12. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
13. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
14. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
15. R. J. Mural *et al.*, *Science* **296**, 1661 (2002).
16. A mate pair is a set of two sequence reads derived from either end of a clone insert such that their relative orientation and distance apart are known.
17. Unitigs are sets of sequence reads that have been uniquely assembled into a single contiguous sequence such that no fragment in the unitig overlaps a fragment not in the unitig. The depth of reads in a unitig and the mate pair structure between it and other unitigs are used to determine whether a given unitig has single or multiple copies in the genome. We define contigs as sets of overlapping unitigs. Unlike scaffolds, which comprise ordered and oriented contigs, unitigs and contigs do not have internal gaps.
18. A nucleotide position was considered to be a SND if the respective column of the multialignment satisfied the following three criteria. First, two different bases (A, C, G, T, or unknown) had to be observed, each in at least two fragments. Second, the total number of fragments covering the column had to be ≤15 [halfway between single (10×) and double (20×) coverage] to reduce the frequency of false positives resulting from overcollapsed repeats. Third, we eliminated all but one of a run of adjacent SND columns so that block mismatches or (more likely) block indels (insertions/deletions) were counted only once.
19. SND "balance" is the ratio of the number of fragments showing the second most frequent character in a column to the number showing the most frequent character.
20. SND "association" shows, for a sliding window of 100 kb, the fraction of polymorphic columns that can be partitioned into two consistent haplotypes. For an SND column *A* of the multiple sequence alignment and the previous such column *B*, each fragment might have one of four possible haplotype phases: *AB*, *Ab*, *aB*, or *ab*, where the upper- and lowercase letters indicate alternative nucleotides. We say that columns *A* and *B* are consistent if only two of these four haplotypes are present. For the test to be nontrivial, we require that at least two fragments be observed with each of the two haplotype phases.
21. C. F. Aquadro, A. L. Weaver, S. W. Schaeffer, W. W. Anderson, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 305 (1991).
22. R. Wang, L. Zheng, Y. T. Touré, T. Dandekar, F. C. Kafatos, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10769 (2001).
23. D. J. Begun, P. Whitley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5960 (2000).
24. P. Andolfatto, *Mol. Biol. Evol.* **18**, 279 (2001).
25. D. Thomasová *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8179 (2002).
26. N. J. Besansky, J. R. Powell, *J. Med. Entomol.* **29**, 125 (1992).
27. M. Ashburner, Drosophila: *A Laboratory Handbook* (Cold Spring Harbor Laboratory Press, Plainview, NY, 1989), p. 74.
28. F. H. Collins, unpublished data.
29. J. M. Comeron, *Curr. Opin. Genet. Dev.* **1**, 652 (2001).
30. D. L. Hartl, *Nature Rev. Genet.* **1**, 145 (2000).
31. C. Rizzon, G. Marais, M. Gouy, C. Biemont, *Genome Res.* **12**, 400 (2002).
32. A. J. Cornel, F. H. Collins, *J. Hered.* **91**, 364 (2000).
33. On the basis of empirical tests, homologous proteins were required to be one of the five best mutual Blast hits within the entire genome, to fall within 15 gene calls of the closest neighboring pair, and to consist of three or more spatial matches.
34. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
35. S. Aparicio *et al.*, *Science* **297**, 1302 (2002).
36. S. L. Salzberg, R. Wides, unpublished data.
37. E. M. Zdobnov *et al.*, *Science* **298**, 149 (2002).
38. R. Apweiler *et al.*, *Nucleic Acids Res.* **29**, 37 (2001).
39. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
40. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
41. The complete hierarchy of InterPro entries is described at www.ebi.ac.uk/interpro; the hierarchy for GO is described at www.geneontology.org.
42. M. J. Gorman, S. M. Paskewitz, *Insect Biochem. Mol. Biol.* **31**, 257 (2001).
43. G. Dimopoulos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6619 (2000).
44. M. Matsushita, T. Fujita, *Immunol. Rev.* **180**, 78 (2001).
45. R. Le Borgne, Y. Bellaiche, F. Schweisguth, *Curr. Biol.* **12**, 95 (2002).
46. C. H. Lee, T. Herman, T. R. Clandinin, R. Lee, S. L. Zipursky, *Neuron* **30**, 437 (2001).
47. S. Ansieau, A. Leutz, *J. Biol. Chem.* **277**, 4906 (2002).
48. D. K. Yeates, B. M. Wiegmann, *Annu. Rev. Entomol.* **44**, 397 (1999).
49. A. N. Clements, *Biology of Mosquitoes, Vol. I: Development, Nutrition, Reproduction* (Chapman & Hall, Wallingford, UK, 1992).
50. G. Dimopoulos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8814 (2002).
51. H. Ranson *et al.*, *Insect Mol. Biol.* **9**, 499 (2000).
52. H. Ranson *et al.*, *Science* **298**, 179 (2002).
53. H. Ranson *et al.*, *Biochem. J.* **359**, 295 (2001).
54. C. A. Hill *et al.*, *Science* **298**, 176 (2002).
55. G. Dimopoulos, H. M. Muller, E. A. Levashina, F. C. Kafatos, *Curr. Opin. Immunol.* **13**, 79 (2001).
56. G. K. Christophides *et al.*, *Science* **298**, 159 (2002).
57. F. H. Collins *et al.*, *Science* **234**, 607 (1986).
58. J. Ito, A. Ghosh, L. A. Moreira, E. A. Wimmer, M. Jacobs-Lorena, *Nature* **417**, 452 (2002).
59. L. Zheng *et al.*, *Science* **276**, 425 (1997).
60. J. F. Abril, R. Guigó, *Bioinformatics* **16**, 743 (2000).
61. Supported in part by NIH grant U01AI50687 (R.A.H.) and grants U01AI48846 and R01AI44273 (F.H.C.) on behalf of the *Anopheles gambiae* Genome Consortium, and by the French Ministry of Research. We thank K. Aultman (NIAID) for her insights and effective coordination, D. Lilley (Celera) for competent financial and administrative management, and all members of the sequencing and support teams at the sequencing centers Celera, Genoscope, and TIGR.

Supporting Online Material

www.sciencemag.org/cgi/content/full/298/5591/129/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S5

15 July 2002; accepted 6 September 2002

Comparative Genome and Proteome Analysis of *Anopheles gambiae* and *Drosophila melanogaster*

Evgeny M. Zdobnov,[1]* Christian von Mering,[1]* Ivica Letunic,[1]* David Torrents,[1] Mikita Suyama,[1] Richard R. Copley,[2] George K. Christophides,[1] Dana Thomasova,[1] Robert A. Holt,[3] G. Mani Subramanian,[3] Hans-Michael Mueller,[1] George Dimopoulos,[4] John H. Law,[5] Michael A. Wells,[5] Ewan Birney,[6] Rosane Charlab,[3] Aaron L. Halpern,[3] Elena Kokoza,[7] Cheryl L. Kraft,[3] Zhongwu Lai,[3] Suzanna Lewis,[8] Christos Louis,[9] Carolina Barillas-Mury,[10] Deborah Nusskern,[3] Gerald M. Rubin,[8] Steven L. Salzberg,[11] Granger G. Sutton,[3] Pantelis Topalis,[9] Ron Wides,[12] Patrick Wincker,[13] Mark Yandell,[3] Frank H. Collins,[14] Jose Ribeiro,[15] William M. Gelbart,[16] Fotis C. Kafatos,[1] Peer Bork[1]

Comparison of the genomes and proteomes of the two diptera *Anopheles gambiae* and *Drosophila melanogaster*, which diverged about 250 million years ago, reveals considerable similarities. However, numerous differences are also observed; some of these must reflect the selection and subsequent adaptation associated with different ecologies and life strategies. Almost half of the genes in both genomes are interpreted as orthologs and show an average sequence identity of about 56%, which is slightly lower than that observed between the orthologs of the pufferfish and human (diverged about 450 million years ago). This indicates that these two insects diverged considerably faster than vertebrates. Aligned sequences reveal that orthologous genes have retained only half of their intron/exon structure, indicating that intron gains or losses have occurred at a rate of about one per gene per 125 million years. Chromosomal arms exhibit significant remnants of homology between the two species, although only 34% of the genes colocalize in small "microsyntenic" clusters, and major interarm transfers as well as intra-arm shuffling of gene order are detected.

The fruit fly *Drosophila melanogaster* (in the following, *Drosophila*) and the malaria mosquito *Anopheles gambiae* (in the following, *Anopheles*) are both highly adapted, successful dipteran species that diverged about 250 million years ago (*1*, *2*). They share a broadly similar body plan and a considerable number of other features, but they are also substantially different

in terms of ecology, morphology, life style, and genome size [the *Anopheles* genome is twice the size of that of *Drosophila* (*3–5*)]. A prominent difference is the ability of *Anopheles* to feed on the blood of specific hosts. Hematophagy is essential for the female mosquito to produce eggs and propagate; it also has been exploited by viruses and parasites that use *Anopheles* as a vehicle for transmission among vertebrates. Hematophagy is linked to specific host-seeking abilities as well as to nutritional challenges and requirements distinct from those of *Drosophila*. Here we aim to compare the two genomes as well as the derived proteomes to understand how they reflect the common and distinct features of the species.

Conservation of the Proteomes

Extent of similarity at the protein level. We first compared the genomes at the protein level, considering 12,981 deduced *Anopheles* proteins [out of 15,189 annotated transcripts (*5*), omitting transposon-derived or bacterial-like sequences and alternative transcripts]. The proteins were classified into four categories, according to their evolutionary relationships (Fig. 1). The first includes *Anopheles* proteins with one clearly identifiable counterpart in *Drosophila* and vice versa [1:1 orthologs (*6*)]. The function of these proteins is most likely conserved (*6*, *7*). We used two different approaches (reciprocal best matches and derivation of orthologous groups; see materials and methods) that produced similar results, identifying 6089 protein pairs as clear orthologs (that is, 47% of the *Anopheles* and 44% of the *Drosophila* proteins). The second category includes 1779 *Anopheles* proteins (Fig. 1) that belong to orthologous groups (*7*) in which gene duplication has occurred in one or both species after divergence [that is, paralogy (*6*)], resulting in "many-to-many" orthologs. The third category includes 3590 *Anopheles* predicted proteins (Fig. 1) that have homologs in *Drosophila* and/or other species but without easily discernable orthologous relationships (for example, homologs might only share a domain or be divergent members of larger families). A subset of this group in *Anopheles* consists of 1283 proteins (Fig. 1) that show little or no homology in *Drosophila* but instead have best matches to other species. Finally, for the remaining proteins (1437 in *Anopheles* and 2570 in *Drosophila*), no detectable homologs were found in any other species with a fully sequenced genome; these might be encoded by new or quickly evolving genes. These genes are clearly the shortest when compared to the genes of the other categories (Fig. 1).

