www.sciencemag.org/cgi/content/full/science.1258522/DC1

Supplementary Materials for

Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes

Daniel E. Neafsey,* Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, Bruno Arcà, Peter Arensburger, Gleb Artemov, Lauren A. Assour, Hamidreza Basseri, Aaron Berlin, Bruce W. Birren, Stephanie A. Blandin, Andrew I. Brockman, Thomas R. Burkot, Austin Burt, Clara S. Chan, Cedric Chauve, Joanna C. Chiu, Mikkel Christensen, Carlo Costantini, Victoria L. M. Davidson, Elena Deligianni, Tania Dottorini, Vicky Dritsou, Stacey B. Gabriel, Wamdaogo M. Guelbeogo, Andrew B. Hall, Mira V. Han, Thaung Hlaing, Daniel S. T. Hughes, Adam M. Jenkins, Xiaofang Jiang, Irwin Jungreis, Evdoxia G. Kakani, Maryam Kamali, Petri Kemppainen, Ryan C. Kennedy, Ioannis K. Kirmitzoglou, Lizette L. Koekemoer, Njoroge Laban, Nicholas Langridge, Mara K. N. Lawniczak, Manolis Lirakis, Neil F. Lobo, Ernesto Lowy, Robert M. MacCallum, Chunhong Mao, Gareth Maslen, Charles Mbogo, Jenny McCarthy, Kristin Michel, Sara N. Mitchell, Wendy Moore, Katherine A. Murphy, Anastasia N. Naumenko, Tony Nolan, Eva M. Novoa, Samantha O'Loughlin, Chioma Oringanje, Mohammad A. Oshaghi, Nazzy Pakpour, Philippos A. Papathanos, Ashley N. Peery, Michael Povelones, Anil Prakash, David P. Price, Ashok Rajaraman, Lisa J. Reimer, David C. Rinker, Antonis Rokas, Tanya L. Russell, N'Fale Sagnon, Maria V. Sharakhova, Terrance Shea, Felipe A. Simão, Frederic Simard, Michel A. Slotman, Pradya Somboon, Vladimir Stegniy, Claudio J. Struchiner, Gregg W. C. Thomas, Marta Tojo, Pantelis Topalis, José M. C. Tubio, Maria F. Unger, John Vontas, Catherine Walton, Craig S. Wilding, Judith H. Willis, Yi-Chieh Wu, Guiyun Yan, Evgeny M. Zdobnov, Xiaofan Zhou, Flaminia Catteruccia, George K. Christophides, Frank H. Collins, Robert S. Cornman, Andrea Crisanti, Martin J. Donnelly, Scott J. Emrich, Michael C. Fontaine, William Gelbart, Matthew W. Hahn, Immo A. Hansen, Paul I. Howell, Fotis C. Kafatos, Manolis Kellis, Daniel Lawson, Christos Louis, Shirley Luckhart, Marc A. T. Muskavitch, José M. Ribeiro, Michael A. Riehle, Igor V. Sharakhov, Zhijian Tu, Laurence J. Zwiebel, Nora J. Besansky*

*Corresponding author. E-mail: neafsey@broadinstitute.org (D.E.N.); nbesansk@nd.edu (N.J.B.)

Published 27 November 2014 on *Science* Express
DOI: 10.1126/science.1258522

This PDF file includes:

Table of Contents:

Author Details

Author Contributions

Project Leadership
Daniel E. Neafsey*, Robert M. Waterhouse*, Nora J. Besansky
*These individuals contributed equally

Project coordination
George K. Christophides, Frank H. Collins, Scott J. Emrich, Michael C. Fontaine, William Gelbart, Matthew W. Hahn, Paul I. Howell, Fotis C. Kafatos, Daniel Lawson, Marc A.T. Muskavitch

Resource generation
Paul I. Howell (primary supplier of tissue), Mohammad R. Abai, James E. Allen, James Amon, Hamidreza Basseri, Nora J. Besansky, Thomas R. Burkot, Austin Burt, Mikkel Christensen, Frank H. Collins, Carlo Costantini, Wamdaogo M. Guelbeogo, Thaung Hlaing, Daniel S.T. Hughes, Maryam Kamali, Petri Kemppainen, Lizette L. Koekemoer, Njoroge Laban, Nicholas Langridge, Daniel Lawson, Neil F. Lobo, Ernesto Lowy, Gareth Maslen, Charles Mbogo, Samantha O'Loughlin, Mohammad A. Oshaghi, Anil Prakash, Lisa J. Reimer, Tanya L. Russell, N'Fale Sagnon, Frederic Simard, Michel A. Slotman, Pradya Somboon, Catherine Walton, Guiyun Yan

Non-coding RNA gene annotation
James E. Allen, Elena Deligianni, Vicky Dritsou, Adam M. Jenkins, Daniel Lawson, Christos Louis, Marc A.T. Muskavitch, Pantelis Topalis, Robert M. Waterhouse

Automated genome annotation
Daniel S.T. Hughes, Daniel Lawson, Ernesto Lowy

Manual gene model curation
Bruno Arcà, Stephanie A. Blandin, Mikkel Christensen, Martin J. Donnelly, Adam M. Jenkins, Irwin Jungreis, Evdoxia Kakani, Daniel Lawson, Sara N. Mitchell, Marc A.T. Muskavitch, David P. Price, José M. Ribeiro, Robert M. Waterhouse, Craig S. Wilding, Judith H. Willis, Xiaofan Zhou, Laurence J. Zwiebel

Genome sequencing, genome assembly, and assembly improvement
Sergey S. Aganezov, Max A. Alekseyev, Aaron Berlin, Andrew B. Hall, Xiaofang Jiang, Daniel E. Neafsey (theme leader), Ashley N. Peery, Igor V. Sharakhov, Terrance Shea, Robert M. Waterhouse

Assessment of assemblies & annotations
Sergey S. Aganezov, Max A. Alekseyev, Scott J. Emrich, Michael C. Fontaine, Daniel Lawson, Ernesto Lowy, Daniel E. Neafsey (theme leader), Felipe A. Simão, Robert M. Waterhouse, Yi-Chieh Wu

Global gene family analyses
Matthew W. Hahn (theme leader), Gregg W.C. Thomas

Global orthology delineation
Robert M. Waterhouse

Introns
Felipe A. Simão, Robert M. Waterhouse (theme leader)

Gene fusion/fission
Robert M. Waterhouse, Yi-Chieh Wu (theme leader)

Codon bias
Eva M. Novoa

Whole genome alignments
Michael C. Fontaine, Irwin Jungreis, Robert M. Waterhouse (theme leader)

Global selection analyses
Mara K.N. Lawniczak, Daniel E. Neafsey (theme leader), Robert M. Waterhouse

Stop codon readthrough
Clara S. Chan, Irwin Jungreis (theme leader), Robert M. Waterhouse

Chemosensation
David C. Rinker, Antonis Rokas, Xiaofan Zhou, Laurence J. Zwiebel (theme leader)

Neuropeptide hormones and signaling proteins
Joanna C. Chiu, Shirley Luckhart (theme leader), Wendy Moore, Katherine A. Murphy, Chioma Oringanje, Nazzy Pakpour, Michael A. Riehle (theme leader)

Transporters
Immo Hansen (theme leader), David P. Price

Reproduction
Flaminia Catteruccia (theme leader), Tania Dottorini, Paul I. Howell, Evdoxia Kakani, Sara N. Mitchell

Immunity
Stephanie A. Blandin, Andrew I. Brockman, George K. Christophides (theme leader), Victoria L. M. Davidson, Ioannis K. Kirmitzoglou, Robert M. MacCallum, Kristin Michel, Michael Povelones, Robert M. Waterhouse (theme leader)

Insecticide resistance
Martin J. Donnelly (theme leader), Manolis Lirakis, Christos Louis, Pantelis Topalis, John Vontas, Craig S. Wilding

Sex-biased gene expression
Andrea Crisanti (theme leader), Tania Dottorini, Mara K.N. Lawniczak, Tony Nolan, Philippos A. Papathanos

Sialome
Bruno Arcà, José M. Ribeiro (theme leader), Claudio J. Struchiner

Chromosomal evolution
Sergey S. Aganezov, Max A. Alekseyev, Gleb Artemov, Lauren A. Assour, Cedric Chauve, Scott J. Emrich, Andrew B. Hall, Mira V. Han, Xiaofang Jiang, Anastasia N. Naumenko, Ashley N. Peery, Ashok Rajaraman, Igor V. Sharakhov (theme leader), Maria V. Sharakhova, Vladimir Stegniy, Robert M. Waterhouse

