M, Terence E, Wondji MJ, Tchoupo M, et al. A
1036 combination of metabolic resistance and high frequency of the 1014F kdr mutation is driving
1037 pyrethroid resistance in Anopheles coluzzii population from Guinea savanna of Cameroon.
1038 Parasites & Vectors. 2019;12(1):263.
1039 93. Santolamazza F, Calzetta M, Etang J, Barrese E, Dia I, Caccone A, et al. Distribution of
1040 knock-down resistance mutations in Anopheles gambiae molecular forms in west and west-
1041 central Africa. Malaria Journal. 2008;7(1):74.
1042 94. Jones CM, Liyanapathirana M, Agossa FR, Weetman D, Ranson H, Donnelly MJ, et al.
1043 Footprints of positive selection associated with a mutation (N1575Y) in the voltage-gated
1044 sodium channel of Anopheles gambiae. Proceedings of the National Academy of Sciences of the
1045 United States of America. 2012;109(17):6614-9.
1046 95. Essandoh J, Yawson AE, Weetman D. Acetylcholinesterase (Ace-1) target site mutation
1047 119S is strongly diagnostic of carbamate and organophosphate resistance in Anopheles gambiae
1048 s.s. and Anopheles coluzzii across southern Ghana. Malaria Journal. 2013;12(1):404.
1049 96. Wilson TG. Transposable Elements as Initiators of Insecticide Resistance. Journal of
1050 Economic Entomology. 1993;86(3):645-51.
1051 97. ffrench-Constant R, Daborn P, Feyereisen R. Resistance and the jumping gene.
1052 BioEssays. 2006;28(1):6-8.
1053 98. Rostant WG, Wedell N, Hosken DJ. Chapter 2 - Transposable Elements and Insecticide
1054 Resistance. In: Goodwin SF, Friedmann T, Dunlap JC, editors. Advances in Genetics. 78:
1055 Academic Press; 2012. p. 169-201.
1056 99. Weedall GD, Riveron JM, Hearn J, Irving H, Kamdem C, Fouet C, et al. An Africa-wide
1057 genomic evolution of insecticide resistance in the malaria vector Anopheles funestus involves

1058 selective sweeps, copy number variations, gene conversion and transposons. PLOS Genetics.
1059 2020;16(6):e1008822.
1060 100. Main BJ, Everitt A, Cornel AJ, Hormozdiari F, Lanzaro GC. Genetic variation associated
1061 with increased insecticide resistance in the malaria mosquito, Anopheles coluzzii. Parasites &
1062 vectors. 2018;11(1):225-.
1063 101. Adolfi A, Poulton B, Anthousi A, Macilwee S, Ranson H, Lycett GJ. Functional genetic
1064 validation of key genes conferring insecticide resistance in the major African malaria vector,
1065 Anopheles gambiae. Proceedings of the National Academy of Sciences.
1066 2019;116(51):25764.
1067 102. Bamou R, Sonhafouo-Chiana N, Mavridis K, Tchuinkam T, Wondji CS, Vontas J, et al.
1068 Status of Insecticide Resistance and Its Mechanisms in Anopheles gambiae and Anopheles
1069 coluzzii Populations from Forest Settings in South Cameroon. Genes. 2019;10(10):741.
1070 103. Cassone BJ, Kamdem C, Cheng C, Tan JC, Hahn MW, Costantini C, et al. Gene
1071 expression divergence between malaria vector sibling species Anopheles gambiae and
1072 An. coluzzii from rural and urban Yaoundé Cameroon. Molecular Ecology. 2014;23(9):2242-59.
1073 104. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, et al.
1074 Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes.
1075 Science (New York, NY). 2007;316(5832):1738-43.
1076 105. Clayton AM, Dong Y, Dimopoulos G. The Anopheles Innate Immune
1077 System in the Defense against Malaria Infection. Journal of Innate Immunity. 2014;6(2):169-81.
1078 106. Volz J, Müller H-M, Zdanowicz A, Kafatos FC, Osta MA. A genetic module regulates
1079 the melanization response of Anopheles to Plasmodium. Cellular Microbiology. 2006;8(9):1392-
1080 405.