All the above numbers and derived estimates are necessarily approximations. It must be emphasized that the annotation of genomes and proteomes is an ongoing effort and that various limitations here and elsewhere can lead to over- or underestimates affecting genes and particular biological systems (table S1). It is likely that, as in other animal genomes, some *Anopheles* genes have not been sequenced yet (they might be located in highly polymorphic regions or in highly repetitive contexts); the assembly has some errors [in *Anopheles*, two different and sometimes very divergent haplotypes caused considerable assembly difficulties (*5*)]; gene predictions are subject to a considerable error rate, in particular at the exon level; and homology-based analysis methods sometimes lack the sensitivity and selectivity required for precise statements (table S1). Yet current data and methods do produce results exceeding the 70% accuracy level (*8*), and thus general conclusions should be reasonably reliable.

The core of conserved proteins. The 1:1 orthologs (6089 pairs) can be considered the conserved core. Although automated gene predictions may sometimes be imperfect and incomplete [for instance, because of the presence of unannotated small exons (fig. S1 and table S1, footnote d)], identities are usually distributed throughout much of the length of the orthologs' sequence. The average sequence identity is 56%, as compared to 61% for the 7350 orthologs shared by the genomes of humans (*9*) and pufferfish (*10*), which diverged approximately 450 million years age (*10*). This indicates that insect proteins diverge at a higher rate than vertebrate proteins, possibly because insects have a substantially shorter life cycle, a different reproductive strategy, and a larger effective population size, and may experience different selective pressures.

Putative effects of selection are also evident in the wide range of sequence similarities among the 6089 orthologs of *Anopheles* and *Drosophila* (Fig. 2A). Differences in average sequence similarity are observed among 11 functional classes based on Gene Ontology (*11*) classification and manual assignment; proteins involved in immunity

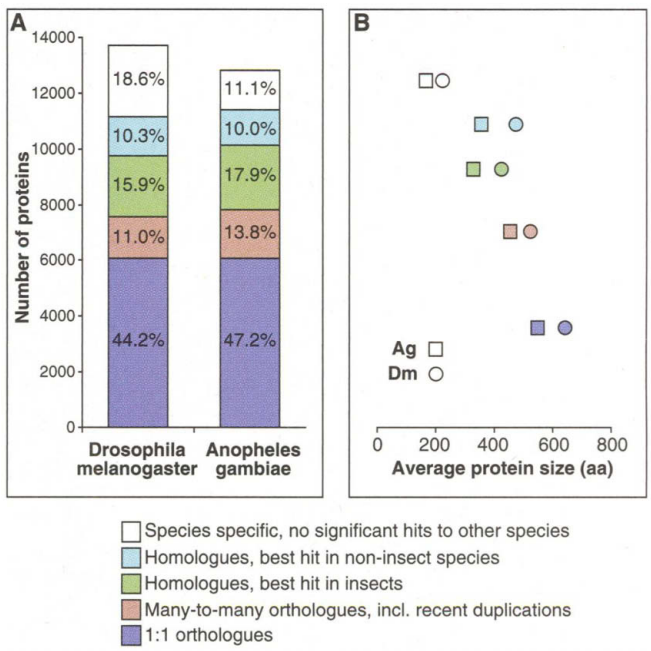

Fig. 1. Classification of proteins in *Anopheles* and *Drosophila* according to their evolutionary relationships. 13,885 *Drosophila* proteins from a preliminary version of FlyBase release 3 were compared to 12,981 proteins from the *Anopheles* sequencing project [in both species, only the best-matching transcript per gene was chosen and all entries flagged as "likely transposon" or "bacterial-like" (*5*) were omitted]. Orthology (*6*) was assigned by testing for triangles of reciprocal best matches in Smith-Waterman searches (*61*), aided by the information in other fully sequenced eukaryotic genomes and allowing for recent duplications (see materials and methods). (**A**) Classification of the proteins according to their conservation. (**B**) For each class, the average protein length is plotted (separately for the two species).

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Wellcome Trust Center for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [3]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [4]Department of Biological Sciences, Center for Molecular Microbiology and Infection, Imperial College, London SW7 2AZ, UK. [5]University of Arizona, Tucson, AZ 85721–0088, USA. [6]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [7]Institute of Cytology and Genetics, Sibirian Division of the Russian Academy of Sciences, 630090 Novosibirsk, Russia. [8]University of California, Berkeley, CA 94720–3200, USA. [9]Institute for Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, Post Office Box 1527, GR-711 10 Heraklion, Crete, Greece, and University of Crete, GR-711 10 Heraklion, Crete, Greece. [10]Colorado State University, Fort Collins, CO 80523–1671, USA. [11]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [12]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. [13]Genoscope/Centre National de Sequencage and CNRS-UMR 8030, 2 rue Gaston Cremieux, 91057 Evry Cedex 06, France. [14]Center for Tropical Disease Research and Training, University of Notre Dame, Notre Dame, IN 46556–0369, USA. [15]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, 4 Center Drive, MSC 0425, Bethesda, MD 20892–0425, USA. [16]Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.

show the highest divergence rates [see also (*12*)], and structural proteins are the most conserved (Fig. 2B).

Notwithstanding these indications of rapid gene divergence, the orthologous proteins constitute a core of conserved functions and contribute to basic biological processes. An example is genes involved in early embryonic development. Recent descriptions in *Anopheles albitarsis* (*13*) indicate that the basic events in early embryogenesis are conserved between *Drosophila* and *Anopheles*. In a compilation of 315 early developmental genes in *Drosophila* (fig. S2 and materials and methods), 251 genes showed a clear single ortholog in *Anopheles*, and manual processing added another 14 single matches. Thus, ~85% of the developmental genes have single orthologs: a much higher percentage than the 47% noted for the genome as a whole. The conservation of gene content is also seen in specific signaling pathways. For example, almost all members of the *decapentaplegic* signaling pathway are represented by individual orthologous genes: the upstream regulator (*dl*), the ligand (*dpp*), extracellular accessory proteins for shaping the ligand gradient (*sog* and *tok*), the receptors (*put*, *tkv*, and *sax*), the intracellular signaling partners (*mad* and *med*), and a downstream target (*shn*). Only two elements of the pathway appear to be missing in *Anopheles*: the negative regulator *brinker* and the ligand *scw*. *brinker* is dispensable for some instances of *dpp* signaling in *Drosophila* (*14*), so it might be a relatively recent addition to the pathway.

Family expansions and reductions. Differences in functions are suggested by increases and decreases in protein family sizes. They can indicate adaptations to environment and life strategies, leading to changes in cellular and phenotypic features. Family expansions can be measured in several ways depending on how narrowly a protein family is defined and what resolution is required. At a low resolution, an established measure is the difference in domain content of the genomes, as reflected in the InterPro resource (*15*, *16*), which contains manually curated domain collections such as PFAM (*17*) and SMART (*18*). A complementary approach is cluster analysis of homologous protein families in both *Anopheles* and *Drosophila*, which does not require the existence of annotated domains. A higher resolution is provided by the analysis of the many-to-many orthologs; these are less strictly defined than the "one-to-one" orthologs but can still be assigned to a single ancestral gene, thus implying duplications after speciation. An example of such an orthologous group is the epsilon subunit of the adenosine triphosphate–synthase complex. This subunit is encoded by two genes in both *Anopheles* and *Drosophila*; a phylogenetic tree of the protein sequences supports the interpretation that they shared a single-copy ancestral gene that was present at the time of speciation and was duplicated independently later (fig. S3).

The many-to-many orthologous groups reveal many uneven expansions or reductions; often a single protein in one of the two organisms has several counterparts in the other. By this measure, recent gene duplications seem to have occurred considerably more often in *Anopheles* than in *Drosophila* (fig. S4). Although this observation can partly be explained by assembly artifacts due to the two haplotypes in *Anopheles* (table S1, footnote b), numerous family expansions are unequivocal. Arthropod-specific genes encoding cuticular proteins, for example, are particularly dynamic in terms of duplications: A few of the genes present in the common ancestor sometimes gave rise to groups of 10 or more genes in one of the two species, partially balanced by losses in other branches of the family; however, overall the number of cuticle genes in *Anopheles* as compared to *Drosophila* is higher by one-third (Table 1).

