Trends in Parasitology February 2015, Vol. 31, No. 2

Evolving the world's most dangerous animal

David Weetman and Christopher S. Clarkson

Vector Biology Department, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA UK

The devastating consequences of malaria are well known but many mysteries remain about its key protagonists, a handful of *Anopheles* species. New work provides a framework for solving such puzzles, by generation and analysis of whole genome assemblies for 16 *Anopheles* species, with genomic flexibility a key emergent theme.

The question of what makes a malaria vector is fundamental to man's unfortunately intimate relationship with the disease. There are roughly 500 species of *Anopheles*, over 80% of which cannot or do not transmit malaria. Unravelling the complex interplay between morphological, physiological, life history, and behavioural differences that separate major, minor, and non-malaria vectors is a daunting task, but one which could provide fundamental information for long-term elimination. This question is at the heart of the *Anopheles* 16 genomes project, the first major works from which have recently been published [1,2].

Genomes and transcriptomes of *Anopheles* from across the world have been sequenced and assembled, including some of the most efficient malaria vectors. To facilitate comparative analyses, species were chosen to occupy progressively greater phylogenetic distances from the primary African vector species pair *Anopheles gambiae s.s.* and *Anopheles coluzzii* (formerly the S and M forms of *Anopheles gambiae* [3]), overall representing 100 million years of evolution. Why is this project important when the genome of *A. gambiae* (PEST strain) was fully sequenced and near-fully assembled over a decade ago [4]? The answer becomes clear when comparing the newly sequenced genomes to those of the existing PEST assembly. *A. gambiae* is a species complex of morphologically identical, relatively closely-related species. Genome sequence data from each of the five members of the *A. gambiae* complex in the project align well to the PEST assembly (with >85% alignment success). However, this concordance drops dramatically once sequences of the more distantly related species are considered; down to <15% for the major Central/South American malaria vectors *Anopheles albimanus* and *Anopheles darlingi* [1]. In other words, because major malaria vectors are often not closely related, a single reference genome provides a poor general template. Indeed, even for species within the *A. gambiae* complex, and despite gaps in sequence coverage, inevitable in newly sequenced genomes, alignment of additional sequenced specimens to their species-specific assemblies was significantly better in every case than to the fully-assembled PEST reference [2].

The first paper by Neafsey, Waterhouse *et al.* [1] presents an in-depth description of the methodologies involved in the project, including a very useful glossary of the many software tools applied (pages 123-4 of the supplementary materials). Assembly qualities were evaluated by searching each reference genome and paired RNA transcripts for almost 3000 evolutionary conserved single-copy orthologous genes from *Drosophila*, that is, those which should be present in all of the genomes. Assembly qualities generally appear impressive and surprisingly consistent, given the highly variable number and length of the sequence scaffolds. Variation in contiguity of assemblies was highlighted however when analysing cuticular genes, which comprise 2% of protein coding genes and tend to occur in clusters. All new assemblies show fewer genes per cluster than the fully assembled *A. gambiae* PEST genome, which is most likely a result of assembly difficulties for highly similar genes in clusters, toward the end of scaffolds [1]. This highlights an easily overlooked issue with genome sequence data now obtained relatively quickly and inexpensively. Meaningful comparative analyses require extensive work on assembly and gene annotations, and rigorous quality checks, of the kind performed by the 16 genomes team, other major sequencing consortia (http://www.malariagen.net/projects/vector/ag1000g), and the community as a whole. To this end the 16 genomes consortium are to be congratulated for a release policy which has allowed substantial deposition of data in VectorBase far in advance of publication.

Intriguing insights emerge from comparing variation across the *Anopheles* species sequenced to results from the *Drosophila* 12 genomes project [5]. For example, there was a five-fold higher rate of gene gain/loss among *Anopheles* than among *Drosophila* species, along with higher rates of intron loss, and a lack of the constraining influence of the codon usage bias that is evident in *Drosophila*. All hint at flexibility in *Anopheles* which might be a key to the multiple occurrences of adaptation to human hosts. A distinct evolutionary profile of the X chromosome is highlighted by elevated rates of gene gain and loss relative to autosomes [1] and comes to the fore in the second paper. Fontaine, Pease *et al.* [2] conduct an in-depth analysis of species relationships and introgression (inter-specific gene flow) within the *A. gambiae* complex. The true species tree within the complex has been unresolved for many years but compelling evidence now shows that the major vector *Anopheles arabiensis* appears closely related to *A. gambiae* and *A. coluzzii* throughout most of the genome as a result of extensive introgression. In fact the true species tree is reconstructed against a vast genomic majority rule, and

Corresponding author: Weetman, D. (david.weetman@lstmed.ac.uk)
Keywords: *Anopheles*; phylogenomics; introgression.

1471-4922/
 http://dx.doi.org/10.1016/j.pt.2015.01.001

is recovered most accurately from a section of the X chromosome retardant to introgression. Presumably therein lie the key genes maintaining reproductive isolation. These results beautifully advertise the power of phylogenomics, not via obtaining an overwhelming amount of data from which an average consensus can be obtained, but rather by comparative analysis across species and different parts of the genome. Some long-held views of *A. gambiae* history also appear to have been overturned, notably the timing of speciation in relation to the expansion of human agriculture (which it now seems to pre-date) and the transfer of a huge, ecologically important chromosomal inversion from *A. arabiensis* to *A. gambiae* (the reverse now seems probable).

The ability of *Anopheles* genomes to accept and incorporate large sections of autosomes, even if highly distinct, without loss of species integrity, could be a key to successful adaptation [6]. Such flexibility may result from interplay of properties of the genomes themselves as suggested [1], and the strong selective pressures incurred by adaptation to human hosts. The importance of genetic transfer long after species split will be a familiar story to plant and microbial geneticists, but was long thought rare in animals [7]. However, results from both the *Anopheles* 16 genomes consortium and the *Heliconius* genome consortium [8] suggest that interspecific hybridization plays a major role in adaptive phenotypic change. A chilling difference is that whilst butterflies exchange genetic variants controlling their often beautiful colour patterns, *Anopheles* species may transfer adaptive variation which enhances their efficiency as malaria vectors.

References

1 Neafsey, D.E. *et al.* (2015) Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* 347, 1258522
2 Fontaine, M.C. *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347, 1258524
3 Coetzee, M. *et al.* (2013) *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa* 3619, 246–274
4 Holt, R.A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298, 129–149
5 Clark, A.G. *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218
6 Clarkson, C.S. *et al.* (2014) Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. *Nat. Commun.* 5, 4248
7 Hedrick, P.W. (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22, 4606–4618
8 The *Helconius* Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 94–98