Repetitive elements
Peter Arensburger, Xiaofang Jiang, Ryan C. Kennedy, Chunhong Mao, Jenny McCarthy, José M. Ribeiro, Marta Tojo, Zhijian Tu (theme leader), José M.C. Tubio, Maria F. Unger, Robert M. Waterhouse

Cuticular proteins
Robert S. Cornman (theme leader), Judith H. Willis

Epigenetic modifiers
Adam M. Jenkins, Marc A.T. Muskavitch (theme leader)

Advisors to participant(s)
Bruce W. Birren, George K. Christophides, Frank H. Collins, Michael C. Fontaine, Stacey B. Gabriel, William Gelbart, Manolis Kellis, Kristin Michel, Zhijian Tu, Robert M. Waterhouse, Evgeny M. Zdobnov

Author Acknowledgements

AMJ was supported by Boston College Research Fund. ANP and IVS were supported in part by the Institute for Critical Technology and Applied Science (ICTAS) and the NSF award 0850198. AP acknowledges the staff and field teams of RMRC, Dibrugarh. AR was supported by PIMS International Graduate Training Centre in Mathematical Biology. BA was supported by funds from the EU grant INFRAVEC (228421) and from MIUR (PRIN 2010-2011, SKINFLAM, 2010C2LKKJ_004). CC was supported by NSERC Discovery Grant 249834-2011. CJS was supported by Brazilian Research Council (CNPq) and FAPERJ. CL and ED were supported in part by funds from the EU grant INFRAVEC (228421). CW was supported by grant 089229/Z/09/Z from the Wellcome Trust. DCR acknowledges the National Institute on Deafness and Other Communication Disorders, National Research Service Award F31 DC012991. DPP, IAH was supported by grant # 1SC1AI109055-01 from the National Institute of Health. ED was supported in part by the Infravec project (FP7). EMN was supported by Human Frontier Science Program postdoctoral fellowship. FAS was supported by Swiss National Science Foundation award 31003A-143936 to EMZ. FC was supported by a European Research Council FP7 ERC Starting Grant project Anorep (grant ID: 260897) and an NIH grant (grant ID: NIH 1R01AI104956-01A1). GA was supported by The President of Russia Federation Grant for young scientists MK-4158.2012.4. IJ acknowledges Mike Lin for supplying a PhyloCSF training utility. IVS was supported by grants from National Institutes of Health R21AI094289 and R21AI099528. JHW was supported by NIH R01AI055624. JMCT was supported by Marie Curie IEF Fellowship. JMR was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, USA. LA was supported by a Naughton Family fellowship. LJZ acknowledges the resources of the Advanced Computing Center for Research and Education at Vanderbilt University and was supported by grants from the Innovation and Discovery in Engineering And Science (IDEAS) program of Vanderbilt University. MAA was supported by the National Science Foundation under Grant No. IIS-1253614. MAO acknowledges Dr Mohammad M. Sedaghat for help collecting *Anopheles culicifacies* specimens. MAR was supported by NIH grants AI073745, AI107263 & AI080799. MJD and CSW acknowledge Lien Tran for help collecting *Anopheles epiroticus* specimens. MKNL was supported by MRC CDA

G1100339. MWH was supported by National Science Foundation grant DBI-0845494. NJB was supported by the National Institutes of Health (R01AI076584, R21AI101459). NJB, AB, AC, MCF, TN and SO'L were supported by the FNIH through the VCTR program of the Grand Challenges in Global Health Initiative. RCK and MFU acknowledge the Notre Dame Center for Research Computing. RMW was supported by Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312. RSC was supported by NIH grant A155624 awarded to JHW. SJE, LA, CL, PT, DSTH, GM, JEA, NL, DL, FHC, RMM, MC, RCK, and MFU were supported in part by National Institutes of Health/National Institute for Allergy and Infectious Diseases [grant number HHSN272200900039C; VectorBase]. SAB laboratory received support from Inserm, CNRS and University of Strasbourg and from the French Agence Nationale de la Recherche through the program Investissement d'Avenir ANR-11-LABX-0024, it is supported by the European Research Council starting grant N_260918. SL was supported by NIH grants AI073745, AI107263 & AI080799. SSA was supported by the National Science Foundation under Grant No. IIS-1253614. TD was supported by Marie Curie Intra-European Fellowship for Career Development (IEF) PIEFGA-273268. TH acknowledges the staff and field team of DMR. VD and CL were supported by i-Move fellowships (FP7). VLMD and KM were supported by National Institutes of Health grant R01AI095842. WM was supported by NSF grants DEB-0908187 and DEB-1256976. ZT was supported by R01AI077680.

Supplementary Materials, Methods, and Information

Samples, Sequencing, and Assembly

Biological samples

The *Anopheles* 16 genomes project (*8*) applied both evolutionary and eco-ethological criteria to the choice of species and strains for whole genome sequencing. Four species were selected from the *An. gambiae* sibling species complex followed by sampling at increasing evolutionary distances within the three main anopheline subgenera with particular emphasis on the subgenus *Cellia* (Table S1). A subset of species was also selected for SNP-discovery sequencing (Table S2). The selected species also cover the spectra of geography - old to new world, and vectorial capacity - from highly efficient to non-vectors, and most are available from colonies housed by the Malaria Research and Reference Reagent Resource Center.

Table S1. Sample sources for reference genome assemblies.

Subgenus, strain, vector status, and sample sources for each of the 16 newly-sequenced anopheline species.

Species	Subgenus	Strain	Vector Status	Sample Source
A. albimanus	*Nyssorhynchus*	STECLA¤	Minor	El Salvador, isofemale subcolony
A. arabiensis	*Cellia [Pyretophorus†]*	DONGOLA¤	Major	Sudan, isofemale subcolony 2Rb/b homokaryotype
A. atroparvus	*Anopheles*	EBRO¤	Major	Spain, isofemale subcolony
A. christyi	*Cellia [Pyretophorus]*	No colony	Non	Kenya, wild collected
A. culicifacies	*Cellia [Myzomyia]*	No colony	Major	Iran, wild collected
A. dirus	*Cellia [Neomyzomyia]*	WRAIR2¤	Major	Thailand, isofemale subcolony
A. epiroticus	*Cellia [Pyretophorus]*	No colony	Minor	Vietnam, wild collected
A. farauti	*Cellia [Neomyzomyia]*	FAR1¤	Major	Papua New Guinea, isofemale subcolony
A. funestus	*Cellia [Myzomyia]*	FUMOZ¤	Major	Mozambique, Matolo Province
A. maculatus	*Cellia [Neocellia]*	MACULATUS3	Major	Malaysia, preserved females
A. melas	*Cellia [Pyretophorus†]*	No colony	Minor	Cameroon, wild collected
A. merus	*Cellia [Pyretophorus†]*	MAF¤	Minor	South Africa, isofemale subcolony
A. minimus	*Cellia [Myzomyia]*	MINIMUS1¤	Minor	Thailand, Mae Sot, Tak Province
A. quadriannulatus	*Cellia [Pyretophorus†]*	SANGWE	Non	South Africa, isofemale subcolony heterokaryotype X+f/f
A. sinensis	*Anopheles*	SINENSIS¤	Major	Korea, isofemale subcolony
A. stephensi	*Cellia [Neocellia]*	SDA-500	Major	Pakistan, isofemale subcolony

† Member of *Anopheles gambiae* species complex.

¤ Colony available from the Malaria Research and Reference Reagent Resource Center.

Table S2. Sample sources for SNP discovery.

Numbers and sources of SNP-discovery samples for 12 anopheline species.

Species	Samples	Sample Sources
A. arabiensis	12	Burkina Faso, Cameroon, Kenya.
A. culicifacies	8	Species A, species D, and undefined taxon "D-like" from localities in Iran.
A. dirus	4	Three pools of Species D (baimaii) from Myanmar distinguished by large geographic distances, inversion frequency differences, and larval habitats (wells in villages versus forest), with outgroup sample from N. India.
A. epiroticus	12	Vietnam, samples from 2 localities that differ in levels of insecticide resistance.
A. farauti	16	Queensland, Haleta, Madang, Tanna.
A. funestus	10	Karyotyped forms "Folonzo" and "Kiribina" collected in sympatry in Burkina Faso.
A. melas	11	Individual samples from Cameroon, Equatorial Guinea, plus three pooled samples from Cameroon, Equatorial Guinea, and The Gambia.
A. merus	10	Kenya, South Africa.
A. minimus	2	Pools from Assam, India and Lamphun Province, Thailand
A. quadriannulatus	10	Zimbabwe (sp. A).
A. sinensis	2	China.
A. stephensi	5	Southern Iran, Southeast Iran, India.