For an unbiased view of broadly defined protein families and their expansions, we have tabulated differences in family sizes derived via single-linkage clustering (table S2). The most notable difference is a family of 27 hypothetical *Anopheles* proteins with no counterpart in *Drosophila*. Only a manual search for homology provided evidence for a distant similarity to helicases of the *DexD* subfamily (*19*). Although cluster analysis reveals that several uncharacterized protein families contribute to the phenotypic variations, a clearer picture of functional differences emerges through the comparison of known domain families (Table 1) (*5*). The most obvious one is a large expansion of mosquito proteins containing a domain resembling the COOH-terminus of the beta and gamma chains of fibrinogen (FBN) (Table 1 and table S2). FBN domains were found originally in human blood coagulation proteins but in invertebrates are thought to be involved mostly in the innate immune system (*20*–*24*). In order to quantify the expansion of the FBN family, we reconstructed the genes (many of those predicted appeared to be truncated artificially), identified additional members in genomic DNA (table S1, footnote e), and removed likely pseudogenes and allelic variants. A phylogenetic tree of the resulting 58 *Anopheles* and 13 *Drosophila* FBN genes revealed that they largely belong to two distinct species-specific clades (Fig. 3) and sur-

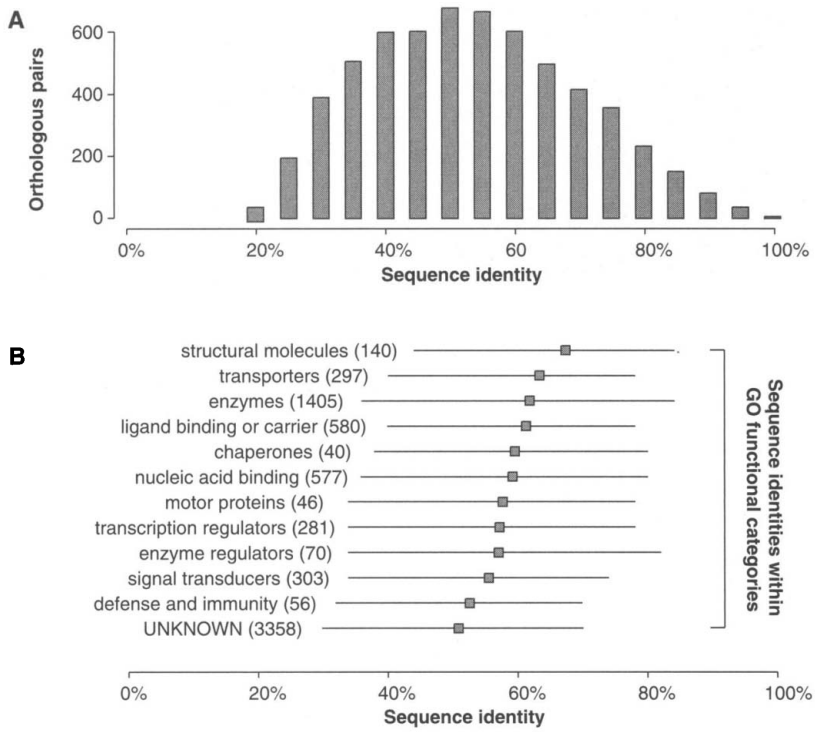

Fig. 2. Properties of 1:1 orthologs. **(A)** Histogram of sequence identities. Identities provide an intuitive estimate of conservation and selective pressure. Only five proteins were virtually identical (allowing for deviations at the termini): two histone proteins, a ribosomal protein, calmodulin, and adenosine diphosphate ribosylation factor. At the other extreme, most of the highly diverged sequences (identities <25%), are not characterized experimentally, indicating a bias in experimental analysis. **(B)** Sequence conservation by functional category. The average identity of orthologous sequences was computed separately for 11 different Gene Ontology categories (related to molecular function; in addition to Gene Ontology annotations, some categories were also populated manually; some proteins are counted in more than one category). Horizontal bars delineate the interval that covers 80% of the orthologous pairs in the category.

prisingly identified only two 1:1 orthologous relationships; the *Drosophila* representatives were the developmental protein encoded by *scabrous* and the uncharacterized CG9593 (which appears to be closely related to horseshoe crab tachylectin 5A). The massive expansion of the *Anopheles* gene family must be associated with particular aspects of the mosquito's biology, possibly hematophagy and exposure to *Plasmodium*. The blood meal imposes challenges associated with proliferation of the microbial flora in the gut and coagulation of ingested blood; the bacteria-binding properties of FBNs (*23*) may be important in controlling and/or aggregating bacteria in the midgut, or the mosquito may use a number of these proteins as anticoagulants (for instance, as competitive inhibitors preventing polymerization of blood FBN). Some mosquito FBN proteins are up-regulated by invading malaria parasites (*12*, *21*), suggesting a possible role in an antimalarial defense system.

Additional differences in gene family sizes are clearly evident (Table 1) [see (*5*) and other companion papers in this issue]. Only 6 of the 200 most frequent InterPro domain families, however, have statistically significant size differences (Table 1), indicating the overall similarity of domain content of the two proteomes. Nevertheless, small differences in family sizes can become biologically significant when a broader context is studied, such as metabolic pathways or multipathway systems such as immunity (*12*) or the management of oxidative stress.

A significant load of reactive oxygen species (ROS) is created by the tracheal respiratory system of insects and their exposure to ionizing ultraviolet radiation. Hematophagy represents an additional challenge, because blood meal–derived heme also results in ROS production. Therefore, we performed a species comparison with special emphasis on three biochemical pathways (fig. S5).

The abundant thiol tripeptide glutathione (GSH) can directly scavenge ROS but also functions as an oxidizable substrate for enzymes such as glutathione-dependent peroxidases and glutaredoxins, permitting efficient neutralization of peroxides and disulfide reduction. GSH also permits detoxification by glutathione *S*-transferases (GSTs) (*25*). Because the key enzyme for regeneration of GSH, glutathione reductase (GR), appears to be absent from both *Drosophila* and *Anopheles* (*26*) (Table 2), thioredoxins (Trx's) take over this role and are themselves regenerated by an NADPH-coupled enzyme, TrxR (NADPH, reduced nicotinamide adenine dinucleotide phosphate). Trx can also more directly reduce peroxides via thioredoxin peroxidase enzymes (TPx's). In this pathway, *Anopheles* not only has a smaller number of Trx genes (three versus seven, as compared to *Drosophila*) but also lacks the ortholog of the mitochondrial-specific TrxR-2 gene (*26*). The underrepresentation of GSH and Trx-utilizing enzymes in *Anopheles* is unexpected given the challenges resulting from hematophagy. However, microarray experiments indicate that certain mosquito TPx and GST enzymes are highly induced in female mosquitoes after the blood meal (*27*). In the second pathway [the conversion of superoxide anions to hydrogen peroxide and hence to O_2 and H_2O by the superoxide dismutase (SOD)/catalase system], differences in enzyme numbers are also minor: *Anopheles* has one less SOD gene and one extra catalase gene.

Major differences in gene numbers were only observed in the peroxidase (Px) system, which serves to nonspecifically catalyze the oxidation of diverse substrates. A Px isolated from the salivary glands of *Anopheles albimanus* has been implicated in blood feeding (*28*), and preliminary analysis indicates that Px's are important during the invasion of the mosquito midgut epithelium by malaria parasites (*29*). We identified 18 Px's in *Anopheles* as compared to only 10 in the *Drosophila* genome. The expanded family members cluster tightly with the salivary Px of *A. albimanus* (fig. S6). It is thus likely that *A. gambiae*, and possibly other mosquitoes, have been selected for additional copies of genes encoding such peroxidases as part of the adaptation to the blood-feeding process.

Gene genesis and gene loss. More remarkable than the expansion or reduction of family sizes is the genesis or loss of entire gene families. A total of 1437 predicted genes in *Anopheles* have no detectable homology with genes of other species; 522 of these have putative paralogs only within *Anopheles*, and 575 are supported by expressed sequence tag (EST) matches, including at least 26 genes expressed in the adult female salivary glands. The category of genes unique to either *Anopheles* or *Drosophila* probably contains a

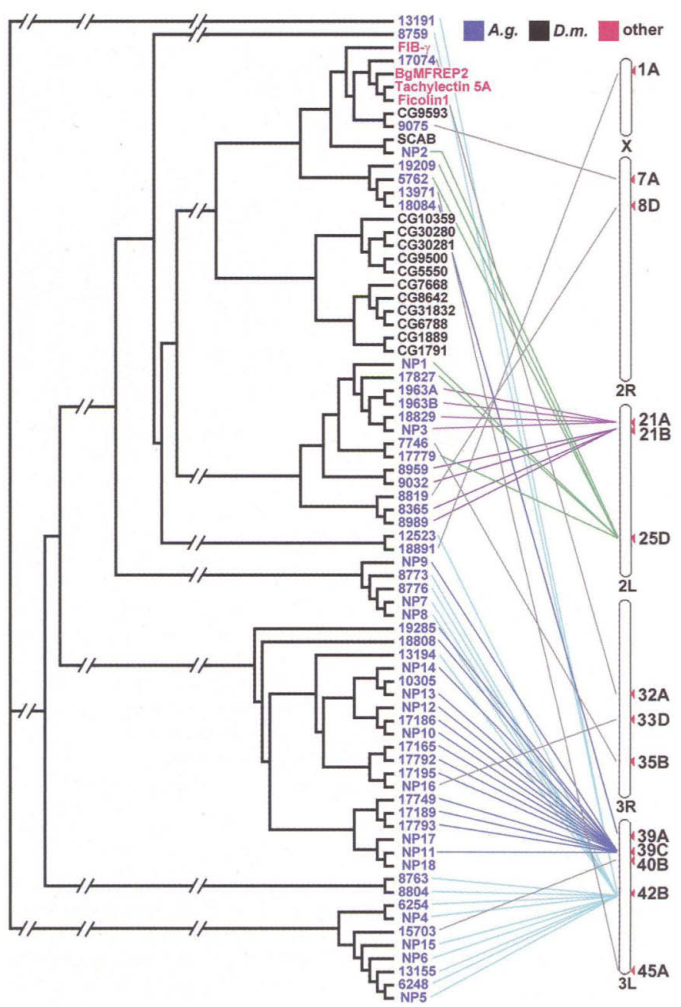

Fig. 3. Expansions of proteins with FBN-like domains in *Anopheles*. Annotated FBN proteins in *Anopheles* (40 sequences) and 18 additional FBN sequences deduced from the genome (see table S1 for their genomic coordinates) were aligned to 13 *Drosophila* homologs (prefix CG, and Scabrous), human Ficolin1 (protein accession number BAA12120), human FBN gamma (protein accession number P02679), *Tachypleus tridentatus* tachylectin 5A (protein accession number BAA84188), and *Biomphalaria glabratas* BgMFREP2 (protein accession number AAK13550), and a sequence divergence tree was built (*Anopheles* genes of known cytogenetic locations are linked to their respective map positions with lines of the same color for each chromosomal location). For *Anopheles*, only the last five digits of the ENSEMBL gene IDs are indicated. Correlation of sequence divergence, exon/intron organization, and chromosomal location is apparent for the *Anopheles* members. Of the 20 annotated FBN genes mapping to 39C and 42B of the third chromosome, only two appear to have introns, whereas the majority of the other annotated members have introns. Of the 13 *Drosophila* members, only two pairs of closely related genes, CG30280/CG30281 and CG1889/CG1791, map to the same position, 2R-Div 58D2 and X-Div 9A3, respectively.

mixture of previously existing genes mutated beyond recognition [numerous cases have been reported for *Drosophila* (*30*)] as well as genes arising through an ongoing high rate of gene genesis (*31*). Only 84 of the 2570 genes unique to *Drosophila* have functional annotations mapped to Gene Ontology terms; among those are small quickly evolving proteins such as neuropeptides or antibacterial peptides but also a number of proteins implicated in the formation of the chorion or the puparium.

