Introduction

Transposable elements are ubiquitous component of eukaryotic genomes and, besides their mutagenic role [1], they are considered as the major source of variability that can change genomes and their expression, either considering short term or large evolutionary scale time. The action exerted by transposable elements on genomes is predominantly described in studies performed in insect where the abundance of both active and inactive forms of mobile elements have shaped their genomes structurally, functionally and evolutionarily.

The post-genomic era offers a great opportunity to shed light on the evolution of mobile genetic elements with respect to eukaryotic genome. The results obtained from several genomic studies allow the comparison of related sequences from different organisms. In addition, the great amount of sequence data produced have led to the identification of novel families of mobile genetic elements and posed a problem concerning their classification [2,3]. Looking at their transposition mechanism, transposons can be classified into two main classes [4]. Class I elements, or retrotransposons, reverse transcribe a RNA intermediate into cDNA molecules, which is then inserted in the genome. Class I elements can be further categorized in LTR- and non-LTR retrotransposons depending on the presence or absence of direct terminal repeats. Retrotransposons are major components of eukaryotic genomes; they are among the strongest evolutionary driving force acting on the genomes [5], and are potentially able to change gene expression patterns [6] [7]. Their ability to inflate eukaryotic genome size [8] is also at the basis for their use as molecular markers in organisms of socio-economic interest [9].

In the last years the rising interest in the field of mosquitoes’ genomics is demonstrated by the completion of three genome sequences, and this mainly comes from their role as vectors of virus-borne diseases.

Three mosquitoes’ genomes have been sequenced and assembled to date. The first mosquito genome to be sequenced was the Anopheles gambiae [10] followed by the sequencing of the Aedes aegypti’s genome [11].

Culex quinquefasciatus is the main vector of the nematode W. bancrofti, one of the known causes of the lymphatic filariasis, and its genome (about 540 Mbp) [12] has been recently sequenced [13]. Among the Culicidae family, the Anophelinae and the Culicinace subfamilies have diverged about 145–200 Mya, while within the Culicinae subfamily, Aedes and Culex genera have diverged about 52–54 Mya [13]. With this effort, a solid genomic platform for mosquito comparative genomics has been established.

Few Culex transposon families have been described in reports published before the publication of the Culex genome paper, being limited to few DNA transposon [14] [15] and retrotransposon [16] [17] [18] families.

The genomic sequence analysis performed by Arensburger et al. [13] has revealed that nearly 30% of the Culex genome is composed of TEIs. Compared with the TEIs content in the genomes of A. gambiae (16%) and A. aegypti (50%), this appears to be an
intermediate value, as well as intermediate is the genome size of \( C. \) quinquefasciatus compared to the above mentioned genomes (286 Mbp and 1.3 Gbp respectively). The LTR retrotransposons identified and described in the genome sequencing paper have been deposited in the TEfam database [19], a specialized database for transposable elements retrieval and analyses, which focus on mosquito species. In its Culex quinquefasciatus section TEfam contains 81 families of Bel/Pao elements, 32 families of Ty1/copia elements and 57 families of Ty3/gypsy elements in addition to 179 families of non-LTR retrotransposons, 32 families of “cut and paste” transposons families, 3 helitrons families and 100 MITEs families.

A novel class of mobile elements with striking features has been previously described in \( C. \) quinquefasciatus, \( \text{Twin} \) is a family of atypical SINE elements with a dimer-like structure similar to a tRNA gene. It has been proposed that \( \text{Twin} \) is a family of a moderately repetitive sequence specific of the genus Culex, as it is absent in the genome of Aedes species [20].

Furthermore, we have recently described a family of \( \text{Osvaldo} \)-like elements with peculiar structure of the LTRs [17].

Here, we report the presence of twenty-nine families of LTR retrotransposons in the genome of \( C. \) quinquefasciatus, identified using the LTR_STRUC program [21] and not reported in the TEfam database. One of these elements has an atypical Primer Binding Site region probably generated by the insertion of a tRNA dimer immediately downstream the 5’ LTR. Furthermore, we have identified a group of 38 families probably composed of non-autonomous elements, apparently unrelated to any known retrotransposon family, which contain tandem repeated sequences between the LTRs.

The results of the genomic distribution analysis show that the novel retrotransposons identified in this paper are preferentially located in intergenic regions or in intron sequences in the genome of \( C. \) quinquefasciatus. Several insertions that may potentially contribute to the organization of protein-coding genes have been identified. The possible functional role of these insertions on the host gene organization is discussed.

Materials and Methods

**LTR_STRUC analysis and classification of LTR retrotransposons**

The genome sequence of \( C. \) quinquefasciatus was downloaded from the Broad Institute website (http://www.broad.mit.edu/index.html) and scanned with the LTR_STRUC program [21] using the default parameters. 1179 putative retrotransposon sequences obtained as output were subjected to an “all against all” BLAST in order to group sequences with % identity greater than 98% over a sequence of at least 1 Kb. 157 groups containing at least one sequences were obtained after this step. The final subset of LTR retrotransposons was then BLAST-searched against the TEfam database in order to define families and to highlight previously not annotated sequences. In order to confirm the results obtained by LTR_STRUC we have performed a LTR-retrotransposon search using the LTRharvest program [22]. The results obtained were compared to the TEfam database and the LTR_STRUC output.

Criteria for defining LTR-retrotransposons were identical to the previously described criteria adopted during \( A. \) aegypti TE analysis [17]. Briefly, sequences of the Ty3/gypsy LTR retrotransposons are considered as belonging to the same element if they share at least 85% nucleotide identity along at least 400 bp in their coding region. Ty1/Copia sequences that share at least 85% identity at the nucleotide level over at least 1000 bp are considered belonging to the same element. Copies of \( \text{Pao} / \text{Bel} \) retrotransposons are considered as belonging to the same element if they show at least 70% identity at the nucleotide level in their coding sequences.

The names assigned to the newly discovered retrotransposons follow the nomenclature adopted in the Repbase database [23] and contain the prefix “Cq” for species (\( C. \) quinquefasciatus), the specification of the family (namely Ty3/gypsy, Ty1/copia, BEL, etc.) and a number suffix.

**Analysis of insertions**

The ORF finder program (http://www.ncbi.nlm.nih.gov/orffor/gorf/gorf.html) was used to determine the ORF number of each element detected.

The TSD (Target Site Duplicated upon insertion) and the length of the LTRs of each element obtained were determined by visual inspection of sequences. In absence of a reported list of the tRNA gene sequences in \( C. \) quinquefasciatus the PBS sequences were determined by comparison of a tRNA dataset of \( A. \) gambiae at the http://lowelab.ucsc.edu/GtRNAdb/Agamb/ website. The tRNA genes of \( A. \) gambiae are highly similar (if not identical in most of the cases) to the tRNA of \( C. \) quinquefasciatus, as demonstrated by BLAST analysis (not shown). This data ensure that a good PBS prediction has been done using the \( A. \) gambiae tRNA dataset.

To detect retrotransposon insertions near (or overlapping) host genes, a BLAST search at the Vectorbase database (http://www. vectorbase.org/) was performed using the following arbitrary criteria: 1) only insertions with average similarity greater than 85% were counted; 2) insertions shorter than 180 bp were not taken in account; 3) the E value was lower than 1E-40. These criteria allow the detection of full-length elements and defective elements without missing solo LTR and preventing misleading results coming from low quality alignments. The analyses of tandem repeats contained into retrotransposon were performed with the Tandem Repeat Finder program [24] using the basic option.

**RepeatMasker analysis**

RepeatMasker software (version 3.2.9) [25] was used to estimate the retrotransposons occupancy as percent of the genome fraction. Repeats search was performed using Cross_Match as sequence search engine. A repeats library was built starting from the LTR retrotransposon group described in this paper (file S1), and it was used to scan the genome sequence. Scanning was carried out using a cutoff value of 250.

