The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*

Robert A. Holt,[1]*† G. Mani Subramanian,[1] Aaron Halpern,[1] Granger G. Sutton,[1] Rosane Charlab,[1] Deborah R. Nusskern,[1] Patrick Wincker,[2] Andrew G. Clark,[3] José M. C. Ribeiro,[4] Ron Wides,[5] Steven L. Salzberg,[6] Brendan Loftus,[6] Mark Yandell,[1] William H. Majoros,[1,6] Douglas B. Rusch,[1] Zhongwu Lai,[1] Cheryl L. Kraft,[1] Josep F. Abril,[7] Veronique Anthouard,[2] Peter Arensburger,[8] Peter W. Atkinson,[8] Holly Baden,[1] Veronique de Berardinis,[2] Danita Baldwin,[1] Vladimir Benes,[9] Jim Biedler,[10] Claudia Blass,[9] Randall Bolanos,[1] Didier Boscus,[2] Mary Barnstead,[1] Shuang Cai,[1] Angela Center,[1] Kabir Chatuverdi,[1] George K. Christophides,[9] Mathew A. Chrystal,[11] Michele Clamp,[12] Anibal Cravchik,[1] Val Curwen,[12] Ali Dana,[11] Art Delcher,[1] Ian Dew,[1] Cheryl A. Evans,[1] Michael Flanigan,[1] Anne Grundschober-Freimoser,[13] Lisa Friedli,[8] Zhiping Gu,[1] Ping Guan,[1] Roderic Guigo,[7] Maureen E. Hillenmeyer,[11] Susanne L. Hladun,[1] James R. Hogan,[11] Young S. Hong,[11] Jeffrey Hoover,[1] Olivier Jaillon,[2] Zhaoxi Ke,[1,11] Chinnappa Kodira,[1] Elena Kokoza,[14] Anastasios Koutsos,[15,16] Ivica Letunic,[9] Alex Levitsky,[1] Yong Liang,[1] Jhy-Jhu Lin,[1,6] Neil F. Lobo,[11] John R. Lopez,[1] Joel A. Malek,[6]‡ Tina C. McIntosh,[1] Stephan Meister,[9] Jason Miller,[1] Clark Mobarry,[1] Emmanuel Mongin,[17] Sean D. Murphy,[1] David A. O'Brochta,[13] Cynthia Pfannkoch,[1] Rong Qi,[1] Megan A. Regier,[1] Karin Remington,[1] Hongguang Shao,[10] Maria V. Sharakhova,[11] Cynthia D. Sitter,[1] Jyoti Shetty,[6] Thomas J. Smith,[1] Renee Strong,[1] Jingtao Sun,[1] Dana Thomasova,[9] Lucas Q. Ton,[11] Pantelis Topalis,[15] Zhijian Tu,[10] Maria F. Unger,[11] Brian Walenz,[1] Aihui Wang,[1] Jian Wang,[1] Mei Wang,[1] Xuelan Wang,[11]§ Kerry J. Woodford,[1] Jennifer R. Wortman,[1,6] Martin Wu,[6] Alison Yao,[1] Evgeny M. Zdobnov,[9] Hongyu Zhang,[1] Qi Zhao,[1] Shaying Zhao,[6] Shiaoping C. Zhu,[1] Igor Zhimulev,[14] Mario Coluzzi,[18] Alessandra della Torre,[18] Charles W. Roth,[19] Christos Louis,[15,16] Francis Kalush,[1] Richard J. Mural,[1] Eugene W. Myers,[1] Mark D. Adams,[1] Hamilton O. Smith,[1] Samuel Broder,[1] Malcolm J. Gardner,[6] Claire M. Fraser,[6] Ewan Birney,[17] Peer Bork,[9] Paul T. Brey,[19] J. Craig Venter,[1,6] Jean Weissenbach,[2] Fotis C. Kafatos,[9] Frank H. Collins,[11]† Stephen L. Hoffman[1]‖

Anopheles gambiae is the principal vector of malaria, a disease that afflicts more than 500 million people and causes more than 1 million deaths each year. Tenfold shotgun sequence coverage was obtained from the PEST strain of *A. gambiae* and assembled into scaffolds that span 278 million base pairs. A total of 91% of the genome was organized in 303 scaffolds; the largest scaffold was 23.1 million base pairs. There was substantial genetic variation within this strain, and the apparent existence of two haplotypes of approximately equal frequency ("dual haplotypes") in a substantial fraction of the genome likely reflects the outbred nature of the PEST strain. The sequence produced a conservative inference of more than 400,000 single-nucleotide polymorphisms that showed a markedly bimodal density distribution. Analysis of the genome sequence revealed strong evidence for about 14,000 protein-encoding transcripts. Prominent expansions in specific families of proteins likely involved in cell adhesion and immunity were noted. An expressed sequence tag analysis of genes regulated by blood feeding provided insights into the physiological adaptations of a hematophagous insect.

The mosquito is both an elegant, exquisitely adapted organism and a scourge of humanity. The principal mosquito-borne human illnesses of malaria, filariasis, dengue, and yellow fever are at this time almost exclusively restricted to the tropics. Malaria, the most important parasitic disease in the world, is thought to be responsible for 500 million cases of illness and up to 2.7 million deaths annually, more than 90% of which occur in sub-Saharan Africa (*1*).

Anopheles gambiae is the major vector of *Plasmodium falciparum* in Africa and is one of the most efficient malaria vectors in the world. Its blood meals come almost exclusively from humans, its larvae develop in temporary bodies of water produced by human activities (e.g., agricultural irrigation or flooded human or domestic animal footprints), and adults rest primarily in human dwellings. During the 1950s and early 1960s, the World Health Organization (WHO) malaria eradication campaign succeeded in eradicating malaria from Europe and sharply reduced its prevalence in many other parts of the world, primarily through programs that combined mosquito control with antimalarial drugs such as chloroquine. Sub-Saharan Af-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]Genoscope/Centre National de Sequencage and CNRS-UMR 8030, 2 rue Gaston Cremieux, 91057 Evry Cedex 06, France. [3]Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA. [4]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases (NIAID), Building 4, Room 126, 4 Center Drive, MSC-0425, Bethesda, MD 20892, USA. [5]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. [6]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [7]Grup de Recerca en Informatica Biomedica, IMIM/UPF/CRG, Barcelona, Catalonia, Spain. [8]Department of Entomology, University of California, Riverside, CA 92521, USA. [9]European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. [10]Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. [11]Center for Tropical Disease Research and Training, University of Notre Dame, Galvin Life Sciences Building, Notre Dame, IN 46556, USA. [12]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [13]Center for Agricultural Biotechnology, University of Maryland Biotechnology Institute, College Park, MD 20742, USA. [14]Institute of Cytology and Genetics, Lavrentyeva ave 10, Novosibirsk 630090, Russia. [15]Institute of Molecular Biology and Biotechnology of the Foundation of Research and Technology–Hellas (IMBB-FORTH), Post Office Box 1527, GR-711 10 Heraklion, Crete, Greece, and University of Crete, GR-711 10 Heraklion, Crete, Greece. [16]Department of Biology, University of Crete, GR-711 10 Heraklion, Crete, Greece. [17]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [18]Dipartimento di Scienze di Sanità Pubblica, Sezione di Parassitologia, Università degli Studi di Roma "La Sapienza," P.le Aldo Moro 5, 00185 Roma, Italy. [19]Unité de Biochimie et Biologie Moléculaire des Insectes, Institut Pasteur, Paris 75724 Cedex 15, France.

*Present address: Canada's Michael Smith Genome Science Centre, British Columbia Cancer Agency, Room 3427, 600 West 10th Avenue, Vancouver, British Columbia V5Z 4E6, Canada.

‡Present address: Agencourt Bioscience Corporation, 100 Cummings Center, Suite 107J, Beverly, MA 01915, USA.

§Present address: Department of Pharmacology, Sun Yat-Sen Medical School, Sun Yat-Sen University #74, Zhongshan 2nd Road, Guangzhou (Canton), 510089, P. R. China.

‖Present address: Sanaria, 308 Argosy Drive, Gaithersburg, MD 20878, USA.

†To whom correspondence should be addressed. E-mail: robert.holt@celera.com, rholt@bcgsc.ca (R.A.H.), frank.h.collins.75@nd.edu (F.H.C.).

rica, for the most part, did not benefit from the malaria eradication program, but the widespread availability of chloroquine and other affordable antimalarial drugs no doubt helped to control malaria mortality and morbidity. Unfortunately, with the appearance of chloroquine-resistant malaria parasites and the development of resistance of mosquitoes to the insecticides used to control disease transmission, malaria in Africa is again on the rise. Even control programs based on insecticide-impregnated bed nets, now widely advocated by WHO, are threatened by the development of insecticide resistance in *A. gambiae* and other vectors. New malaria control techniques are urgently needed in sub-Saharan Africa, and to meet this challenge we must grasp both the ecological and molecular complexities of the mosquito. The International *Anopheles gambiae* Genome Project has been undertaken with the hope that the sequence presented here will serve as a valuable molecular entomology resource, leading ultimately to effective intervention in the transmission of malaria and perhaps other mosquito-borne diseases.