A simple strategy for identifying gene losses is to search for genes that are present in only one of the two insects but that do have orthologs in other species. Although it is difficult to prove loss at this stage of the sequencing and assembly (table S1, footnote f), many of the observed cases seem biologically relevant. For example, four *Anopheles* paralogs without a counterpart in *Drosophila* are similar to a human gene encoding leukotriene B4 12-hydroxy dehydrogenase, an enzyme that can inactivate the proinflammatory leukotriene B4. It is tempting to speculate that *Anopheles* has retained or acquired this gene to interfere with inflammatory reactions in the human host. Other genes found in *Anopheles* have been entirely limited to vertebrates so far (they are absent from *Drosophila*, *Caenorhabditis elegans*. and *Saccharomyces cerevisiae*); a specific example is the human *cabin1* gene, which encodes a calcineurin-binding intracellular regulatory protein implicated in controlling T cell apoptosis (*32*, *33*), a process limited to vertebrates. The presence of a clear *Anopheles* ortholog to *cabin1* implies that the functional spectrum and phylogenetic breadth of this gene family are probably much wider than initially reported.

Multiple losses and gains of genes can also be revealed by analyzing orthology across several species. The observed phylogenetic distribution of the orthologs (Fig. 4) is largely in agreement with the current consensus on eukaryotic phylogeny, with deviations indicating the prevalence of gene loss in the various species. In particular, any widespread orthologs missing from both *Anopheles* and *Drosophila* (Table 2) are putative insect-specific gene losses and may be associated with distinct features of insect physiology. For example, the absence of several enzymes involved in sterol metabolism (Table 2) reflects the known inability of insects to synthesize sterols (*34*). Similarly, the requirement for niacine/nicotinic acid (*35*) is reflected by the absence of three enzymes needed in a pathway leading to nicotinate (Table 2). Another intriguing finding is the absence of the DNA repair enzyme uracil-DNA glycosylase. This enzyme is required in organisms in which genomic DNA is methylated at cytosine residues, because methylation can lead to spontaneous deamination of cytosine to uracil, which then needs to be removed. *Drosophila* has long been known to have no or only very little DNA methylation (*36*), and it would seem that it shares this feature with *Anopheles*, suggesting that either DNA methylation is absent in most if not all insects or that another enzyme family in insects took over the role of uracil-DNA glycosylase.

A total of 579 orthologs are restricted to *Anopheles* and *Drosophila* (they do not even share domains or short motifs with genes in other organisms), and these should help determine insect-specific features. So far, only about 100 of these have been functionally annotated in *Drosophila*. Many are predicted to code for specific odorant and taste receptors, cuticle proteins, pheromone and pheromone-binding proteins, and insect-specific defense molecules (such as prophenoloxidase and antibacterial proteins and peptides).

Comparison of pseudogene content. The dynamics of gene content evolution also are

Table 1. The 20 most significantly differing InterPro families. The 20 most significant expansions or reductions of *Anopheles* families as compared to *Drosophila* families are indicated (out of the 200 largest families), sorted by significance. Statistically significant expansions at a P value level of 10^{-3} (bold text) and 10^{-1} (bold italic text) are indicated. The significance is estimated by means of a chi square test with respect to the total number of genes in the genomes and Dunn-Sidak corrections (*66*). The background of human, pufferfish, and *C. elegans* family sizes is given. Shown are the total numbers of genes matching a signature, the percent of the total number of genes in that genome, and the rank of the family size as compared to others (in parentheses). Families with considerable fractions of proteins that are viral or transposon-derived are marked in italics at left (see also table S1, footnote e).

InterPro name	*A. gambiae*	*D. melanogaster*	*H. sapiens*	*F. rubripes*	*C. elegans*
IPR000477: RNA-directed DNA polymerase (reverse transcriptase)	**87/0.7% (17)**	**13/0.1% (163)**	165/0.7% (19)	256/0.8% (14)	63/0.3% (42)
IPR001878: Zn-finger, CCHC type	**89/0.7% (15)**	**28/0.2% (62)**	43/0.2% (86)	66/0.2% (62)	44/0.2% (62)
IPR002181: FBN, beta/gamma chain, COOH-terminal globular	**46/0.4% (36)***	**10/0.1% (218)***	24/0.1% (152)	39/0.1% (119)	5/0.0% (453)
IPR001254: serine protease, trypsin family	**305/2.3% (2)**	**206/1.5% (4)**	110/0.5% (29)	125/0.4% (27)	13/0.1% (202)
IPR004822: histone-fold/TFIID-TAF/NF-Y domain	***48/0.4% (33)***	***14/0.1% (155)***	105/0.4% (30)	53/0.2% (82)	88/0.5% (26)
IPR005135: endonuclease/exonuclease/ phosphatase family	***41/0.3% (39)***	***14/0.1% (147)***	140/0.6% (24)	43/0.1% (109)	27/0.1% (105)
IPR002126: cadherin domain	41/0.3% (38)	17/0.1% (125)	114/0.5% (27)	168/0.5% (23)	16/0.1% (161)
IPR001584: integrase, catalytic domain	18/0.1% (114)	4/0.0% (490)	10/0.0% (412)	72/0.2% (61)	20/0.1% (134)
IPR000301: CD9/CD37/CD63 antigen	15/0.1% (145)	36/0.3% (39)	27/0.1% (135)	48/0.2% (93)	20/0.1% (135)
IPR001969: eukaryotic/viral aspartic protease, active site	27/0.2% (62)	11/0.1% (187)	17/0.1% (246)	19/0.1% (233)	22/0.1% (122)
IPR003006: immunoglobulin/major histocompatibility complex	177/1.4% (6)	135/1.0% (6)	675/2.8% (3)	542/1.7% (2)	80/0.4% (32)
IPR002893: Zn-finger, MYND type	35/0.3% (47)	17/0.1% (118)	14/0.1% (298)	25/0.1% (175)	10/0.1% (245)
IPR000618: insect cuticle	133/1.0% (9)	99/0.7% (12)	0/0.0% (–)	0/0.0% (–)	0/0.0% (–)
IPR001599: alpha-2-macroglobulin	17/0.1% (126)	6/0.0% (331)	14/0.1% (294)	15/0.0% (293)	1/0.0% (1426)
IPR002890: alpha-2-macroglobulin, NH_2-terminal	15/0.1% (138)	5/0.0% (383)	13/0.1% (324)	16/0.1% (278)	1/0.0% (1192)
IPR001611: leucine-rich repeat	151/1.2% (7)	117/0.9% (9)	218/0.9% (13)	288/0.9% (10)	60/0.3% (47)
IPR000863: sulfotransferase	22/0.2% (93)	10/0.1% (197)	23/0.1% (155)	43/0.1% (110)	5/0.0% (452)
IPR003662: general substrate transporter	68/0.5% (28)	95/0.7% (14)	56/0.2% (59)	80/0.3% (51)	84/0.4% (28)
IPR002085: zinc-containing alcohol dehydrogenase	19/0.1% (106)	10/0.1% (199)	23/0.1% (161)	21/0.1% (207)	13/0.1% (211)
IPR001594: Zn-finger, DHHC type	11/0.1% (194)	21/0.2% (87)	20/0.1% (196)	28/0.1% (157)	16/0.1% (175)

*Further manual analysis (see Fig. 3) reveals that the actual numbers are 58 and 13 in *Anopheles* and *Drosophila*, respectively.

evident in the number, kind, and location of detectable pseudogenes. Searches in all predicted intergenic regions identified 4163 segments in *Anopheles* with significant sequence similarity to known proteins but only 1075 in *Drosophila*. These segments correspond to overlooked (parts of) genes or to pseudogenes. Among them are 166 and 176 sequences in *Anopheles* and in *Drosophila*, respectively, that appear to be clear pseudogenes because they present detectable open reading frame (ORF) disruptions (stop codons and/or frame shifts). The counts suggest a roughly similar pseudogene content in both genomes, despite an approximately twofold difference in genome size. This would deviate from the general belief that the rate of noncoding DNA loss (expected to be negatively correlated to pseudogene content) is determined by genome size constraints (*37*).

For a more reliable estimate of the prevalence of pseudogenes in both species, we analyzed the ratio of nonsynonymous (K_a) to synonymous (K_s) substitutions in all intergenic regions with similarity to known proteins (see materials and methods). K_a/K_s ratios tend to be around one for pseudogenes and are lower for functional genes, because mutations leading to amino acid replacements with functional consequences are selected against (*38*). Our estimates of the number of neutrally evolving pseudogenes range from 439 to 1319 in *Anopheles* and from 162 to 396 in *Drosophila*. Although from these ranges we cannot postulate a twofold difference in pseudogene content between the two genomes, the number of pseudogenes in both species is clearly higher than indicated by the ORF disruption counts. In any case, the pseudogene content in *Anopheles* and *Drosophila* seems considerably lower than those of the mouse and human genomes, in which the same method identifies (with an associated error margin of <5%) far more than 10,000 pseudogenes (*39*).

