

Inversions and Gene Order Shuffling in *Anopheles gambiae* and *A. funestus*

Igor V. Sharakhov,^{1†} Andrew C. Serazin,^{1†} Olga G. Grushko,¹
 Ali Dana,¹ Neil Lobo,¹ Maureen E. Hillenmeyer,¹
 Richard Westerman,² Jeanne Romero-Severson,³
 Carlo Costantini,^{4,5} N'Fale Sagnon,⁵ Frank H. Collins,¹
 Nora J. Besansky^{1*}

In tropical Africa, *Anopheles funestus* is one of the three most important malaria vectors. We physically mapped 157 *A. funestus* complementary DNAs (cDNAs) to the polytene chromosomes of this species. Sequences of the cDNAs were mapped in silico to the *A. gambiae* genome as part of a comparative genomic study of synteny, gene order, and sequence conservation between *A. funestus* and *A. gambiae*. These species are in the same subgenus and diverged about as recently as humans and chimpanzees. Despite nearly perfect preservation of synteny, we found substantial shuffling of gene order along corresponding chromosome arms. Since the divergence of these species, at least 70 chromosomal inversions have been fixed, the highest rate of rearrangement of any eukaryote studied to date. The high incidence of paracentric inversions and limited colinearity suggests that locating genes in one anopheline species based on gene order in another may be limited to closely related taxa.

Malaria morbidity and mortality in tropical Africa remain disproportionately high relative to other malaria-endemic areas of the world, in part because of three efficient vectors in subgenus *Cellia*: *A. gambiae*, *A. arabiensis*, and *A. funestus*. These species co-occur geographically across sub-Saharan Africa and can inhabit the same villages, shelter in the same houses, and feed on the same individuals. Yet *A. funestus* has evolved unique breeding site preferences, mating behavior, relative seasonal abundance, and degree of specialization on humans. The genetic basis of these differences is unknown. This species has received far less attention than *A. gambiae*, in part because obligatory swarming behavior associated with mating has been an obstacle to establishing laboratory colonies. However, it is obvious that successful malaria control strategies for Africa must take this and other species into account. The full genome sequence of *A. gambiae* (1) allows detailed comparative genomic analysis of closely related anopheline species. To the

extent that synteny and colinearity are conserved between model organisms such as *A. gambiae* and more poorly characterized species such as *A. funestus*, comparative gene mapping is a powerful tool for candidate positional cloning. These comparisons could locate the genes responsible for ecological adaptations, speciation, insecticide resistance, host preference, and parasite defense before their signal is lost in a background of accumulated mutational noise.

Using fluorescence in situ hybridization (FISH), we mapped *A. funestus* cDNA clones on the five arms of the polytene chromosome complement (2). Incorporation of different fluorescent labels allowed us to probe simultaneously with two different cDNAs (Fig. 1). The cDNAs were isolated from a library prepared from *A. funestus* larval, pupal, and adult mRNA. The clones chosen for mapping were a subset of a larger collection of 3233 clones, whose sequence was determined from the 5'-end as part of an ongoing expressed sequence tag (EST) project (3). Based on the results of BLASTN (Basic Local Alignment Search Tool) searches against the *A. gambiae* genome, we selected *A. funestus* cDNAs whose scores had statistical significance thresholds of less than 10^{-6} (2). Of 157 cDNAs used as probes, 116 mapped to single chromosomal locations on the *A. funestus* cytogenetic map (111 of which are given in table S1), and the remainder hybridized in multiple locations (table S2). These constitute the only available physical map of this species. No genetic linkage map exists be-

cause controlled crosses cannot be performed. The cDNAs mapped not only to euchromatic bands but also to apparent interbands [for example, 24_G09 in subdivisions 10C, 12B, and 18A (Fig. 1B)]. Four others hybridized to a diffuse region of β -heterochromatin on 3L:38C-39A that serves as an attachment site to the nuclear envelope (Fig. 1C) (4).

