

The structure of the 2La inversion breakpoints in *Anopheles gambiae* complex

Igor Sharakhov^{1,2}, Bradley White¹, Maria Sharakhova^{1,2}, Jonathan Kayondo¹, Neil Lobo¹, Federica Santolamazza³, Alessandra della Torre³, Frank H. Collins¹, Nora J. Besansky¹

¹Center for Tropical Disease Research and Training, University of Notre Dame, Notre Dame, IN 46556, USA

²Department of Entomology, Virginia Tech, Blacksburg, VA 24061, USA

³Sezione di Parassitologia, Dipartimento di Scienze di Sanita Pubblica, Universita di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy

Fixed paracentric inversions play a major role in reshaping genome architecture in evolution of malaria mosquitoes and have been used as robust markers for inferring mosquito phylogenies. Polymorphic inversions contribute to ecological differentiation within species. The African malaria mosquito *An. gambiae* represents a complex of seven morphologically similar species. The 2La inversion is polymorphic in *An. gambiae* s.s. and is associated with susceptibility to *Plasmodium* and with adaptation to an arid environment. The 2La arrangement is fixed in *An. merus* and *An. arabiensis* (as 2La/a). The remaining species of the complex are fixed for the alternative arrangement (2L^{+a/+a}) which was assumed to be ancestral. Previous DNA sequence analyses of regions within or near the 2La inversion breakpoints suggested that the sister taxa, *An. merus* and *An. gambiae*, have two different inversions, 2La' and 2La. It was also suggested that non-sister species, *An. arabiensis* and *An. gambiae*, share the same 2La inversion through introgression.

To investigate the actual structure of the 2La inversion breakpoints we have cloned and sequenced DNA fragments that span both breakpoints of these three species. The correspondence between the *An. gambiae* PEST (2L^{+a/+a}) genome sequence and the polytene chromosome complement allowed us to identify BAC clones that cross the inversion breakpoints. Subsequent analysis identified candidate segments of the BAC clones which were used as probes to screen genomic libraries of the SUA (2La/a) and Bamako (2La/a) strains of *An. gambiae*, and the V12 strain of *An. merus*. Primers designed based on the SUA sequence were used to obtain the breakpoints from *An. arabiensis* by PCR from genomic DNA. Sequencing, *in situ* and *in silico* analyses of the positive clones and PCR products confirmed that the proximal and distal breakpoints were recovered from SUA, Bamako, *An. merus*, and *An. arabiensis*. The ClustalX alignments of the 2La breakpoint sequences have revealed identical molecular organization among these species and strains.

Detailed comparison of the SUA 2La/a sequences spanning proximal (4,369 bp) and distal (3,409 bp) breakpoints with the corresponding sequences of *An. gambiae* PEST 2L^{+a/+a} (the reference *An. gambiae* genome) identified insertions/deletions, putative genes and transposable elements, and repetitive DNA. The SUA proximal breakpoint sequence has two genes that encode sulfatase and zinc finger protein, respectively. Both genes in PEST are present at both breakpoints; a full-length copy at one breakpoint (sulfatase at the proximal, zinc finger at the distal) and a truncated copy in opposite orientation at the other breakpoint (sulfatase at the distal; zinc finger at the proximal). Put another way, a ~750 bp fragment of the SUA 2La/a proximal breakpoint is represented twice in PEST 2L^{+a/+a}, in opposite

orientations at both breakpoints. The PEST distal breakpoint has a 241 bp insertion flanked with 45 bp inverted repeats, one of which is adjacent to the sulfatase pseudogene. The PEST proximal breakpoint has a 3,972 bp insertion of clustered and scrambled transposable element fragments, flanked at one end with the same 45 bp repeat. This repeat is adjacent to the zinc finger pseudogene.

The presence of full-length genes and their pseudogene copies at opposite breakpoints of the standard ($2L^{+^a}/+^a$, PEST) arrangement strongly suggests that the inverted ($2La/a$) arrangement common to *An. arabiensis* and *An. merus* is ancestral. Our data indicate that the generation of the $2L^{+^a}/+^a$ arrangement was associated with the insertion of a repetitive element before or after inversion happened. The precise molecular mechanism and order of the events for the inversion formation is not apparent, but the outcome is inconsistent with a simple cut-and-paste mechanism. One possible model assumes that the same DNA breaks participate in inversions and unequal recombination between slightly different segments of two homologous chromosomes in meiosis. If these two processes coincide spatially and temporarily, the $2L^{+^a}/+^a$ arrangement can be generated from $2La/a$ within the same meiotic prophase.