**Multiple sequence alignment and phylogenetic analysis**

As previously described [26] the best way to reconstruct phylogeny of retroelements is to perform multiple alignment of RT-RnaseH-INT domains. These domains encoded by each putatively active element were extracted from the translated ORF encoding the POL polyprotein and used to reconstruct the phylogenetic history of \( C. \) quinquefasciatus Ty3/gypsy like retrotransposons. We have no evidence of domain swapping by performing multiple alignment using \( RnaseH \), RT or INT domains (data not shown) at least for the elements analyzed in this paper. Either MUSCLE [27] or ClustalX [28] were used to perform multiple alignments. After a manual check of the alignments Neighbor-joining tree with bootstrap analyses were generated using MEGAS [29]. As reference, previously described elements in other species [19] [17] [26] were used to establish relationships between \( C. \) quinquefasciatus retroelements.

Multiple alignments are available as file S2.

**Results**

The genome sequence of \( C. \) quinquefasciatus (assembly version CqpiJ1) was analyzed using the LTR_STRUC program, in order to obtain LTR-retrotransposon sequences.
The 1179 insertions obtained were clustered into groups of nearly identical sequences (see Material and Methods section). This allowed the identification of 157 families of elements containing at least one retrotransposon copy. The DNA sequence of representative elements of each family was BLAST-searched against the TEfam database. Only sequences that did not match any of the elements reported in TEfam were further analyzed. This led to the identification of 29 previously not described and potentially active elements (i.e. containing the generic specification for the transposition machinery and the required cis-acting sequences). A single representative element of each family was used in the phylogenetic analysis. Representative elements were chosen among those having the best match between the two LTRs, the longest sequence and the simplest ORF structure, coding for the entire set of protein domain typically found in the family. Furthermore, elements with such features could be potentially functional and transpositionally active. Although we have identified Ty1-copia and Bel-Pao elements, they were not further analyzed due to the presence of identical sequences in the TEfam database.

A phylogenetic analysis was performed in order to identify the origin of each group of sequences extrapolated from the LTR_STRUC output. The RT-RNaseH-INT domains of the POL polyproteins were aligned along with the corresponding domains of reference elements. This multiple alignment was then used to generate a NJ tree. As can be observed in figure 1, all the novel elements identified fall into the Ty3-gypsy superfamily of LTR-retrotransposons. Furthermore, the results reported in figure 1 clearly show that the new elements reported belong to five distinct lineages (namely gypsy, Osvaldo, Mag mdg3 and mdg1). No novel GRon1-like elements were detected despite they are well represented in the genome of C. quinquefasciatus, as demonstrated by the presence of nine GRon1-like elements in the TEfam database.

The structural features of the retrotransposons identified in this study were also analyzed and reported in table 1. Except for few cases that will be discussed below, the main features of these elements (namely PBS type and LTR mean length) are in agreement with those of known elements belonging to the same lineage and described in other species. In table 1 is also reported the percent nucleotide identity between the LTR of each insertion detected. This value gives an approximate idea of the age of the insertions. To the best of our knowledge, the synonymous substitution rate has not been estimated for C. quinquefasciatus; consequently we are not able to make more precise estimations of the age of insertions.

No target site preference was observed for any of the retrotransposons analyzed.

A closer view of the phylogenetic analysis results indicates that eleven elements can be classified as Osvaldo-like, two fall in the gypsy lineage, one in the Mdg1 and Mdg3 lineages respectively. The analysis performed was aimed to dissect the structural properties for each family detected, and to compare them with those of known elements of the same phylogenetic lineage.

**Gypsy lineage**

Two novel gypsy-like elements have been identified in this study. The structural analyses have revealed that the first base of the putative PBS overlaps the last base of the 5′ LTR in these elements; this was also observed for the gypsy element of D. melanogaster [30]. It can be assumed that this a general rule for the members of the gypsy lineage identified in other organisms.

Several members of the gypsy lineage identified so far in other organisms contain an ORF that could potentially encode for the envelope protein (ENV), a typical retrovirus like protein reported to be important in the horizontal transmission process [31]. The two gypsy-like elements detected in this study also contain an ORF that potentially encodes an ENV-like protein. The conceptual translation of these putative env-coding regions reveals typical domains of ENV proteins (not shown).

**Mag lineage**

Members of this lineage have been previously identified in several insect genomes such as B. mori [32], A. gambiae, D. melanogaster [33] and C. elegans [34] [35].

Thirteen families are phylogenetically related to the Mag element. The PBS of the Mag-like elements is complementary either to the tRNA\(^{\text{Lys}}\) or to the tRNA\(^{\text{Ser}}\). A single element (cqgypsy_23) with an atypical PBS sequence, complementary to the tRNA\(^{\text{Ser}}\), has been identified. Three elements (namely cqgypsy_24, cqgypsy_25 and cqgypsy_60) contain tandem repeated sequences in the 5′ UTR. The unusual size of the cqgypsy_24 element (greater than 10 Kbp) is due to the size of a repeated region (about 5 Kbp).

The phylogenetic analysis shows that the Mag clade is formed by two subgroups strongly supported by high bootstrap values. Four elements of C. quinquefasciatus co-cluster with the Mag element, while 9 elements fall into the second cluster with five elements from A. aegypti used as reference elements.

**Mdg1 and Mdg3 lineages**

Two elements identified in this paper belong to the Mdg1 and Mdg3 clades respectively. cqgypsy_47 belongs to the Mdg1 clade while cqgypsy_29 belongs to the Mdg3 lineage. Looking at the TEfam database, eight Mdg1-like elements and three Mdg3-like elements can be retrieved. This suggests the possibility that these two clades could be poorly represented in the genome of C. quinquefasciatus.

**Osvaldo lineage**

Existing data in the TEfam database, suggest that these elements are abundant in the family Culicidae. Twenty-nine Osvaldo-like elements are annotated in the genome of A. aegypti and five elements in the genome of C. quinquefasciatus.