The chromosomal positions of uniquely hybridizing *A. funestus* cDNAs are shown in Fig. 2, along with their corresponding locations in *A. gambiae* (2). Not shown are three cDNAs for which no significant BLASTN match was found to the *A. gambiae* genome: 04_F09, a putative inhibitor of apoptosis (2R:8E); 66_C12, similar to an *A. gambiae* protein with a C-type lectin domain (3R:35E); and 07_D02, function unknown (2R:13C). Recognizing that inferences about synteny, gene order, and sequence relationships require comparison of orthologous genes, we also did not include in Fig. 2 sequences that were dispersed to multiple locations in either species, because the hypothesis of gene orthology (related by speciation) versus paralogy (related by gene duplication) would be less firm. The relative positions of sequences with unique map locations in both species support the hypothesized chromosome arm homologies and the reciprocal whole arm translocation between 2L and 3R, postulated previously on the basis of relative length and banding pattern (4). Correspondence between chromosome arms was contradicted by only two of the cDNAs examined in this study. Clones 11_G12 and 04_A10 (not shown in Fig. 2) hybridized to 3R:32D and 3L:45A in *A. funestus*, yet the corresponding sequences in the *A. gambiae* genome are located on 3R:29A and 2R:18A, rather than on 2L and 3L, which is consistent with a transposition event. Aside from these two exceptions, the data reveal perfect conservation of synteny at the whole-arm level.

Within corresponding arms, paracentric inversions have had a major impact on genome architecture since the divergence of these species. Gene order has not been preserved along the length of any chromosome arm, although there are segments with conserved gene order (hereafter, "conserved segments"). Clear examples of conserved segments are located in regions near centromeres and telomeres where the rate of meiotic recombination may be reduced (Fig. 2). However, this generalization does not hold without exception, as 2R sequences near the telomere in *A. funestus* (08_H11, 01_H04) are located more proximally in *A. gambiae*, on the opposite side of a flanking marker (21_F03). This represents one of three small inversions that can be inferred at the distal end of 2R, including a rearrangement involving the 8C region in *A. gambiae* that contains

¹Center for Tropical Disease Research and Training, University of Notre Dame, Notre Dame, IN 46556-0369, USA. ²Horticulture Department, Purdue University, West Lafayette, IN 47907-1159, USA. ³Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907-1165, USA. ⁴Dipartimento di Scienze di Sanità Pubblica, Sezione di Parassitologia, Università "La Sapienza," 00185 Roma, Italy. ⁵Centre National de Recherche et de Formation sur le Paludisme, Ouagadougou, Burkina Faso.

*To whom correspondence should be addressed. E-mail: besansky.1@nd.edu

†These authors contributed equally to this work.

the major *Plasmodium*-refractoriness locus *Pen1* (5). From these data, it is apparent that inversions have involved large as well as relatively small chromosomal segments.

What has been the extent of rearrangement of gene order between these species? The number of inversion events can be estimated by considering the mean length of conserved segments, because this length decreases with each inversion fixed since the divergence of *A. gambiae* and *A. funestus* from a common ancestor. The method of Nadeau and Taylor (6) was applied to estimate mean lengths of all conserved segments in the genome, based on the nucleotide distance in *A. gambiae* between the outermost markers that defined the segments observed in our sample. An assumption of the method, that rearrangements fixed during evolution are randomly distributed in the genome, seems unlikely given the extraordinary concentration of polymorphic inversions on 2R in both lineages. Of eight polymorphic inversions described in *A. gambiae*, seven occur on chromosome 2R (7). Similarly, 11 of 15 polymorphic inversions found in *A. funestus* involve 2R (8). Accordingly, we assessed each arm independently. The estimated mean lengths of all conserved segments on each arm, defined with respect to *A. gambiae*, were X, 2.0 ± 0.2 megabases (Mb); 2R, 0.9 ± 0.2 Mb; 2L, 2.2 ± 0.4 Mb; 3R, 2.2 ± 1.0 Mb; and 3L, 1.1 ± 0.4 Mb. In a slight departure from Nadeau and Taylor (6), each rearrangement was assumed to be an inversion requiring two disruption events. Therefore, n inversions result in $2n + 1$ conserved segments. We cannot discount the possibility that some disruptions of gene order were caused by intrachromosomal transpositions rather than inversions, events that are impossible to distinguish at this level of resolution. However, the contribution of transposition events was considered negligible.