Strain Selection

Populations of *A. gambiae sensu stricto* are highly structured into several morphologically indistinguishable forms. Paracentric inversions of the right arm of chromosome 2 define five different "cytotypes" or "chromosomal forms" (Mopti, Bamako, Bissau, Forest, and Savanna), and variation in the frequencies of these forms correlates with climatic conditions, vegetation zones, and human domestic environments (*2, 3*). An alternative classification system based on fixed differences in ribosomal DNA recognizes two "molecular forms" (M and S) (*4*). The S and M molecular forms were initially observed in the Savanna and Mopti chromosomal forms, respectively. However, analysis of *A. gambiae* populations from many areas of Africa has shown that the molecular and chromosomal forms do not always coincide. This can be explained if it is assumed that inversion arrangements are not directly involved in any reproductive isolating mechanism and therefore do not actually specify different taxonomic units. Indeed, laboratory crossing experiments have failed to show evidence of any premating or postmating reproductive isolation between chromosomal forms (*5*).

The *A. gambiae* PEST strain was chosen for this genome project because clones from two different PEST strain BAC (bacterial artificial chromosome) libraries had already been end-sequenced and mapped physically, in situ, to chromosomes. Further, all individuals in the colony have the standard chromosome arrangement without any of the paracentric inversion polymorphisms that are typical of both wild populations and most other colonies (*6*), and the colony has an X-linked pink eye mutation that can readily be used as an indicator of cross-colony contamination (*7*). The PEST strain was originally used in the early 1990s to measure the reservoir of mosquito-infective *Plasmodium* gametocytes in people from western Kenya. The PEST strain was produced by crossing a laboratory strain originating in Nigeria and containing the eye mutation with the offspring of field-collected *A. gambiae* from the Asembo Bay area of western Kenya, and then reselecting for the pink eye phenotype (*8*). Outbreeding was repeated three times, yielding a colony whose genetic composition is predominantly derived from the Savanna form of *A. gambiae* found in western Kenya. This colony, when tested, was fully susceptible to *P. falciparum* from western Kenya (*9*). The PEST strain is maintained at the Institut Pasteur (Paris), and *A. gambiae* strains with various biological features can be obtained from the Malaria Research and Reference Reagent Resource Center (www.malaria.mr4.org).

Sequencing and Assembly

Plasmid and BAC DNA libraries were constructed with stringently size-selected PEST strain DNA. Two BAC libraries were constructed, one (ND-TAM) using DNA from whole adult male and female mosquitoes and the other (ND-1) using DNA from ovaries of PEST females collected about 24 hours after the blood meal (full development of a set of eggs requires ~48 hours). Plasmid libraries containing inserts of 2.5, 10, and 50 kb were constructed with DNA derived from either 330 male or 430 female mosquitoes. For each sex, several libraries of each insert size class were made, and these were sequenced such that there was approximately equal coverage from male and female mosquitoes in the final data set. DNA extraction, library construction, and DNA sequencing were undertaken by means of standard methods (*10–12*). Celera, the French National Sequencing Center (Genoscope), and TIGR contributed sequence data that collectively provided 10.2-fold sequence coverage and 103.6-fold clone coverage of the genome, assuming the indicated genome size of 278 million base pairs (Mbp) (tables S1 and S2). Electropherograms have been submitted to the National Center for Biotechnology Information trace repository (www.ncbi.nlm.nih.gov/Traces/trace.cgi) and are publicly available as a searchable data set.

The whole-genome data set was assembled with the Celera assembler (*8*), which has previously been used to assemble the *Drosophila*, human, and mouse genomes (*12–15*). The whole-genome assembly resulted in 8987 scaffolds spanning 278 Mbp of the *Anopheles* genome (table S2). The largest scaffold was 23.1 Mbp and the largest contig was 0.8 Mbp. Scaffolds are separated by interscaffold gaps that have no physical clones spanning them, although small scaffolds are expected to fit within interscaffold gaps. The sequence that is missing in the intrascaffold gaps is largely composed of (i) short regions that lacked coverage because of random sampling, and (ii) repeat sequences that could not be entirely filled using mate pairs [sequence reads from each end of a plasmid insert (*16*)]. Most intrascaffold gaps are spanned by 10-kbp clones that have been archived as frozen glycerol stocks. These clones have been submitted to the Malaria Research and Reference Reagent Resource

Fig. 1 (foldout). Annotation of the *Anopheles gambiae* genome sequence. The genome sequence is displayed on a nucleotide scale of about 200 kb/cm. Scaffold order along chromosomes was determined with the use of a physical map constructed by in situ hybridization of PEST strain BACs to salivary gland polytene chromosomes. Scaffold placement is shown in the track directly below the nucleotide scale. Individual scaffolds are identified by the last four digits of their GenBank accession number (e.g., scaffold AAAB01008987 is represented by 8987). For purposes of illustration, all scaffolds are separated by the average length of an interscaffold gap (317,904 bp, which is the total length of the unmapped scaffolds divided by the number of mapped scaffolds). Gaps between scaffolds are shaded gray in the scaffold track. The remainder of the figure is organized into three main groups of tracks: forward strand genes, sequence analysis, and reverse strand genes (from top to bottom, respectively). For each DNA strand (forward and reverse), each mapped gene is shown at genomic scale and is color-coded according to the automated annotation pipeline that predicted the gene (see Gene Authority panel on figure key). In addition, genes that are shorter than 10 kb and have two or fewer exons are shown in a separate track near the central sequence analysis section. All genes that are greater than 10 kb or have three or more exons are shown in an additional pair of tracks, expanded to a resolution close to 25 kb/cm. In these expanded tiers, exons are depicted as black boxes and introns are color-coded according to a set of Gene Ontology categories (GO, www.geneontology.org), as shown in the corresponding panel in the figure key. Three sequence analyses appear between the gene tracks: G+C content, sequence similarity to *Drosophila melanogaster*, and SNP density. The natural logarithm of the number of SNPs per 10 kb of sequence is used to color-code the SNP density analysis; G+C content is depicted by a nonlinear scale described in the figure key. Blocks of sequence with similarity to *D. melanogaster* genomic contigs are shown between the G+C and SNP tracks. Genes that have matching *A. gambiae* ESTs are shown directly flanking the central sequence analysis tracks, and are color-coded according to changes in EST density induced by a blood meal (see Post-Blood-Meal EST Density panel in figure key). This figure was generated with gff2ps (www1.imim.es/software/gfftools/GFF2PS.html), a genome annotation tool that converts General Feature Formatted records (www.sanger.ac.uk/Software/formats/GFF) to a Postscript output (*60*).

Center (www.malaria.mr4.org). Although there are many scaffolds, 8684 short scaffolds account for only 9% of the sequence data; the remaining 91% of the genome is organized into just 303 large scaffolds.

As the final step of the assembly process, scaffolds were assigned a chromosome location and orientation according to a physical map constructed by in situ hybridization of nearly 2000 PEST strain end-sequenced BACs to salivary gland polytene chromosomes (*8*). Scaffolds constituting about 84% of the genome have been assigned (table S3), and chromosome arms X, 2L, 2R, 3L, and 3R are represented by 10, 13, 49, 42, and 28 large scaffolds, respectively. Efforts are continuing to map many of the small scaffolds and to increase the density of informative BACs in large scaffolds to approximately one per Mb.

The entire *Anopheles* genome assembly has been submitted to GenBank. Accession numbers for the 8987 genome scaffolds are AAAB01000001 through AAAB01008987. The entire scaffold set in Fasta format can be downloaded from ftp://ftp.ncbi.nih.gov/genbank/genomes/Anopheles_gambiae/Assembly_scaffolds.

The assembly was screened computationally for contaminating sequence (*8*) and evaluated for integrity of pairing of mate pairs. Abnormal mate pairs, either with incorrect orientations or with distances that differ from the mean plasmid library insert size by several standard deviations, can be diagnostic of local misassembly. Of 1,644,078 total mate pairs, only 27,703 have distance violations and only 10,166 have orientation violations. However, we identified 726 regions that have high-density mate pair violations (more than six violations per 10 kbp), 639 of which are distance violations with correct orientation. The cause of these violations appears to be separation of divergent genotypes, as discussed below (*8*). The mean length of these regions is 28 kbp, and in total they constitute 21.3 Mbp or 7.7% of the assembly. These obvious trouble spots have been flagged in our GenBank accessions according to scaffold coordinates and are illustrated as pink bands in Fig. 1.