Querying the TEfam dataset for Osvaldo-like elements in C. quinquefasciatus results in five annotated elements. We have identified 11 unreported Osvaldo-like elements in the genome of C. quinquefasciatus. Their LTRs length ranges from 997 to 2055 bp, a feature that characterizes members of the Osvaldo lineage. In
| Lineage | Family  | copies | Element | length | LTRs | %LNI  | ORFs | PBS  | TSD | supercont |
|---------|---------|--------|---------|--------|------|-------|------|------|-----|----------|
| Mag     | cqgypsy_8 | 3      | cqgypsy_8.1 | 4993 | 179 | 100 | 1 | Leu | gtcac | 3.1653 |
| Mag     | cqgypsy_9 | 1      | cqgypsy_9.1 | 4568 | 145 | 100 | 1 | Leu | ttagt | 3.1194 |
| Mag     | cqgypsy_11 | 11     | cqgypsy_11.1 | 5129 | 169/181 | 99 | 2 | Ser | ataa | 3.429 |
| Mag     | cqgypsy_15 | 4      | cqgypsy_15.1 | 4851 | 197 | 99 | 1 | Ser | tccgt | 3.1361 |
| Mag     | cqgypsy_21 | 2      | cqgypsy_21.1 | 6184 | 287 | 99 | 2 | Ser | tcctt | 3.770 |
| Mag     | cqgypsy_24 | 6      | cqgypsy_24.1 | 10446 | 304 | 99 | 2 | Ser | accag | 3.163 |
| Mag     | cqgypsy_25 | 6      | cqgypsy_25.1 | 7859 | 198 | 99 | 2 | Arg | ggaag | 3.176 |
| Mag     | cqgypsy_27 | 4      | cqgypsy_27.1 | 5260 | 196 | 97 | 2 | Ser | gtgcc | 3.790 |
| Mag     | cqgypsy_32 | 3      | cqgypsy_32.1 | 4918 | 190 | 99.5 | 1 | Leu | ggaat | 3.540 |
| Mag     | cqgypsy_37 | 1      | cqgypsy_37.1 | 7479 | 182 | 97.8 | 3 | Leu | attac | 3.1290 |
| Gypsy   | cqgypsy_13 | 7      | cqgypsy_13.1 | 7544 | 208 | 100 | 2 | Ser | ctott | 3.7 |
| Gypsy   | cqgypsy_20 | 3      | cqgypsy_20.1 | 7438 | 357 | 100 | 3 | Ser | atata | 3.1285 |
| Mdg3    | cqgypsy_29 | 4      | cqgypsy_29.1 | 5316 | 264 | 100 | 2 | Leu | gttg  | 3.462 |
| Mdg1    | cqgypsy_47 | 25     | cqgypsy_47.1 | 6771 | 431 | 100 | 2 | Arg | cttc  | 3.2173 |
| Osvaldo | cqgypsy_1 | 3      | cqgypsy_1.1 | 11926 | 2137 | 99 | 2 | Lys | ggtt  | 3.349 |
| Osvaldo | cqgypsy_2 | 21     | cqgypsy_2.1 | 12138 | 2055/2056 | 99 | 2 | Lys | actt  | 3.1399 |
| Osvaldo | cqgypsy_3 | 7      | cqgypsy_3.1 | 10049 | 1591/1596 | 99 | 2 | Lys | aagttg | 3.7 |
| Osvaldo | cqgypsy_4 | 6      | cqgypsy_4.1 | 10581 | 1742 | 99 | 2 | Lys | caac  | 3.169 |
| Osvaldo | cqgypsy_5 | 5      | cqgypsy_5.1 | 6914 | 997 | 99 | 1 | Lys | aagt  | 3.38 |
| Osvaldo | cqgypsy_6 | 29     | cqgypsy_6.1 | 10354 | 1340 | 99.6 | 1 | Lys | caaa  | 3.458 |
| Osvaldo | cqgypsy_14 | 14     | cqgypsy_14.1 | 9473 | 1384 | 100 | 2 | Lys | aat  | 3.568 |
| Osvaldo | cqgypsy_15 | 15     | cqgypsy_15.1 | 9393 | 1369/1368 | 95.1 | 2 | Lys | ttagt  | 3.83 |
| Osvaldo | cqgypsy_61 | 48     | cqgypsy_61.1 | 9196 | 1151 | 99.6 | 2 | Lys | caaag  | 3.1285 |
| Osvaldo | cqgypsy_60 | 48     | cqgypsy_60.1 | 10172 | 1308 | 99.3 | 2 | Lys | acaac  | 3.133 |
| Osvaldo | cqgypsy_64 | 41     | cqgypsy_64.1 | 12479 | 2045/2046 | 99.8 | 2 | Lys | acgt  | 3.82 |
| Osvaldo | cqgypsy_64 | 41     | cqgypsy_64.2 | 12368 | 2038/2037 | 98.6 | 2 | Lys | aagc  | 3.191 |
| Osvaldo | cqgypsy_64 | 41     | cqgypsy_64.3 | 8014 | 2047/234 | 98 | frag | Lys | agat  | 3.141 |
Figure 2 shows the phylogenetic relationships of the Osvaldo-like elements identified in A. aegypti and C. quinquefasciatus. No species-specific cluster was observed in the distribution of these elements. Copy number varies among different families of Osvaldo-like elements (see Table 1) and the PBS is invariably complementary to the 3′ end of the tRNA 499. This is also the initiator tRNA used by Osvaldo [36]. As reported in our previous analyses, both genomes of A. aegypti and C. quinquefasciatus contain retrotransposons that are strictly related to the woot element of T. castaneum [17], but containing unusually short LTRs. The CPGYPSY5 element identified by Minervini et al by BLAST similarity search was also identified during the course of this analysis by the LTR_STRUC program.

cqgypsy_1 is a peculiar element of the Osvaldo lineage. It has been detected as single copy retrotransposon by LTR_STRUC analysis, but probably present in multiple copies in the genome of C. quinquefasciatus as revealed by BLAST analyses on the trace archive (not shown). The structural analysis of its PBS region shows that it has a non-canonical PBS. Instead of a short nucleotide stretch complementary to the 3′ end of a tRNA, we have found a 149 bp long sequence identical to two tRNA arranged in a head to head fashion. The 149 bp sequence is recognized by the tRNA Scan program, which in turn gives two perfectly folded tRNA molecules as output (figure 3B).

The unusual configuration of the PBS region of cqgypsy_1 has been analyzed in details. As can be observed in figure 3, both tRNA-like sequences have the terminal CCA sequences. Furthermore a direct duplication of 26 nucleotides of the 5′LTR has been found at both sides of the tandem tRNA copies. The tandem copies of tRNA identified in cqgypsy_1 are somehow reminiscent of the structure of Tiew elements described by Feschotte and co-authors [20]. Tiew has been described as a novel type of SINE element consisting of two tRNA related regions separated by a 39 bp spacer.

We have also analyzed in detail the phylogenetic relationships of cqgypsy_1 with other elements of the Osvaldo lineage belonging to different mosquito genomes. Its closest relative is the Ty3_gypsy_Ele185 and Ty3_gypsy_Ele180 elements annotated in the TEfam database (TEfam ID TF000935 and TF000939 respectively). None of the related elements of A. aegypti contain such tandem copy of tRNA. We do not expect to observe significant sequence similarity at the nucleotide level when Culex and Aedes elements were compared in a pair-wise alignment, despite the strict relationship observed at the protein level. By comparing the three Osvaldo-like elements, cqgypsy_1, Ty3_gypsy_Ele185 and Ty3_gypsy_Ele180, we have detected a similarity region in a 29–30 nucleotides region encompassing the boundary between the 5′LTR and the PBS region (see figure 3C), suggesting an unusually strong cross-species conservation of the LTR sequence flanking the PBS. This conservation across the 5′LTR boundary and the PBS was not observed after comparison of any of the retrotransposons analyzed in this paper with their relatives in Aedes aegypti. Taken together, these results confirm the phylogenetic relationship among these elements and indicate a strong conservation of the 30-nucleotide long sequence across the 5′ LTR shared by Culex and Aedes elements, which is probably under functional constrains.

### Non-autonomous elements

Non-autonomous elements are important to understand the evolutionary dynamics of transposable elements in the genomic context [37]. Non-autonomous elements were also detected and analyzed in this work. The LTR_STRUC program is also able to find aberrant retrotransposon sequences (i.e. LTR-retrotransposon with internal deletions of various size); in this case a lower score is assigned to a potentially active retrotransposon. However, most of the defective LTR retrotransposons detected are false positives resulting from a couple of direct repeats (mimicking the LTRs) but lacking PBS, PPT, the target site duplication and the coding sequences. A certain number of low scoring sequences extracted by LTR_STRUC are bona fide defective elements. Several nested elements were also found in the output of LTR_STRUC, but no significant bias of nesting was observed. In general, truncated retrotransposons are related to at least one putatively active element, in the TEfam dataset or in our output, thus falling into a specific family of elements that, for this reason, will be composed by autonomous and non-autonomous elements.

Notably, we have found a group of non-autonomous elements lacking coding sequences, and that cannot be related to any of the known putatively active elements annotated in TEfam, nor to any of the elements identified in this work. The features of these elements are summarized in Table 2.

Elements belonging to this group are featured by highly similar LTR sequences (>98% identity), a sharply definable PBS sequence immediately downstream the 5′LTR, a PPT upstream the 3′LTR and a duplicated sequence at the insertion site. We were unable to classify these elements using phylogenetic criteria, due to the lack of coding sequences that would enable common RT-based phylogenetic analyses.

In addition, a common feature of all these elements is the presence of tandemly repeated sequences bracketed by the retrotransposon LTRs.

The presence of repeated sequences into a retrotransposon seems to be a nearly exclusive feature of this group of elements. The exception is represented by three putatively active elements belonging to the Mag lineage (cqgypsy_24, cqgypsy_25 and cqgypsy_66), carrying tandemly repeated sequences, identified during the genome wide screening in C. quinquefasciatus. Moreover, the exceptional size of these three Mag-like elements is due to the presence of repeats.

| Lineage | Family | copies | Element | length | LTRs | %LNI | ORFs | PBS | TSD | supercont |
|---------|--------|--------|---------|--------|------|------|------|-----|-----|----------|
| Osvaldo | cqgypsy_65 | 20 | cqgypsy_65.1 | 10447 | 1565 | 99% | 2 | Lys | aacc | 3.254 |

Note that two values are reported in the LTRs column if the two LTRs of an element differ in size. "frag" indicates fragmented coding regions. **copy number** indicates the number of copies of tRNA identified in tRNA-like sequences have the terminal CCA sequences. Further-
The repeated region sequence varies among families, and constitutes as much as 95% of the entire length of a given element. Tandem repeats Finder [24] allows the estimation of the entropy value for a given DNA sequence, a parameter based on the percent base composition and whose value is comprised between zero (indicating low sequence complexity) and two (indicating high sequence complexity). A base composition analysis of the repeated sequences in these LTR-retrotransposons suggests that in only a few cases they are composed by simple di-nucleotide iterations (i.e. cqUNK_3, first repeat), while in most of the cases repeats are complex stretches of DNA as demonstrated by entropy values very close to two (table 2).