1081 107. Oliveira GdA, Lieberman J, Barillas-Mury C. Epithelial nitration by a peroxidase/NOX5
1082 system mediates mosquito antiplasmodial immunity. Science (New York, NY).
1083 2012;335(6070):856-9.
1084 108. Dennison NJ, BenMarzouk-Hidalgo OJ, Dimopoulos G. MicroRNA-regulation of
1085 Anopheles gambiae immunity to Plasmodium falciparum infection and midgut microbiota.
1086 Developmental & Comparative Immunology. 2015;49(1):170-8.
1087 109. Ruiz JL, Yerbanga RS, Lefèvre T, Ouedraogo JB, Corces VG, Gómez-Díaz E. Chromatin
1088 changes in Anopheles gambiae induced by Plasmodium falciparum infection. Epigenetics &
1089 Chromatin. 2019;12(1):5.
1090 110. Fernández-Medina RD, Ribeiro JMC, Carareto CMA, Velasque L, Struchiner CJ. Losing
1091 identity: structural diversity of transposable elements belonging to different classes in the
1092 genome of Anopheles gambiae. BMC genomics. 2012;13.
1093 111. Ou S, Liu J, Chougule KM, Fungtammasan A, Seetharam AS, Stein JC, et al. Effect of
1094 sequence depth and length in long-read assembly of the maize inbred NC358. Nature
1095 Communications. 2020;11(1):2288.
1096 112. Kofler R, Nolte V, Schlötterer C. Tempo and Mode of Transposable Element Activity in
1097 Drosophila. PLOS Genetics. 2015;11(7):e1005406.
1098 113. Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL,
1099 Horváth V, et al. Stress response, behavior, and development are shaped by transposable
1100 element-induced mutations in Drosophila. PLOS Genetics. 2019;15(2):e1007900.
1101 114. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms
1102 recapitulate human evolution. Mobile DNA. 2015;6:21-.

1103 115. Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-
1104 specific chromatin looping and gene regulation in mammalian genomes. Nature
1105 Communications. 2020;11(1):1796.
1106 116. Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, et al. European maize
1107 genomes highlight intraspecies variation in repeat and gene content. Nature Genetics.
1108 2020;52(9):950-7.
1109 117. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh
1110 JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. eLife.
1111 2016;5:e15716.
1112 118. Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of
1113 long terminal repeat retrotransposons in the model grass Brachypodium distachyon. New
1114 Phytologist. 2020;227(6):1736-48.
1115 119. Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of
1116 transposable elements in two Drosophila QTL mapping resources. Molecular biology and
1117 evolution. 2013;30(10):2311-27.
1118 120. Johnson MTJ, Munshi-South J. Evolution of life in urban environments. Science.
1119 2017;358(6363):eaam8327.
1120 121. Mateo L, Ullastres A, González J. A Transposable Element Insertion Confers Xenobiotic
1121 Resistance in Drosophila. PLOS Genetics. 2014;10(8):e1004560.
1122 122. Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. Transposable elements
1123 contribute to the genomic response to insecticides in Drosophila melanogaster. Philosophical
1124 Transactions of the Royal Society B: Biological Sciences. 2020;375(1795):20190341.

1125 123. Kamgang B, Tchapga W, Ngoagouni C, Sangbakembi-Ngounou C, Wondji M, Riveron
1126 JM, et al. Exploring insecticide resistance mechanisms in three major malaria vectors from
1127 Bangui in Central African Republic. Pathogens and global health. 2018;112(7):349-59.
1128 124. Grau-Bové X, Tomlinson S, O'Reilly AO, Harding NJ, Miles A, Kwiatkowski D, et al.
1129 Evolution of the Insecticide Target Rdl in African Anopheles Is Driven by Interspecific and
1130 Interkaryotypic Introgression. Molecular Biology and Evolution. 2020;37(10):2900-17.
1131 125. The Anopheles gambiae Genomes Consortium. Genome variation and population
1132 structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and
1133 Anopheles coluzzii. Genome Research. 2020;30(10):1533-46.
1134 126. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of
1135 transposable elements to the innovation of gene regulatory networks. Genome research.
1136 2014;24(12):1963-76.
1137 127. Jiang J-C, Upton KR. Human transposons are an abundant supply of transcription factor
1138 binding sites and promoter activities in breast cancer cell lines. Mobile DNA. 2019;10:16-.
1139 128. Villanueva-Cañas JL, Horvath V, Aguilera L, González J. Diverse families of
1140 transposable elements affect the transcriptional regulation of stress-response genes in Drosophila
1141 melanogaster. Nucleic Acids Research. 2019;47(13):6842-57.
1142 129. Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-
1143 regulatory sequences in mammalian genomes. Philosophical Transactions of the Royal Society
1144 B: Biological Sciences. 2020;375(1795):20190347.
1145 130. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable
1146 and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome
1147 research. 2017;27(5):722-36.