Dynamics of Gene Structure

Intron gain and loss. Pairwise alignment of the 6089 1:1 orthologous genes provides unequivocal support for the conclusion that *Drosophila* has experienced a reduction of noncoding regions (*5*, *40*); equivalent introns in *Drosophila* have only half the length of *Anopheles*, whereas exon lengths and intron frequencies are roughly similar (Table 3). There are also considerable differences in intron positions. In only 394 out of 5196 orthologous gene pairs having one or more introns, the positioning of all the introns agrees down to the base pair (half of these genes have only one intron). In total, 11,007 out of 20,161 *Anopheles* introns in 1:1 or-

Table 2. Gene losses in insects. The genes shown are absent in both *Anopheles* and *Drosophila* but are present in other eukaryotes (in a pattern that implies losses in the insect lineage, or earlier, as opposed to gains in other lineages: Genes must be present in at least one animal but also in fungi or plants). Only genes with functional annotations are shown, limited to clear cases. Eukaryotic genomes are indicated as follows: D, fruit fly (*D. melanogaster*); A, mosquito (*A. gambiae*); P, plant (*Arabidopsis thaliana*); Y, yeast (*S. cerevisiae*); W, worm (*C. elegans*); H, human (*Homo sapiens*); M, mouse (*Mus musculus*). Enzyme Commission (EC) numbers are indicated in parentheses.

	D	A	P	Y	W	H	M
Sterol metabolism							
Squalene monooxygenase (EC:1.14.99.7)	–	–	x	x	–	x	x
7-Dehydrocholesterol reductase (EC:1.3.1.21)	–	–	x	x	x	x	x
Farnesyl-diphosphate farnesyltransferase (EC:2.5.1.21)	–	–	x	x	–	x	x
Lanosterol synthase (EC:5.4.99.7)	–	–	x	x	–	x	x
Lanosterol synthase (EC:5.4.99.7)	–	–	x	x	–	x	x
3-Oxo-5-alpha-steroid 4-dehydrogenase 1 (EC:1.3.99.5)	–	–	x	–	x	x	x
C-5 sterol desaturase (EC:1.3.3.2) Ergosterol biosynthesis	–	–	x	x	–	x	x
Cytochrome P450 P51, sterol 14-alpha demethylase	–	–	x	x	–	x	x
Diminuto/24-dehydrocholesterol reductase ("seladin1")	–	–	x	–	x	x	x
Biosynthesis of NAD							
Kynureninase (EC:3.7.1.3)	–	–	–	x	x	x	x
3-Hydroxyanthranilate 3,4-dioxygenase (EC:1.13.11.6) synthesis of excitotoxin quinolinic acid	–	–	–	x	x	x	x
Quinolinate phosphoribosyltransferase (EC:2.4.2.19)	–	–	x	x	–	x	x
DNA methylation and repair							
DNA (cytosine-5)-methyltransferase 1*	–	–	x	–	–	x	x
Uracil-DNA glycosylases	–	–	x	–	x	x	x
DNA-(apurinic or apyrimidinic site) lyase (EC:4.2.99.18)	–	–	–	x	x	–	–
Others							
Histidine ammonia-lyase (EC:4.3.1.3)†	–	–	x	–	x	x	x
Guanidinoacetate *N*-methyltransferase (EC:2.1.1.2) creatine biosynthesis‡	–	–	x	x	–	x	x
Threonine synthase (EC:4.2.3.1)§	–	–	–	x	–	x	x
Glutathione reductase (EC:1.6.4.2)	–	–	x	x	x	x	x
Putative aspartyl aminopeptidase (EC:3.4.11.21)	–	–	x	x	x	x	x
Dihydroxyacetone kinase 1 (EC:2.7.1.29) (glycerolipid metabolism)	–	–	x	x	x	x	x
Peroxisomal acyl-coenzyme A thioester hydrolase (EC:3.1.2.2)	–	–	x	x	x	x	x
Cockayne syndrome WD-repeat protein CSA	–	–	–	x	–	x	x
Indoleamine 2,3-dioxygenase (EC:1.13.11.42)	–	–	–	x	–	x	x
3-mercaptopyruvate sulfurtransferase (EC:2.8.1.2) (yeast 2.8.1.1)	–	–	x	x	–	x	x
Hypoxanthine-guanine-phosphoribosyl transferase (EC:2.4.2.8)	–	–	–	–	x	x	x
Aquaporin	–	–	–	x	x	x	x
Ornithine carbamoyltransferase (EC:2.1.3.3)	–	–	x	x	–	x	x
Glutamate carboxypeptidase 2	–	–	x	x	x	x	x
Malonyl-CoA decarboxylase	–	–	x	–	x	x	x

*An atypical form of DNA methyltransferase is present in both insects [with similarity to mammalian Dnmt2, for which methyltransferase activity has not yet been detected (*36*)]. †Apparent lack of a main catabolic route of histidine. Accordingly, mosquitoes excrete a large amount of histidine in the feces after a blood meal (*67*). In addition, insects may not need to degrade much histidine because they use it heavily: histidine constitutes the main pH buffer in the hemolymph of insects. ‡Phosphoarginine, and not phosphocreatine, is the principal reserve of high-energy phosphate compounds in insect muscle (*68*). The normal pathway for creatine synthesis appears absent; alternative routes may remain to be discovered. §Threonine is an essential amino acid in insects.

thologs have equivalent positions in *Drosophila*; conversely, almost 10,000 introns have either been lost or gained. The exact number depends on the extent of lateral intron movement ["intron sliding" (*41–43*)]. Our analysis reveals this effect to be smaller than 1% [when allowing introns to slide up to 10 base pairs (bp) (fig. S7 and table S3)]. Thus, because about 5000 genes were considered in species with 250-million-year (My) divergence time, it follows that about one intron has been gained or lost per gene per 125 My.

Table 3. Gene structure comparison of 6089 orthologous gene pairs.

	Anopheles	*Drosophila*
Average protein length (amino acids)	548	649
Average intron size (bp)*	1,061	628
Average coding exon size (bp)	366	443
Total number of exons	27,380	30,762
Total number of introns	21,279	24,605
Total coding exon length (bp)	10,009,635	13,635,856
Total intron length (bp)*	22,572,174	12,861,230
Average number of introns per gene	3.47	4.67

*Only introns between coding exons are considered.

The intron/exon structure appears to be more conserved when alternative splicing is involved. An example is the *Drosophila Dscam* gene, which has been reported to encode up to 38,000 proteins through extensive alternative splicing (*44*). This is possible because there are three different cassettes of duplicated exons that can generate exponential combinations of splice variants (*44*) (Fig. 5). Because only one gene product is annotated in both species compared, we used an algorithm for the detection of exon duplications (*45*) to confirm that the numbers of exons within the cassettes are at least similar in *Anopheles*. The intervening nonduplicated exons (black in Fig. 5) show a larger degree of intron gain or loss. Although a large-scale study is required, alternative splicing seems to be conserved in both species in several examined cases. For example, all 15 known splice forms in the myosin heavy chain (*46*) have counterparts in their *Anopheles* orthologs, as revealed by genomic structure comparison, alignment of each splice variant, and EST mapping (fig. S8).

Variability of noncoding regions. Introns are expected to diverge rapidly (*47*), and indeed only 160 (1.7%) of the 9632 introns in equivalent positions showed significant sequence similarity (below the default BLAST threshold of $E = 0.01$). Similarly, an analysis of 5' and 3' untranslated regions (UTRs) of all 6089 1:1 orthologs (operationally defined as 10,000 bp to the 5' or 3' ends of terminal exons) only revealed sequence homologies in 228 5' UTR regions (3.74%) and 243 3' UTR regions (3.99%). We also searched for homology in 547 intergenic regions between pairs of orthologs that remained closely linked (see below). Of these regions, 57 (10.42%) had sequence similarity that had not been detected in the searches mentioned above.

Altogether, only 687 matches between corresponding potentially noncoding genomic regions have been observed. However, as many as 55 of these (8%) are similar to proteins; that is, they are likely to encode parts of genes or pseudogenes. In the remaining 632 matches, additional coding sequences and noncoding RNAs are likely to be contained. Thus, less than 3% of the areas compared contain conserved noncoding regions (most of which are short), supporting the fast divergence of noncoding DNA. Overall, fewer noncoding regulatory regions are conserved between the two diptera than between pufferfish and mammals (*10*, *48*, *49*).

Extent of Genome Rearrangements

Microsynteny. At this evolutionary distance it can be expected that, in addition to changes in gene (intron/exon) structure, genome structure may vary greatly, to the extent that only small regions of conserved gene neighbor-

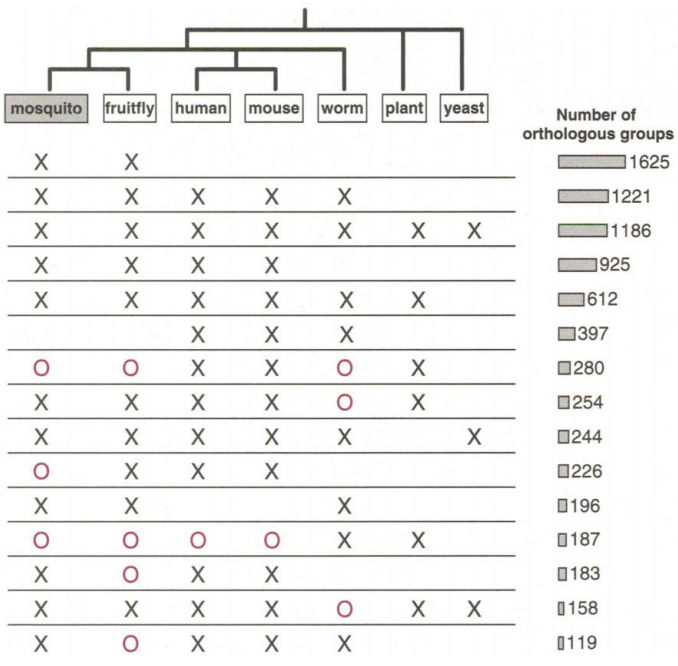

Fig. 4. Ortholog taxonomy. The 15 most frequent phylogenetic distributions of orthologous groups (with relevance to insects) are shown. The figure accounts for more than 85% of orthology assignments in the *Anopheles* proteome. Red circles indicate phylogenetic distributions that deviate from the common consensus of eukaryotic phylogeny, indicating putative losses and/or genes missed during the sequencing and annotation process. The tree shown on top of the figure has some nonbifurcating areas, because the exact location of *C. elegans* (or yeast/plants) is still under debate (*63*). Closer inspection of the patterns of losses provides some support for grouping *C. elegans* with arthropods [as stated by the ecdysozoan theory (*64*)]: For example, the most frequent loss pattern (row 7) is more parsimonious when placing *C. elegans* with insects, as this requires only single losses, whereas otherwise double losses are required. However, when less emphasis is placed on the parsimony of losses and more emphasis on the amount of shared genes, *C. elegans* does not group with arthropods (*65*) (Note that both approaches might be dominated by niche or life-style adaptations that do not always correspond to common ancestry.)