LTR-retrotransposons containing repeats have been so far identified in other species [38]. Such repeats are usually located in the 5' UTR or in the 3' UTR of these retroelements. It has been demonstrated that tandem repeated sequences carried by retrotransposons of Drosophila melanogaster could behave as powerful regulatory sequences, such as enhancers of the gene expression or genetic insulators. As an example, the tandem repeat in the 5'UTR of gyph is a powerful insulator [39].

Retrotransposon lacking coding sequences and not relatable to any known master copy have been also identified in A. gambiae in previous genome wide searches (Marsano RM unpublished results). Unlike the non-autonomous elements identified in C. quinquefasciatus and described above, those identified in A. gambiae do not contain tandemly repeated sequences.

Distribution of the retrotransposons in the genome of C. quinquefasciatus

We have performed distribution analysis at the genomic level using BLAST and RepeatMasker [25]. RepeatMasker allows a rapid estimation of the genomic fraction occupied by the sequences analyzed. The analysis was performed using a custom library of repeats identified in this paper.

The genome fraction occupied by the retrotransposon sequences showed in table 1 is 0.82% (4.75 Mbp/579 Mbp). This is likely to be an underestimation due to the criteria used (see materials and methods section). Furthermore, we have intentionally excluded from this analysis the defective retrotransposons described in the previous paragraph, as they could inflate the genomic fraction due to the presence of tandem repeats, which can be found as part of complex satellite rather than retrotransposons.

The BLAST search was performed against C. quinquefasciatus genomic database in order to discriminate among insertions in gene free (or intergenic) genomic regions. A great number of insertions are represented by rearranged elements and by solo-LTRs that can be generated by homologous recombination events between the 5' and 3' LTRs.

It has been reported that several families of gyph-like elements are loaded with potent regulatory elements such as enhancers [40], and insulators [41]. Such cis-regulatory elements, when brought in proximity of genes by mean of novel insertions, are able to modify their original expression pattern, in a way that is dependent of the strength of the regulatory element carried by the retrotransposon and of the distance from the endogenous gene. In order to define the distance occurring between LTR-retrotransposons and nearby genes, we performed our analysis using an arbitrary window length of 5 Kb upstream the transcriptional start site or 5 Kbp downstream the termination of transcription of genes annotated in Vectorbase and in which insertions have been detected. This analysis also enables to know if there is a contribution in the gene organization and evolution in C. quinquefasciatus.

Due to the large number of BLAST hits (more than 7000) obtained by searching non-autonomous elements against the genomic sequence, we have performed the BLAST search against the transcripts database and considering only insertions in the coding region of predicted genes. The results of these analyses are reported in table 3 and table 4.

The results summarized in table 3 have been obtained using the elements listed in table 1 as query for BLAST analyses; 84% (313 out of 371) of the insertions detected lay in intergenic regions (i.e. outside the 5 Kbp window upstream/downstream the genes). The remaining 16% (58 insertions) lay in genomic loci where also genes reside (i.e. within 5 Kbp upstream/downstream of validated mosquito genes). It is possible that such insertions could contribute to define the spatial and temporal pattern of expression of strictly linked genes.

Among the insertions in proximity of annotated genes, nineteen insertions (5% of the insertions detected) hit genes, and, among them, six insertions (less than 2%) are localized in introns. Standing to the exon-intron organization reported in Vectorbase, the remaining insertions contribute to entire exons or part of them or are localized at exon-intron boundaries. These data suggest that at least a fraction of the LTR retrotransposon insertions that we have considered, could contribute to define the protein-coding regions of genes.

The results obtained using the elements listed in table 2 as query for BLAST analyses indicate that such non-autonomous elements can also be found in genes. Similarly they seem to contribute at the same strength in the building of protein-coding regions of genes in the genome of C. quinquefasciatus (table 4). However, after extensive searches against the ESTs databases, we have not been able to find evidences supporting that the retrotransposons analyzed are recruited as exons in the mature transcripts of the genes in which they are inserted. Furthermore, the comparison (not shown) of the genes reported in tables 3 and 4 with the respective orthology in Aedes aegypti suggests that, such insertions are probably recent, and have occurred specifically in the evolutionary lineage of C. quinquefasciatus.

These results could be an underestimation, because we intentionally excluded from the BLAST output insertions into, or in proximity to hypothetical protein coding genes that, with the ongoing annotation of the genome could be classified as C. quinquefasciatus genes.