1148 131. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for
1149 third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
1150 132. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly
1151 from long uncorrected reads. Genome research. 2017;27(5):737-46.
1152 133. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using
1153 only nanopore sequencing data. Nature Methods. 2015;12(8):733-5.
1154 134. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
1155 Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly
1156 Improvement. PLOS ONE. 2014;9(11):e112963.
1157 135. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies.
1158 F1000Research. 2017;6(1287).
1159 136. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO:
1160 fast and accurate reference-guided scaffolding of draft genomes. Genome Biology.
1161 2019;20(1):224.
1162 137. Shumate A, Salzberg SL. Liftoff: an accurate gene annotation mapping tool. bioRxiv.
1163 2020:2020.06.24.169680.
1164 138. Okonechnikov K, Golosova O, Fursov M, the Ut. Unipro UGENE: a unified
1165 bioinformatics toolkit. Bioinformatics. 2012;28(8):1166-7.
1166 139. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.
1167 Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-6.
1168 140. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
1169 sequencing data. Bioinformatics (Oxford, England). 2012;28(23):3150-2.

1170 141. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An
1171 Automatic Transposable Element Classification Tool. PLOS ONE. 2014;9(5):e91929.
1172 142. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
1173 architecture and applications. BMC bioinformatics. 2009;10:421-.
1174 143. Fernández-Medina RD, Struchiner CJ, Ribeiro JMC. Novel transposable elements from
1175 Anopheles gambiae. BMC Genomics. 2011;12(1):260-.
1176 144. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in
1177 eukaryotic genomes. Mobile DNA. 2015;6(1):4-9.
1178 145. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with
1179 Gene Gain, Loss and Rearrangement. PLOS ONE. 2010;5(6):e11147.
1180 146. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
1181 2018;34(18):3094-100.
1182 147. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo
1183 assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with
1184 dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito
1185 (Aedes aegypti). Genome biology and evolution. 2015;7(4):1192-205.
1186 148. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software
1187 for Computing and Annotating Genomic Ranges. PLOS Computational Biology.
1188 2013;9(8):e1003118.
1189 149. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-
1190 access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research.
1191 2004;32(suppl_1):D91-D4.

1192 150. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif.
1193 Bioinformatics. 2011;27(7):1017-8.
1194 151. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE:
1195 tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-8.
1196 152. Sloutskin A, Danino YM, Orenstein Y, Zehavi Y, Doniger T, Shamir R, et al. ElemeNT:
1197 a computational tool for detecting core promoter elements. Transcription. 2015;6(3):41-50.