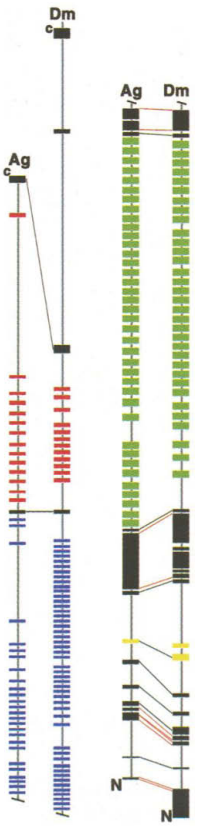

Fig. 5. Comparison of the gene structures of *Dscam* genes. *Dscam* has been reported to encode up to 38,000 distinct proteins in *Drosophila* through alternative splicing from cassettes of duplicated exons [blue, red, and yellow (*44*)]. The structure of the noncassette exons (black) has been modified by intron losses and insertions, but the alternative splicing cassettes are conserved. Thus, *Anopheles Dscam* is probably able to code for the same or a highly similar number of proteins. The trend of longer introns in *Anopheles* also applies to *Dscam*.

hood will be retained [this is referred to as microsynteny (*50*)]. Although almost intuitive by manual inspection (*51*), microsynteny is difficult to define and any assignment is operational. For the detection of conserved gene order within a species, triples of homologous genes have been used previously (*5*, *52*, *53*). For comparisons between the two genomes, we chose a set of criteria that should be both more sensitive and more selective at this evolutionary distance. In brief, we first required neighborhood conservation of two homologs, allowing no more than five unrelated genes in the intervening DNA; this resulted in 7992 candidate microsynteny regions. The additional requirement of having at least two orthologous groups (1:1 or many-to-many orthologs) within such a region reduced the number to 948 confirmed microsynteny blocks. The largest of these contained 8 and 31 homologous genes in *Anopheles* and *Drosophila*, respectively, including 7 orthologous groups; others contained up to 12 orthologous groups (Fig. 6). Most of the microsynteny blocks are much smaller and show substantial variation in gene content as well as evidence of numerous local inversions, translocations, and gene duplications (Fig. 6). In total, 4099 *Anopheles* genes (2962 orthologs) and 4244 *Drosophila* genes (2866 orthologs) were assigned to the 948 confirmed microsynteny blocks. We consider as the best measure of partially retained local neighborhood the fraction of orthologs that remain within confirmed microsynteny blocks; this amounts to about 34% in *Anopheles*, representing a significant level of highly local neighborhood conservation. Again, the fraction is considerably lower than the corresponding one for pufferfish and humans and supports the faster radiation of insects as compared to vertebrates. At the microsynteny level, we did not detect any obvious, recent segmental duplications within the *Anopheles* genome that would involve more than two orthologous groups (but see table S1, footnote h, for artificial duplications).

Chromosome mapping. We examined the similarity of chromosomal arms in the two species and the degree of long-range conservation of gene arrangements within corresponding arms (macrosynteny). Both *Anopheles* and *Drosophila* have five major chromosomal arms (X, 2L, 2R, 3L, and 3R, plus a small chromosome 4

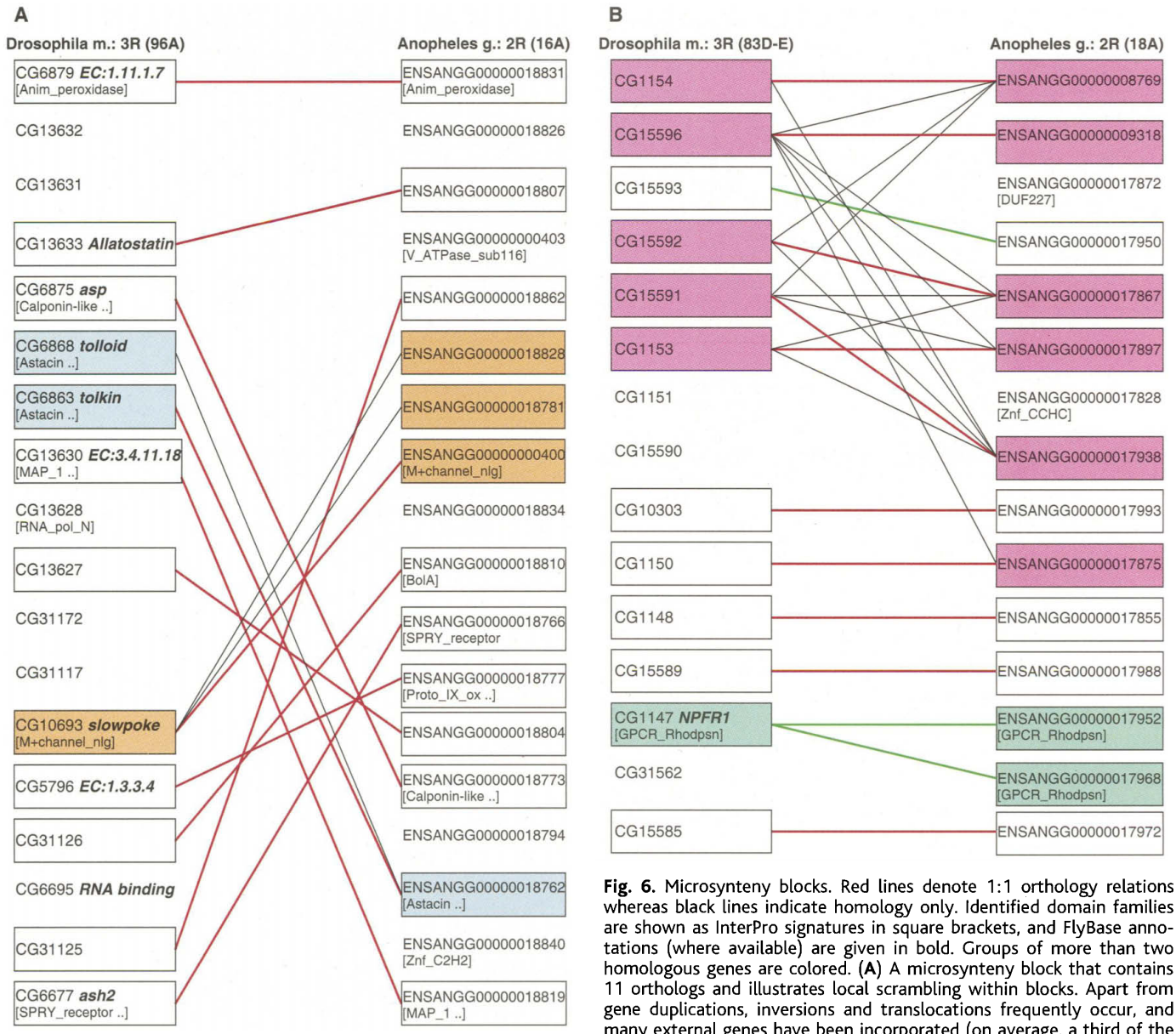

Fig. 6. Microsynteny blocks. Red lines denote 1:1 orthology relations whereas black lines indicate homology only. Identified domain families are shown as InterPro signatures in square brackets, and FlyBase annotations (where available) are given in bold. Groups of more than two homologous genes are colored. (**A**) A microsynteny block that contains 11 orthologs and illustrates local scrambling within blocks. Apart from gene duplications, inversions and translocations frequently occur, and many external genes have been incorporated (on average, a third of the genes in a block have no local correspondence). (**B**) One of the longest and most conserved (in terms of gene order) microsynteny blocks containing 13 orthologs in *Anopheles* chromosome 2R corresponding to a section of *Drosophila* chromosome 3R.

in *Drosophila melanogaster*). In the genus *Drosophila*, reassortment of recognizable chromosomal arms occurs by fission and fusion at the centromeres (*53*). To study the degree of common ancestry among the *Anopheles* and *Drosophila* chromosomes, we mapped the 6089 1:1 orthologs and the 948 microsynteny blocks onto the chromosomal arms. The statistical significance of the mapping (Fig. 7) permitted clear assignments, most of which were confirmed by both data sets, although the microsynteny mapping showed less significance because of fewer data points (Fig. 7). The predominant 1:1 homologies between the chromosomal arms of the two diptera have been inferred previously (*54*) and, with both genomes completed, can now be confirmed by analysis of homologous protein sequences. In addition, remnants of synteny and

Ag \ Dm	2L (2453)	2R (2692)	3L (2616)	3R (3405)	4 (82)	X (2260)
2L (2672)	63 (227)	501 (251)	674 (260)	82 (320)	4 (8)	91 (188)
	0 (40)	91 (37)	137 (42)	6 (51)	0 (0)	5 (16)
2R (3590)	106 (305)	136 (337)	218 (349)	1049 (430)	10 (11)	332 (252)
	4 (53)	6 (50)	23 (56)	199 (68)	0 (0)	35 (22)
3L (2105)	59 (179)	353 (198)	270 (205)	91 (252)	8 (6)	115 (148)
	4 (31)	51 (29)	39 (33)	5 (40)	0 (0)	4 (13)
3R (2523)	832 (214)	180 (237)	54 (245)	106 (302)	4 (8)	65 (177)
	182 (37)	28 (35)	5 (40)	5 (48)	0 (0)	0 (15)
UNKN (971)	14 (82)	24 (91)	19 (94)	54 (116)	2 (3)	19 (68)
	2 (14)	6 (13)	1 (15)	8 (18)	0 (0)	1 (6)
X (1071)	27 (91)	22 (100)	24 (104)	168 (128)	13 (3)	288 (75)
	2 (16)	1 (15)	0 (16)	25 (20)	3 (0)	36 (6)

Fig. 7. Mapping of orthologs and microsynteny blocks to chromosomal arms in *Anopheles* and *Drosophila*. Significant assignments are indicated in pink, and significant avoidances are in yellow (the increasing intensity of the colors marks *P* value cutoffs at 10^{-1} and 10^{-3}). Significance is conservatively estimated by the chi square test with respect to the number of genes on the smallest chromosomal arm. Shown are numbers of observations and random expectations in brackets of shared 1:1 orthologs at the top and microsynteny blocks below.

