Brief Communication

A whole genome gene content phylogenetic analysis of anopheline mosquitoes.

Jeffrey A. Rosenfeld^{1,2}, Sara Oppenheim¹ and Rob DeSalle¹

1. Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024
2. Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ, USA

Abstract

Construction of stringent gene content matrices was accomplished for 21 Anopheline mosquito species and strains and four outgroups species. The presence absence matrix using e^{-75} as a cutoff in single linkage clustering had over 17,000 ortholog groups. We used the gene content matrix to generate a phylogenetic hypothesis that is in general agreement with gene sequence based phylogenies. In addition to establishing a congruent gene content phylogeny we examined the consistency of three methods for analyzing presence absence data – unweighted parsimony, dollo parsimony and maximum likelihood using a BINGAMMA model. An examination of the chromosomal location of the gains and losses in the presence absence matrix revealed a low frequency of gains and losses at centromeres and tips of chromosomes.

The availability of whole genomes for a wide array of organisms allows for the establishment of ortholog groups whose presence and absence can be used to infer phylogenetic relationships (Snel and Huynen, 1999; Snel et al., 2005). While at deep phylogenetic levels the approach has some problems (Rosenfeld and DeSalle, 2012; Rosenfeld et al., 2016), at more fine grained levels, such as within the family Drosophilidae, gene content phylogenetic analysis recovers

branching orders relatively accurately (Rosenfeld et al., 2008; Rosenfeld and DeSalle, 2012). In addition, gene content analysis can also allow for the quantification of patterns of gene gains and losses over evolutionary time. A recent sequencing analysis of several species of anopheline mosquitoes (Neafsey et al., 2015) allows for the application of this approach to a recently diverged and medically important insect group, and lends an opportunity to corroborate sequence based phylogenetic hypotheses.

Genome sequences for *Drosophila melanogaster* were obtained from FlyBase (<https://www.flybase.org>) (Attrill et al., 2016), sequences for all other species were obtained from VectorBase (<https://www.vectorbase.org>; Giraldo-Calderón et al., 2014). Presence/absence matrices were generated using a range of e-value cutoffs as in Rosenfeld and DeSalle (2012). In this approach, the protein sequences from all of the species included in the study are subjected to reciprocal all-against-all BLASTP searches. The BLAST hits are then filtered iteratively based upon e-value (specifically e^{-5} , e^{-25} , e^{-50} , e^{-75} , e^{-100} , e^{-125} , e^{-150} , e^{-175} and e^{-200}). The resulting matches are single-linkage clustered (Lienau et al., 2006). When a clustered gene had multiple closely related sequences we simply scored the gene as present and ignored the duplicates. In addition, any gene or ortholog group not present in only one taxon or absent in only one taxon was removed from the analysis as phylogenetically uninformative. This procedure resulted in nine presence absence matrices, one for each e-value iteration. Because the quality of assemblies and annotations of the species in this data set (Neafsey et al., 2015) are of concern, we suggest that exploring this problem through exploring stringent e-value cutoffs in the single linkage clustering approach, overcomes some of the problems with the annotations. Matrices are available as supplemental information (Supplemental File 1). We first compared the character information for the different e-value matrices. Supplemental Figure 1 shows a comparison of the counts of ortholog groups as determined by the single linkage clustering method we used (Lienau et al., 2006). This analysis suggests that the e^{-50} and e^{-75} counts give the maximum number of informative characters for all species and strains. We have also graphed the actual gene number counts as determined from their whole genome sequences for these species in this analysis. The gene count numbers indicate differences in

number of genes (from 12,000 genes in some anophelines to 18,000 in our outgroup species) across the species of this group (Supplemental Figure 1), while the ortholog group number stays somewhat constant at around 9500. There are some exceptions to this general statement as the following species and strains dip down to around 7000 ortholog groups: *darlingi*, *albimanus*, *dirus*, *atroparvus* and *maculatus*. The outgroups (*Aedes*, *Culex*, *Drosophila* and *Glossina*) all show low numbers of ortholog groups when placed into the context of anopheline genomes.