Discussion

In this paper we present data from the LTR_STRUC scan of the Culex quinquefasciatus genome. We have been able to identify, by the use of an alternative in silico approach, the presence of 67 novel LTR-retrotransposons in the Culex genome. These results contribute to increase the already large dataset of retrotransposons present in the TEfam database. The first consideration to be done is that, in order to identify the repeats complement of a euukaryotic genome the implementation of different methods is necessary.
Figure 3. Organization of the LTR-PBS region of *cqgypsy_1*. A) The tRNA sequences inserted into the 5’LTR of the *cqgypsy_1* element. The LTR sequence is colored in red, while the PBS sequence is colored in blue. The red bar indicates the duplicated sequence surrounding the putative Twin element. Each of the tRNA halves of the putative Twin is highlighted in turquoise (tRNALys) or in yellow (tRNA_Glu). The PBS is depicted in blue. ACATTGCTATCATAGTTGGAGTAGTTGGGGAATTTAGTTGTTATAGTTGAAAACCTTTGTT. B) **tRNA^Glu** and **tRNA^Lys** cloverleaf secondary structures. C) Local alignment results of *cqgypsy_1* with the *gypsy_Ele180* and *gypsy_Ele185*. The aligned region correspond to the 5’LTR (red)/PBS(black) boundary. doi:10.1371/journal.pone.0030770.g003
| Element     | supercont | length | LTR | PBS | TSD | Rep Position | Period | Copies | Entropy | %     |
|-------------|-----------|--------|-----|-----|-----|--------------|--------|--------|---------|-------|
| cqUNK_1     | 3.2       | 2994   | 261 | Arg | caagg | 583–1963    | 155    | 8.9    | 1.96    | 52.9  |
|             |           |        |     |     |      | 2188–2226   | 16     | 2.4    | 1.40    |       |
|             |           |        |     |     |      | 2206–2380   | 22     | 8.0    | 1.40    |       |
| cqUNK_3     | 3.322     | 14602  | 315/337 | Leu | attcc | 3252–3314   | 2      | 31.5   | 1.10    | 0.9   |
|             |           |        |     |     |      | 13294–13367 | 33     | 2.2    | 1.63    |       |
| cqUNK_4     | 3.720     | 2990   | 389 | Arg | ttct  | 540–625     | 34     | 2.5    | 1.94    | 2.8   |
| cqUNK_5     | 3.49      | 6237   | 931 | Asn | nd    | 2701–2763   | 12     | 5.3    | 1.28    |       |
|             |           |        |     |     |      | 2933–3723   | 164    | 4.8    | 1.97    | 14.2  |
|             |           |        |     |     |      | 5366–5398   | 17     | 1.9    | 1.94    |       |
| cqUNK_6     | 3.403     | 1629   | 125 | Pro | nd    | 323–375     | 16     | 3.3    | 1.50    |       |
|             |           |        |     |     |      | 536–1337    | 164    | 4.9    | 1.97    |       |
|             |           |        |     |     |      | 2086–2447   | 129    | 2.8    | 1.75    | 94.9  |
|             |           |        |     |     |      | 4245–4281   | 18     | 2.1    | 1.50    |       |
|             |           |        |     |     |      | 7914–7960   | 16     | 3.0    | 1.86    |       |
| cqUNK_7     | 3.506     | 2267   | 182 | Arg | ggtgc | 608–1551    | 117    | 8.1    | 1.96    | 41.6  |
| cqUNK_9     | 3.65      | 4795   | 193 | Ser | gatc  | 3252–3314   | 2      | 31.5   | 1.10    |       |
| cqUNK_10    | 3.176     | 3540   | 194 | Ser | agaag | 1278–1780   | 144    | 4.5    | 1.91    | 14.2  |
| cqUNK_11    | 3.710     | 2554   | 170 | Arg | catt  | 1040–2015   | 298    | 3.3    | 2.00    | 38.2  |
| cqUNK_12    | 3.654     | 5450   | 280 | Leu | acaag | 3411–4140   | 88     | 8.3    | 1.94    | 20.1  |
|             |           |        |     |     |      | 4176–4541   | 50     | 7.3    | 1.95    |       |
|             |           |        |     |     |      | 3114–3334   | 51     | 4.3    | 1.96    | 14.2  |
|             |           |        |     |     |      | 3345–3646   | 48     | 6.3    | 1.92    |       |
| cqUNK_19    | 3.258     | 3639   | 360/337 | Tyr | nd   | 1481–1731   | 82     | 3.0    | 1.93    | 5.2   |
|             |           |        |     |     |      | 1226–1561   | 154    | 2.2    | 1.95    |       |
| cqUNK_20    | 3.707     | 6208   | 519 | Ser | actg  | 2130–2843   | 39     | 18.2   | 1.97    | 11.5  |
| cqUNK_21    | 3.144     | 5520   | 269 | Ser | acgac | 642–734     | 44     | 2.1    | 1.80    |       |
|             |           |        |     |     |      | 1286–1714   | 114    | 3.7    | 1.96    | 20.2  |
|             |           |        |     |     |      | 1960–2554   | 88     | 6.8    | 1.92    |       |
| cqUNK_22    | 3.589     | 8814   | 182 | Ser | tact  | 1373–3671   | 164    | 14.0   | 1.97    | 40.1  |
|             |           |        |     |     |      | 5792–6779   | 164    | 6.0    | 1.97    |       |
|             |           |        |     |     |      | 7247–7500   | 27     | 9.7    | 1.91    |       |
|             |           |        |     |     |      | 7526–7852   | 160    | 2.0    | 1.80    |       |
| cqUNK_23    | 3.1311    | 1909   | 193 | Arg | gtaac | 1003–1633   | 76     | 8.3    | 1.98    | 33.0  |
| cqUNK_24    | 3.393     | 4023   | 185 | Ser | ttcat | 401–444     | 9      | 5.1    | 1.40    | 9.2   |
|             |           |        |     |     |      | 3370–3696   | 41     | 8.0    | 1.84    |       |
| cqUNK_25    | 3.2077    | 2965   | 276 | Met | ttggg | 1389–1758   | 68     | 5.4    | 1.98    | 12.4  |
| cqUNK_26    | 3.220     | 3444   | 337 | Ser | cagcc | 593–2143    | 150    | 10.3   | 1.89    | 51.4  |
|             |           |        |     |     |      | 2332–2552   | 73     | 3.0    | 1.64    |       |
| cqUNK_27    | 3.124     | 4080   | 219 | Arg | gcctt | 1118–2136   | 217    | 4.7    | 1.98    | 31.1  |
|             |           |        |     |     |      | 2160–2411   | 64     | 3.9    | 1.97    |       |
| cqUNK_28    | 3.537     | 5520   | 197 | Arg | cacc  | 942–1803    | 72     | 12.0   | 1.99    | 15.6  |
| cqUNK_29    | 3.172     | 3300   | 176 | Arg | caagc | 968–1807    | 78     | 10.7   | 1.97    | 25.4  |
| cqUNK_31    | 3.1198    | 3124   | 371 | Ser | gtcac | 1001–1586   | 159    | 3.7    | 1.93    | 31.4  |
Until now several criteria for the identification of transposable elements have been successfully applied in sequenced genomes. As for the prediction of protein coding genes, two different approaches can be considered for predicting sequences related to transposable element: intrinsic and extrinsic methods. Intrinsic methods allow the identification of transposable elements through identification of genomic sequences having structural properties typical of mobile genetic elements. In contrast extrinsic methods are based on the identification of transposable elements by sequence similarity. It is evident that the latter methods rely on the use of a known transposable element’s sequence as query sequence. This is the case for 38 LTR retrotransposon sequences identified in this study that contain tandemly repeated sequences. This constitutes the main limitation of these methods, which makes difficult the identification of novel elements with low sequence similarity respect to the queries. This problem is overcome by the use of intrinsic methods, which look for structures rather than sequence similarity. LTR_STRUC is a program designed for the identification of LTR-retrotransposons [21]. It has been successfully used to identify LTR retrotransposons in mammalian [42] as well as in insect genomes [21]. It has been successfully used to identify LTR retrotransposons in mammalian [42] as well as in insect genomes [21].

The results obtained integrate the considerably large amount of data existing for mosquitoes’ genomes. Indeed, our analyses have uncovered the existence of an additional fraction of the C. quinquefasciatus genome related to LTR retrotransposons. This fraction accounts for the 0.8% of the genome occupied by only 29 out of the 67 LTR retrotransposon families detected in this study. In fact, if the non-autonomous elements were also taken in account then this value would have been considerably greater (about 8%).

Our results suggest that a number of LTR retrotransposons insertions could contribute to the built the exon-intron structure of genes in Culex quinquefasciatus. Standing to the predicted exon-intron structures of genes in Culex some of the insertions detected could potentially give a contribution in term of exons or parts of them, to the mature form of mRNA expressed from endogenous genes, underlining the importance of retrotransposons and, in general, of mobile elements in shaping the eukaryotic genomes. This aspect could be particularly important for organisms of social general, of mobile elements in shaping the eukaryotic genomes. This aspect could be particularly important for organisms of social relevance, like C. quinquefasciatus, because polymorphic TE insertion sites can be at the basis of the resistance emergence that characterize some populations [44]. However we were not able to find ESTs in support of this hypothesis, as well as no homologous genes in related species, such as Aedes aegypti, contain retrotransposon related sequences.

Among the novel element identified the vast majority can be classified using conventional criteria, such as combination of phylogenetic clustering and structural features. Unfortunately, these criteria are not sufficient to classify elements lacking coding sequences. This is the case for 38 LTR retrotransposon sequences identified in this study that contain tandemly repeated sequences between LTRs.

Table 2. Cont.

| Element   | supercont | length | LTR | PBS | TSD | Rep Position | Period | Copies | Entropy | %     |
|-----------|-----------|--------|-----|-----|-----|--------------|--------|--------|---------|-------|
| cqUNK_32  | 3.496     | 5091   | 318 | Ser | nd  | 1633–2047    | 40     | 11.3   | 1.76    | 13.4  |
| cqUNK_33  | 3.1048    | 5047   | 694 | Tyr | nd  | 835–1520     | 62     | 11.0   | 1.92    | 6.0   |
| cqUNK_35  | 3.492     | 6910   | 224 | Gln | nd  | 915–1217     | 167    | 1.8    | 2.00    | 39.6  |
| cqUNK_37  | 3.343     | 5389   | 189 | Met | actgg | 2243–5096   | 44     | 67.3   | 1.95    | 3.5   |
| cqUNK_38  | 3.1148    | 7609   | 246 | Arg | ggtat | 558–812     | 74     | 3.4    | 1.94    | 7.5   |
| cqUNK_39  | 3.820     | 5080   | 573/581 | Tyr | tggag | 914–1124   | 31     | 6.8    | 1.88    | 1.4   |
| cqUNK_41  | 3.773     | 10955  | 223 | Ala | gtggt | 1200–1303  | 31     | 3.4    | 1.88    | 17.0  |
| cqUNK_42  | 3.7       | 7544   | 208 | Ser | ctatt | 475–5140   | 45     | 7.6    | 1.91    | 1.4   |
| cqUNK_43  | 3.2654    | 6161   | 224 | Thr | caagg | 9068–9146  | 39     | 1.9    | 1.64    | 17.0  |
| cqUNK_45  | 3.590     | 3246   | 315 | Gln | nd  | 9288–9395   | 48     | 2.2    | 1.65    | 1.4   |
|           |           |        |     |     |     | 9464–9533   | 42     | 2.1    | 1.49    | 1.4   |
|           |           |        |     |     |     | 691–1175    | 127    | 3.9    | 1.90    | 6.4   |
|           |           |        |     |     |     | 1199–3865   | 46     | 58.1   | 1.93    | 43.3  |
|           |           |        |     |     |     | 2382–2901   | 278    | 1.9    | 1.99    | 16.0  |

