Impact of transposable elements on the genome of the urban malaria vector *Anopheles coluzzii*
Carlos Vargas-Chavez et al

GENERAL COMMENTS

In this work, authors have analyzed the TE content of the malaria vector *An. coluzzii*, a sister species to the well characterized species *An. gambiae.* Their work has one major originality: the use of long-read sequencing strategies that allow to span full TE insertions. Authors have used these data to annotate *An. coluzzii* TEs, including the discovery of new TE families, have studied their distribution throughout the genome, and shown their presence at chromosomal inversion breakpoints. They have also studied TEs associations with genic regions as well as the presence of transcription factor binding sites and promoters in some TEs, and shown that some insertions were associated with insecticide-associated or -responsive genes, or with genes putatively involved in immunity.

While this work has thus provided a wealth of data that are potentially very useful and open roads for further studies, I fear that the manuscript contains too many flaws to be recommended as it stands for a support by PCI Genomics.

My major criticism is that authors constantly overstate the meaning of their results. The work is presented from the start as an analysis of the role of TEs in *An. coluzzii* adaptative processes and of their impact on genic functions, and sentence like "To better understand the role of TEs in rapid urban adaptation, we sequenced..." as well as "we found that TEs have an impact in ...the regulation of functionally relevant genes", are even found in the Abstract.
Yet nothing in the results provides any hints on the role of TEs in rapid urban adaptation, nor demonstrate any impact on gene regulation.

More specifically:
1) Authors have sequenced different larvae from two sampling sites, and interpret differences as true genomic differences and TE variability throughout the manuscript. For instance, authors have found large differences in the gypsy content. However authors have not evaluated whether these differences could simply be due to variabilities during the sequencing process (such as the large coverage differences between samples or other technical aspects). Some sequencing replicates for individual larva samples should have been performed to adress this issue, but are mentioned nowhere. In the absence of such replicates, these inter-sample differences in global TE contents or in specific superfamilies have little meaning.
More generally, all inter-sample quantitative aspects of the results, including TE insertions distribution throughout the genomes should be considered very cautiously. In the Discussion, authors merely mention the possibility of heterochromatic TE differences resulting from of possible differences in the quality of the genome assembly, but discard this possibility for euchromatic TEs, for no clear reason.
Exemples of TE variations identified between sequenced genomes are provided as references, but many of those variations were actually analyzed *via* split-reads strategies, allowing to securely determine that insertions are truly absent or present from determined genomic positions.
Similarly, what is the point of analyzing the TE landscape separately for each larva ?

2) Furthermore, the study of TEs association with genic regions, and the presence of transcription factor binding sites and promoters in these elements, is merely descriptive and rather anecdotical, and does not really provide any evidence of any impact on genes. A true first step towards such evidence would have been the demonstration of an enrichment of TEs and their TFBS/promoter regions in insecticide-associated or immunity genes, which is not the case in this paper. Only the presence of TEs near such genes is shown, but these TEs could also be present or even enriched near other gene classes. Such larger scale enrichment analysis could be done, as authors have the genomic data.

In summary, the characterization of TE families from long-read sequencing is quite interesting, authors mostly need to water down excessive claims, make a better use of their data and reorientate their paper.

In addition, I have noted a number of other flaws that will need adressing (some more technical comments at the bottom).

OTHER COMMENTS

3) Most primary Figures are very poor quality (letterings are fuzzy and barely readable), better quality figures should be uploaded in the manuscript. Supplementary Figures are OK.

4) The material sampling is quite confusing. As I understand from Material and Methods, in addition to a single larva (LVB11) used for PacBio, six larvae from each of the five other breeding sites were used for Nanopore sequencing.
- Why then mention that 25 larvae were initially sampled at each site if only six (or one in the case of LVB11), plus one used for complementary Illumina, were actually used from each sampling site?

- line 535, were the 5 larvae from each site barcoded separately? This would agree with the fact that DNA was extracted separately from each larva. Yet results are only presented per sampling sites. Or were the 5 larvae DNA pooled, with each sampling site barcoded for pooled Nanopore sequencing (in this case, why extract DNA separately from each larvae ?)
Clearly some clarity is needed here.

