Genome mapping revealed scaffold misassemblies and elevated gene shuffling on the X chromosome in malaria mosquitoes

Igor V. Sharakhov, Ashley Peery, Anastasia N. Naumenko, Xiaofang Jiang, A. Brantley Hall, Maria V. Sharakhova, Zhijian Tu

Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA, igor@vt.edu

Gleb Artemov, Vladimir N. Stegniy

Tomsk State University, Tomsk 634050, Russia, g-artemov@mail.ru

The recently sequenced and assembled genomes and transcriptomes of 16 anophelines from Africa, Asia, Europe, and Latin America allowed researchers to investigate the genomic basis of vectorial capacity [1, 2]. However, unmapped scaffolds may conceal gaps and misassemblies, which could cause incorrect or incomplete annotation of genomic sequences. Anchoring scaffolds onto chromosomes can correct these mistakes and can facilitate the development of high-quality reference genome assemblies that could be used for studying chromosome evolution and gene-order phylogeny. Toward this task, we mapped 50-90% of the *Anopheles arabiensis, An. stephensi*, *An. funestus*, *An. atroparvus*, and *An. albimanus* Illumina-based genome assemblies to chromosomal arms by fluorescent *in situ* hybridization (FISH). Each of the species was individually compared to *An. gambiae*. The comparative positions of genes within mapped scaffolds based on orthology relationships were plotted using the R program genoPlotR [3]. Orientation of scaffolds on chromosomes was obtained from physical mapping data. The combination of cytogenetic and bioinformatics approaches identified and corrected multiple cases of scaffold misassemblies in *An. arabiensis, An. stephensi*, and *An. albimanus*.

Both mosquito and fruit fly genomes have conserved gene membership on chromosome arms (elements). Unlike *Drosophila* [4], chromosome elements in *Anopheles* reshuffle between chromosomes via translocations as intact elements, and do not show fissions or fusions. Analysis of genes in anchored regions showed that synteny at the whole-

arm level is highly conserved, despite several whole-arm translocations. In contrast, small-scale rearrangements disrupt gene colinearity within arms over time, leading to extensive shuffling of gene order over a timescale of 100 Myr ago. We estimated the number of chromosomal inversions between *An. gambiae* and each of the species using the Genome Rearrangements in Mouse and Man (GRIMM) program [5]. As in *Drosophila* [6], rearrangement rates are higher on the X chromosome than on autosomes. However, the difference is significantly more pronounced in *Anopheles*, where X chromosome rearrangements are 2.7-fold more frequent than autosomal rearrangements; in *Drosophila*, the corresponding ratio is only 1.2 [6] (*t* test, $t_{10} = 7.3$, $P < 1 \times 10^{-5}$). The fast rate of X chromosome rearrangements is in sharp contrast with the paucity of polymorphic inversions on the X in anopheline species. This finding could be indicative of a greater role of the X chromosome rearrangements in speciation of malaria mosquitoes. Previous studies indicated that the X chromosome has a disproportionately large effect on male and female hybrid sterility and inviability in *An. gambiae* and *An. arabiensis* [7]. Such dynamic gene shuffling may be facilitated by the multiple families of DNA transposons and LTR and non-LTR retroelements found in mosquito genomes, as well as a weaker dosage compensation phenotype in *Anopheles* compared to *Drosophila* [8].

The high rate of rearrangements on the X chromosome gave us the opportunity to reconstruct inversion phylogeny of the *An. gambiae* complex from ancestral and derived gene orders at the breakpoints of fixed chromosomal inversions. Based on the banding pattern of polytene chromosomes, 10 fixed inversions in the *An. gambiae* complex have been observed [9], of which five are on the X chromosome. Based on these five fixed inversions, the *An. gambiae* complex can be divided into three groups: 1) *An. merus* and *An. gambiae*, which share the compound Xag inversion; 2) *An. quadriannulatus*, *An. bwambae* and *An. melas*, which carry the arbitrary standard X arrangement; and 3) *An. arabiensis,* which carries the compound Xbcd inversion. We identified the genomic coordinates for breakpoints of fixed inversions, using the newly available *Anopheles* genome assemblies [1, 2]. We determined

the ancestral (Xa, Xg, 2Ro, $2R^{+p}$, 2La) and derived (Xb, Xc, Xd, X^{+ag}, $2R^{+o}$, 2Rp, $2L^{+a}$) arrangements for fixed chromosomal inversions in the *An. gambiae* complex based on comparisons to outgroup species. The calculation of inversion distances among the included species of the *An. gambiae* complex and the outgroup species was performed using the Multiple Genome Rearrangement (MGR) program available at http://grimm.ucsd.edu/MGR/. Our phylogenetic approach rejects the majority autosomal topology caused by introgression and produces an inversion phylogeny consistent with the species phylogeny determined from X chromosome sequences [10]. We also estimated the divergence time in the *An. gambiae* complex using rates of X chromosome evolution estimated for the genus [1]. Accordingly, divergence between the *An. gambiae* and *An. arabiensis* lineages occurred about 1.88±0.05 Myr ago, an estimate very close to the approximate divergence time of 1.85±0.47 Myr inferred independently from sequence divergence on the X chromosome [10].

Our study demonstrates that a number of analyses including rearrangement phylogeny and chromosome evolution depend of the availability of chromosome-based genome assemblies. In addition to increasing the value of genome sequence data to the research community, chromosome mapping can potentially identify gaps and misassembled scaffolds within assemblies. Therefore, the development of high-resolution physical maps in combination with computational approaches is an important framework for improving the quality of the genome assembly, annotation, and analysis.

Acknowledgements

This work is supported by the Fralin Life Science Institute and by NIH grants AI094289 and AI099528 to IVS.

References

1. Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., et al.

(2015). Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science *347*, 1258522.

2. Jiang, X., Peery, A., Hall, A., Sharma, A., Chen, X.G., Waterhouse, R.M., Komissarov, A., et al. (2014). Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. Genome Biol *15*, 459.
3. Guy, L., Kultima, J.R., and Andersson, S.G. (2010). genoPlotR: comparative gene and genome visualization in R. Bioinformatics *26*, 2334-2335.
4. Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.
5. Tesler, G. (2002). GRIMM: genome rearrangements web server. Bioinformatics *18*, 492-493.
6. von Grotthuss, M., Ashburner, M., and Ranz, J.M. (2010). Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila. Genome Res *20*, 1084-1096.
7. Slotman, M., Della Torre, A., and Powell, J.R. (2005). Female sterility in hybrids between Anopheles gambiae and A. arabiensis, and the causes of Haldane's rule. Evolution *59*, 1016-1026.
8. Baker, D.A., and Russell, S. (2011). Role of testis-specific gene expression in sex-chromosome evolution of Anopheles gambiae. Genetics *189*, 1117-1120.
9. Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M.A., and Petrarca, V. (2002). A polytene chromosome analysis of the Anopheles gambiae species complex. Science *298*, 1415-1418.
10. Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., et al. (2015). Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science *347*, 1258524.