For each non-autonomous element is reported the supercontig in which a representative element can be found, the overall length, the LTR size, the tRNA complementary to the PBS. It is also indicated the position, the period and the copies of the repeated DNA contained in the elements listed. The entropy value gives an estimation of the complexity of the repeats (see main text). The portion occupied by repeats in terms of % of the total size of the element is also indicated (column %). doi:10.1371/journal.pone.0030770.t002
Table 3. The contribution of LTR-retrotransposons to *C. quinquefasciatus* gene organization.

| Element | Interaction | Description               | GENE ID  | Supercont:position |
|---------|-------------|---------------------------|----------|--------------------|
| Cqgypsy_2 | Within intron | Dual specificity tyrosine-phosphorylation-regulated kinase | CPU004687 | 3.72: 556,787–583,090 |
|         | Exon-Intron junction | 5′-3′ exoribonuclease, putative | CPU016423 | 3.74: 154,787–166,602 |
|         | 1–2 Kbp upstream | fimbrin/plastin | CPU004008 | 3.57: 387,682–393,342 |
|         | 2–3 Kbp downstream | allergen, putative | CPU018993 | 3.1504: 58,993–71,374 |
| Cqgypsy_3 | 0–1 Kbp downstream | Adenylyltransferase and sulfurrtransferase MOC53 | CPU016212 | 3.19: 246,047–247,568 |
| Cqgypsy_5 | 1–2 Kbp downstream | chaperonin | CPU013429 | 3.475: 239,574–242,207 |
|         | Exon-Intron junction | 40 S ribosomal protein S2 | CPU012693 | 3.480: 10,936–19,727 |
|         | 4–5 Kbp upstream | serine threonine-protein kinase | CPU018896 | 3.1443: 17,857–21,694 |
| Cqgypsy_7 | Within intron | Brahma associated protein 170 kDa, putative | CPU002241 | 3.30: 656,859–677,280 |
|         | Overlap first exon | ribosomal protein L23a | CPU016489 | 3.385: 52,166–53,573 |
|         | 2–3 Kbp downstream | allergen, putative | CPU018993 | 3.1504: 58,993–71,374 |
|         | 0–1 Kbp downstream | Adenylyltransferase and sulfurrtransferase MOC53 | CPU016212 | 3.19: 246,047–247,568 |
| Cqgypsy_8 | 1–2 Kbp upstream | suppressor of ty3 | CPU014381 | 3.539: 307,272–309,528 |
|         | 1–2 Kbp downstream | suppressor of ty3 | CPU014381 | 3.539: 307,272–309,528 |
|         | exon-intron junction | suppressor of ty3 | CPU014381 | 3.539: 307,272–309,528 |
| Cqgypsy_13 | 0–1 Kbp downstream | transcription factor IIIB 90 kDa subunit | CPU008270 | 3.167: 169,563–182,965 |
| Cqgypsy_15 | Within intron | dystrophin major muscle isoform | CPU013032 | 3.423: 50,593–185,416 |
| Cqgypsy_20 | 0–1 Kbp downstream | histone-lysine n-methyltransferase | CPU000732 | 3.1443: 17,857–21,694 |
| Cqgypsy_21 | 0–1 Kbp upstream | flotillin-2 | CPU007626 | 3.148: 169,798–180,783 |
| Cqgypsy_25 | Exon-Intron junction | phd finger protein | CPU014131 | 3.545: 75,916–92,572 |
| Cqgypsy_29 | 3–4 Kbp downstream | protein phosphatase-1 | CPU008212 | 3.168: 687,712–708,602 |
|         | 0–1 Kbp downstream | helicase | CPU019431 | 3.1585: 41,788–51,610 |
| Cqgypsy_37 | 1–2 Kbp downstream | sphingomyelin synthetase | CPU002233 | 3.30: 341,656–342,335 |
|         | 2–3 Kbp downstream | DEAD box ATP-dependent RNA helicase | CPU006204 | 3.118: 567,640–586,793 |
| Cqgypsy_47 | 2–3 Kbp downstream | sphingomyelin synthetase | CPU002233 | 3.30: 341,656–342,335 |
|         | 4–5 Kbp upstream | serine protease inhibitor, serpin | CPU012013 | 3.346: 385,437–397,533 |
|         | 3–4 Kbp downstream | pre-mrna splicing factor prp17 | CPU011807 | 3.365: 424,340–426,137 |
|         | 1–2 Kbp upstream | uridine cytidine kinase i | CPU016204 | 3.736: 44,070–45,560 |
|         | 1–2 Kbp downstream | uridine cytidine kinase i | CPU016204 | 3.736: 44,070–45,560 |
|         | 1–2 Kbp downstream | coatomer | CPU014834 | 3.606: 201,807–202,272 |
|         | 1–2 Kbp downstream | poly A polymerase | CPU014835 | 3.606: 201,807–202,272 |
| Cqgypsy_51 | 1–2 Kbp upstream | zinc finger protein | CPU009854 | 3.243: 201,807–202,272 |
|         | 2–3 Kbp upstream | DNA replication licensing factor MCM7 | CPU009855 | 3.243: 201,807–202,272 |
|         | 2–3 Kbp downstream | mitochondrial 39 S ribosomal protein L3 | CPU017407 | 3.941: 86,591–87,964 |
|         | 1–2 Kbp downstream | 26 S protease regulatory subunit 6a | CPU017405 | 3.941: 78,697–79,159 |
| Cqgypsy_53 | 3–4 Kbp downstream | esterase B1 precursor | CPU016336 | 3.777: 170,027–172,027 |
|         | 4–5 Kbp upstream | ATP synthase D chain, mitochondrial | CPU011691 | 3.328: 179,980–180,887 |
|         | 2–3 Kbp downstream | Eftud2 protein, putative | CPU000064 | 3.1: 1,221,757–1,228,227 |
|         | 1–2 Kbp downstream | cell division protein kinase 5 | CPU000065 | 3.1: 1,233,232–1,234,222 |
|         | Exon-Intron junction | sarcomembral associated protein-2, putative | CPU013133 | 3.310: 56,964–69,006 |
| Cqgypsy_56 | Exon-Intron junction | semaphorin | CPU001593 | 3.17: 1,304,933–1,339,494 |
|         | Exon-Intron junction | microbrillar-associated protein, putative | CPU020039 | 3.2342: 8,606–7,972 |
|         | Exon-Intron junction | male-specific doublesex protein | CPU004057 | 3.59: 681,384–685,772 |
|         | 1–2 Kbp downstream | polypeptide of 976aa, putative | CPU018525 | 3.1222: 1,304,933–1,339,494 |
| Cqgypsy_59 | 1 Kbp upstream | negative elongation factor E | CPU000025 | 3.1: 727,264–728,238 |
|         | 4–5 Kbp downstream | semaphorin | CPU000027 | 3.1: 744,915–761,147 |
|         | 4–5 Kbp downstream | superoxide dismutase, putative | CPU005173 | 3.91: 822,776–823,755 |
|         | Within intron | enhancer of polycomb | CPU018246 | 3.1131: 70,175–79,159 |
| Cqgypsy_60 | Overlaps last exon | serine protease inhibitors putative | CPU007021 | 3.133: 747,858–748,701 |
|         | 0–1 Kbp upstream | serine protease inhibitor | CPU007023 | 3.133: 758,924–759,336 |
|         | 1 Kbp downstream | nucleoporin | CPU013031 | 3.426: 362,618–363,727 |
Novel Mosquitoes’ Retrotransposons

Non-autonomous elements lacking ORFs have been well documented especially in plant genomes [45]. Typically, these elements lack all coding sequences but have retained the LTRs, the primer-binding site and the polyurinic tract. These are the minimal features required for replication, because the LTRs contain the promoter needed to produce a template RNA, and the primer-binding site and the polyurinic tract are needed to prime the reverse transcription steps. They are extremely heterogeneous in size varying from few hundreds base pairs (TRIM retrotransposons [46] to few Kilobase pairs (LARDS retrotransposons [47]). In mammalian genomes MaLRs retrotransposons (Mammalian Apparent LTR Retrotransposons) [48] have also been described with similar features.