Under these assumptions, the number of inversions on each arm was 5 ± 1 , 36 ± 9 , 11 ± 3 , 11 ± 3 , and 19 ± 5 , respectively. Assuming a divergence time of 5 million years (My) (2), the rate of fixation per My for each chromosome arm can be estimated as 0.5, 3.6, 1.1, 1.1, and 1.9, respectively (or 7 when estimated across the genome). When normalized to account for differences in chromosome length, the number of inversions per Mb per My for X, 2R, 2L, 3R, and 3L was estimated as 0.023, 0.057, 0.022, 0.021, and 0.044, respectively (0.031 genome-wide). This rapid rate, even more extreme than the genome-wide estimate for *Drosophila* (9), is the highest reported for any eukaryotic species. Moreover, these results suggest that 2R has a higher rate of rearrangement than other arms. Higher resolution comparative studies of 2R are needed to provide insights about the mechanism and dynamics of paracentric chromosomal inversions.

It is clear that tightly linked genes in *A.*

gambiae are unlikely to be similarly linked in *A. funestus*, particularly on 2R. The estimate of mean conserved segment length derived for each arm can be used to predict the probability of linkage in *A. funestus*, given the known distance between genes in *A. gambiae* and the assumption of random distribution of breakpoints (6). As an example, the probability that genes 1 Mb apart on 2R in *A. gambiae* are linked on 2R in *A. funestus* is only 0.31. The probability of linkage conservation across this distance estimated over all five chromosome arms was higher, but still only 0.54. Both species possess polymorphic inversions on homologous arms that have been associated with differences in *Plasmodium* infection rates and resting behavior after feeding. The standard arrangement of inversion 2La in *A. gambiae* and the inverted arrangement of 3Rab in *A. funestus* are associated with higher infection rates and endophily (10, 11). The limited colinearity between species leaves no doubt that the inversions associated with these behaviors have not captured identical sets of genes between the breakpoints. Only by sequencing within inversions 3Rab will it be possible to define smaller segments in common with 2La and study their function and possible contribution to bionomic heterogeneities.

Has the rapid rearrangement of gene order been accompanied by a similarly rapid pace of evolution at the nucleotide level? The extent of sequence divergence can be estimated by comparison of samples of orthologous genes. Forty *A. funestus* ESTs were chosen that mapped to unique sites in the *A. funestus* genome and for which clear orthologs were available from *A. gambiae* in the dBEST database (2). Levels of divergence varied considerably among ESTs (table S3), which is consistent with the well-known heterogeneity among nuclear genes in other organisms. Across a total of 5832 aligned codons, sequence difference ranged from 2.25 to 29.41% at the nucleotide level, averaging $13.03 \pm 6.05\%$. Corresponding amino acid divergence ranged even more widely, from 0 to 32.5% (average, $7.46 \pm 7.59\%$).