1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214

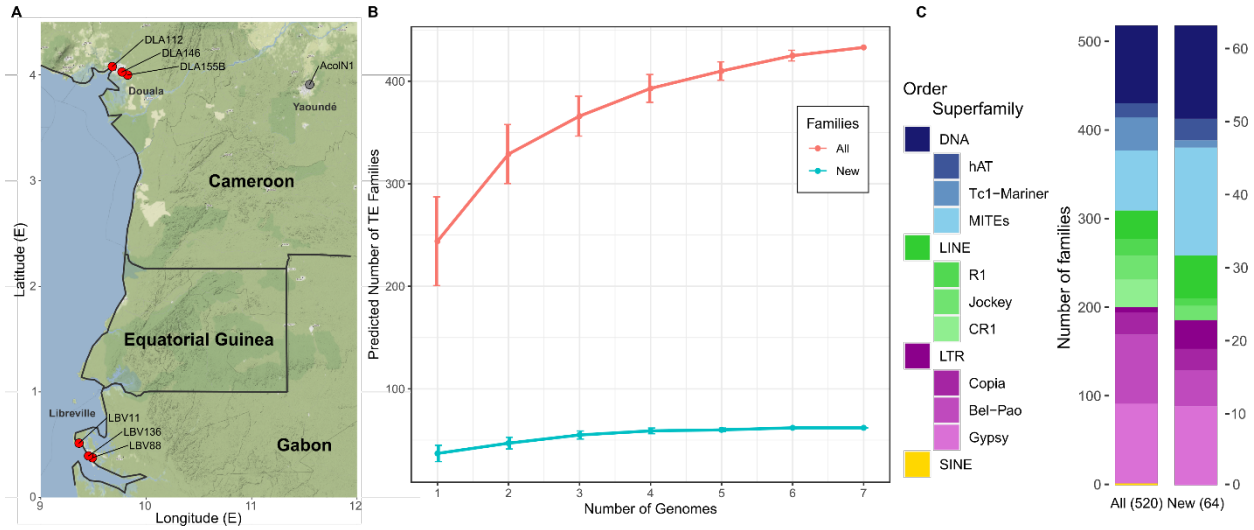
1215 **FIGURES**

1216

1217 **Figure 1. Transposable elements in *An. coluzzii*.**

1218 A) Geographic location of the six breeding sites analyzed (in red) and of the place of origin of
1219 the Ngousso colony (in grey) which was used to generate the *AcolN1* genome. B) Number of TE
1220 families identified when using a single genome or when using all possible combinations of more
1221 than one genome. The red line shows the total number of TE families and the blue line shows the
1222 number of newly described families. Note that 76% of all the TE copies where already identified
1223 when analyzing a single genome (Additional File 2). C) Classification of all TE families and
1224 newly described families in *An. coluzzii*. The three most abundant superfamilies from each order
1225 are shown.

1226



1227

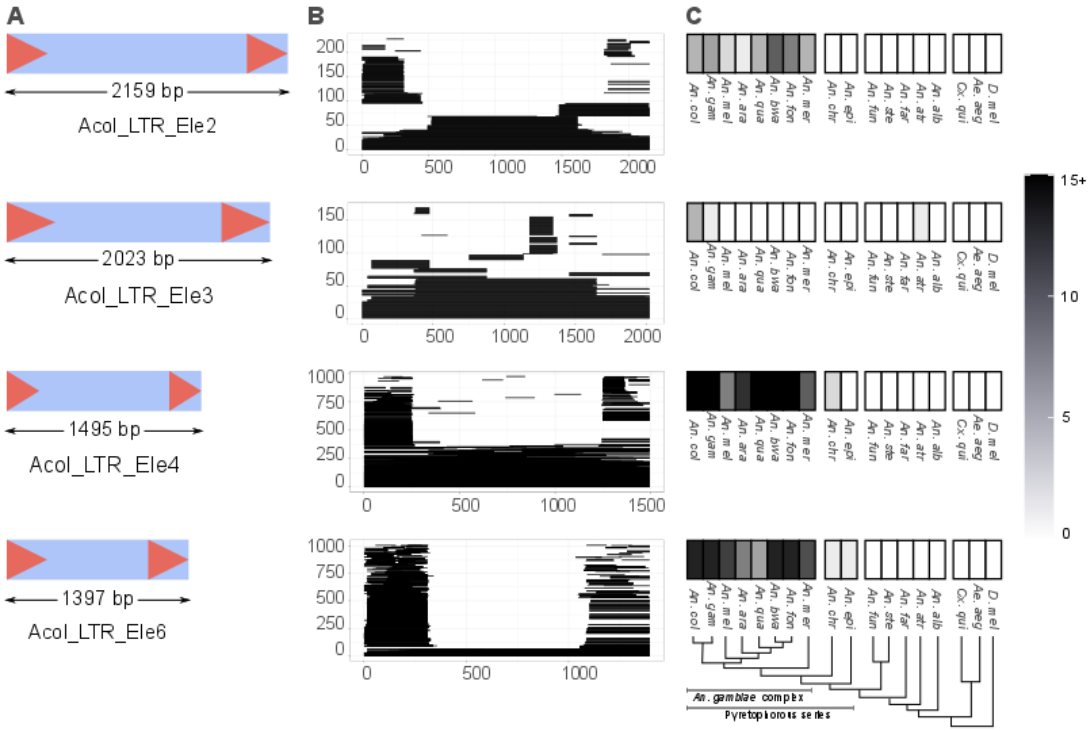
1228

1229

1230

1231 **Figure 2. Structure, abundance and phylogenetic distribution of novel TE families.**