Very interestingly, Arensburger et al. [13] have detected a single element resembling in structure a LARDS retrotransposon in the genome of C. quinquefasciatus. In this paper we have also reported the identification of C. quinquefasciatus contains LTR-retroelements with peculiar features. This was also evident from previous works, which have demonstrated the presence of the Twin elements in this genome [13] and have allowed the identification of Osvaldo-like elements with a non-canonical structure of the LTRs [17]. In this paper we have also reported the identification of cgypsy_1, an Osvaldo-like element with an atypical PBS with a tRNA-dimer structure. The tRNA-dimer is somehow reminiscent of the target site duplication occurring upon integrase-mediated insertion, suggesting that the tRNA-dimer has been integrated by a transpositional mechanism.

As can be observed from figure 3A both tRNA like sequence halves have the terminal CCA sequences. This structural feature would suggest that the mature form of endogenous tRNA molecules have been incorporated into the retrotransposon backbone after a reverse transcription process. As far as we know, dimerization or aggregation of tRNAs in vivo is a known phenomenon, but it typically occurs under non-physiological conditions [51] [52]. On the other hand differences can be highlighted between Twin elements and the head to head tRNA repeat found in cgypsy_1. The target site duplication, where it was found, of Twin is an AT rich sequence. A poly-A tract derived from the retroposition event is located downstream Twin elements. These features are absent in the Twin-like structure that we have detected, suggesting a different origin of the insertion detected in cgypsy_1 element.

In conclusion, findings from this and previous reports make C. quinquefasciatus a potential niche-genome in which the evolution of transposable elements occurs and generates strong genomic diversity.

The importance of studying the mosquito’s mobilome also resides in the possibility to use such DNA sequences as molecular biomarkers [53], or for the control of insecticide resistance populations in order to contrast the spread of virus borne diseases

Table 3. Cont.

| Element   | Interaction       | Description                              | GENE ID         | Supercont:position |
|-----------|-------------------|------------------------------------------|-----------------|--------------------|
| Cagypsy_61| Overlaps last exon| cytochrome c oxidase subunit I            | CPU016836       | 3.816: 17,796–33,791|
| Cagpsy_63 | Within intron     | Dual specificity tyrosine-phosphorylation-regulated kinase | CPU004687       | 3.72: 556,787–583,090|
|           | Exon-Intron junction | 5'-3' exoribonuclease, putative          | CPU016423       | 3.746: 154,787–166,602|
| Cagpsy_64 | exon-intron junction | alanyl-glyoxylate aminotransferase      | CPU06409        | 3.128: 202,128–218,875|
|           | Exon-Intron junction | arsenite inducible RNA associated protein aip-1 | CPU005006      | 3.82: 520,914–524,059|
|           | exon-intron junction | nk homeobox protein                      | CPU019260       | 3.1511: 40,639–44,852|
|           | 2–3 Kbp downstream | 60 S ribosomal protein L7               | CPU017548       | 3.961: 25,320–28,131|

For each insertions detected in proximity (+/- 5 Kbp) or into genes are reported the kind of interaction (upstream, downstream, exon, intron), the Vectorbase identifier of the gene, its description and its position in the supercontig.

doi:10.1371/journal.pone.0030770.t003
In this view our results could be helpful for future studies concerning such topics.

Supporting Information

File S1 DNA sequences of the 67 LTR-retrotransposons identified in this paper. Each sequence contains 50 bases upstream and downstream allowing unique identification of a reference copy in the genome of *C. quinquefasciatus*.

File S2 Multiple alignment file used to obtain the phylogenetic tree in figure 1.

Table 4. Contribution of the non-autonomous elements identified in this paper to the formation of mature mRNAs of *C. quinquefasciatus* genes.

| Element | Description | gene ID | Supercontig:position |
|---------|-------------|---------|----------------------|
| CqUNK_3 | ATP-dependent RNA helicase DHX8 | CPUI011263 | 3.322: 64334–68170 |
| CqUNK_3 | cell cycle control protein ccw6 | CPUI011261 | 3.322: 54937–56710 |
| CqUNK_3 | tRNA methyltransferase | CPUI011262 | 3.322: 62963–64258 |
| CqUNK_3 | carboxylesterase-6 | CPUI006908 | 3.137: 13520–18988 |
| CqUNK_3 | bombesin receptor subtype-3 | CPUI017637 | 3.980: 85049–96960 |
| CqUNK_3 | N-acetylgalactosaminyltransferase 7 | CPUI014647 | 3.660: 5557–7406 |
| CqUNK_5 | saposin | CPUI014133 | 3.545: 130173–138797 |
| CqUNK_9 | bombesin receptor subtype-3 | CPUI017637 | 3.980: 85049–96960 |
| CqUNK_11 | dopamine beta hydroxylase | CPUI019622 | 3.1797: 2222–17524 |
| CqUNK_11 | midasin | CPUI010145 | 3.251: 518823–535901 |
| CqUNK_11 | sterol desaturase, putative | CPUI009637 | 3.227: 524825–527033 |
| CqUNK_19 | malate dehydrogenase | CPUI015299 | 3.661: 78722–91118 |
| CqUNK_19 | igf2 mRNA binding protein, putative | CPUI011349 | 3.312: 412154–436319 |
| CqUNK_20 | regulator of chromosome condensation | CPUI019976 | 3.2094: 4450–9107 |
| CqUNK_23 | midasin | CPUI010145 | 3.251: 518823–535901 |
| CqUNK_23 | centaurin-alpha 2 | CPUI019112 | 3.1516: 939–9235 |
| CqUNK_23 | sterol desaturase, putative | CPUI009637 | 3.227: 524825–527033 |
| CqUNK_28 | zinc finger protein 40 | CPUI018875 | 3.1321: 31848–32423 |
| CqUNK_28 | allatostatin receptor | CPUI016163 | 3.734: 81478–87202 |
| CqUNK_28 | aldehyde oxidase 2 | CPUI016888 | 3.821: 149733–154740 |
| CqUNK_28 | bombesin receptor subtype-3 | CPUI017637 | 3.980: 85049–96960 |
| CqUNK_28 | defective proboscis extension response, putative | CPUI017715 | 3.897: 12891–19932 |
| CqUNK_32 | 40 S ribosomal protein S14 | CPUI010397 | 3.291: 13809–144395 |
| CqUNK_32 | choline O-acetyltransferase | CPUI001609 | 3.19: 128455–132041 |
| CqUNK_33 | phosphatidylinositol-4-phosphate 5-kinase type i | CPUI006826 | 3.145: 280313–292406 |
| CqUNK_42 | Transcription factor Ken 2 | CPUI012629 | 3.427: 256203–270029 |
| CqUNK_42 | laminin gamma-3 chain | CPUI005194 | 3.96: 903006–924273 |
| CqUNK_42 | trypsin | CPUI004641 | 3.70: 338878–341919 |
| CqUNK_42 | f-actin capping protein alpha | CPUI01271 | 3.322: 247301–257388 |
| CqUNK_42 | malate dehydrogenase | CPUI008123 | 3.169: 118442–121449 |
| CqUNK_43 | elongation factor tu | CPUI002277 | 3.30: 1195429–1198055 |
| CqUNK_43 | zinc finger protein | CPUI002883 | 3.27: 1030192–1037437 |
| CqUNK_43 | kakapo | CPUI003239 | 3.41: 533346–586665 |
| CqUNK_45 | monocarboxylate transporter | CPUI008119 | 3.184: 607328–610272 |
| CqUNK_45 | pol-like protein | CPUI018514 | 3.1248: 73743–87450 |
| CqUNK_45 | elongation factor 1 alpha | CPUI009557 | 3.231: 370372–372795 |
| CqUNK_45 | olfactory receptor, putative | CPUI013754 | 3.526: 317670–324857 |

doi:10.1371/journal.pone.0030770.t004

Acknowledgments

We thank anonymous Reviewers and the Editor for their comments, which significantly contributed to improving the quality of the manuscript.