A better reflection of the pattern of mutation is obtained by consideration of sites that are synonymous (where substitutions do not result in an amino acid change) separately from those that are nonsynonymous (where substitutions do result in an amino acid change). Nonsynonymous sites may be evolving under the constraint of purifying selection or the force of positive selection, depending on the structural and functional

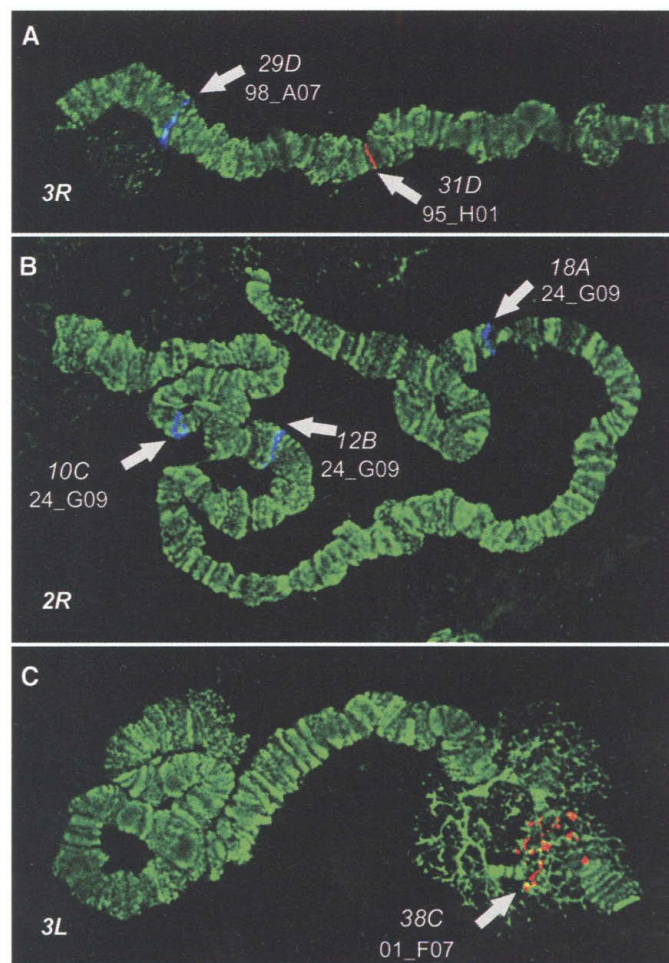


Fig. 1. (A to C) FISH performed on the chromosomes of *A. funestus*. Chromosomes counterstained with the fluorophore YOYO-1 and hybridized with fluorescently labeled probes Cy5 (blue) and Cy3 (red) are shown. The chromosome arm featured in each panel is identified at bottom left; the probe and numbered/lettered subdivision to which it hybridized are given beside arrows pointing to the corresponding signal.

THE MOSQUITO GENOME: *ANOPHELES GAMBIAE*

requirements of the protein. Synonymous sites, although not evolving neutrally if there is codon usage bias (12), can be used to approximate the neutral rate. The mean num-

bers of substitutions per synonymous site (K_s) and nonsynonymous site (K_a) were 0.612 ± 0.392 and 0.041 ± 0.044 , although there was considerable heterogeneity among

genes for both estimates, particularly for K_a , and a large degree of uncertainty surrounding the estimates (table S3). Genes involved in adaptive processes, pathogen defense, and