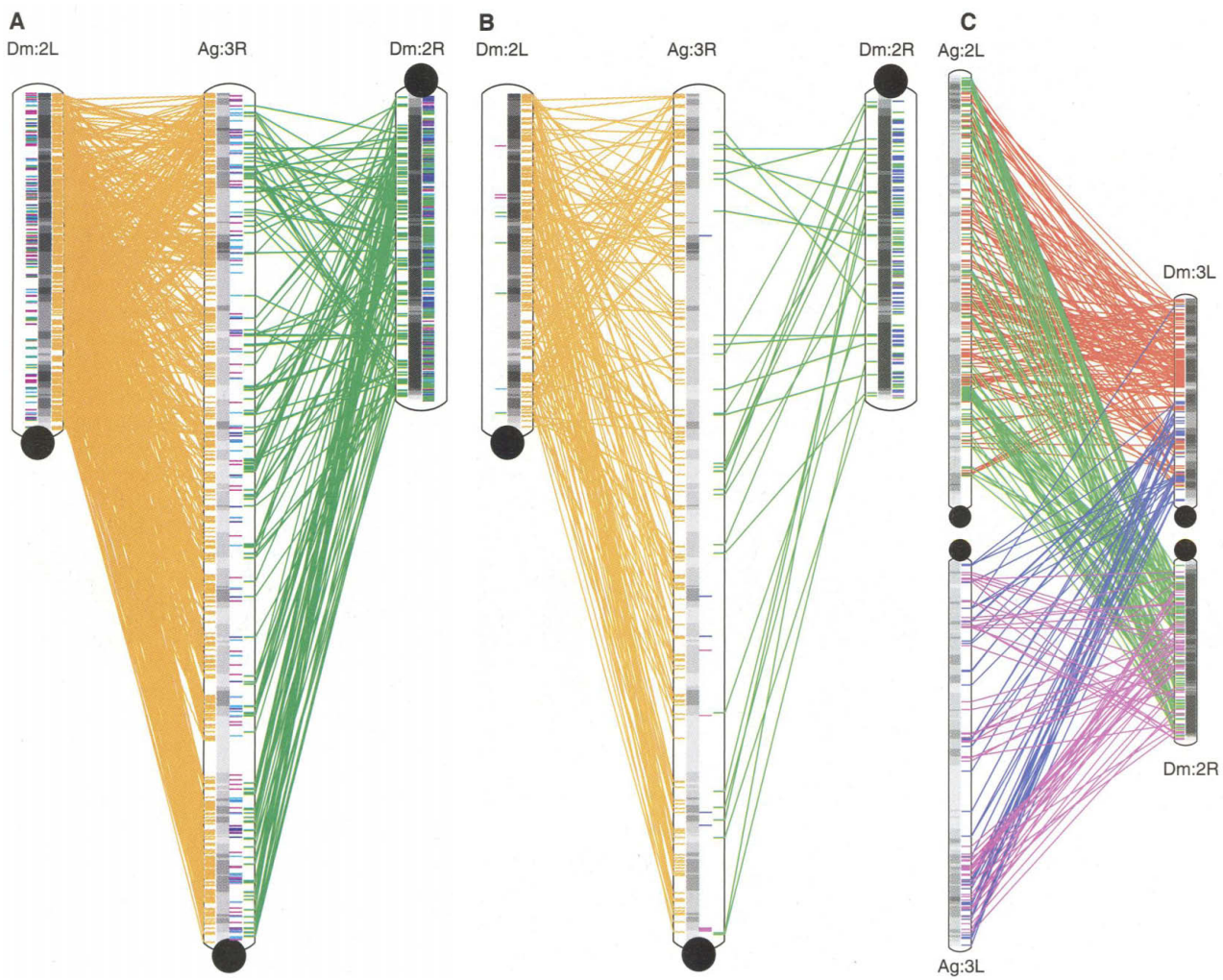

Fig. 8. Chromosome mapping. Significant mapping of *Anopheles* 3R and *Drosophila* 2L chromosomes in comparison to a nonsignificant mapping of *Drosophila* 2R based on (**A**) 1:1 orthologs and (**B**) microsynteny blocks. The gene density along the chromosomal arms is shown by the intensity of gray, calculated with a sliding window of 1 Mb. (**C**) One of two complex chromosome mappings involving four chromosomal arms. It illustrates a large segment of *Anopheles* chromosome 3L that corresponds to parts of *Drosophila* chromosome 3L and is probably the most recent segmental shuffling between the chromosomes of both species. Centromeres are illustrated by black dots.

the distribution of orthologs (Fig. 7) reveal a more detailed and complex relationship.

The most conserved pair of chromosomal arms is *Dm*2L and *Ag*3R, with 76% of the orthologs and 95% of microsynteny blocks in *Dm*2L mapping to *Ag*3R (table S4 and Fig. 8, A and B). The remaining genes and blocks represent exchanges with other arms (Fig. 9), but none of these show a statistically significant signal above a random expectation. The opposite is also significant, in that 67% of the *Ag*3R orthologs and 83% of its microsynteny blocks map onto *Dm*2L. For other chromosomal arms, dual correspondences are detected, each with two arms of the other species (Figs. 7 and 8, figs. S9 and S10, table S4). Thus, judging by the content of orthologous pairs, the *Anopheles* 2L chromosome arm harbors approximately 42 and 54% of the gene contents of the *Drosophila* 2R and 3L chromosome arms, respectively. Other relationships are *Ag*2R to *Dm*3R (70%) and *Dm*X (37%), as well as *Ag*3L to *Dm*2R (30%) and *Dm*3L (22%).

Significant portions of the *Anopheles* X chromosome appear to have been derived from what are presently autosomal *Drosophila* chromosome segments: the largest representing 11% of *Dm*3R and 33% of *Dm*4. (However, smaller fractions from each of the other *Drosophila* autosomal arms are also found on the *Anopheles* X chromosome; conversely, some of the *Drosophila* X chromosomal genes are found dispersed on the various *Anopheles* autosomal arms.) Such translocations between autosomes and chromosome X are not easy to explain, as the originally autosomal genes need to come under the control of the necessary dosage compensation system to equalize their activity in the homogametic and heterogametic sexes. However, studies in *Drosophila* have shown that the protein-RNA dosage compensation complex has fewer than 100 entry sites on the X chromosome and spreads from there in cis to "paint" the hyperactivated chromosome (*56*, *57*). If this mechanism of dosage compensation has been conserved in the *Anopheles* lineage, it would explain the apparent acceptability of gene migration between the X chromosome and autosomes, because X-inserted autosomal segments would acquire dosage compensation due to neighboring nucleation sites, whereas X chromosome sequences that have translocated to an autosome would lose dosage compensation unless the translocation included one of these sites. We examined the *Anopheles* genome for the presence of all the components known to be necessary for dosage compensation in *Drosophila*, namely the five proteins MLE, MOF, MSL-1, MSL-2, and MSL-3 and two noncoding RNAs, roX1 and roX2 [reviewed in (*58*, *59*)]. Single orthologs for four of the five protein components were readily identified within the *Anopheles* predicted proteome, and an ortholog of the fifth component was identified by homology searches at the level of genomic DNA. Neither noncoding RNA gene from *Drosophila* showed any evidence of similarity within the *Anopheles* genome. It remains to be determined whether noncoding RNA components are also present (either highly diverged versions of roX1 and/or roX2 or components of independent origin). However, the basic protein machinery of the dosage compensation complex is conserved between *Drosophila* and *Anopheles*, presumably facilitating flexibility in the evolution of the sex chromosome.

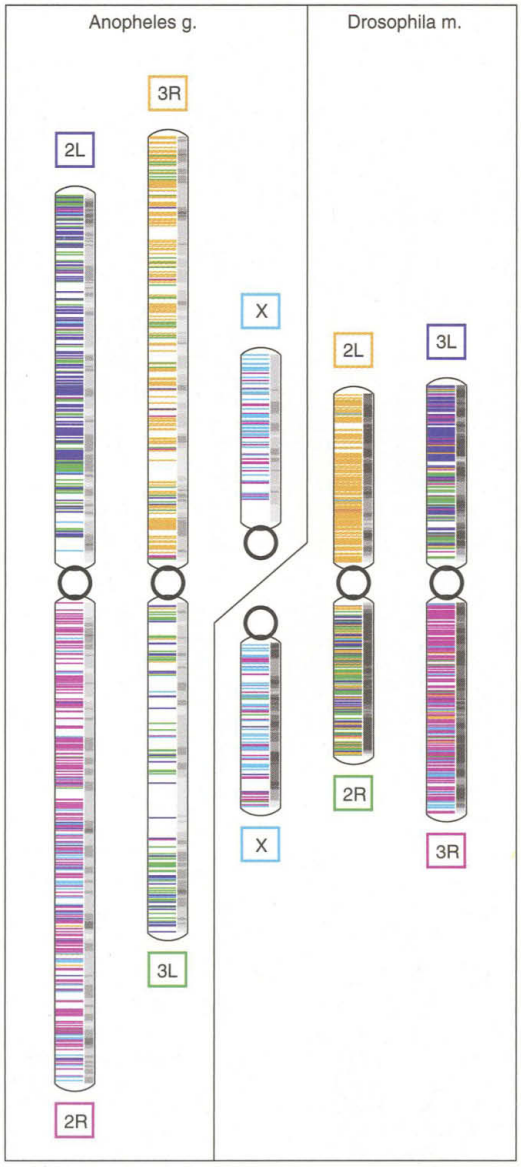

Fig. 9. Homology of chromosomal arms. Each chromosomal arm is marked by a color shown around its name (pairs of chromosomes with significant homology, such as *Dm*2L/*Ag*3R, use the same color). Coloring inside the schematic chromosome arms denotes microsynteny matches to a region in the other species; the color shown is the color of the chromosome containing the matching region in the other species.