Geographic distributions

The distribution map presented in Figure 1 of the main text was compiled using predicted distributions from The Malaria Atlas Project (www.map.ox.ac.uk). For each species, the georeferenced distribution map was retrieved from the repository predicted as detailed in (*55-58*) and publicly available at www.map.ox.ac.uk/explore/mosquito-malaria-vectors. These were imported into ArcGIS 10.2 (ESRI) and colored across eight global regions, where species for which predictions were not available (*An. quadriannulatus*, *An. christyi*, *An. epiroticus*) have distributions that are completely embedded in those of other species.

Sequencing strategies

All species were assembled from 101 base pair (bp) paired-end reads generated by an Illumina HiSeq2000 platform. A single female mosquito of each species was used to generate two sequencing libraries with different insert sizes: a 180 bp insert 'fragment' library and a 1.5 kb 'jump' library. Template was limited to a single individual per species in order to minimize heterozygosity in the sequencing data. The fragment libraries were generated from native genomic DNA. The jump libraries were generated from whole genome amplified (Qiagen REPLI-g) genomic DNA from the same individual, as the amount of input DNA required by the jump library protocol exceeded

the typical DNA extraction yield from an individual mosquito. The fragment and jump libraries were sequenced to a depth of approximately 100-fold coverage (fragment range = 32-170; jump range = 54-145) (Table S3). For 11 species where abundant template was available (laboratory colonies; *An. albimanus*, *An. arabiensis*, *An. atroparvus*, *An. dirus*, *An. farauti*, *An. funestus*, *An. merus*, *An. minimus*, *An. quadriannulatus*, *An. stephensi*, *An. sinensis*), a third large insert 'fosill' sequencing library (*59*) was constructed from high molecular weight DNA extracted from several hundred females to improve scaffolding. Fosill libraries were lightly sequenced to a mean depth of 3-fold coverage (range = 1-11). RNAseq was performed for the 11 species with laboratory colonies described above as well as from RNA-later preserved wild-caught adult carcasses from *An. epiroticus*. Tissue for RNA extraction for the colonized species was extracted as a single pool of males and females from various life stages (larvae, pupae, and adults). Illumina RNAseq libraries were generated in a strand-agnostic fashion and sequenced to a depth of at least 75-fold coverage.

Table S3. Sequencing coverage statistics.

Sequencing coverage statistics of fragment, jump, and fosill libraries for each of the 16 newly-sequenced anopheline species.

Species	Fragments	Jumps	Fosills
An. albimanus	74	119	11
An. arabiensis	62	119	8
An. atroparvus	30	85	7
An. christyi	77	100	-
An. culicifacies	100	145	-
An. dirus	165	56	7
An. epiroticus	142	67	-
An. farauti	84	100	5
An. funestus	170	76	1
An. maculatus	32	71	-
An. melas	150	88	-
An. merus	74	111	5
An. minimus	162	54	7
An. quadriannulatus	77	95	3
An. sinensis	50	50	5
An. stephensi	153	56	1

Assembly strategies

Assemblies were generated with ALLPATHS-LG (*11*), with the "HAPLOIDIFY" parameter set to "True" for all assemblies to minimize separate assembly of homologous haplotypes due to heterozygosity. K-mer normalization was applied to *An. farauti* and *An. epiroticus* read data using the ALLPATHS tool ReadFilterByKmerFreq to generate down-sampled read data sets normalized by a factor of 0.6 and 0.8, respectively. Assemblies for species belonging to the *An. gambiae* species complex (*An. arabiensis*, *An. quadriannulatus*, *An. merus*, *An. melas*) were reference-assisted (ALLPATHS-LG "ASSISTED_PATCHING=1" parameter) using the *Anopheles gambiae* P3 assembly as a reference. Reference assistance reduces, but does not obviate, the threshold of read-pair data required to support contig joins, and is therefore not expected to introduce reference bias into assemblies. The post-assembly improvement tool Pilon was run on assembly scaffolds to close captured gaps and identify assembly errors (www.broadinstitute.org/software/pilon). Assemblies were analyzed using the GAEMR assembly evaluation package and manually reviewed for quality (www.broadinstitute.org/software/gaemr). Assembly contigs were screened against an NCBI mitochondrial database to identify and remove mitochondrial contigs. Additionally, host and bacterial contamination were removed by filtering contigs with significant alignments to host and bacteria sequences in the nucleotide (nt) database.

De novo transcriptome assemblies were generated using Trinity (*60*). Trinity was run in strand-specific paired-end mode (--SS_lib_type set to "RF") with the --min_kmer_cov parameter set to 2. Assembly contigs were screened against an NCBI vector database to identify and remove Illumina adapter contamination.

Assembly statistics

Genome assembly statistics (Table S4) were computed with custom Perl scripts to compare the properties of the 16 new anopheline genomes with five previously sequenced anopheline genomes - *An. gambiae* PEST (*4*) and Pimperena S (*61*), *An. coluzzii* (*61*) (formally *An. gambiae* M molecular form), *An. darling* (*5*), and *An. stephensi* INDIAN (*7*) - and those of three other Dipterans - *Aedes aegypti* (*62*), *Culex quinquefasciatus* (*63*), *Drosophila melanogaster* (*64*). The new anopheline assemblies range in size from 142Mb for *An. maculatus* to 376Mb for the second version of the *An. sinensis* assembly. The most contiguous assembly is that of *An. albimanus*, assembled into 204 scaffolds with an N50 of 18Kb, which contrasts the fragmented assembly of *An. maculatus* with 47,797 scaffolds and an N50 of only 4Kb.

Table S4. Genome assembly statistics.

Genome assembly statistics for each of the 16 newly-sequenced anopheline genomes, five previously-sequenced anopheline genomes, and three other dipteran genomes.

Species	Assembly Version	GenBank Assembly	Assembly Size (Mb)	Gaps (Kb)	Scaffold N25‡ (Kb)	Scaffold N50‡ (Kb)	Scaffold N75‡ (Kb)	Number of Scaffolds	%GC
A. albimanus	AalbS1	GCA_000349125.1	170.5	6,961	24,066	18,068	8,610	204	49.21
A. arabiensis	AaraD1	GCA_000349185.1	246.6	35,125	10,035	5,604	2,699	1,214	44.68
A. atroparvus	AatrE1	GCA_000473505.1	224.3	35,338	12,071	9,207	4,026	1,371	46.35
A. christyi	AchrA1	GCA_000349165.1	172.7	2,671	18	9	4	30,369	42.74
A. coluzzii	AcolM1	GCA_000150765.1	224.5	14,926	7,194	4,437	1,263	10,521	44.38
A. culicifacies	AculA1	GCA_000473375.1	203.0	15,840	42	22	11	16,162	42.68
A. darlingi	AdarC2	-	134.7	27	272	115	54	2,160	48.39
A. dirus	AdirW1	GCA_000349145.1	216.3	18,566	13,137	6,906	3,666	1,266	46.18
A. epiroticus	AepiE1	GCA_000349105.1	223.5	20,855	743	367	162	2,673	43.95
A. farauti	AfarF1	GCA_000473445.1	181.0	5,030	1,777	1,197	541	550	44.69
A. farauti	AfarF2	GCA_000473445.2	183.1	7,280	22,739	12,895	6,025	310	44.69
A. funestus	AfunF1	GCA_000349085.1	225.2	35,208	1,127	672	380	1,392	41.59
A. gambiae	AgamP3	GCA_000005575.1	273.1	20,655	53,201	49,364	42,390	7†	44.27
A. gambiae	AgamS1	GCA_000150785.1	236.4	8,363	6,249	3,801	1,822	13,042	44.33
A. maculatus	AmacM1	GCA_000473185.1	141.9	10,063	7	4	2	47,797	44.21
A. melas	AmelC1	GCA_000473525.1	227.4	20,678	31	18	10	20,281	44.94
A. melas	AmelC2	GCA_000473525.2	224.2	20,733	31	18	10	20,229	44.84
A. merus	AmerM1	GCA_000473845.1	251.8	33,614	658	342	134	2,753	44.64
A. merus	AmerM2	GCA_000473845.2	288.0	70,729	2,833	1,490	538	2,027	44.64
A. minimus	AminM1	GCA_000349025.1	201.8	15,387	21,278	10,313	5,773	678	42.70
A. quadriannulatus	AquaS1	GCA_000349065.1	283.8	74,862	2,616	1,641	622	2,823	44.76
A. sinensis	AsinS1	GCA_000472065.1	241.4	49,229	151	81	35	11,270	43.93
A. sinensis	AsinS2	GCA_000472065.2	375.8	185,483	1,104	579	208	10,448	43.95
A. stephensi	AsteI2	GCA_000300775.2	221.3	11,843	2,767	1,591	597	23,371	44.80
A. stephensi	AsteS1	GCA_000349045.1	225.4	29,185	1,402	837	450	1,110	45.02
Aedes aegypti	AaegL1	GCA_000004015.1	1384.0	73,881	2,717	1,547	742	4,758	38.27
Culex quinquefasciatus	CpipJ1	GCA_000209185.1	579.0	39,083	949	487	221	3,171	37.42
Drosophila melanogaster	BDGP5	GCA_000001215.2†	168.7	67,764	27,905	23,012	22,423	14†	41.74