Assembly of the Y chromosome is ongoing but has been complicated because Y appears to be composed largely of regions containing transposons or transposon fragments that are also found at autosomal centromeres. No scaffolds have yet been assigned to the Y chromosome.

Genetic Variation

Genetic variation within the PEST strain posed a particular challenge to assembling the genome, by making it difficult to distinguish diverged haplotypes from repeats (*8*). The effect of genetic variation is illustrated in Fig. 2, where correlation among ease of assembly [measured by unitig (*17*) length], internal consistency of the assembly (measured by mate pair integrity), and genetic variation [measured by single-nucleotide discrepancies (SNDs) (*18*)] can be clearly seen. The challenges to assembly introduced by this variation exceed those encountered in *D. melanogaster* or mouse, whose genomes were virtually entirely homozygous, or human, whose genome has a much lower level of polymorphism.

The most highly variable regions in the genome appeared to consist of two haplotypes of roughly equal abundance ("dual haplotypes"), as revealed by strong concordance among SND rate, SND balance (*19*), and SND association (*20*) (Fig. 2). The most likely explanation is that recombination among the *A. gambiae* cytotypes that contributed genetically to the PEST strain resulted in a mosaic genome structure. The underlying polymorphic differences between the Savanna and Mopti cytotypes may reflect important differences in their biologies. Two other possible causes for dual haplotypes are the widespread presence of genomic inversions that suppress recombination [as in *Drosophila pseudoobscura* (*21*)], and real duplications in the genome that were erroneously collapsed in the assembly.

Details of the assembly make each of these alternative explanations unlikely. First, the PEST strain was specifically selected to lack large, cytologically visible inversions. If its genome still contained numerous small inversion polymorphisms, one would expect the assembly to display a characteristic pattern of mate pair misorientations. For example, suppose that there were a previously undetected inversion that defined the major alleles in a given region, and that the assembly integrated both copies of the inversion into a single contig that was placed in a scaffold also containing the flanking single-haplotype regions. In this situation, mate pairs straddling an inversion breakpoint would

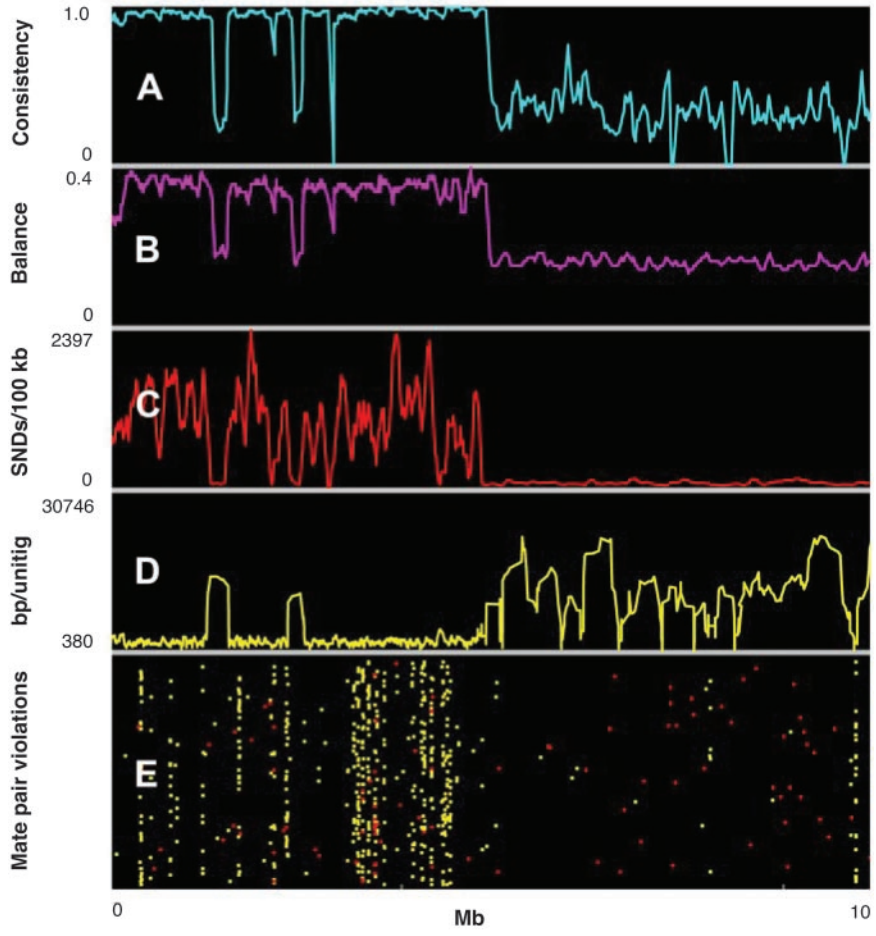

Fig. 2. Large-scale correlation of single-nucleotide discrepancies (SNDs) and assembly characteristics over a 10-Mb section from a single scaffold. (**A**) SND "association" for a sliding window of 100 kb shows the fraction of polymorphic columns whose partitioning is consistent with the partitioning at the previous polymorphic columns (*20*). (**B**) SND "balance" for a sliding window of 100 kb compares the ratio of fragments in the second most frequent character in a column to fragments in the most frequent character (*19*). (**C**) SND rate shows counts of polymorphic columns in a sliding window of 100 kb (*18*). (**D**) Unitig size is shown as the mean size of 21 adjacent unitigs. (**E**) Mate pair violations are shown by drawing a yellow line segment for each mate pair that is correctly oriented but has its fragments separated by more than three standard deviations from the library mean. A red segment corresponds to each incorrectly oriented mate pair.

include one of the sequenced ends in inverted orientation (fig. S1). Such misorientations were not detected. Second, the collapse of two duplicated regions of the genome as the basis for the observation of dual haplotypes can be similarly dismissed, as this explanation would imply that fragment coverage in the dual-haplotype regions should be approximately twice that of single-haplotype regions. In fact the reverse is true: Fragment coverage tends to be lower in the dual-haplotype regions. A final possibility that remains to be fully tested is a prevalence of balanced lethal mutations. If there were tightly linked balanced lethal alleles in the PEST strain, then all viable individuals would be heterozygous in regions of the genome surrounding the lethal alleles. Sampling of the two alternative haplotypes in the shotgun sequence therefore ought to be binomial with a 50:50 chance of either haplotype. Although haplotypes do appear to be approximately balanced in dual-haplotype regions (Fig. 2), we have been unable to confirm a statistical fit of allele frequency to such a model. A direct test for SNP heterozygosity among individuals of the PEST strain is under way and should resolve the issue of genotypic frequencies in these regions.

Many of the SNDs occurred in regions having small unitigs (*17*) and other attributes suggesting difficulties with the assembly. Although there is a co-clustering of small unitigs, mate pair violations, and SNDs, not all regions with a high density of SNDs have problematic assemblies. The breeding history of the PEST strain of *A. gambiae* (*8*) led us to predict that the strain would not be totally inbred, which suggested that the genome would also harbor a large number of polymorphic nucleotides (single-nucleotide polymorphisms or SNPs). High-quality discrepancies of base calls in regions where the assembly is strongly supported ought to be considered as SNPs, allowing a genome-wide analysis of polymorphism.

Celera designed and implemented a SNP pipeline for identifying SNPs on the basis of high–sequence quality mismatches in the human whole-genome assembly (*8, 12*). With some parameter tuning, the same pipeline was adapted to identify SNPs in the *Anopheles* genome and produced a conservative inference of 444,963 SNPs.

The distribution of SNPs along the chromosomes was highly variable, with some regions having only a few SNPs per 100 kb and others having more than 800 SNPs per 100 kb (Fig. 3), despite a nearly homogeneous power to detect them. The overall estimate of mean heterozygosity at the nucleotide level of this strain is 1.6×10^{-3}, but the distribution has high variance and skew, with 45% of the 100-kb intervals having heterozygosity below 5.0×10^{-5} and 10% of the 100-kb intervals having a heterozygosity above 4.7×10^{-2}. The X chromosome has a markedly lower average level of polymorphism, and overall the X-linked nucleotide heterozygosity is 1.2×10^{-4}, markedly below that of the autosomes (discussed below).

It appears that the genome of the PEST strain has resulted from a complex introgression of divergent Mopti and Savanna chromosomal forms (cytotypes). If this is so, then we would expect that some genomic regions may be derived only from one or the other cytotype, yielding a low density of SNPs, whereas other genomic regions may continue to segregate both divergent cytotypes. Microsatellite surveys suggest that the degree of sequence divergence between haplotypes derived from the Mopti and Savanna cytotypes exceeds the variability within each (*22*), so genomic regions with both cytotypes segregating might be expected to have unusually high SNP density. As predicted by this model, the resulting SNP density distribution is markedly bimodal (Fig. 4), with one mode at roughly one SNP every 10 kb, and another mode at one SNP every 200 bp. SNP rates along the X chromosome for the most part do not show this bimodal pattern; we take this to imply a lower rate of introgression on this chromosome, possibly due to male hemizygosity. Although experimental work is required for confirmation, relative lack of introgression seems the most promising explanation for the lower overall SNP rate in the X chromosome, as compared to population genetic explanations based on smaller effective population size of the X chromosome (*23, 24*). In addition, heterozygosity of the X chromosome is expected to be depressed because of the selection for homozygosity of the X-linked pink eye mutation.