1232 The four newly identified TRIMs families are shown, for the remaining 60 novel families see
1233 Additional file 3: Figure S2. A) The structure of each new family is displayed: the light blue box
1234 represents the full extension of the TE and the red arrows represent LTRs. B) All insertions for
1235 each TE family were identified and are shown as a coverage plot in which each lines represents a
1236 copy in the genome. Note that the large number of stacked horizontal lines in the extremes of the
1237 plot represent and abundance of solo LTRs. C) Phylogenetic distribution of the TE family
1238 insertions in 15 members of the *Anopheles* genus, *Culex quinquefasciatus, Ae. Aegypti* and *D.*
1239 *melanogaster*. The number of insertions with more than 80% identity and spanning at least 80%
1240 of the consensus, in each species is shown using a black and white gradient. Species with no
1241 insertions are shown in white while species with 15 or more insertions are shown in black.



1242

1243

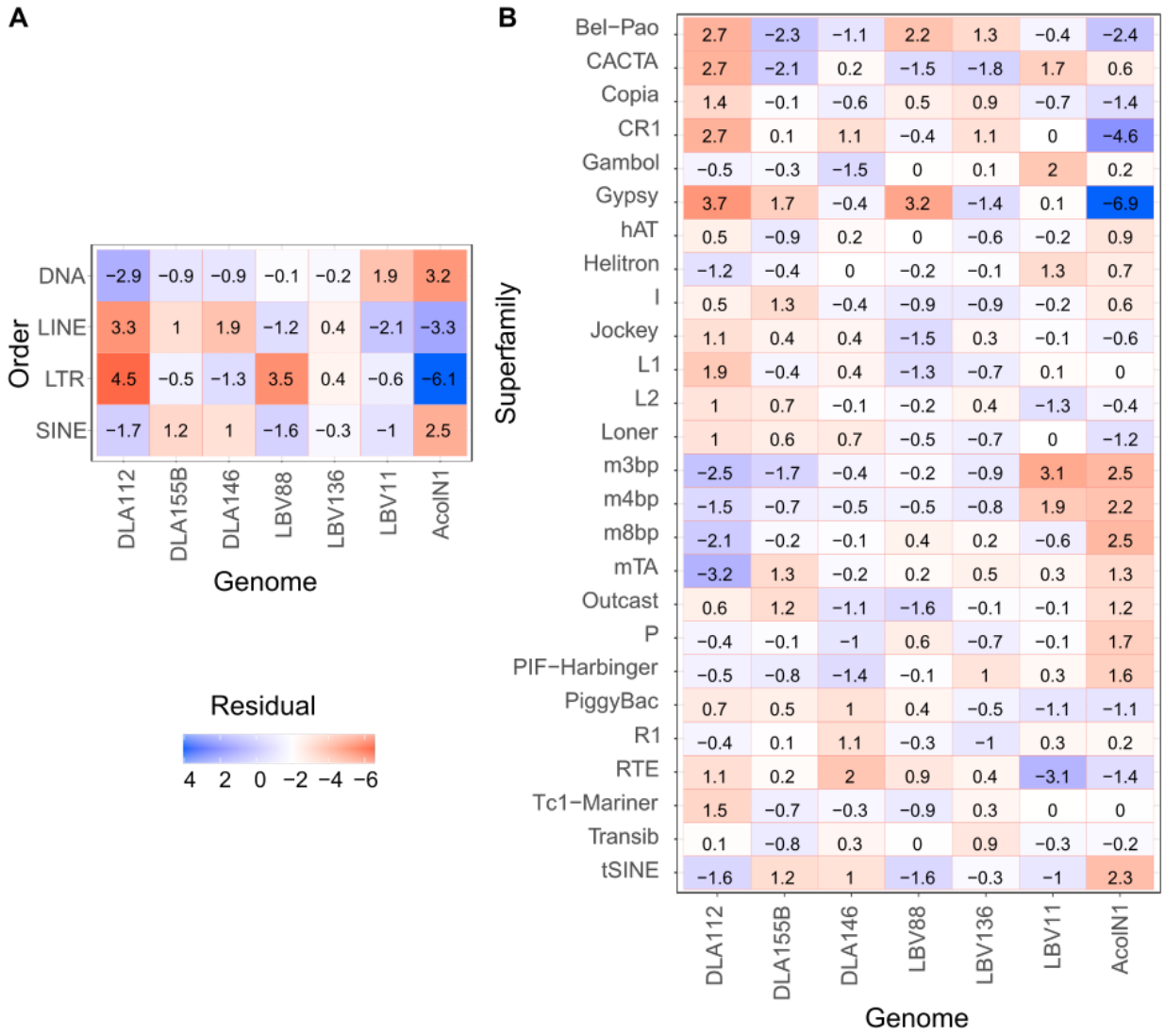
1244 **Figure 3. Differences in TE content between the seven *An. coluzzii* genomes.**