The evidence that significant portions of present-day *Anopheles* chromosomal arms correspond to an originally nonhomologous arm of *Drosophila* raises the questions of how such gene migrations were achieved and what the fate of transferred chromosomal segments is. Multicolor mappings and sliding window plots of orthologs (or microsynteny regions) according to their current association in the other species give a visual indication that genes may have predominantly translocated in large segments (Figs. 8C and 9 and fig. S10). From these, genes or blocks of genes then seem to diffuse within the new arm by the normal process of interarm reshuffling.

For example, the *Drosophila* chromosomal arm 3L appears to be largely homologous to *Ag*2L (Fig. 7); its telomeric half has only one larger region (72 orthologs) with correspondence to another chromosome (*Ag*2R). The relations in the centromeric half are more complex, however, with two regions of 124 and 106 orthologs matching to *Ag*3L; in total, the centromeric half contains roughly equal numbers of orthologs matching *Ag*3L and *Ag*2L (fig. S10D). The current picture might be the result of two independent translocation events from *Ag*3L or a single event followed by an interarm translocation of *Ag*2L orthologs.

Within the genus *Drosophila*, extensive reorganization can be observed in the polytene chromosome complements, although a conserved 1:1 homology between the chromosomal

arms of the different species had already been noticed in the 1940s (*54*). Most of the interspecies rearrangements can be attributed to the occurrence of paracentric inversions (pericentric inversions degrade the integrity of the chromosomes). Additional processes such as simple or Robertsonian translocations (although occurring much less frequently than inversions in *Drosophila*) presumably would most easily explain major exchanges between chromosomal arms, which our analysis indicated. Finally, transposon-mediated rearrangements involving large chromosomal segments (*60*, *61*) could also have led to the extensive recombinations observed in our interspecies comparisons. The sequencing of additional insect genomes in the future will certainly help elucidate some of these evolutionary consequences.

References

1. M. W. Gaunt, M. A. Miles, *Mol. Biol. Evol.* **19**, 748 (2002).
2. D. K. Yeates, B. M. Wiegmann, *Annu. Rev. Entomol.* **44**, 397 (1999).
3. N. J. Besansky, J. R. Powell, *J. Med. Entomol.* **29**, 125 (1992).
4. M. Ashburner, in *Drosophila: A Laboratory Handbook* (Cold Spring Harbor Laboratory Press, Plainview, NY, 1989) p. 74.
5. R. A. Holt *et al.*, *Science* **298**, 129 (2002).
6. W. M. Fitch, *Syst. Zool.* **19**, 99 (1970).
7. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
8. P. Bork, *Genome Res.* **10**, 398 (2000).
9. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
10. S. Aparicio *et al.*, *Science* **297**, 1301 (2002).
11. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
12. G. K. Christophides *et al.*, *Science* **298**, 159 (2002).
13. A. T. Monnerat *et al.*, *Mem. Inst. Oswaldo Cruz* **97**, 589 (2002).
14. M. Affolter, T. Marty, M. A. Vigano, A. Jazwinska, *EMBO J.* **20**, 3298 (2001).
15. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
16. R. Apweiler *et al.*, *Bioinformatics* **16**, 1145 (2000).
17. A. Bateman *et al.*, *Nucleic Acids Res.* **30**, 276 (2002).
18. I. Letunic *et al.*, *Nucleic Acids Res.* **30**, 242 (2002).
19. S. R. Schmid, P. Linder, *Mol. Microbiol.* **6**, 283 (1992).
20. G. Dimopoulos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6619 (2000).
21. G. Dimopoulos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8814 (2002).
22. C. M. Adema, L. A. Hertel, R. D. Miller, E. S. Loker, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 8691 (1997).
23. S. Gokudan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10086 (1999).
24. N. Kairies *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13519 (2001).
25. H. Ranson *et al.*, *Science* **298**, 179 (2002).
26. S. M. Kanzok *et al.*, *Science* **291**, 643 (2001).
27. E. M. Zdobnov *et al.*, data not shown.
28. J. M. Ribeiro, J. G. Valenzuela, *J. Exp. Biol.* **202**, 809 (1999).
29. C. Barillas-Mury, unpublished results.
30. K. J. Schmid, D. Tautz, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 9746 (1997).
31. P. Green *et al.*, *Science* **259**, 1711 (1993).
32. H. D. Youn, J. O. Liu, *Immunity* **13**, 85 (2000).
33. H. D. Youn, L. Sun, R. Prywes, J. O. Liu, *Science* **286**, 790 (1999).
34. A. J. Clark, K. Bloch, *J. Biol. Chem.* **234**, 2578 (1959).
35. R. H. Dadd, in *Comprehensive Insect Physiology, Biochemistry and Pharmacology*, L. I. Gilbert, Ed. (Pergamon, Oxford, 1985), vol. 4, pp. 313–390.
36. F. Lyko, *Trends Genet.* **17**, 169 (2001).
37. D. A. Petrov, *Trends Genet.* **17**, 23 (2001).
38. W. H. Li, T. Gojobori, M. Nei, *Nature* **292**, 237 (1981).
39. D. Torrents *et al.*, data not shown.
40. D. A. Petrov, E. R. Lozovskaya, D. L. Hartl, *Nature* **384**, 346 (1996).
41. A. Stoltzfus, J. M. Logsdon Jr., J. D. Palmer, W. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10739 (1997).
42. W. Gilbert, S. J. de Souza, M. Long, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7698 (1997).
43. I. B. Rogozin, J. Lyons-Weiler, E. V. Koonin, *Trends Genet.* **16**, 430 (2000).
44. D. Schmucker *et al.*, *Cell* **101**, 671 (2000).
45. I. Letunic, R. R. Copley, P. Bork, *Hum. Mol. Genet.* **11**, 1561 (2002).
46. E. L. George, M. B. Ober, C. P. Emerson Jr., *Mol. Cell. Biol.* **9**, 2957 (1989).
47. W. Dietmaier, S. Fabry, *Curr. Genet.* **26**, 497 (1994).
48. S. Aparicio *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1684 (1995).
49. C. Hardison, *Trends Genet.* **16**, 369 (2000).
50. D. Thomasova *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8179 (2002).
51. I. Bancroft, *Trends Genet.* **17**, 89 (2001).
52. S. Wong, G. Butler, K. H. Wolfe, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9272 (2002).
53. K. H. Wolfe, D. C. Shields, *Nature* **387**, 708 (1997).
54. H. J. Muller, in *The New Systematics*, J. H. Huxley, Ed. (Clarendon, Oxford, 1940), pp. 185–268.
55. V. N. Bolshakov *et al.*, *Genome Res.* **12**, 57 (2002).
56. R. L. Kelley *et al.*, *Cell* **98**, 513 (1999).
57. V. H. Meller, *Trends Cell Biol.* **10**, 54 (2000).
58. A. Pannuti, J. C. Lucchesi, *Curr. Opin. Genet. Dev.* **10**, 644 (2000).
59. C. Schutt, R. Nothiger, *Development* **127**, 667 (2000).
60. R. Paro, M. L. Goldberg, W. J. Gehring, *EMBO J.* **2**, 853 (1983).
61. P. Hatzopoulos, M. Monastirioti, G. Yannopoulos, C. Louis, *EMBO J.* **6**, 3091 (1987).
62. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
63. J. E. Blair, K. Ikeo, T. Gojobori, S. B. Hedges, *Biomed. Central Evol. Biol.* **2**, 7 (2002).
64. A. M. Aguinaldo *et al.*, *Nature* **387**, 489 (1997).
65. J. O. Korbel, B. Snel, M. A. Huynen, P. Bork, *Trends Genet.* **18**, 158 (2002).
66. R. R. Sokal, F. J. Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, New York, 1995).
67. A. N. Clements, *The Biology of Mosquitoes* (Chapman & Hall, London, 1992), vol. 1.
68. G. Wegener, *Experientia* **52**, 404 (1996).

Supporting Online Material

www.sciencemag.org/cgi/content/full/298/5591/149/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 to S5
References

6 August 2002; accepted 6 September 2002

Immunity-Related Genes and Gene Families in *Anopheles gambiae*

George K. Christophides,[1]* Evgeny Zdobnov,[1]* Carolina Barillas-Mury,[2] Ewan Birney,[3] Stephanie Blandin,[1] Claudia Blass,[1] Paul T. Brey,[4] Frank H. Collins,[5] Alberto Danielli,[1] George Dimopoulos,[6] Charles Hetru,[7] Ngo T. Hoa,[8] Jules A. Hoffmann,[7] Stefan M. Kanzok,[8] Ivica Letunic,[1] Elena A. Levashina,[1] Thanasis G. Loukeris,[9] Gareth Lycett,[1] Stephan Meister,[1] Kristin Michel,[1] Luis F. Moita,[1] Hans-Michael Müller,[1] Mike A. Osta,[1] Susan M. Paskewitz,[10] Jean-Marc Reichhart,[7] Andrey Rzhetsky,[11] Laurent Troxler,[7] Kenneth D. Vernick,[12] Dina Vlachou,[1] Jennifer Volz,[1] Christian von Mering,[1] Jiannong Xu,[12] Liangbiao Zheng,[8] Peer Bork,[1] Fotis C. Kafatos[1]†

We have identified 242 *Anopheles gambiae* genes from 18 gene families implicated in innate immunity and have detected marked diversification relative to *Drosophila melanogaster*. Immune-related gene families involved in recognition, signal modulation, and effector systems show a marked deficit of orthologs and excessive gene expansions, possibly reflecting selection pressures from different pathogens encountered in these insects' very different life-styles. In contrast, the multifunctional Toll signal transduction pathway is substantially conserved, presumably because of counterselection for developmental stability. Representative expression profiles confirm that sequence diversification is accompanied by specific responses to different immune challenges. Alternative RNA splicing may also contribute to expansion of the immune repertoire.

Malaria transmission requires survival and development of the *Plasmodium* parasite in two invaded organisms: the human host and the mosquito vector. Interactions between the immune system of either organism with the parasite can hinder or even abort its development. The mosquito is known to mount robust immune reactions (*1*), accounting in part for the major parasite losses that occur within the vector. For example, melanotic encapsulation in a refractory strain of *A. gambiae*, the major vector of