Because BAC clones provide clear information on the organization of SNPs into haplotypes, analysis of BAC sequences is more informative than a random shotgun for inferring the population history of these regions of high SNP density. Recent BAC-by-BAC sequencing of a 528-kb chromosomal region in the PEST strain identified two alternative haplotypes that

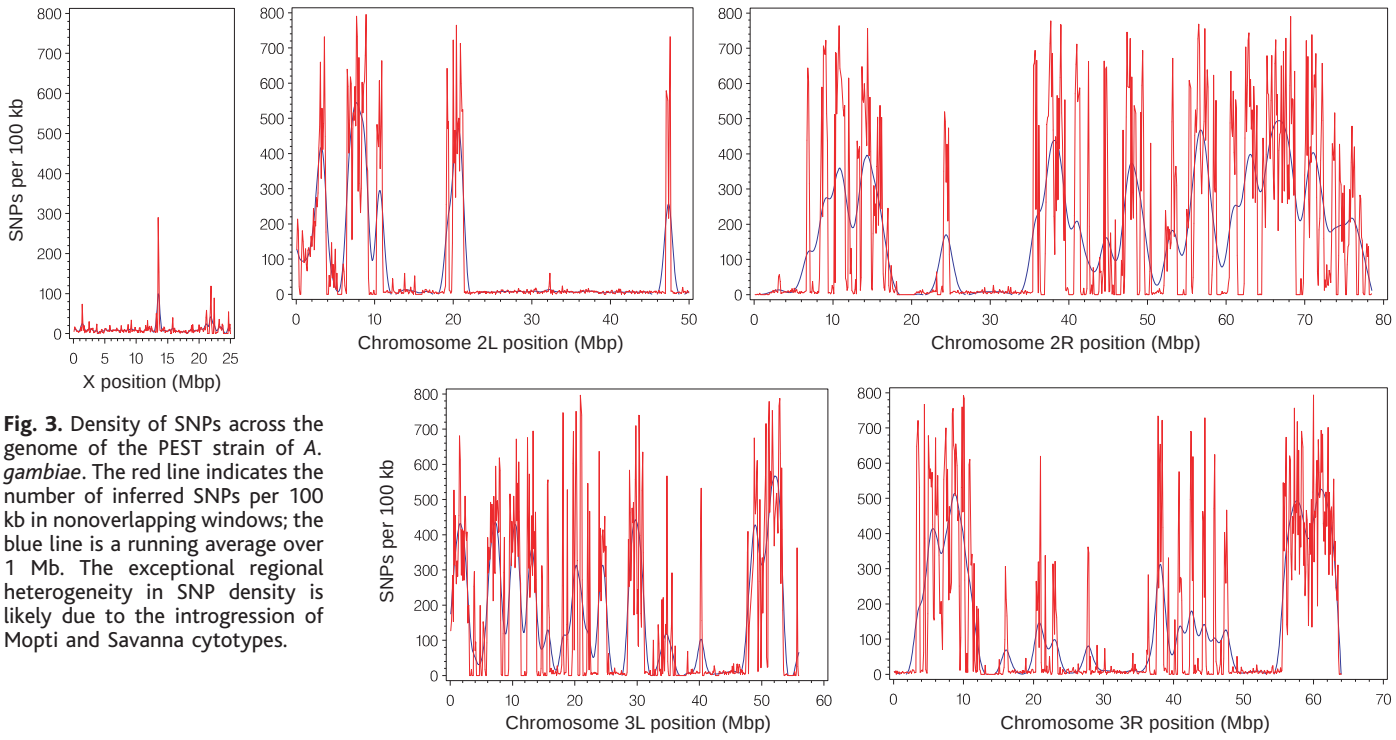

Fig. 3. Density of SNPs across the genome of the PEST strain of *A. gambiae*. The red line indicates the number of inferred SNPs per 100 kb in nonoverlapping windows; the blue line is a running average over 1 Mb. The exceptional regional heterogeneity in SNP density is likely due to the introgression of Mopti and Savanna cytotypes.

differ by 3.3% in sequence and extended for at least 122 kb; reverse-transcription polymerase chain reaction analysis revealed their existence in additional strains, indicating that this phenomenon is not unique to the PEST strain (*25*).

By aligning the SNP calls with predicted genes (gene prediction results are described below), it was possible to place the SNPs into functional categories on the basis of their predicted propensity to alter gene function (e.g., whether they are in intergenic regions, promoter regions, nonsynonymous coding, introns, etc.). Table 1 shows the total count of each functional class and the estimated heterozygosity for the 444,060 SNPs for which this inference could be made. As was the case for the SNPs in the human genome, the overwhelming majority were in intergenic regions, but there was still an abundance of SNPs within functional genes. Introns and intergenic regions had virtually identical heterozygosities, but the silent coding positions appear to have more than twofold enrichment of variability. In general, silent coding sites are considered as having more stringent constraints than introns or intergenic regions because of biased codon usage, and this is reflected in a lower diversity of silent sites in most organisms. The reason for elevated silent variation in *A. gambiae* is at present unknown. Nucleotides with strong functional constraints, such as splice donors, splice acceptors, and stop codons, had the lowest heterozygosity, and nonsynonymous (missense) positions were also evidently low in heterozygosity. All *A. gambiae* SNP data discussed here are available at ftp://ftp.ncbi.nih.gov/genomes/Anopheles_gambiae/SNP.

Annotation

Automated annotation pipelines established by Celera and the Ensembl group at the European Bioinformatics Institute/Sanger Institute were used to detect genes in the assembled *A. gambiae* sequence. Both pipelines use ab initio gene-finding algorithms and rely heavily on diverse homology evidence to predict gene structures (*8*).

We manufactured a "consensus set" of Celera ("Otto") and Ensembl annotations by first populating a graph wherein each node represented an annotated transcript. For each set, an edge was placed between two transcripts if any of their exons overlapped. By this procedure we found that the 9896 transcripts annotated by Ensembl reduced to 7465 distinct genes, and that the 14,564 Otto transcripts reduced to 14,332 distinct genes. Combining the 9896 Ensembl and 14,564 Otto annotations and subjecting them to the same procedure collapsed the combined 24,460 transcripts to 15,189 genes. Of these, 1375 genes were represented solely by Ensembl and 7840 genes solely by an Otto annotation; 5974 genes were identified by both Ensembl and Otto. We then chose the annotation containing the largest number of exons to represent each gene. In cases where a gene was represented by Otto and Ensembl annotations with equal numbers of exons, we chose the Otto annotation to represent the gene. Results of annotation of the *A. gambiae* genome are presented in Fig. 1 and Table 2.

We screened the 15,189 *Anopheles* gene predictions for transposable element sequences that may not have been adequately masked during the automated annotation process. We also screened for contaminating bacterial gene predictions because the genomic libraries used for sequencing were constructed from whole adult mosquitoes and some level of sequence contamination from commensal gut bacteria was expected. We found 1506 putative transposable elements and 663 genes of possible bacterial origin (*8*). Analysis of transposable elements in *A. gambiae* is ongoing, and experimental efforts are currently under way to further characterize bacterial contaminants and to explore the possibility of real horizontal transfer events. Putative transposable elements and bacterial contaminants were flagged before submission to GenBank and, where appropriate, were excluded from further genome analysis either before an automated analysis step was run or during manual interpretation of results.

As a more rigorous quality assurance exercise, we randomly selected 100 annotations from the unflagged portion of the consensus set to manually assess the accuracy of the predicted gene structures. Of these, 35 were predicted correctly, 40 were incompletely annotated (they lacked start and/or stop codons), 4 were merged, 1 was split, and 4 were identified as transposable elements that escaped earlier detection. A further 16 annotations presented various problems with gene structure and needed exon edge adjustment. The large proportion of partial annotations is likely due to lower sequence conservation in gene termini and thus a reduced likelihood of recognition of these regions by similarity-based automated annotation systems.

To estimate the number of genes that may have been missed by the automated annotation process, we examined FgenesH and Grailexp predictions that showed similarity to known proteins but were not represented in the consensus set. We also examined regions where an *A. gambiae* expressed sequence tag (EST) matched the genomic sequence across a putative splice junction and no gene call was made. On the basis of these analyses, we expect that as many as 1029 genes may have escaped automated annotation and therefore are not displayed in Fig. 1 or included in our analysis of the proteome. The *Anopheles* annotation described herein should be considered a first approximation, providing a framework for future improvement by manual curation.