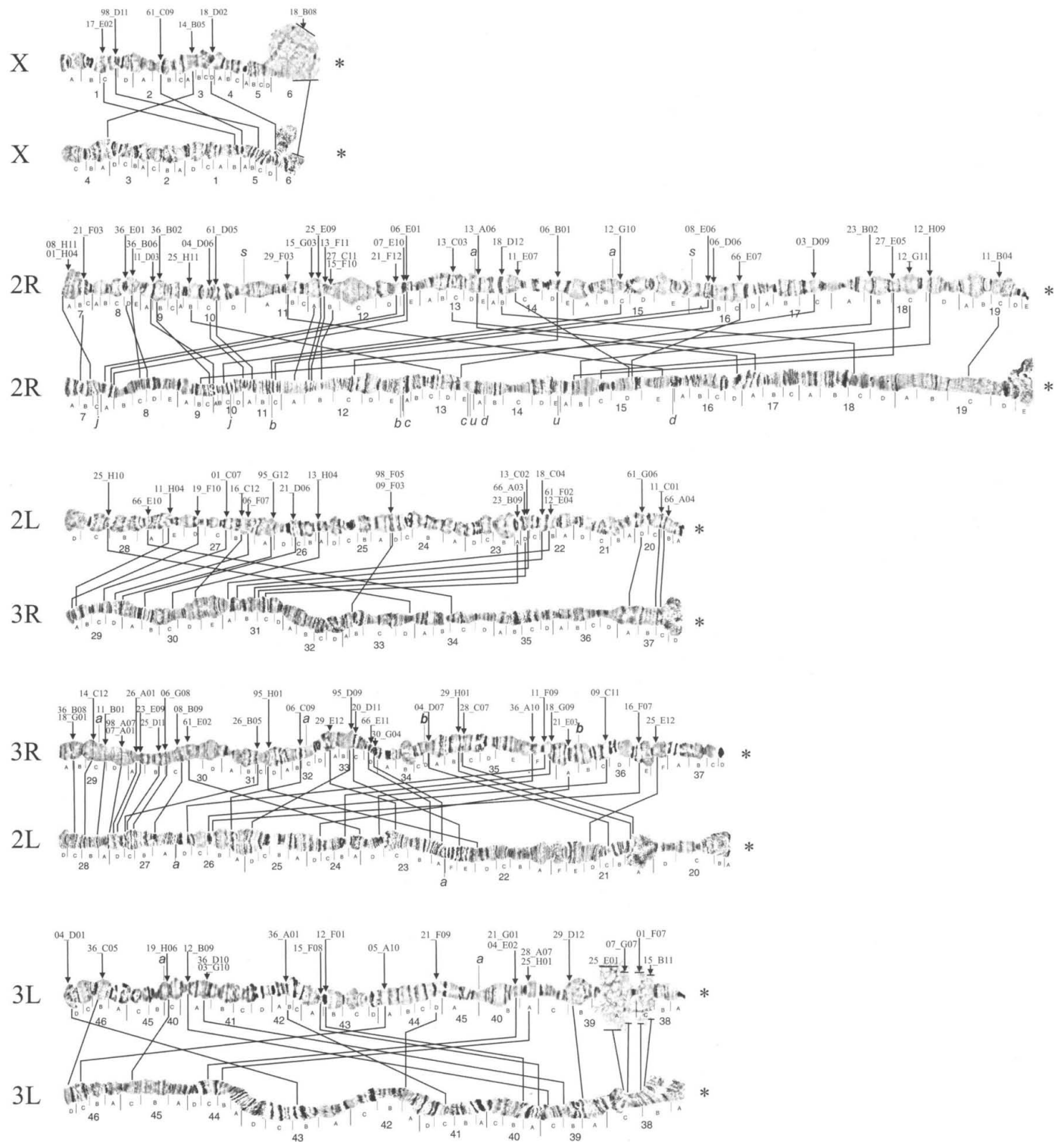


Fig. 2. Physical location of *A. funestus* cDNA clones and location of the putative *A. gambiae* orthologs, given with respect to the polytene chromosome photomaps of both species. Homologous chromosome arms are juxtaposed, with *A. funestus* chromosomes shown above those of *A. gambiae*, oriented with centromeres on the right (indicated by asterisks). Only

cDNAs that localized to one cytological position in both species are shown, with their relative positions on homologous arms indicated by interconnecting lines. Also indicated are the approximate breakpoints of common polymorphic inversions (identified by lowercase italicized letters), shown in their standard (uninverted) and presumed ancestral orientation (15, 16).

speciation are expected to show a higher nonsynonymous rate of substitution than those involved in "housekeeping" functions. The extreme K_a values for the immune-related genes *gambicin* and *lysozyme* (0.144 and 0.124, respectively), as compared with the K_a values for ribosomal proteins L8 and S10 (0.008 and 0.020, respectively), are consistent with this expectation.