We next implemented three optimality criteria to generate phylogenetic hypotheses using the presence/absence matrices: unweighted parsimony (uMP), dollo parsimony (dMP) and maximum likelihood (ML). MP trees were generated in PAUP* (Swofford, 2001) using unweighted and dollo weighting, and ML trees were generated using the BINGAMMA model in RAxML (Stamatakis, 2014). All trees were generated using bootstrapping and only nodes with 100% bootstrap support were retained in the comparisons we discuss below. The phylogenetic hypotheses we generated with the various criteria are summarized in Supplemental Figure 2 where we show the phylogenetic results for the matrix generated with e^{-75} as the cutoff and the different phylogenetic analyses mentioned above. We also include a modified tree based on the sequence analysis in Neafsey et al., (2015). While the relationships within the *gambiae* group is controversial (Fontaine et al., 2015), we used the topology in Supplemental Figure 2 to assess congruence of the phylogenetic hypotheses using the various methods, the three trees we generated using uMP, dMP, and ML approaches and the modified “true” tree from Neafsey et al., (2015) were compared using the RhoI consistency index. Figure 1 shows the results of this analysis. In general, for e-values $> e^{-125}$, the methods fare well in comparison to each other and to the “true” tree. At lower e-values, dMP outperforms the other two approaches resoundingly, showing a high degree of consistency to all four trees, “true”, uMP, dMP and ML. It is clear from Figure 1 that dMP and ML perform equally well at e-values in the range of e^{-50} to e^{-100} . If the topology of the *gambiae* complex is altered in the true tree by using other topologies in the literature (such as the *gambiae* complex in Fontaine et al., 2015) the results of comparison with the RhoI index do not change.

Figure 2 shows the “true” tree based on Neafsey et al., (2015) compared to the three presence absence approaches. In this figure we tallied whether the nodes in the presence absence trees agreed with the nodes in the “true” tree. The dMP and ML approaches agree with nearly every node in the “true” tree, while the uMP approach disagrees at nearly 40% of the nodes in the “true” tree at the e^{-75} cutoff we used. It should be noted though that the dollo parsimony approach performed slightly better than the other two approaches (uMP and ML). We used a simple model of evolution for the ML analysis, and other more complex models might improve the performance of ML. Again resolving the nodes within the *gambiae* complex differently such as in the Fontaine et al., (2015) analysis does not alter the gist of these comparisons.

With a well-resolved and robust phylogenetic hypothesis in hand for presence/absence data for this group, we next examined the chromosomal location of gene gains and losses over the evolution of the group. We performed a very coarse-grained determination of the location of losses and gains on the *A. gambiae* reference chromosomes to test whether the gains and losses are evenly distributed across the chromosome elements for *A. gambiae*. We should point out that there has been considerable rearrangement of the chromosomes in the species of this study, and that in the present study we are using *A. gambiae* as a frame of reference for locating the gains and losses. Chromosomal locations of the 10,000 genes we compiled for *A. gambiae* were established using VectorBase (Giraldo-Calderón et al., 2014). The gains and losses of genes at the nodes in the “true” tree were then determined using the “apolist” command in PAUP* (Swofford, 2001). Next the genes that are gained and lost across the phylogeny were compiled and their locations on the *A. gambiae* chromosomes were determined using the coordinates established for the *A. gambiae* chromosomal elements. We restricted our analysis to the *A. gambiae* chromosome elements because of the multiple inversions that have occurred in other anophelines. While this approach restricts correlation of the gain/loss events with chromosomal location across all anophelines, it does give a preliminary picture of the localization of these events in one species.

Since gene numbers are low in peri-centromeric and peri-telomeric regions of the genomes of animals, and this could skew estimates of loss and gain frequency, we used 500 gene sliding windows to normalize the relative frequency of gene losses and gains on the *A. gambiae* chromosome elements. The chromosomal location of the ends of each 500 gene windows were used to position the 500 gene window onto the chromosomal elements. While the exact location of the genes in this analysis are not represented, the sequential pattern that emerges from one end of the chromosomal elements to the other is clear (Figure 3). Some generalizations that can be made from this analysis are, first, both gains and losses, at least at the degree of resolution we show here, are not distributed evenly across the *A. gambiae* karyogram. Second, it appears that the midsections of chromosome arms are focal points for gains and losses. Third, and related to the other two points, gains and losses largely appear to be localized away from chromosome tips and centromeres (though some losses for the centromere of chromosome 3 do not follow this trend).

Summary

Phylogenetic analysis using presence absence gene content matrices is accurate at close phylogenetic distances as shown by the inability of the approach to recover reasonable hypotheses at the level of metazoan relationships and eukaryotic relationships (Rosenfeld and DeSalle, 2015). The ability of genome content data to resolve the relationships of the taxa in this study is in line with previous studies using genome content to resolve relationships in other closely related taxonomic groups (Rosenfeld and DeSalle, 2012) including the family Drosophilidae (Rosenfeld et al., 2008) and relationships at the level of insect orders (Rosenfeld et al., 2016). It is interesting to note that the disagreement amongst these trees is mostly near the tips of the tree, in contrast to trees generated using sequences that mostly show weak and conflicting inferences at basal nodes. This result might indicate that gene presence/absence information used in combination with sequence level information could potentially resolve basal nodes in phylogenetic analysis in closely related taxonomic groups. There also appears to be variation in the chromosomal location of gene gains and losses. Specifically, the middle of chromosome arms appears to accrue more gains and losses than other regions. However, this

result may be due to the inability to align other mosquito genomes sequences with the *A. gambiae* genome. Neafsey et al., (2015) showed that regions near the centromeres and on the X chromosome experience “reduced alignability”. Our scan using the chromosome of *A. gambiae* is consistent with the reduced alignability observation and suggests that detecting uninterpretable gain/loss events near centromeres is unavoidable.