Features of the Genome Landscape

The sizes of the *Anopheles* and *Drosophila* genomes have been predicted by CoT analy-

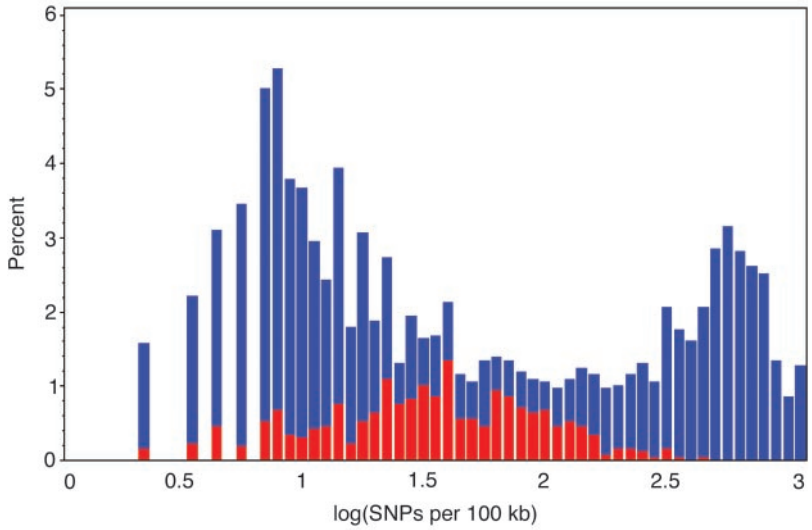

Fig. 4. SNP density on autosomes. The red bars represent X-linked SNPs; the lack of bimodality of X-linked SNPs suggests that there was less successful introgression on the X chromosome.

Table 1. Distribution of SNPs in the *A. gambiae* genome, and their characteristics and heterozygosity per category.

Genomic attribute	SNP count	Heterozygosity ($\times 10^{-3}$)
Intergenic	348,332	1.613
Intron	67,210	1.563
Missense	5,886	0.413
Nonsense	96	0.106
Silent	18,645	3.721
Splice site acceptor	24	0.626
Splice site donor	34	0.886
3′-untranslated region	2,382	0.313
5′-untranslated region	1,451	0.191
Total	444,060	1.596

sis to be 260 Mb (*26*) and 170 Mb (*27*), respectively, and the sizes of their genome assemblies are 278 Mb and 122 Mb (*13*). The discrepancy between estimated and assembled genome size in *Drosophila* is thought to be due to the nature of *Drosophila* heterochromatin, which consists of long tandem arrays of simple repeats that cannot be readily cloned and sequenced with existing technology (*13*). Regarding *Anopheles*, there are several immediate possibilities as to why the assembly is slightly larger than the predicted genome size. The CoT analysis could be slightly inaccurate, or, because it was done with DNA of a different strain, the estimate could simply reflect a real strain difference in genome size. In addition, we know that segregation of haplotypes during the assembly process has led to overrepresentation of the size of the genome by about 21.3 Mb (*8*), and it appears that the *Anopheles* assembly has captured much of the heterochromatic DNA. Unlike *Drosophila*, genomic DNA from *Anopheles* does not show a prominent heterochromatic satellite band when separated on a cesium chloride gradient (*28*), which suggests that the heterochromatin is of higher complexity and thus more amenable to sequencing and assembly. In fact, in the *Anopheles* assembly, there are many scaffolds that exist entirely within known heterochromatic regions or extend into centromeres.

The difference in absolute genome size between *Anopheles* and *Drosophila* could be due to gain in *Anopheles*, loss in *Drosophila*, or some combination thereof. Given that the numbers of genes, numbers of exons, and total coding lengths vary by less than 20% (Table 3), the size difference between the two genomes is due largely to intergenic DNA. The exact nature of *Anopheles* intergenic DNA is unclear, but as discussed above, much of it may consist of moderately complex heterochromatic sequence. By counting the number of times each 20-nucleotide oligomer in the *Anopheles* and *Drosophila* assemblies appeared in its corresponding whole-genome shotgun data, we confirmed that simple repeats are not expanded in *Anopheles* (*8*). However, there does appear to be greater representation of transposons in *Anopheles* heterochromatin than in *Drosophila* heterochromatin, as discussed below.

A likely explanation for the size difference of the two genomes is that *D. melanogaster* has lost noncoding sequence during divergence from *A. gambiae*. All mosquitoes in the Culicidae family have larger genomes, with estimates of 240 to 290 Mb for *Anopheles* species and 500 Mb or larger for all others. *Drosophila* species groups other than *D. melanogaster* and *D. hydei* have genomes of 230 Mb or larger (Center for Biological Sequence Analysis, Database of Genome Sizes, www.cbs.dtu.dk/databases/DOGS). This suggests that the two clusters with smaller genome sizes experienced genome reductions during recent evolutionary time. The fact that most other families of the dipteran order have species with genomes at least as large as that of *A. gambiae* further supports this conjecture. Mechanisms for this relatively rapid loss of noncoding DNA have been modeled and analyzed in insect species (*29, 30*).

Table 2. Features of *A. gambiae* chromosome arms. Known and unknown genes are defined as genes with an assigned versus unassigned/unclassified GO molecular function. Gaps between scaffolds are included in the chromosome length estimate. Each gap has the arbitrary value of 317,904 bp, which is the total length of the unmapped scaffolds divided by the number of mapped scaffolds. There are 602 known genes, 1017 unknown genes, and 22,123 SNPs on unmapped scaffolds.

Chromosome	Length (bp)	Number of scaffolds	Number of known genes	Number of unknown genes	Number of SNPs
X	24,902,716	10	584	500	2,955
2R	78,412,669	49	2166	1461	162,335
2L	52,393,056	13	1615	1078	44,604
3R	64,548,413	28	1541	1000	102,203
3L	56,406,562	42	1278	841	110,743

Table 3. Characteristics of the *A. gambiae* genome. Fractions of total genome size are shown in parentheses.

Genome features	*Anopheles*	*Drosophila*
Total genome size	278,244,063 bp	122,653,977 bp
Percent of G+C in genome	35.2%	41.1%
Total coding size	19,274,180 bp (7%)	23,826,134 bp (19%)
Total intron size	42,991,864 bp (15%)	27,556,733 bp (22%)
Total intergenic size	215,978,019 bp (78%)	71,271,110 bp (58%)
Number of genes	13,683*	13,472
Number of exons	50,609	54,537
Average gene size (±SD)	4,542 ± 10,802 bp	3,759 ± 9,864 bp

*The number of annotated *Anopheles* genes after removal of putative transposable elements.

About 40 different types of transposons or transposon-related dispersed repeats have been identified in the *A. gambiae* genome (*8*) (Table 4). The most abundant are class I repeats, particularly the long terminal repeat (LTR) retrotransposons, small interspersed repeat elements (SINEs), and miniature inverted repeat transposable elements (MITEs), but all major families of class II transposons are also represented. Overall, transposable elements constitute about 16% of the eukaryotic component and more than 60% of the heterochromatic component of the *A. gambiae* genome (*8*), as compared to 2% and 8%, respectively, for *D. melanogaster* (*31*). Transposons present in heterochromatin are highly fragmented in *A. gambiae*, so 60% is likely an underestimate. Because heterochromatin appears to be largely derived from transposons, there must be a mechanism that promotes transposon loss from these regions at a rate that balances the insertion of new copies.

Within the euchromatic part of the genome, repeat density is highest near the centromeres, lowest in the middle of chromosome arms, and somewhat elevated near the telomeres. Moreover, transposon densities differ by arm. Transposon density is highest on the X chromosome (59 transposons per Mb), with chromosome arms 2R, 2L, 3R, and 3L having 37, 46, 47, and 48 transposons per Mb, respectively. Transposon distribution is consistent with the hypothesis that densities are highest in parts of the genome where recombination rates are lowest. The observation that 2R has the lowest overall repeat density may be related to the large number of paracentric inversions on this arm whose frequencies are known to be associated with population structuring (*32*).

A protein-based method developed to identify genomic duplications (*15*) was modified to search for segmental chromosomal duplications in the *A. gambiae* genome. Briefly, at least three proteins within a small interval along a chromosome were required to align with three homologous proteins on a separate genomic interval in order to be considered a potential duplication segment (*33*). A total of 102 duplication blocks, containing 706 gene pairs, were identified by this method.