‡ N25/N50/N75 is the scaffold length, *x*, such that 25%/50%/75% of the genome is assembled on scaffolds of length *x* or longer.
† The AgamP3 and BDGP5 assemblies are mapped to chromosomes, and the BDGP5 assembly from FlyBase (dmel-r5.57) additionally contains heterochromatin.

The relative completeness in terms of expected gene content of the genome assemblies was assessed by two complementary approaches: mapping of RNAseq-based transcripts and sequence searches with conserved arthropod orthologs. Mapping the 12 unfiltered assembled transcriptomes to each respective genome assembly with exonerate (*65*) recovered between 73% and 85% of transcripts (Table S5) indicating good assembly completeness by this measure.

Table S5. Mapped RNAseq reads.

Statistics of assembled RNAseq reads mapped to each of the 12 newly-sequenced anopheline genomes with RNAseq samples.

Species	Assembly	RNA samples	Number of transcripts‡	Mapped transcripts‡	%Mapped
An. albimanus	AalbS1	1	46,888	39,515	84.28
An. arabiensis	AaraD1	1	52,981	44,918	84.78
An. atroparvus	AatrE1	1	63,677	50,364*	79.09
An. dirus	AdirW1	1	55,210	43,091	78.05
An. epiroticus	AepiE1	-	60,857	51,023	83.84
An. farauti	AfarF1	4	71,177	58,527*	82.23
An. funestus	AfunF1	1	60,242	46,809	77.70
An. merus	AmerM1	4	59,732	47,251*	79.11
An. minimus	AminM1	1	64,422	49,970	77.57
An. quadriannulatus	AquaS1	1	64,784	51,627	79.69
An. sinensis	AsinS1	3	87,412	64,045*	73.27
An. stephensi	AsteS1	1	67,144	53,464	79.63

‡ Trinity assembled contigs mapped to reference assemblies (exonerate)
* MAKER gff based counts

Assessment of assembly completeness using Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs) (*66*) reveal generally very few missing BUSCOs and 60-70% near-complete gene recoveries from the assembled genomes (Figure S1). For this analysis, 3,369 *Drosophila melanogaster* BUSCOs selected from OrthoDB7 (www.orthodb.org) Arthropod BUSCOs, 2,966 of these were fully recovered (at >95% length cutoff) from the *D. melanogaster* genome and were then used to search the mosquito genomes. Some, e.g. those found in none of the mosquito genomes could be real losses from the mosquito ancestor, or they could be too fast-evolving or too full of low-complexity sequences to be detected. Thus, a strict filter was imposed such that for a BUSCO to be considered missing from any genome it had to have been found in all the other mosquito assemblies.

This left a total of 2,898 *D. melanogaster* BUSCOs found in all, or all-but-one, of the 23 mosquito genome assemblies. In terms of missing BUSCOs, the new anopheline assemblies all perform very well, except for the *An. maculatus* assembly that shows a much higher number of missing BUSCOs. In terms of apparently duplicated BUSCOs, the elevated numbers in *An. maculatus* are explained by its fragmented assembly (fragments on different scaffolds appear as multiple hits for the same BUSCO), but further investigation of the high number in *An. melas* revealed several scaffolds or parts of scaffolds that were highly similar, suggesting that some haplotypes had not been successfully collapsed during the assembly procedure - these were subsequently identified and removed to produce a new assembly (AmelC2).

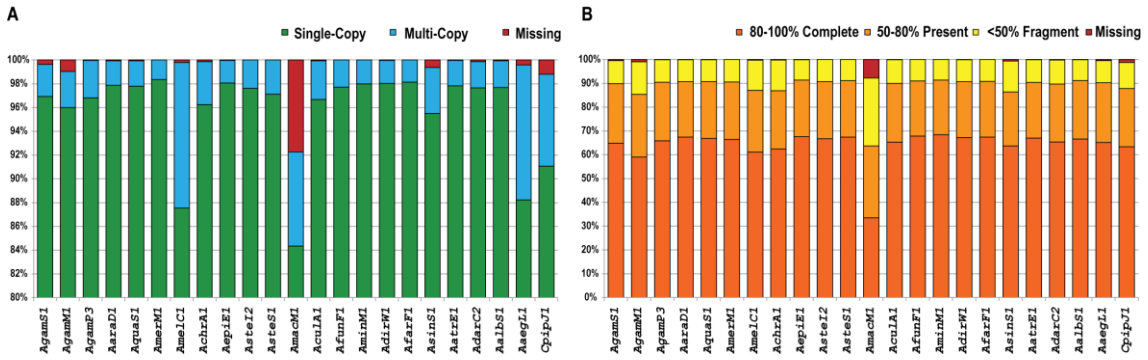


Figure S1. Assembly assessments with universal single-copy orthologs.

Recovery of 2,898 *Drosophila melanogaster* Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs) from the mosquito genome assemblies **A**. Presence: Single-copy, one good hit region found where the orthologue can be predicted. Multi-copy, 2 or more highly similar orthologues (or fragments of orthologs) found ~ recent duplications or assembly errors. Missing, no significant BLAST hits (uniquely missing). **B**. Fragmentation: Complete, more than 80% length recovered. Present, 50-80% length recovered. Fragment, <50% length recovered. Missing, no significant BLAST hits (uniquely missing).

Genome Annotation

Protein-coding gene annotations

VectorBase RNA gene prediction pipeline: Genome annotation was undertaken using both *ab initio* and similarity based methods with subsequent aggregation using the MAKER software (*13*). The annotation process can be broken down into the following phases: (1) Identification of *de novo* repeat sequences using RepeatScout (*67*) and RECON (*68*). These were supplemented with publicly available repeat sequences from GenBank and mapped to the genome assembly using RepeatMasker (*69*). Repetitive regions where excluded from further analyses with regard to the prediction of protein-coding loci. Repeat regions were checked for protein similarities to avoid over prediction which may mask valid protein-coding genes in downstream steps of the annotation process. (2) Training of *ab initio* gene prediction programs SNAP (*70*) and Augustus (*71*) using in the first instance transcript consensus sequences from Trinity assemblies of RNAseq datasets (*60*). Subsequent rounds of re-training were based on the output of the

prediction programs themselves. (3) Similarity based gene predictions were generated using exonerate alignments of EST sequences and Trinity-based assemblies of RNAseq datasets (*65*), alignment and merging of transcriptional fragments (transfrags) from RNAseq datasets using tophat & bowtie (*72*) and protein-based predictions using taxonomically constrained subsets of the non-redundant (nr) protein database UniProt (*73*). (4) Gene predictions from both the *ab initio* and similarity approaches were aggregated into a final set using three rounds of the MAKER software (*13*). The first two rounds were designed to iteratively improve the training of the *ab initio* gene predictions and a final round informed with protein similarities to all metazoan sequences in the nr protein database to guide the final predictions. (5) The candidate gene sets were assessed for completeness, screened for potential transposable elements and filtered based on comparative analysis with the other anopheline datasets. The resulting data sets formed the basis for community led prediction appraisal and improvement.