1245 Differences are shown at the (A) order and (B) superfamily levels. χ^2 tests were performed for

1246 the number of insertions and the Person's residuals are shown. Note that MITEs are divided into

1247 the m3bp, m4bp, m8bp and mTA superfamilies.

1248



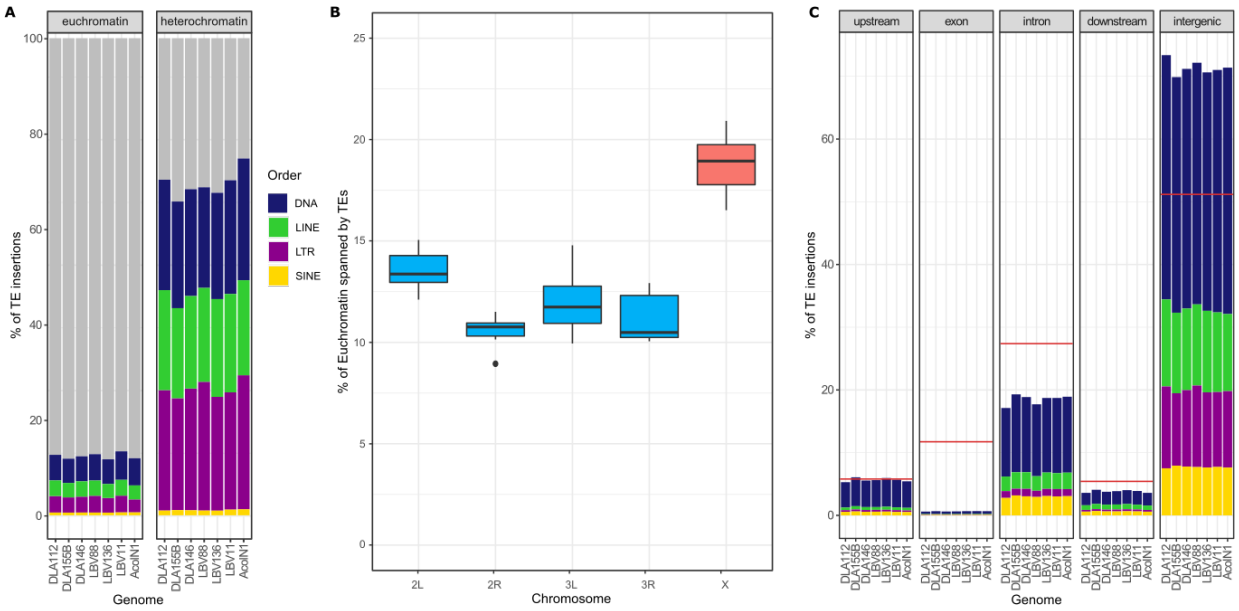
1249

1250

1251 **Figure 4. TE insertions distribution throughout the genomes.**

1252 A) Percentage of euchromatin and heterochromatin occupied by TEs in each of the seven
1253 analyzed genomes. Each order is shown in a different color. B) Boxplots of the percentage of the
1254 euchromatin of each chromosome covered by TEs. Autosomes are shown in blue and the X
1255 chromosome in red. C) Percentage of TE insertions in each genome that fall in a specific
1256 genomic region. A red line is used to display the expected percentage that should be covered by
1257 TEs taking in consideration the size of the genomic region. Each order is shown in a different
1258 color as in A).