We detected only a few large duplicated segments that contain paralogous expansions of a single family distributed in two distinct blocks in the *Anopheles* genome. These could be the result of a single or limited number of gene duplications to a distinct second chromosomal site, followed by further local tandem duplications at the two sites. Alternatively, such distributions could result from a tandem duplication of a given gene, followed by segmental duplication of the tandem block of paralogous genes. These possibilities can only be distinguished by extensive phylogenetic analyses, and we therefore analyzed the 21 largest tandem cluster

pairs in relation to *Drosophila*. Figure S3 illustrates an example in the glutathione S-transferase gene family. The absence of clear segregation of the *Drosophila* and *Anopheles* members, along with other suggestive features of the tree structure, is consistent with tandem gene duplications in the *Anopheles/Drosophila* common ancestor followed by segmental duplication after *Anopheles/Drosophila* divergence.

These results should be contrasted with results from other animal genomes. Although the *Caenorhabditis elegans* (worm) and *Fugu rubripes* (pufferfish) genomes showed minimal evidence of block duplications (*34, 35*), there was a markedly higher frequency of segmental duplications observed in the human and mouse genomes. Analysis of the human protein set revealed 1077 duplicated blocks containing 10,310 gene pairs, including some blocks encompassing >200 genes (*12*). Thus, the human analysis revealed more than 10 times the number of potential segmental duplication blocks found in the mosquito, despite a proteome that is only about twice as large. Many of these duplications were mirrored in the mouse genome (*15*). This contrasts greatly with the observed paucity of segmental duplications in *Anopheles*; moreover, these duplications are not clearly discernible in *Drosophila* (*36*). Thus, the large segmental and chromosome-sized duplications described in vertebrate genomes are not observed in the two insect genomes examined. However, given the limitations of the methods used, ancient large segmental duplications that subsequently underwent massive rearrangement ("scrambling") would not be detected in this analysis.

A broader comparison of the entire predicted protein sets of *A. gambiae* and *D. melanogaster* revealed clear relationships across chromosomes in the two genomes, and in most cases indicated a one-to-one relationship between proteins across the two species. Chromosome 2 of *Anopheles* shares a common ancestor with chromosome 3 of *Drosophila*, and chromosome 3 of *Anopheles* has a common ancestor—with the left and right arms reversed—with chromosome 2 of *Drosophila*. More details of this comparison are given in a companion article (*37*).

The *A. gambiae* Proteome

Two broad questions were asked: (i) What are the most represented molecular functions of the predicted gene products in *A. gambiae*, and how do these compare with other sequenced eukaryotic species and the closest sequenced evolutionary neighbor, *D. melanogaster*? [Our approach involved analysis at the level of protein domains using the InterPro database (*38, 39*) and clustering protein families using a previously published algorithm called LeK (*12, 40*).] (ii) What are the prominent genes in *Anopheles* that are associated with blood feeding? In a companion article, specific differences between *Anopheles* and *Drosophila* genes are examined further, including complementary analyses of strict orthology (*Anopheles* genes with one clearly identifiable counterpart in *Drosophila*, and vice versa), microsynteny, and dynamics of gene structure (*37*).

The results presented here are preliminary, as the gene predictions and functional assignments were computationally generated, and we expect both false-positive predictions (pseudogenes, bacterial contaminants, and transposons) and false-negative predictions (*Anopheles* genes that were not computationally predicted). We also expect a few errors in delimiting the boundaries of exons and genes. Similar limitations are likely in the automatic functional assignments.

We used InterPro and Gene Ontology (GO) (*41*) to classify the predicted *Anopheles* protein set on the basis of protein domains and their functional categories. Figure 1 provides an overview of protein functional predictions according to broad GO molecular function categories, as well as the genomic coordinates of these proteins on mapped scaffolds. We then defined the 50 most prominent InterPro signatures in *Anopheles* and the representation of these domains in other completely sequenced eukaryotic genomes (table S4). The relative abundance of the majority of proteins containing InterPro do-

Table 4. Repetitive DNA sequences in *A. gambiae*. Elements are identified by a name already in use in *A. gambiae*, by the most similar element in another species [usually *D. melanogaster* (-lk = like)], or by commonly recognized family designators (e.g., mariner, piggyBac, or hAT family elements).

Class	Element type	Euchromatic copies	Density per Mb	Heterochromatic copies	Density per Mb
		Class I			
LTR retrotransposons	Beagle	4	0.018	2	0.323
	Copia-lk	743	3.259	65	10.484
	Cruiser	63	0.276	27	4.355
	Gypsy-lk	1184	5.193	106	17.097
	Moose	970	4.254	85	13.710
	Osvaldo	29	0.127	7	1.129
	Pao-lk	886	3.886	88	14.194
	Springer	52	0.228	20	3.226
Non-LTR retrotransposons	Jam2	27	0.118	7	1.129
	Juan-lk	16	0.070	12	1.935
	Lian2	15	0.066	4	0.645
	RT1*	1	0.004	1	0.161
	RT2*	1	0.004	0	0
	RTE-lk	115	0.504	18	2.903
	LINE-lk	12	0.053	7	1.123
	R4-lk	1	0.004	1	0.161
	I-lk	17	0.075	2	0.323
	T1	39	0.171	15	2.407
	Q	69	0.303	29	4.677
	Tx1-lk	4	0.018	0	0
SINEs	Sine200	2389	10.478	132	21.290
		Class II			
DNA transposons	Crusoe	51	0.224	3	0.484
	hAT	10	0.044	5	0.806
	PIF-lk	8	0.035	5	0.806
	P	12	0.053	0	0
	piggyBac	5	0.022	1	0.161
	mariner	157	0.689	16	2.567
	DD34E	227	0.996	69	11.129
	DD37D	144	0.632	8	1.290
	DD37E	12	0.053	0	0
	Pogo-lk	8	0.035	0	0
	Tiang	11	0.048	2	0.323
	Topi	45	0.197	16	2.581
	Tsessebe	14	0.061	6	0.968
MITEs	3bp(I-XII)	807	3.539	51	8.226
	8bp-I	145	0.636	10	1.613
	Ikirara	54	0.237	2	0.323
	Joey	384	1.684	18	2.903
	Pegasus	43	0.189	1	0.161
	TA(I-V)	1671	7.329	76	12.258
	TAA(I-II)	115	0.504	22	3.548

*RT1 and RT2 elements have specific insertion target sites found almost exclusively in the rDNA large subunit coding region. Because rDNA of *A. gambiae* is organized in a long tandem array that does not appear in the assembled genome, these elements are underrepresented in Table 4.

mains was similar between the mosquito and fly, with insect-specific cuticle and chitin-binding peritrophin A domains and the insect-specific olfactory receptors being similarly overrepresented. However, there are several classes of proteins that contain domains that are overrepresented in mosquito compared to fly, and comparison of the representation of these domains in other organisms (table S4) suggests that the representational difference is due to expansion in *Anopheles* rather than loss in *Drosophila*.

The serine proteases, central effectors of innate immunity and other proteolytic processes (*42, 43*), are well represented in both insect genomes, but *Anopheles* has nearly 100 additional members. The presence of additional members in *Anopheles* is perhaps reflective of differences in feeding behavior and its intimate interactions with both vertebrates and parasites.

We observed expansions of specific extracellular adhesion domain–containing proteins in *Anopheles*. There are 36 more fibrinogen domain–containing proteins and 24 more cadherin domain–containing proteins in *Anopheles* than in *Drosophila*. The fibrinogen domain–containing proteins are similar to ficolins, which represent animal carbohydrate-binding lectins that participate in the first line of defense against pathogens by activating the complement pathway in association with serine proteases (*44*). As discussed below, several of these members were up-regulated in response to blood feeding. Expansion of cadherin domain–containing proteins is of interest given their prominent role in cell-cell adhesion in the context of morphogenesis and cytoskeletal and visual organization (*45, 46*). The observed differential expression of some of the members of this family with blood feeding may suggest an unexplored role in regulating the cytoskeletal changes in the mosquito gut to accommodate a blood bolus.

Finally, although there is relative conservation of most of the transcription factor proteins between the two insect genomes and other sequenced organisms (for example, the C2H2 zinc finger, POZ, Myb-like, basic helix-loop-helix, and homeodomain-containing proteins), we observed overrepresentation of the MYND domain–containing nuclear proteins in mosquito. This protein interaction module is predominantly found in chromatinic proteins and is believed to mediate transcriptional repression (*47*).

Building on a previously published procedure, we used the graph-theoretic algorithm LeK (*15, 40*) to simultaneously cluster the protein complements of *Anopheles* and *Drosophila*. Unlike the above InterPro analysis, which grouped proteins on the basis of domain content, LeK sorted homologous proteins (orthologs plus paralogs) into clusters on the basis of sequence similarity (*8*). The variance of each organism's contribution to each cluster was calculated, allowing an assessment of the relative importance of organism-specific expansion and contraction of protein families that have occurred since divergence from their common dipteran ancestor about 250 million years ago (*48*).