Total protein-coding gene counts range from 10,738 to 16,149 with a mean of 13,377, slightly more than the 12,810 genes of *An. gambiae* (AgamP3.8) (Table S6). The mean gene counts are lower and their standard deviations (sd) are higher for all 19 anophelines (mean 13,110, sd. 1,397) and the 16 newly-sequenced anophelines (mean 13,377, sd. 1,294) compared to the 12 drosophilids (mean 15,361, sd. 852). The total coding sequence lengths of the protein-coding exons, ranging from 40,000-57,000, make up 7-12% of the assembled genome length, with median lengths of 200-320 bp. Automated genome annotation must balance many different sources of evidence such as *ab initio* gene predictions and alignments of homologs from closely-related species or RNAseq transcripts from experimental samples. Some of the variation in total gene counts likely stems from the different amounts of supporting RNAseq data, e.g. four species had no RNAseq data. Particular annotation problems exist for members of gene families which have undergone recent duplication to form tandem arrays of loci and with the merging/splitting of predictions based on repeated alignments and/or poor interpretation of transfrag data from RNAseq. The resulting annotations are therefore not perfect or complete (see section on the assessment of protein-coding gene annotations) and benefit greatly from ongoing manual curation to confirm or correct the automated predictions. This is facilitated by the VectorBase Community Annotation Portal which has already received more than 2,000 submissions (as at July 1st 2014) from the research community.

Table S6. Gene annotation statistics.

A summary of annotated protein-coding genes and non-coding RNA genes from each of the 16 newly-sequenced anopheline genomes and six previously-sequenced mosquito genomes.

Species	Gene Set	Protein-Coding Genes					Non-Coding RNA Genes		
		Number of			Med. CDS Len.† (bp)	% of Genome	Number of		
		Genes [1 exon]	Transcripts	Exons‡			tRNAs*	miRNAs*	sn(o)RNAs*
16 Newly Sequenced Anopheline Genomes									
A. albimanus	AalbS1.1	11,911 [1,384]	11,994	52,166	303.0	12.28	306	78	31
A. arabiensis	AaraD1.1	13,162 [1,592]	13,220	54,162	301.5	8.65	358	96	43
A. atroparvus	AatrE1.1	13,776 [1,033]	13,776	57,495	201.0	9.72	337	63	39
A. christyi	AchrA1.1	10,738 [1,375]	10,815	42,199	213.0	10.39	300	83	21
A. culicifacies	AculA1.1	14,335 [1,804]	14,335	54,663	204.0	10.51	285	90	46
A. dirus	AdirW1.1	12,781 [1,547]	12,892	53,534	304.5	9.80	345	72	31
A. epiroticus	AepiE1.1	12,078 [1,361]	12,130	50,261	209.0	9.13	320	87	36
A. farauti	AfarF1.1	13,217 [980]	13,217	56,282	300.0	11.84	354	74	42
A. funestus	AfunF1.1	13,344 [1,867]	13,485	54,604	300.0	9.48	286	78	32
A. maculatus	AmacM1.1	14,835 [3,735]	14,835	39,693	240.0	11.59	127	59	14
A. melas	AmelC1.1	16,149 [2,322]	16,149	56,320	319.5	10.10	349	98	40
A. merus	AmerM1.1	13,887 [1,029]	13,887	56,135	203.0	8.50	353	103	51
A. minimus	AminM1.1	12,560 [1,499]	12,653	53,567	199.0	10.39	313	76	33
A. quadriannulatus	AquaS1.1	13,349 [1,658]	13,415	53,295	201.0	7.45	357	89	44
A. sinensis	AsinS1.1	14,791 [1,432]	14,791	55,564	307.5	8.72	362	87	29
A. stephensi	AsteS1.1	13,113 [1,774]	13,251	53,671	201.0	9.41	329	76	36
Previously Sequenced Mosquito Genomes									
Aedes aegypti	AaegL2.2	15,784 [2,179]	17,143	62,177	204.0	1.65	962	165	88
Anopheles gambiae	AgamP3.8	12,810 [1,595]	14,667	53,466	310.5	7.71	441	187	50
Anopheles coluzzii	AcolM1.0	14,703 [1,471]	14,703	56,643	203.0	9.05	nd	nd	nd
Anopheles darlingi	AdarC2.2	10,457 [884]	10,457	47,988	283.5	13.47	228	105	30
Anopheles stephensi	AsteI2.2	11,789 [1,304]	11,789	49,154	316.5	9.16	344	75	40
Culex quinquefasciatus	CpipJ1.4	18,955 [1,879]	19,019	70,991	194.0	4.31	nd	134	72

‡ Number of protein-coding exons.
† Median CDS (exon protein-coding sequence) length.
* tRNAs: transfer RNAs; miRNAs: microRNAs; sn(o)RNAs: small nuclear and small nucleolar RNAs.

RNA gene annotations

RNA gene predictions: Detailed analysis of rRNAs, tRNAs, miRNAs, and snoRNAs across the anophelines was performed to complement the automated VectorBase prediction pipeline and are described below.

Ribosomal RNAs: rRNAs: The VectorBase (*74*) anopheline genome assemblies were blasted querying, initially, with the *Drosophila melanogaster* sequences encoding the 5.8S, 18S and 28S genes. To close gaps present in almost all repeats in all genomes analyzed, raw reads found at the Sequence Read Archive (SRA) were blasted using as queries the sequences identified with the BLAST searches previously performed previously for each species; contigs were then manually assembled. The strategy was successful in isolating complete genes in some cases (Figure S2) but failed, in all cases, for the identification of spacer segments, and in many cases, for the assembly of the complete genes. This is most probably due to the presence of repetitive elements interrupting the long genes, see (*75*). Several of the identified segments were used to

perform alignments between the different species (Figure S2). A similar strategy was used for the short rDNA genes (5.8S and 5S) though, in these cases, complete sequences were isolated for all species studied (Figure S2). Alignments were produced using the MUSCLE tool (*76*), while Clustal (*77*) was used to generate cladograms.

Transfer RNAs: tRNAs: All the anopheline genomes were screened with tRNAScan-SE (*78*) to identify tRNA genes and a complementary approach, using multiple whole genome alignments as input to RNAz 2.0 suite (*79*) (Figure S2). The positive predictions were verified as being similar to tRNAs by BLASTing to the Rfam database (*80*). No additional tRNAs, undetected by tRNAScan-SE, were identified. The numbers of tRNA genes were the same as the ones predicted by VectorBase.

MicroRNAs: miRNAs: *Ab initio* miRNA gene predictions were performed using: 1) HHMMiR (*81*), predictions with a score lower than 80% were discarded, and 2) MiRPara (*82*), predictions with a score lower than 80% were discarded (Figure S2). A complementary approach identified miRNAs by similarity: 1) known miRNA genes in miRBase v20 (*83*) from *Aedes aegypti*, *Culex quinquefasciatus*, *Anopheles gambiae* and the invertebrate section of RefSeq at NCBI were used to query all genomes. Hits with a similarity lower that 90% were discarded, and 2) mature miRNAs in miRBase v20 were mapped to the genomic assemblies of the anopheline species with no mismatches allowed. The regions identified were then checked for the presence of miRNA genes. In addition, the RNAz pipeline (*79*) was used to identify genomic regions that may contain non-coding RNA genes. The presence of miRNA genes was detected by blasting positive hits (P-value greater than 90%) to the Rfam database (*80*). Finally, all five lines of evidence were combined. All putative miRNA genes predicted/identified by at least two different pipelines were considered to represent *bona fide* miRNA genes.

Small nucleolar RNAs: snoRNAs: Prediction of box C/D snoRNAs was performed using Snoscan (*84*); candidate sequences returned with a score of >20 were retained. SnoReport (*85*) was subsequently used for an independent prediction of snoRNAs; sequences that had a probability score of pSVM >0.99 using this software were retained. A third, stricter category of potential snoRNA genes consisted of the sequences that satisfied the criteria of both predictions, although with a lower score (pSVM >90) for SnoReport. No box H/ACA snoRNAs were predicted using the pipeline described.