1259

1260 

1261

1262

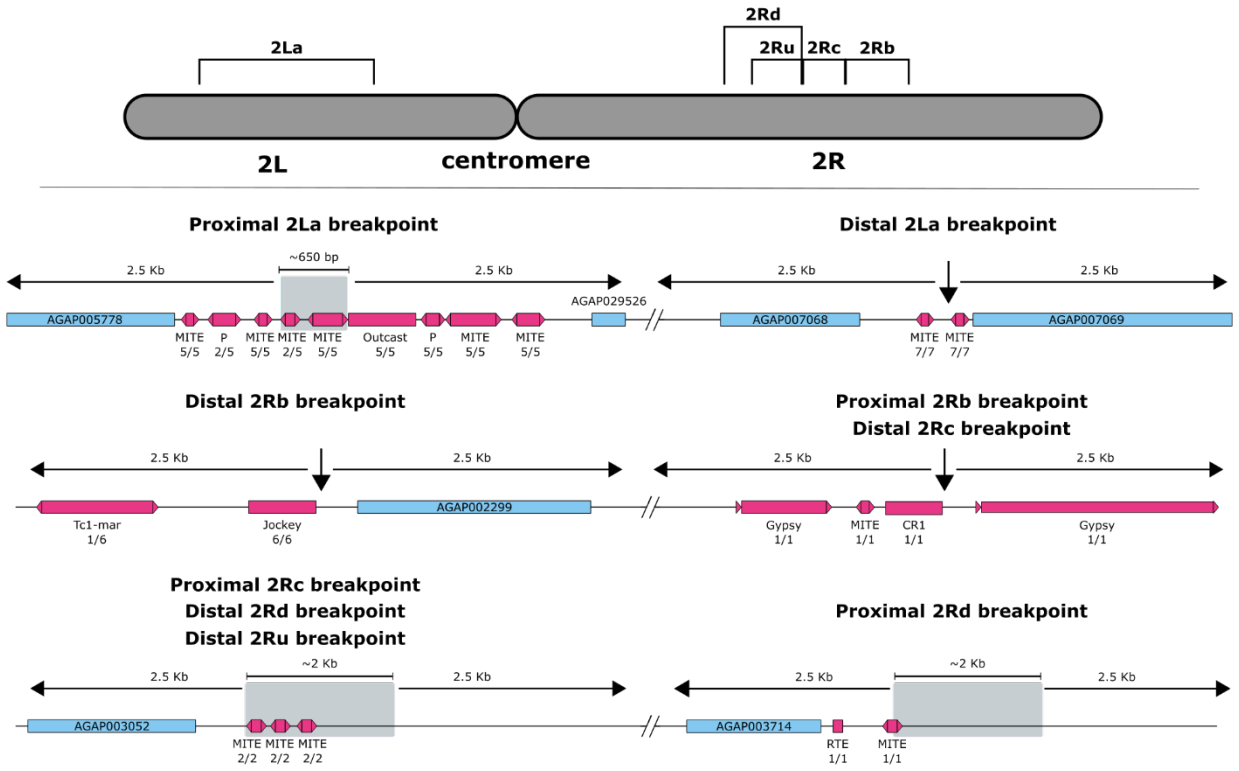
1263

1264

1265 **Figure 5. TE insertions near known inversion breakpoints.**

1266 Diagram of chromosome 2 with the analyzed inversions. For each inversion both breakpoints,
1267 proximal (closer to the centromere) and distal (farther from the centromere), plus 2.5 kb to each
1268 side are shown. When the position of a breakpoint was not identified at the single base pair level,
1269 the interval where the breakpoint is predicted to be is shown in a grey box. Genes are shown as
1270 blue boxes while TEs are shown as pink boxes. Below each TE, the family of the TE is shown
1271 and below the family name the number of genomes where the insertion was found and the
1272 number of genomes where the breakpoint region was identified. Note that breakpoints are shared
1273 among some of the inversions.