The striking degree of evolutionary relatedness between *Anopheles* and *Drosophila* is illustrated in Fig. 5, with a sizable proportion of the *Anopheles* proteome represented by clusters with a 1:1 *Drosophila* ratio. Although there is substantial conservation between *Anopheles* and *Drosophila*, the LeK method of analysis provided 483 clusters that contain only *Anopheles* proteins. Prominent among these is a 19-member odorant receptor family that is entirely absent in *Drosophila*. It is tempting to speculate that this family may be important in mosquito-specific behavior that includes host seeking.

To illustrate some of these prominent differences between the two species, we analyzed protein family clusters that showed at least 50% overrepresentation in *Anopheles*. The degree of overrepresentation and the molecular functions of these proteins are shown in Fig. 5B. In exploring the possible biological relevance of these observed representational differences, we have focused on families with prominent physiological roles (Table 5). These include critical components of the visual system, structural components of the cell adhesion and contractile machinery, and energy-generating glycolytic enzymes that are required for active food seeking. Increased numbers of salivary gland components and anabolic and catabolic enzymes involved in protein and lipid metabolism are consistent with the *Anopheles* blood feeding and oviposition cycle, described below. Of equal interest are protein families that may play a protective role in *Anopheles*. These include determinants of insecticide resistance such as transporters and detoxification enzymes. Although the greater numbers of serine proteases have been described previously in the text and table S4, additional differences (seen here in α2-macroglobulin and hemocyanins) are consistent with a complex innate immune system in *Anopheles*. Finally, representative examples of greater numbers of genes involved in nuclear regulation and signal transduction provide the first glimpse into what perhaps defines a hematophagous dipteran.

After metamorphosis into an adult mosquito, female anopheline mosquitoes take sugar meals to maintain basal metabolism and to energize flight. Flight is needed for mating and finding a host that will provide a blood meal source. The blood meal is a protein-rich diet that the mosquito surrounds after ingestion with the peritrophic matrix (PM), a thin structure containing chitin and proteins. Digestion requires secreted proteases that penetrate the PM. The smaller digestion products are hydrolyzed

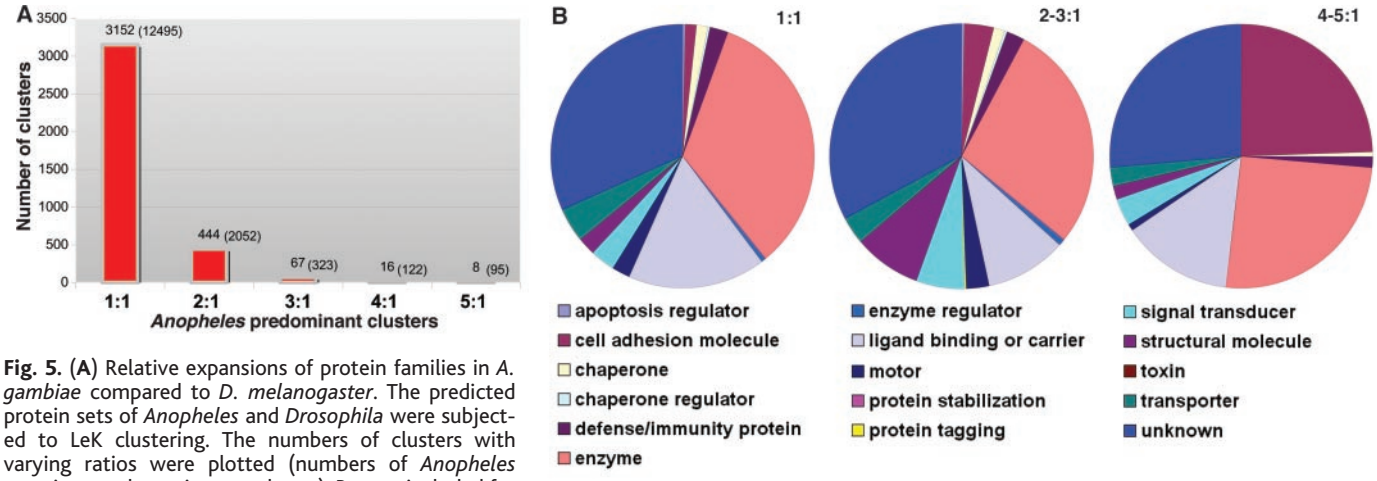

Fig. 5. (**A**) Relative expansions of protein families in *A. gambiae* compared to *D. melanogaster*. The predicted protein sets of *Anopheles* and *Drosophila* were subjected to LeK clustering. The numbers of clusters with varying ratios were plotted (numbers of *Anopheles* proteins are shown in parentheses). Ranges included for each ratio: 1:1 (0.5 to 1.49), 2:1 (1.5 to 2.49), 3:1 (2.5 to 3.49), 4:1 (3.5 to 4.49), and 5:1 (4.5 to 5.49). (**B**) Distribution of the molecular functions of proteins represented in LeK clusters with varying *Anopheles:Drosophila* ratios. Each slice represents the assignment to molecular function categories in the GO.

by microvilli-bound enzymes before absorption by the midgut cells. The blood meal–derived nutrients are processed by the insect fat body (equivalent of the liver and adipose tissue of vertebrates) into egg proteins (vitellogenins) and various lipids associated with lipoproteins. These are exported through the hemolymph to the insect ovaries, where the oocytes develop. The egg development process takes 2 to 3 days, and no further food intake is needed until after oviposition, when a new cycle of active host finding and blood feeding, digestion, and egg development begins (*49*).

We performed an EST-based screen for genes that are regulated differentially in adult female mosquitoes in response to a blood meal (*8*). From a starting set of 82,926 ESTs (43,174 from blood-fed mosquitoes, 39,752 from non–blood-fed mosquitoes), we identified 6910 gene loci with at least one EST hit. Using a binomial distribution and a stringent *P*-value cutoff of 0.001, we identified 97 up-regulated transcripts and 71 that were down-regulated in the blood-fed group (Fig. 6) (table S5). These results are consistent with earlier microarray experiments based on much smaller gene sets (*50*).

After a blood meal, several genes associated with cellular and nuclear signaling, digestive processes, ammonia excretion, lipid synthesis and transport, and translational machinery were overexpressed. In addition, lysosomal enzymes (including proteases found in the fat body and oocytes), genes coding for yolk and oocyte proteins, and genes associated with egg melanization were up-regulated. Conversely, there was down-regulation of genes associated with muscle processes (cytoskeletal and muscle contractile machinery, glycolysis, and ion adenosine triphosphatases) and their associated mitochondrial proteins. Salivary and midgut glycosidases, needed for digestion of a sugar meal, were down-regulated by blood feeding. Four proteins associated with the vision process were also down-regulated, suggesting a degree of detachment of the mosquito from its environment during digestion of a blood meal. Signaling serine proteases of the midgut (important for detection of a protein meal in the gut), peritrophic matrix proteins (matrix components synthesized before the blood meal and accumulated in midgut cell granules), and structural components of the insect cuticle all showed decreased expression after the blood meal. Interestingly, a protein associated with circadian cycle, stress, and feeding behavior was also down-regulated. Finally, the blood meal increased expression of the mitochondrial NADPH-dependent isocitrate dehydrogenase and concomitantly decreased expression of the NAD-dependent form (where NAD is the oxidized form of nicotinamide adenine dinucleotide and NADPH is the reduced form of NAD phosphate). This likely reflects a shift from muscle to fat body metabolism.

Concluding Remarks

Foremost in our minds is how the genomic and EST data can be used to improve control of malaria in the coming decades. Three issues are central to efforts aimed at reducing malaria transmission: reducing the numbers and longevity of infectious mosquitoes, understanding what attracts them to human (as opposed to animal) hosts, and reducing the capacity of parasites to fully develop within them.

Reducing the number of mosquitoes: Anopheline mosquitoes rapidly develop resistance to pesticides. The molecular targets of the major classes of insecticides are known, and mutation of target sites is well understood as a mechanism of resistance (*51*). However, the molecular basis of metabolic resistance is less clear. The *Anopheles* genome provides a near-complete catalog of enzyme families that play an important role in the catabolism of xenobiotics (*52*). Furthermore, the availability of SNPs in these genes will facilitate monitoring of the frequency and spread of resistance alleles and efforts to locate the major loci associated with resistance to DDT and pyrethroids (*51, 53*).

The hematophagous appetite of the female mosquito is exemplified by its remarkable ability to ingest up to four times its own weight in blood. The genome-wide EST expression analysis described here provides evidence that a blood meal results in up-regulation of genes for protein and lipid metabolism, with concomitant down-regulation of genes specific to the musculature and sensory organs. This metabolic reprogramming offers multiple points for intervention. Identification of key pathways that facilitate ingestion of a blood meal provides an opportunity to disrupt the carefully orchestrated host-seeking and concomitant metabolic signals through high-affinity substrate analogs, or by disrupting insect-specific cell signaling pathways.