Acknowledgements: The authors acknowledge the continued support of the Sackler Institute for Comparative Genomics at the AMNH and the Korein Family Foundation. SO acknowledges NSF grant XXXXXXXXXX.

References

Attrill H, *et al.* (2016) FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic acids research* 44(D1):D786-D792.

Fontaine, Michael C., James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang *et al.* "Extensive introgression in a malaria vector species complex revealed by phylogenomics." *Science* 347, no. 6217 (2015): 1258524.

Giraldo-Calderón, Gloria I., Scott J. Emrich, Robert M. MacCallum, Gareth Maslen, Emmanuel Dialynas, Pantelis Topalis, Nicholas Ho *et al.* "VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases." *Nucleic acids research* (2014): gku1117.

Lienau EK, DeSalle R, Rosenfeld JA, Planet PJ. 2006. Reciprocal illumination in the gene content tree of life. [Syst Biol](#) 55(3):441-453

Neafsey, Daniel E., Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon *et al.* Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science* 347, no. 6217 (2015): 1258522.

Rosenfeld JR, DeSalle R, Lee EK, O'Grady P. 2008. Using whole genome presence/absence data to untangle function in 12 *Drosophila* genomes. [Fly](#) 2(6):291-299.

Rosenfeld J, DeSalle R. 2012. E value cutoff and eukaryotic genome content phylogenetics. [Mol Phylogenet Evol](#) 63(2):342-350.

Rosenfeld, J.A., Foox, J. and DeSalle, R., 2016. Insect genome content phylogeny and functional annotation of core insect genomes. *Mol Phylogenets and Evol*, 97, pp.224-232.

Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. *Nat. Genet.* 21, 108–110.

Snel, B., Huynen, M.A., Dutilh, B.E., 2005. Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* 59, 191–209.

Stamatakis, Alexandros. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30, no. 9 (2014): 1312-1313.

Swofford, D., 2001. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), 4.0b7 Beta Version. Sinauer Associates, Sunderland, MA.

Figure Legends

Figure 1. Plots of Rholf consistency index (on Y-axis in all graphs) for the “true,” uMP, dMP, and ML trees for the three phylogenetic approaches used in this study. An index value of “1” indicates perfect consistency between trees. E-values are given on the X-axis. Blue = unweighted parsimony (uMP), orange = Dollo parsimony (dMP), and green = likelihood (ML).

Figure 2. Phylogenetic hypotheses generated with uML, dMP and ML mapped onto the “true” tree of Neafsey et al., (2015). Colored boxes at each node signify that a given method was in agreement with the “true” tree: blue = uMP, orange = dMP, and green = ML. Thus, dMP agrees with the “true” tree at 21 of the 22 nodes, while uMP agrees at only 13 of the 22 nodes in the tree. Numbers in blue indicate the number of genes common to all species below that node (i.e. core genes for the taxa circumscribed by the node). Boxed numbers indicate the anopheline nodes discussed in this study.

Figure 3. Graphical representation of the karyotype of *A. gambiae* with normalized number of gains (in red) and losses (in blue) graphed above the chromosomes. Black dots refer to centromeres. The losses and gains mapped here are ones that occur only a single time on the true tree phylogenetic hypothesis. Five hundred gene sliding windows were used to normalize the relative frequency of gene losses and gains in chromosomal regions over all of the chromosome elements.

Figure 1

Figure 2

Figure 3

A whole genome gene content phylogenetic analysis of anopheline mosquitoes.

Jeffrey A. Rosenfeld^{1,2}, Sara Oppenheim¹ and Rob DeSalle¹

Graphical Abstract

Plot of Rhof consistency index (on Y-axis in all graphs) for the “true,” uMP, dMP, and ML trees for the three phylogenetic approaches used in this study. An index value of “1” indicates perfect consistency between trees. E-values are given on the X-axis.

Phylogenetic hypotheses generated with uML, dMP and ML mapped onto the “true” tree of Neafsey et al., (2015). Colored boxes at each node signify that a given method was in agreement with the “true” tree: blue = uMP, orange = dMP, and green = ML. Boxed numbers indicate the anopheline nodes discussed in this study.

Graphical representation of the karyotype of *A. gambiae* with normalized number of gains (in red) and losses (in blue) graphed above the chromosomes. Black dots refer to centromeres.