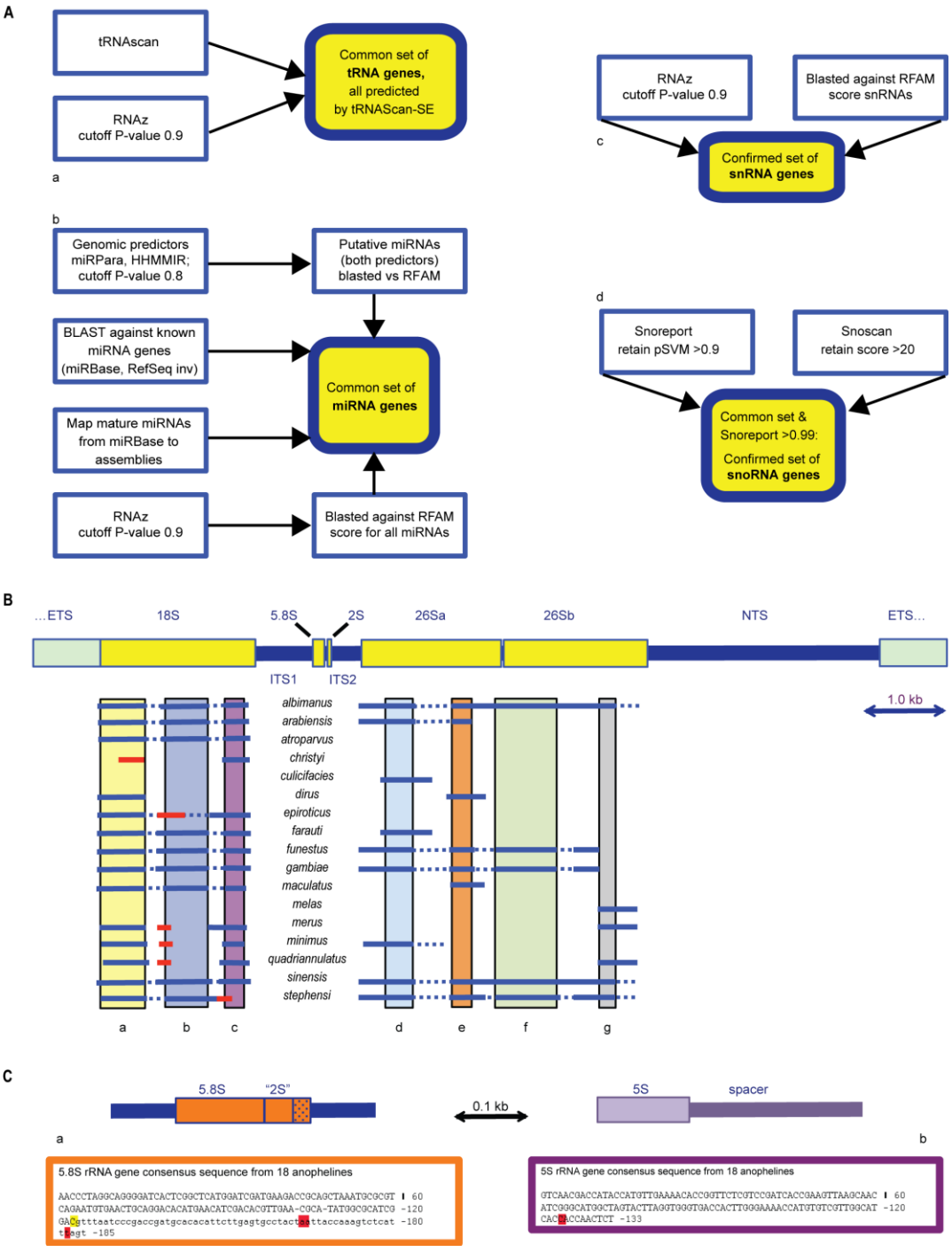


Figure S2. Non-coding RNA genes.

A. Non-coding RNA annotation strategies. Schematic representation of the pipelines used to identify tRNA genes (a), miRNA genes (b), snRNA genes (c) and snoRNA genes (d). **B**. The ribosomal RNA genes of the anophelines. The map at the top shows the canonical repeat of *Drosophila melanogaster* (yellow boxes: mature genes; green boxes: ETS, external transcribed spacer). The lines below show regions that were assembled from each species. Solid lines indicate sequences with a high degree of similarity to the fruit fly, stippled ones show regions that are somewhat dissimilar from their *Drosophila* counterparts. The sequences drawn in red were not used in alignment studies. The

vertical colored boxes below the map show the extent of the sequences used for alignments. Box a extends from nucleotide 27 to 162, box b spans nucleotides 871-1,421 and box c spans nucleotides 1,837-1,942 of the 18S gene. For the 26S gene, box d spans nucleotides 269-413, box e spans nucleotides 972-1,081, box f spans nucleotides 2,452-2,890 and box g spans nucleotides 3,521-3,639. Sizes of the different segments (reference: *D. melanogaster* from PUBMED/EMBL entry M21107 and from (*86*) are: NTS, Non-transcribed spacer or IGS, Intergenic spacer, 3,632 bp; ETS, External transcribed spacer, 864 bp; 18S, rRNA gene, 1,995 bp; ITS1, Internal transcribed spacer 1, 726 bp; 5.8S, rRNA gene, 123 bp; Spacer between 5.8S & 2S gene, 29 bp; 2S, rRNA gene, 30 bp; ITS2, Internal transcribed spacer 2, 385 bp; 26S, rRNA gene (1st part: 1,787 bp, spacer: 46 bp, 2nd part: 2,112 bp), 945 bp. **C**. The 5.8S and 5S rRNA genes of the anophelines. The maps at the top shows a schematic representation of the genes. The 5.8S gene is located within the rDNA repeat while the 5S gene is unlinked and tandemly repeated hundreds of times. a shows the 5.8S consensus sequence obtained from 18 species. In contrast to the fruit fly, the 5.8S gene in anophelines includes the sequence corresponding to the 2S gene of *D. melanogaster*. The two dashes denote indels with respect to the fruit fly gene. The yellow-highlighted nucleotide corresponds to the last nucleotide of the *D. melanogaster* homologue, while the three red-highlighted ones show the last nucleotides of the three different 5.8 species identified among *An. gambiae* transcripts in a ratio of 4:4:1. The predominant species are shown in solid red in the map, while the longest one is stippled. b shows the 5S consensus sequence obtained from 18 species. The red-highlighted nucleotide shows the last base of the molecule based on the analysis of the transcriptomes. In the fruit fly the mature RNA is 130 bases long, 4 shorter than the primary transcript. The spacer has a length of about 245 base pairs (*87*).

The anopheline RNA gene repertoire: The numbers of VectorBase RNA gene predictions range from 127-362 transfer RNAs (tRNAs), 59-103 microRNAs (miRNAs), and 14-51 small nuclear and small nucleolar (sn(o)RNAs), mostly fewer than the 441, 187, and 50 in *An. gambiae* (AgamP3.8) (Table S6), respectively. In depth complementary analyses confirmed the majority of VectorBase predictions and are summarized below.

Ribosomal genes: Complete or "almost" complete sequences (>90% of the length of the gene) were obtained for 10 species and 5 species, respectively, for the 18S and 28S genes (Figure S2). In contrast, full-length sequences were obtained for the small ribosomal genes, 5S and 5.8S. No significantly extended spacer segments were obtained for either external or internal spacers of the ribosomal gene repeats in most species studied. We interpret the failure to obtain complete rDNA units as being due to potential repeated segments that led to poor assemblies in all species examined. Alignments were performed with 3 segments of the 18S gene and 4 of the 28S gene, involving 13 and 9 species respectively. The length of the segments analyzed ranged between ~100 and ~550 bp. The results of the alignments and the cladograms derived from them differed, slightly, from what is known for the relations between the species in question. We attribute these differences to putative polymorphisms that are present in the individual sequences due to the assembly of the NGS reads. Sequence comparisons of two short ribosomal genes (5S and 5.8S) were equally inconclusive. Although, here, the complete sequences were available, the relatively short length of the genes, combined with the high degree of conservation between all species were the reason for the "discrepancies" observed: for example, an additional single polymorphic nucleotide can lead to substantial differences in the computed correlations between the sequences. The number of copies of both 5S

rDNA and the major ribosomal repeat could not be unambiguously determined, but the range computed is within what is known for dipterans.