1274



1275

1276

1277 **Figure 6. TE insertions in the neighborhood of genes involved in insecticide resistance.**

1278 The gene structure is shown in black with arrows representing the exons. TE insertions are

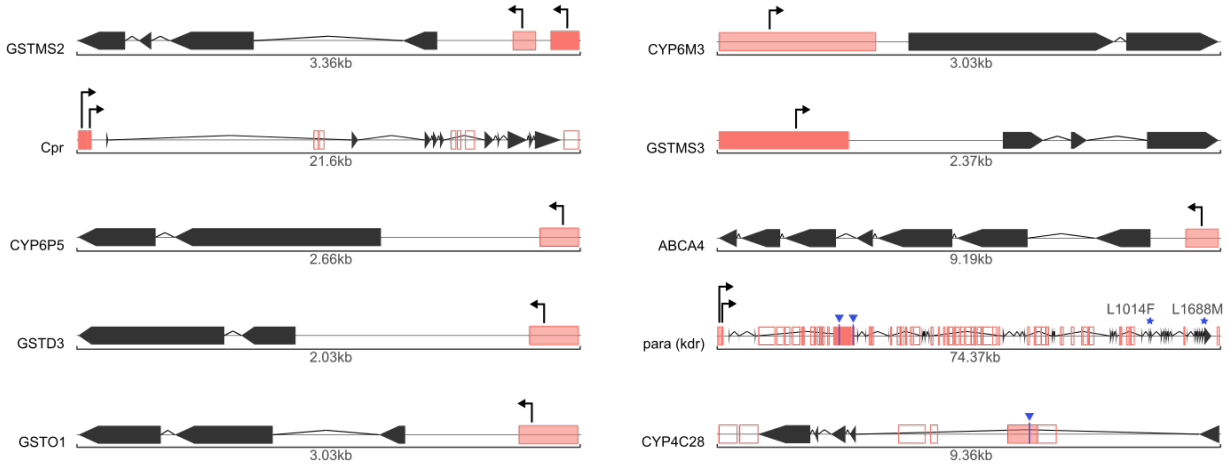
1279 depicted as red boxes. When containing a TFBS for *cnc* or a promoter they are filled in red,

1280 otherwise they are empty. The red color is darker on fixed TEs and lighter on polymorphic TEs.

1281 Promoters are shown as arrows while *cnc* binding sites are shown in blue. Resistance alleles are

1282 shown for *para* (*kdr*).

1283



1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294 **ADDITIONAL FILES**

1295

1296 File name: Additional file 1

1297 File format: Microsoft Excel Binary File format (xls)

1298 Title of data: Supplementary Tables

1299 Description of data: Supplementary Tables

1300

1301 File name: Additional file 2

1302 File format: Portable document format (pdf)

1303 Title of data: Figure S1. Number of TE copies identified when using the TE libraries from an

1304 increasing number of genomes.

1305 Description of data: Number of TE copies identified when using the TE library of a single

1306 genome or when using all possible combinations of more than one genome.

1307

1308 File name: Additional file 3

1309 File format: Portable document format (pdf)

1310 Title of data: Figure S2. Novel TE families

1311 Description of data: Newly described families. A) The structure of each new family is displayed:

1312 the light blue box represents the full extension of the TE and the red arrows represent LTRs. B)

1313 All insertions for each TE family are shown as a coverage plot where each line represents a copy

1314 in a genome. C) Phylogenetic distribution of the TE family insertions in 15 members of the

1315 *Anopheles* genus, *Culex quinquefasciatus, Ae. Aegypti* and *D. melanogaster*. The number of

1316 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 4

File format: Portable document format (pdf)

Title of data: Figure S3. Number of TE insertions vs genome size

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

File name: Additional file 5

File format: Portable document format (pdf)

Title of data: Figure S4. TE landscapes

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

File name: Additional file 6

File format: Portable document format (pdf)

Title of data: Figure S5. 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 7

File format: Portable document format (pdf)

1340 Title of data: Figure S6. Genes associated with insecticide resistance with TE insertions

1341 Description of data: Diagrams of genes associated with insecticide resistance showing the gene

1342 structure and the TE insertions closer than 1 kb to gene.