The Fondazione Cassa di Risparmio di Puglia, which granted the project “Genomica degli elementi trasponibili applicata all’identificazione e caratterizzazione molecolare di specie di Culex presenti sul territorio Barese” is gratefully acknowledged.

Author Contributions

Conceived and designed the experiments: RMM. Performed the experiments: RMM DL PD. Analyzed the data: RMM DL PD LV ET RC. Wrote the paper: RMM.
References

1. Kidwell MG, Liech DR (2002) Transposable elements as sources of genomic variation. In: Craig N, Craigie R, Gellert M, Lambowitz A, eds. Mobile DNA. American Society for Microbiology Press. pp 59–90.

2. Wicker T, Sabet F, Haas-Van A, Benetzen JL, Capp P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973–982.

3. Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9: 411–412; author reply 414.

4. Finnegan DJ (1992) Transposable elements. Curr Opin Genet Dev 2: 861–867.

5. Dergun JM, Casacuberta JM, Panaud O (2000) Plant transposable elements. Genome Dyn 4: 69–82.

6. Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics. TGIG 19: 69–72.

7. Medstrand P, van de Lagaesta LN, Dunn CA, Landry JR, Svenbek D, et al. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenetic and Genome Research 110: 342–352.

8. Kumar A, Benetzen JL. (1999) Plant transposons. Annual Review of Genetics 33: 479–532.

9. Kalendr R, Flavell AJ, Ellis TH, Sjakte T, Moiys C, et al. (2011) Analysis of plant diversity with retrotransposon-based molecular markers. Heredity 106: 520–530.

10. Holt RA, Subramanian GM, Halpern S, Sutton GG, Haas B, et al. (2002) Genome sequence of Aedes aegypti, the principal vector for dengue fever. Science 298: 129–149.

11. Nene V, Wortman JR, Lawson D, Halpern A, Subramanian GM, et al. (2007) The genome sequence of the malaria mosquito Anopheles gambiae. Science 317: 1716–1723.

12. Rao PN, Rai KS (1990) Genome evolution in the mosquitoes and other closely related members of superfamsily Culicoidea. Hereditas 113: 139–144.

13. Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, et al. (2010) Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science 330: 86–88.

14. Feschotte C, Zhang X, Wessler SR (2002) Miniature inverted repeat transposable elements and their relationship to established DNA transposons. In: Craig N, Craigie R, Gellert M, Lambowitz A, eds. Mobile DNA. American Society for Microbiology Press. pp 1147–1158.

15. Coy SB, Tu Z (2007) Genetic and evolutionary analysis of Tanger transposons in Aedes aegypti. Anopheles gambiae and other mosquito species. insect Mol Biol 16: 411–421.

16. Biedler JK, Tu Z (2007) The Juan non-LTR retrotransposon in mosquitoes: genomic impact, vertical transmission and indication of recent and widespread activity. BMC Evol Biol 7: 112.

17. Minervini CF, Viggiano L, Caizzi R, Marsano RM (2009) Identification of novel Mosquitoes’ Retrotransposons. PLoS ONE 4: e520–530.

18. Biedler JK, Tu Z (2007) The Juan non-LTR retrotransposon in mosquitoes: genomic impact, vertical transmission and indication of recent and widespread activity. BMC Evol Biol 7: 112.

19. Minervini CF, Viggiano L, Caizzi R, Marsano RM (2009) Identification of novel LTR retrotransposons in the genome of Aedes aegypti. Gene 440: 42–49.

20. Burgatf-Mercklem N, Cagnes C, Desmons I, Salvado JC, Karana S, et al. (1997) CM-gag, a transposable-like element reiterated in the genome of Culex pipiens mosquito, contains only a gag gene. Genetica 100: 141–148.

21. TEFam database. Available: http://tefm.biochem.vt.edu/tefm/ Accessed 2011 Dec 04.

22. Feschotte C, Forrier N, Desmons I, Mouches C (2001) Birth of a retrotransposon: the twin SINE family from the vector mosquito Culex pipiens may have originated from a dimeric tRNA precursor. Molecular and Evolutionary Biology 18: 74–87.

23. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362–367.

24. Ellingham D, Kurtz S, Willhoefi U (2000) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 162.

25. Jurka J, Kapitonov VV, Pavlacka D, Klonowski P, Kohany O, et al. (2003) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 100: 462-467.

26. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 580–582.

27. Smit AFA, Hubley R, Green P (1996) RepeatMasker Open-3.0. Available: http://www.repeatmasker.org

28. Makal HS, Eickbush TH (1999) Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. Journal of Virology 73: 5186–5190.

29. Edgar RA (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.

30. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25: 4876–4882.

31. Syomin BV, Leonova TY, Byim VV (2002) Evidence for horizontal transfer of the LTR retrotransposon mdg3, which lacks an env gene. Molecular Genetics and Genomics 267: 418–423.

32. Michaila J, Mathiasan S, Guillard J, Garel A (1999) The complete sequence of a new retrotransposon in Bombyx mori. Nucleic Acids Research 18: 674.

33. Tubio JM, Naveira H, Costas J (2005) Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of Anopheles gambiae. Molecular Biology and Evolution 22: 29–37.

34. Maliek HS, Henkel D, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertible retroviruses. Genome Research 10: 1307–1318.

35. Ganko EW, Fielden KT, McDonald JF (2001) Evolutionary history of CER elements and their impact on the C. elegans genome. Genome Research 11: 2006–2017.

36. Fainzilber A, Labrador M, Fontdevila A (1999) The retrotransposon Osvaldo from Drosophila buzzatii displays all structural features of a functional retrovirus. Molecular Biology and Evolution 16: 909–921.

37. Kazarzin HH, Jr. (2004) Mobile elements: drivers of genome evolution. Science 303: 1626–1632.

38. Minervini CF, Marsano RM, Casperi P, Fanti L, Caizzi R, et al. (2007) Heterochromatin protein 1 interacts with 5’LTR of transposable element ZAM in a sequence-specific fashion. Gene 393: 1–10.

39. Sapan C, Harrison DA, Corces VG (1983) The Drosophila melanogaster suppressor of hairy-wing protein binds to specific sequences of the gypsy retrotransposon. Genes & Development 2: 1144–1143.

40. Conte C, Dastugue B, Vaury C (2002) Coupling of enhancer and insulator properties identified in two retrotransposons modulates their mutagenic impact on nearby genes. Molecular and Cellular Biology 22: 1767–1777.

41. Gohula DA, Gerasimova TV, Corces VG (1996) Genetic and molecular analysis of the gypsy chromatin insulator of Drosophila. Proceedings of the National Academy of Sciences of the United States of America 93: 9378–9383.

42. McCarty EM, McDonald JF (2004) Long terminal repeat retrotransposons of Mus musculus. Genome Biology 5: R14.

43. Marsano RM, Caizzi R (2005) A genome-wide screening of BEL-Pao like retrotransposons in Anopheles gambiae by the LTR_STRUC program. Gene 357: 115–121.

44. Darboux I, Charles JF, Pauchet Y, Watot S, Paron D (2007) Transposon-mediated resistance to Bacillus subtilis in a field-evolved population of Culex pipiens (Diptera: Culicidae). Cellular Microbiology 9: 2022–2029.

45. Saltot F, Schulman A (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker’s guide to the genome. Heredity 97: 381–388.

46. Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proceedings of the National Academy of Sciences of the United States of America 98: 13778–13783.

47. Kalendr R, Vicient CPM, Peleg O, Anamthawat-Jonsson K, Bolhoy A, et al. (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166: 1437–1450.

48. Smit AF (1995) Identification of a new, abundant superfamily of mammalian transposable elements implemented in Repbase. Nat Rev Genet 9: 411–412.

49. Novel Mosquitoes’ Retrotransposons