miRNA genes: Our approach revealed, on average, the presence of 64 miRNA genes per species (range 43-98 genes). As was the case for tRNA genes, *An. maculatus* is again the species with the lowest number of miRNA genes observed. We have identified seven miRNA genes that are present in all anopheline species studied. These are let-7, mir-100, mir-279, mir-957, mir-965. Within members of the *An. gambiae* species complex, we found an additional 18 miRNA genes that are present in all members. These are: mir-8, mir-11, mir-12, mir-100, mir-124, mir-133, mir-137, mir-184, mir-252, mir-275, mir-277, mir-281, mir-308, mir-375, mir-932, mir-970, mir-989, mir-1000. The number of genes determined with our pipeline is consistent with what has been described earlier for insects: the latest version of miRBase (version 20, June 2013, (*83*)) reports 65 miRNA genes for *An. gambiae*, 426 miRNA genes for *Drosophila melanogaster*, 222 miRNA genes for *Apis mellifera*, 124 miRNA genes for *Aedes aegypti* and 93 miRNA genes for *Culex quinquefasciatus*. All of these were used as a positive set in our prediction pipeline.

tRNA genes: Our pipeline identified an average of about 340 tRNA genes in all anophelines studied (range: 285-362) with the exception of *An. maculatus* in which only 127 genes were recognized.

snoRNA genes: Running SnoReport yielded between 314 and 573 positives for the genomes analyzed (pSVM>0.99) while snoscan identified, for 18S rRNA, 21 to 111 and, for 28S rRNA, 14-314 putative genes. Combining the output of the two programs, but using a pSVP >0.90, pointed to a number of common positive snoRNA genes in most species. Their number, though, ranged (overall) from 0-19, i.e. lower that what would be expected.

Assessment of protein-coding gene annotations

Near-universal orthologs: Numbers of potentially missing orthologs and rare gene duplications were estimated by analyzing the species composition and gene counts of mosquito and dipteran orthologous groups from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2) as delineated by the OrthoDB (Waterhouse et al. 2013) methodology. Counts of orthologous groups were computed for each of the following phyletic profiles: single-copy (SC) in all species, but missing (mis) from one or two species; present (PR) in all species (with at least one multi-copy), but missing from one or two species; single-copy in all species, but multi-copy (dup) in one or two species.

Although true gene losses can and do occur, counts of orthologous groups with orthologs from almost all considered species can also reflect relative numbers of potentially missed or poorly-annotated genes in each genome. Similarly for gene duplications, when orthologs are found in single-copy in almost all other species, rare duplications could suggest that the assembly contains haplotypes (especially if the

duplicates have a high percent identity). Thus, phyletic analysis of near-universal orthologous groups serves to highlight numbers of potentially missing orthologs or rare duplications and assess the relative quality of the gene annotations.

The anophelines generally have few potentially missing orthologs: averages for mosquito orthologous groups, 53 SC all-but-1-mis, 91 SC all-but-2-mis, 107 PR all-but-1-mis, 113 PR all-but-2-mis, with the notable exceptions of *An. maculatus* and *An. christyi* that have ~3.5 times as many potentially missing orthologs (Figure S3). These two species have the most fragmented assemblies, which makes annotation more difficult and therefore explains the elevated numbers of potentially missing orthologs. The anophelines also generally have few rarely-duplicated orthologs: averages for mosquito orthologous groups, 40 SC all-but-1-dup, 33 SC all-but-2-dup, with the most notable exception of *An. melas* (~7.5 times the anopheline average) and to a lesser extent *An. maculatus*. Further analysis of the rarely-duplicated orthologs in *An. melas* revealed several scaffolds or parts of scaffolds that were highly similar, suggesting that some haplotypes had not been successfully collapsed during the assembly procedure - these were subsequently identified and removed to produce a new assembly (AmelC2).

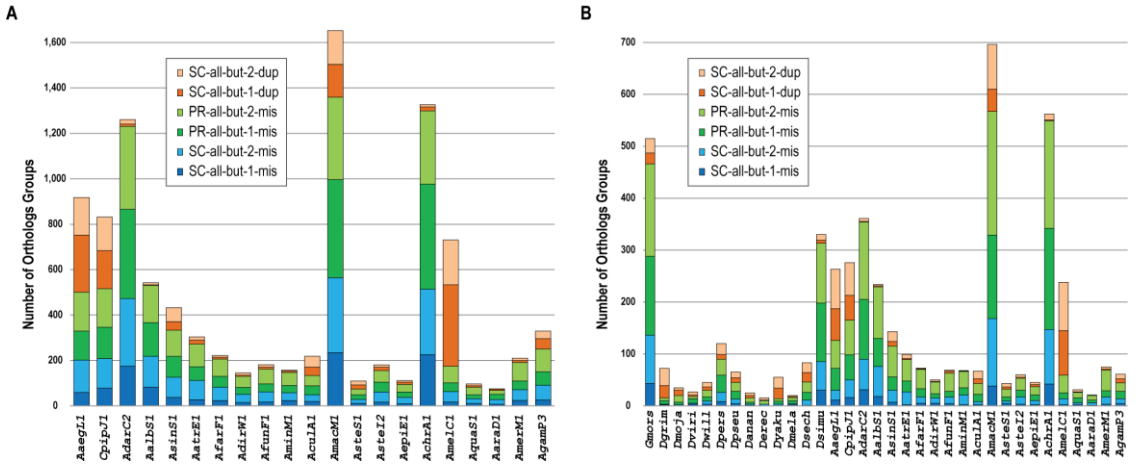


Figure S3. Orthology-based assessment of gene set completeness.

The bar charts show counts of potentially missing orthologs (blue and green) and rare duplications (orange) from near-universal orthologous groups, SC single-copy; PR present; 1/2-mis, missing from one or two species; 1/2-dup, duplicated in one or two species, across **A**. 21 mosquitoes including 19 anophelines, and **B**. across 34 dipterans including 12 *Drosophila*. Species: AgamS1, *An. gambiae* (S); AgamM1, *An. coluzzii*; AgamP3, *An. gambiae* (PEST); AaraD1, *An. arabiensis*; AquaS1, *An. quadriannulatus*; AmerM1, *An. merus*; AmelC1, *An. melas*; AchrA1, *An. christyi*; AepiE1, *An. epiroticus*; AsteI2, *An. stephensi* (INDIAN); AsteS1, *An. stephensi* (SDA-500); AmacM1, *An. maculatus*; AculA1, *An. culicifacies*; AfunF1, *An. funestus*; AminM1, *An. minimus*; AdirW1, *An. dirus*; AfarF1, *An. farauti*; AsinS1, *An. sinensis*; AatrE1, *An. atroparvus*; AdarC2, *An. darlingi*; AalbS1, *An. albimanus*; AaegL1, *Aedes aegypti*; CpipJ1, *Culex quinquefasciatus*; Gmors, *Glossina morsitans*; Dgrim, *D. grimshawi*; Dmoja, *D. mojavensis*; Dviri, *D. virilis*; Dwill, *D. willistoni*; Dpers, *D. persimilis*; Dpseu, *D. pseudoobscura*; Danan, *D. ananassae*; Derec, *D. erecta*; Dyaku, *D. yakuba*, Dmela, *D. melanogaster*; Dsech, *D. sechellia*; Dsimu, *D. simulans*.

Comparative Genomics

Orthology

OrthoDB orthology delineation (*66, 88, 89*) was employed to define orthologous groups of genes descended from each last common ancestor of the species phylogeny across 43 insects including 21 mosquitoes - *Hemipterodea*: *Pediculus humanus* & *Rhodnius prolixus*; *Hymenoptera*: *Apis mellifera*, & *Linepithema humile*; *Coleoptera*: *Tribolium castaneum*, *Lepidoptera*: *Bombyx mori* & *Danaus plexippus*; *Diptera*: *Lutzomyia longipalpis*, *Phlebotomus papatasi*, *Glossina morsitans*, 12 *Drosophila* - *D. grimshawi*, *D. mojavensis*, *D. virilis*, *D. willistoni*, *D. persimilis*, *D. pseudoobscura*, *D. ananassae*, *D. erecta*, *D. yakuba*, *D. melanogaster*, *D. sechellia*, and *D. simulans*; 2 culicine mosquitoes - *Aedes aegypti* & *Culex quinquefasciatus*; and 19 anophelines, *An. darlingi*, *An. albimanus*, *An. sinensis*, *An. atroparvus*, *An. farauti*, *An. dirus*, *An. funestus*, *An. minimus*, *An. culicifacies*, *An. maculatus*, *An. stephensi* (SDA-500), *An. stephensi* (INDIAN), *An. epiroticus*, *An. christyi*, *An. melas*, *An. quadriannulatus*, *An. arabiensis*, *An. merus*, and *An. gambiae* (PEST). Orthology refers to the last common ancestor of the species under consideration, and thus OrthoDB explicitly delineates orthologs at each radiation along the species phylogeny. The database of orthologs presents available protein descriptors, together with Gene Ontology and InterPro attributes, which serve to provide general descriptive annotations of the orthologous groups, and facilitate comprehensive orthology database querying available at: http://cegg.unige.ch/orthodbmoz2.