An average synonymous rate has been estimated for mammalian and *Drosophila* genes. The *Drosophila* rate estimate, 16 per site per 10^9 years, is about four to five times higher than the mammalian rate of 3.5 per site per 10^9 years (13). Taking 5 My as the divergence time between *A. gambiae* and *A. funestus* lineages, the average synonymous rate of substitution in *Anopheles* is ~61 per site per 10^9 years. Several sources of error contribute to the *Anopheles* estimate, including limited sample size, sequencing errors in the EST database, and, most important, uncertainty in the divergence time. Nevertheless, a reasonable conclusion is that the rate of nucleotide substitution at silent sites in *Anopheles* is at least as fast as in *Drosophila*.

Throughout mosquito evolution, chromosome number ($2N = 6$) has remained stable (14), unlike the many changes in chromosome number that have characterized mammalian

evolution. Between *A. gambiae* and *A. funestus*, species that diverged from one another about as recently as humans and chimpanzees, morphological change has been relatively slight. However, since the ancestral lines of these mosquito species split about 5 My ago, both gene order and gene sequences have been evolving at rates at least as high as those estimated for *Drosophila*, which are the highest rates known. Our results suggest that the success of positional cloning or interspecific microarray experiments may be limited to very closely related anopheline species. The availability of the complete *A. gambiae* genome sequence and its use in comparative studies with other anopheline species will greatly improve our understanding of how inversions arise, how they shape variation within species, and how they reshape genome architecture between species.

References and Notes

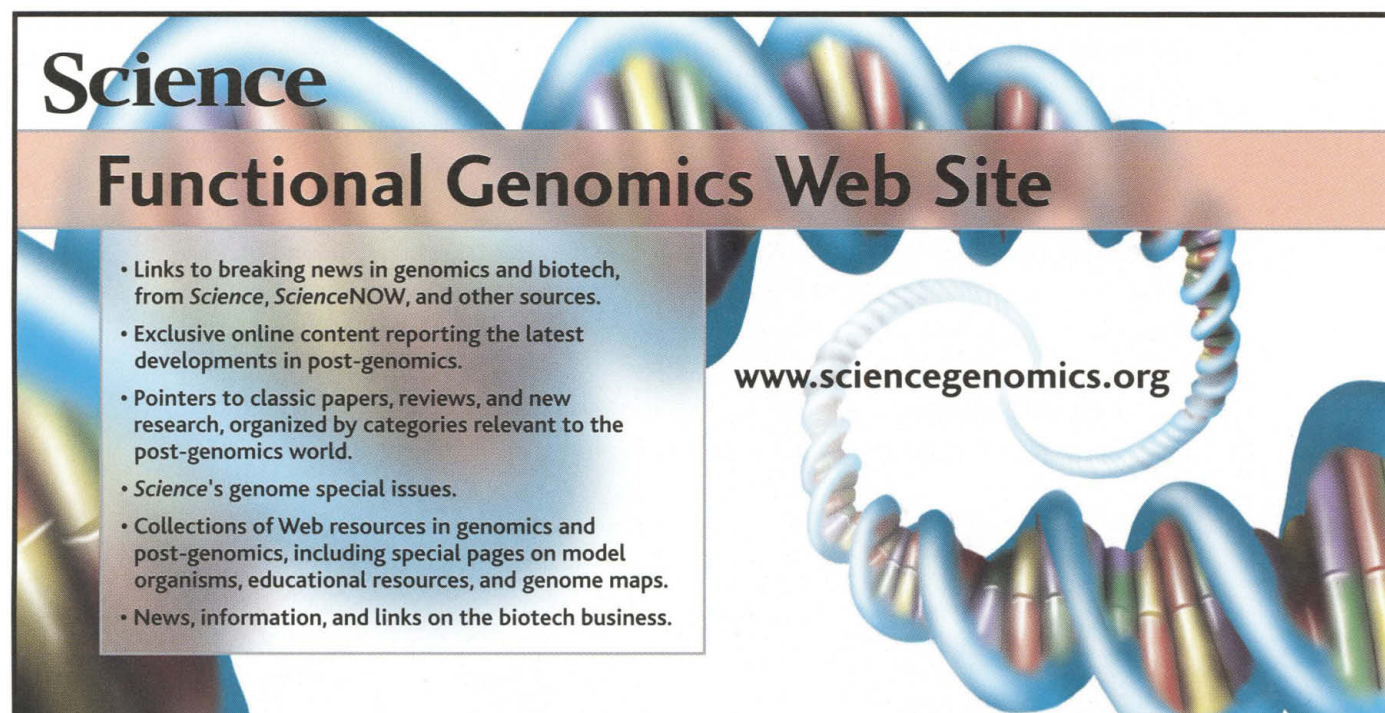
1. R. A. Holt et al., *Science* **298**, 129 (2002).
2. Materials and methods are available as supporting material on Science Online.
3. A. Serazin et al., unpublished data.
4. I. V. Sharakhov, M. V. Sharakhova, C. M. Mbogo, L. L. Koekemoer, G. Yan, *Genetics* **159**, 211 (2001).
5. L. Zheng et al., *Science* **276**, 425 (1997).
6. J. H. Nadeau, B. A. Taylor, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 814 (1984).
7. M. Coluzzi, A. Sabatini, V. Petrarca, M. A. Di Deco, *Trans. R. Soc. Trop. Med. Hyg.* **73**, 483 (1979).