Reducing the anthropophilicity of the mosquito: The molecular basis for the distinct preference for human blood and the ability to find it is unknown, but it almost certainly involves recognition of human-specific odors. *A. gambiae* odorant receptors described here and in a companion report (*54*) may provide insights into what underlies human host preference. This knowledge should be of use in designing safe and effective repellents that reduce the transmission rate of malaria simply by reducing the efficiency with which mosquitoes find and feed on their human prey.

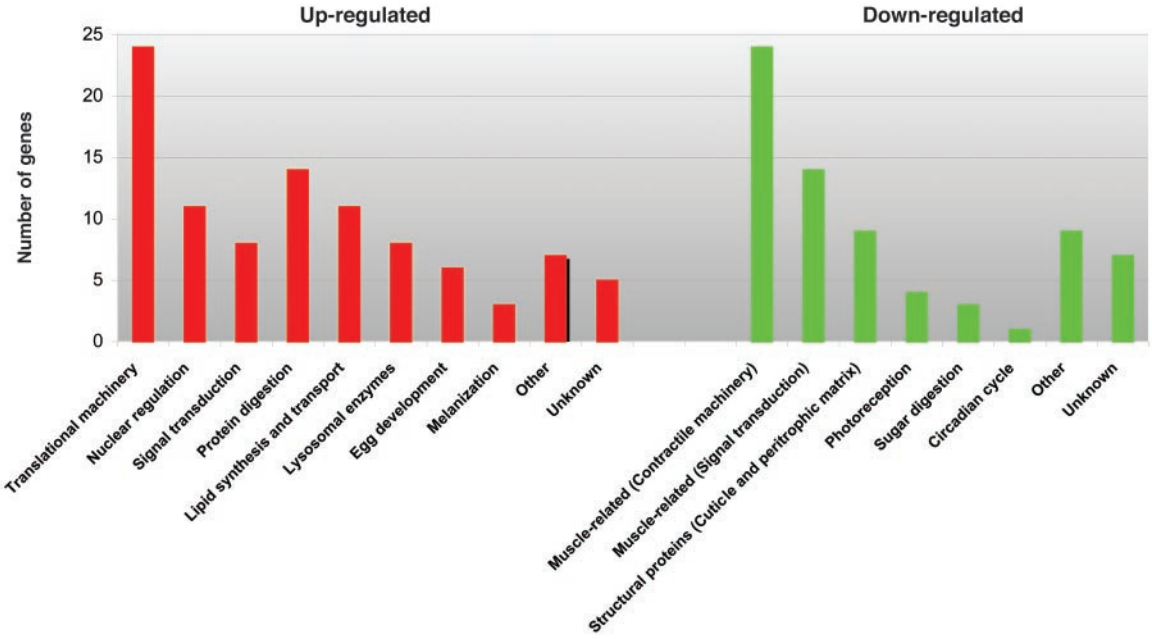

Fig. 6. Functional classes of genes corresponding to ESTs from blood-fed and non–blood-fed *A. gambiae*. The genes that contribute to each functional category are listed in table S5.

Table 5. Representative protein family expansions in *A. gambiae*, as derived from LeK analysis. A/D ratio, *Anopheles/Drosophila* ratio.

Physiological role	LekID; A/D ratio; InterPro domain	Functional assignment
Behavioral	32665; 4/2; DNA_photolyase	Circadian rhythm
(host seeking; blood feeding and malaria transmission)	34088; 7/3	Odorant binding protein
	31848; 19/0; 7tm_6	Olfactory receptor
	29519; 25/12; 7tm_1	Photoreception
Structural	30930; 23/13; Cadherin 29618	Cell adhesion and signaling
(components of cuticle, peritrophic matrix, and extracellular matrix)	6/2; COLFI	Collagen α
	33624; 44/9; FBG	Ficolin-like
	30264; 20/13; Chitin_bind_2	Peritrophic matrix protein
	31643; 73/30; insect_cuticle	Cuticle proteins
	32037; 28/18; insect_cuticle	Cuticle proteins
	30269; 7/0; Chitin_bind_2	Chitinase
	30538; 8/5; Reprolysin	ADAM metalloprotease
Metabolism	30151; 4/2; Aamy	Amylase
(blood and sugar meal digestion; glyconeogenesis; lipid metabolism)	30494; 10/4; Tryp_Tryp_SPC	Serine protease
	30579; 7/0; Tryp_SPC	Serine protease
	33293; 6/2; Lipase_GDS	Lipase
	32295; 14/7; aldo_ket_red	Glycolysis enzyme
	31146; 6/2; Orn_Arg_deC_N	Glycolysis enzyme
	29384; 9/4; aminotran_3	AA catabolism
	30644; 9/2	Cholesterol synthesis
	29690; 8/2	Acyltransferase
	316.67; 17/8; hemocyanin	Hexamerins
	31992; 7/2; lipocalin	Lipid transport apolipoprotein
Proteins found in adult female salivary glands	30799; 4/0	Short D7 protein family
	31625; 4/0	gSG7 protein family
	32969; 4/0	SG1 protein family
	32413; 16/9; peroxidase	Includes salivary peroxidase
	32333; 10/6; nucleotidase	Includes salivary apyrase
Immunity	30476; 9/3; Tryp_SPc	Hemolymph serine protease
(includes hemolymph coagulation, antimicrobial peptide synthesis, and melanization)	31513; 7/4; Glyco_hydro	Immune recognition
	31667; 17/8; hemocyanin	Prophenoloxidase
	32038; 6/3; Cu-oxidase	Monophenoloxidase
	30884; 17/7; A2M	α2-Macroglobulin
	31104; 8/1; GLECT	Galectin
	31703; 5/3; NO_synthase	NO synthase family
	30716; 14/6; Caspases	Cell death after parasite invasion
Detoxification	33859; 15/4	Sulfotransferase
(insecticide resistance)	29536; 5/1; p450	Cytochrome p450
Ion channels	31038; 7/2; CN_hydrolase	Nitrilase
[includes transporters of small molecules (insecticide targets)]	29820; 6/3; Lig_chan	Glutamate receptor
	31803; 9/5; ion_trans	Na^+ transporter (DDT target)
	29625; 8/5; K_tetra	Voltage-sensitive K^+ channel
	33408; 8/5; aa_permeases	Bumetanide-sensitive transporter
Nuclear regulation	30850; 18/4; SET	Protein methyltransferase
	30850; 7/1; MYND	MYND finger
	30322; 5/0; H15	Histone
	29476; 5/1; Rad4	XP-C, DNA repair

Reducing the development of the malarial parasite: The complex orchestration of the *Plasmodium* life cycle in *Anopheles* illustrates several critical points of intervention, such as fusion of gametocytes in the mosquito midgut, penetration of the peritrophic matrix by the ookinete, and migration of sporozooites to the mosquito salivary glands. Likewise, an improved understanding of the *Anopheles* immune response to the parasite can be exploited to disrupt transmission (*55, 56*). Several recent genomic approaches have provided catalogs of genes involved in the response to a wide range of immune stimuli, including infection by *Plasmodium* species (*43, 50, 55, 56*). These strategies provide candidate genes to complement recent developments in generating genetically transformed *A. gambiae* strains that are refractory to *Plasmodium* (*57–59*). Germline transformation thus holds much promise for producing immune-competent, pesticide-susceptible, or zoophilic *A. gambiae*. However, there are serious complicating factors that must be overcome. Knowing the sequence of the *A. gambiae* genome will enable further characterization of candidate genes useful for malarial control, and will allow the characterization of mobile genetic elements that may be used for transformation.

References and Notes

1. J. G. Breman, A. Egan, G. T. Keusch, *Am. J. Trop. Med. Hyg.* **64** (suppl.), 1 (2001).
2. M. Coluzzi, A. Sabantini, V. Petrarca, M. A. Di Deco, *Trans. R. Soc. Trop. Med. Hyg.* **73**, 483 (1979).
3. M. Coluzzi, V. Petrarca, M. A. Di Deco, *Boll. Zool.* **52**, 45 (1985).
4. A. della Torre *et al.*, *Insect Mol. Biol.* **10**, 9 (2001).
5. Y. T. Touré *et al.*, *Parassitologia* **40**, 477 (1998).
6. O. Mukabayire, N. J. Besansky, *Chromosoma* **104**, 585 (1996).
7. G. F. Mason, *Genet. Res.* **10**, 205 (1967).
8. See supporting data on *Science* Online.
9. A. K. Githeko *et al.*, *Trans R. Soc. Trop. Med. Hyg.* **86**, 355 (1992).
10. H. R. Crollius *et al.*, *Genome Res.* **10**, 939 (2000).
11. S. Zhao *et al.*, *Genome Res.* **11**, 1736 (2001).

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