5) Some explanation on the PCR used to identify bona fide *An coluzzi* is required (I understand it is based in the presence/absence of a given SINE insertion, is that right ?)

6) Between 172 and 294 TE families are detected within each genome, but clustering TE libraries from several genomes leads to 435 TE families. This raises many questions and should be better discussed. What is exactly the definition of a TE family in this work? Were families defined as in the Wicker's classification (80-80-80) ? Could additional families be merely slightly divergent versions?

7) Some major data discrepancies are found between Figure 2 and Figure S1: both contain the two new TE families AcoI_LTR_Ele3 and AcoI_LTR_Ele4, but plot coverages differ strongly, as well as species distribution (as far as I can tell, since species abbreviations are not readable on Figure 2). Please check again that all TE identifications and corresponding data are OK.

8) Lines 328-329: what is exactly the meaning of "...the number of genomes where the gene was correctly transferred."? From Table 1 the vast majority of genes (>95% except DLA155B) were detected in the genome assemblies. Yet from Table 3 it seems that the vast majority (all but AGAP002633) of the TE-targeted genes were not detected in all seven assemblies. This actually strenghten the hypothesis that most of the variability between samples is due to technical artefacts.

9) Abundant new TRIM families - Plot coverage patterns in Figure 2 suggest a higher abundance of LTRs versus full copies. Have full length autonomous version of these TRIMs, that would share similar LTRs, been found ? Alternatively, do these plot coverage patterns indicate an abundance of solo-LTRs ?

10) Table S13, Promoter motifs in TE insertions: what is the meaning of "promoter motifs at the correct distance". The correct distance for expression of the gene? Retrotransposons are supposed to contain promoters (and TFBS) for their own expression.
More globally, it is very likely that TEs that are NOT in close proximity to genes also contain various TFBS and promoters.

11) Identification of active families: a more straighforward way of identifying active retrotransposons and their relative age would be to evaluate the divergence between LTRs of the same copy.

12) Lines 353-366: It is unclear why authors focused only in intronic or upstream insertions, as downstream insertions can also affect gene expression, notably RNA stability.

13) Figure S5: out of the 43 genes analyzed, only 23 contained at least one insertion:
- what is the point of showing diagrams for the 20 genes that do not contain any insertion? This is quite superfluous and non informative.
- which one of those genes were differentially expressed in *An. gambiae* when exposed to insecticides ? As I understand, some of the genes are insecticide resistance genes, some are differentially expressed when exposed to insecticides, but are not necessarily insecticide resistance genes.
- this differential expression was described in *An. gambiae*. Was it correlated to any TE insertion in *An. gambiae* ?

OTHER MORE TECHNICAL POINTS

14) Line 299: Additional file 1: Table S ? (Table number is missing)

15) Additional file 2 (Figure S1) is difficult to analyze, especially in C diagrams (TE distributions in other species): species names are indicated only below a few diagrams, and only in an abbreviated form that is nowhere explained in the manuscript. I would suggest to add species abbreviations to each diagram, and to add full species names somewhere in the legend.
Same remark for Figure 2

16) Supplementary Table S12 contain color highlights that are not explained

17) Table S15: "columns with the genome names can be TRUE or FALSE if the insertion is present or not" is quite obscure, replacing TRUE/FALSE by YES/NO would be more clear.

18) Lines 428-429: "four TRIM families previously undescribed in anopheline genomes that are likely to be important players in genome evolution". While it is known that TRIMs are often important players in genome evolution, this particular sentence suggests that this has been demonstrated for the four TRIM families identified here. This is not the case. This turn of phrase needs to be corrected.

19) The number of genes detected in An coluzzii samples is provided only in Table S1 (and thus hard to find) and only a completeness ration is provided in the manuscript core (Table 1). Gene numbers should be added to Table 1.

20) Lines 265-266: It is not very clear what authors mean by "our genomes have the standard conformation for all five inversions" on lines 265-266. It seems to me that from Table S8 some breakpoints are missing from several samples?