8. I. Dia, D. Boccolini, C. Antonio-Nkondjio, C. Costantini, D. Fontenille, *Parassitologia* **42**, 227 (2000).
9. J. González, J. M. Ranz, A. Ruiz, *Genetics* **161**, 1137 (2002).
10. V. Petrarca, J. C. Beier, *Am. J. Trop. Med. Hyg.* **46**, 229 (1992).
11. C. Costantini, N. F. Sagnon, E. Ilboudo-Sanogo, M. Coluzzi, D. Boccolini, *Parassitologia* **41**, 595 (1999).
12. J. R. Powell, E. N. Moriyama, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7784 (1997).
13. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
14. The only known exception is *Chagasia bathana*, $2N = 8$.
15. C. A. Green, R. H. Hunt, *Genetica* **51**, 187 (1980).
16. M. Coluzzi, A. Sabatini, V. Petrarca, M. A. Di Deco, *Trans. R. Soc. Trop. Med. Hyg.* **73**, 483 (1979).
17. We are grateful to the inhabitants of Kuiti and Koubri villages, Burkina Faso, for their cooperation during our ongoing studies and to M. Coulibaly for supplying *A. funestus* from Mali for cDNA library preparation. We thank D. Severson and J. Feder for insightful discussions and J. Powell for critically reading the manuscript. Supported by grants from NIH (AI48842) to N.J.B. and from the Indiana 21st Century Research & Technology Fund to F.H.C. A.C.S. was supported by NSF grant DBI-0139317 to M. Whaley. Financial support to the *A. gambiae* Genome Consortium was provided by NIH grant U01 AI50687 to Celera Genomics, by NIH grant U01 AI48846 to the University of Notre Dame, and by the French government to Genoscope.

Supporting Online Material

www.sciencemag.org/cgi/content/full/298/5591/182/DC1
Materials and Methods
Tables S1 to S3
References and Notes

31 July 2002; accepted 5 September 2002



Science

Functional Genomics Web Site

- Links to breaking news in genomics and biotech, from *Science*, *ScienceNOW*, and other sources.
- Exclusive online content reporting the latest developments in post-genomics.
- Pointers to classic papers, reviews, and new research, organized by categories relevant to the post-genomics world.
- *Science*'s genome special issues.
- Collections of Web resources in genomics and post-genomics, including special pages on model organisms, educational resources, and genome maps.
- News, information, and links on the biotech business.

www.sciencegenomics.org