OrthoDBmoz2 enables searches with relevant identifiers of proteins, genes, orthologous groups, InterPro domains, or Gene Ontology terms, as well as with keywords associated with protein annotations (text search with logical operators). Gene synonyms and recorded phenotypes are integrated in OrthoDBmoz2 for *D. melanogaster* from FlyBase and for *An. gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus* from VectorBase. Queries may also be built to retrieve orthologous groups matching specific copy-number profiles e.g. all single-copy or all multiple-copy (copy-number search). One can use the species copy-number selectors to create a required copy-number profile or one can choose common but more complex profiles from the dropdown profile selector. Orthologous groups may also be retrieved by homology to your query sequence using BLAST (sequence search).

The phylogeny-defined hierarchy of orthologous groups allows one to select a required radiation point by clicking on the nodes of the species tree. Considering many distantly related species delineates fewer, more general (inclusive) orthologous groups containing all the descendants of the ancestral gene, while examining more closely related species produces many fine-grained orthologous groups of mostly one-to-one relations. As well as the many functional annotations, orthologous groups are also annotated with several evolutionary annotations. (1) *Evolutionary Rates*: orthologous

groups that exhibit appreciably higher or lower levels of sequence divergence are highlighted through quantification of the relative divergence among their member genes. These are computed for each orthologous group as the average of inter-species identities normalized to the average identity of all inter-species best reciprocal hits, computed from pairwise Smith-Waterman alignments of protein sequences. (2) *Sibling Groups*: relations among orthologous groups (each level of the phylogeny-defined hierarchy) are defined according to the average and minimum distances from all-against-all Smith-Waterman e-values between all members of an orthologous group to all members of any related groups with a minimum e-value cut-off of 1e-3. (3) *Canonical Gene Architecture*: The table shows mean, median, and standard deviation values of protein lengths and exon counts for each orthologous group, effectively describing a 'consensus' gene architecture. In addition, length (amino acids) and exon counts are listed for each gene, and significant deviation from the 'consensus' is highlighted.

Selection of 1,085 relaxed single-copy orthologs (a maximum of 3 paralogs allowed in no more than 5 species, longest protein selected) across all 43 species included in OrthoDBmoz2 provided a conserved core of orthologs from which to estimate the species phylogeny. The concatenated protein sequence alignments (using MUSCLE (*76*)) followed by alignment trimming with trimAl (*90*) resulted in 720,022 amino acid columns (with 526,151 distinct alignment patterns) from which to estimate the maximum likelihood species phylogeny using RAxML (*91*) with the PROTGAMMAJTT model, rooted with the *Hemipterodea*. The genome-scale quantitative maximum likelihood species phylogeny clearly delineates the evolutionary relationships amongst the anopheline mosquitoes (Figure 1, main text, and Figure S4). Phyletic analyses of orthologous groups identifies about 4,000 orthologs conserved across insects and a further 4,000 to 6,000 orthologs that, although found in mosquitoes, flies, and outgroup species, are not as well-maintained. The mosquitoes appear to share more orthologs with the outgroup species than the flies do, indicating the retention of more ancient orthologs in mosquitoes and/or the faster molecular evolution in flies which obscures distant orthologous relationships (Figure S4). Each of the anopheline species has more than 4,000 genes with orthologs in all of the analyzed mosquitoes, and a further ~2,000 in almost all species. About 2,000 universal single-copy orthologs are identifiable across all the 21 mosquitoes, with a further 2,000 to 3,000 universal multi-copy orthologs.

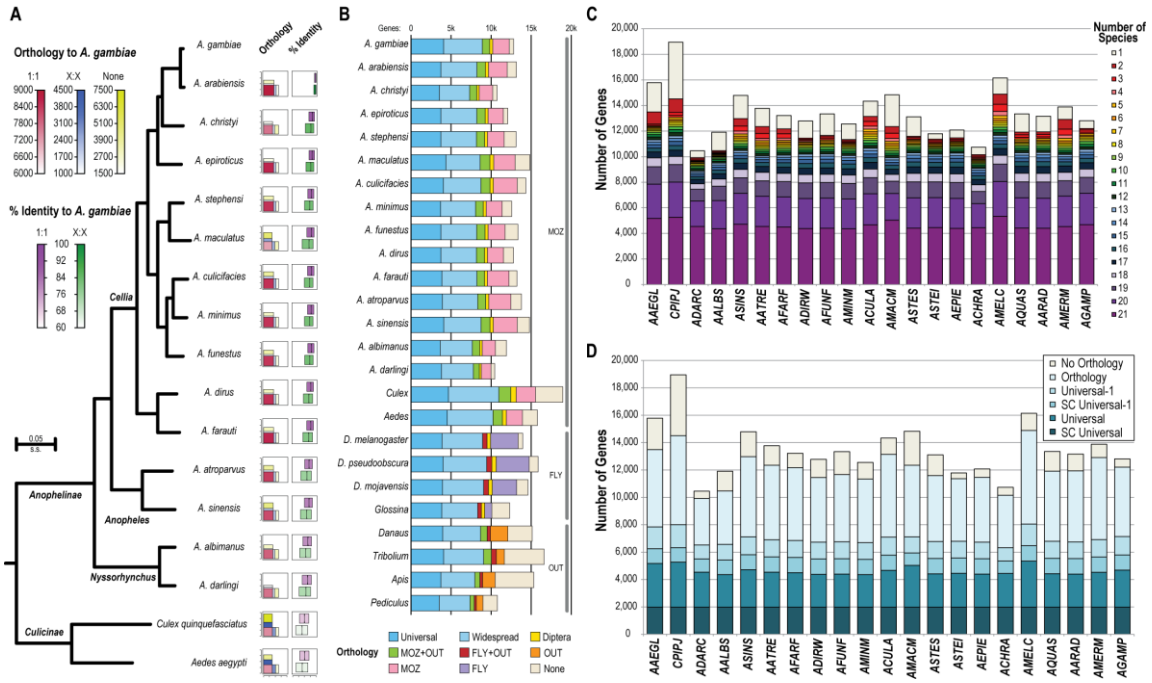


Figure S4. Molecular phylogeny and orthology.

A. The maximum likelihood species phylogeny computed from single-copy orthologs delineates the evolutionary relationships amongst the anopheline mosquitoes rooted with two culicines. Counts of single-copy (1:1, red) and multi-copy (X:X, blue) orthologs with *An. gambiae*, and genes without clear orthologs (None, yellow) are shown (scale 0 to 20,000 genes), along with the percent amino acid identity of 1:1 (pink) and X:X (green) orthologs (scale 40% to 100%). The orthology boxes and % identity boxes show how they decrease with increasing evolutionary distance from *An. gambiae*. **B**. Phyletic analyses of orthologous groups distinguishes among Universal -present in all or all-but-one species, Widespread - found in mosquitoes (MOZ), flies (FLY) and outgroup species (OUT) but not as well-maintained, and various lineage-restricted orthologs, and species-specific genes. **C**. Ortholog sharing details phyletic distributions from species-specific genes (no detectable orthology) to those with orthologs in all other mosquito species. AAEGL, *Aedes aegypti*; CPIPJ, *Culex quinquefasciatus*; ADARC, *An. darlingi*; AALBS, *An. albimanus*; ASINS, *An. sinensis*; AATRE, *An. atroparvus*; AFARF, *An. farauti*; ADIRW, *An. dirus*; AFUNF, *An. funestus*; AMINM, *An. minimus*; ACULA, *An. culicifacies*; AMACM, *An. maculatus*; ASTES, *An. stephensi* (SDA-500); ASTEI, *An. stephensi* (INDIAN); AEPIE, *An. epiroticus*; ACHRA, *An. christyi*, AMELC, *An. melas*; AQUAS, *An. quadriannulatus*; AARAD, *An. arabiensis*; AMERM, *An. merus*; AGAMP, *An. gambiae* (PEST). **D**. Distinguishing between single-copy (SC) and other universal (all species) or near-universal (Universal-1, all-but-one) mosquito orthologs. Species codes same as for panel C.

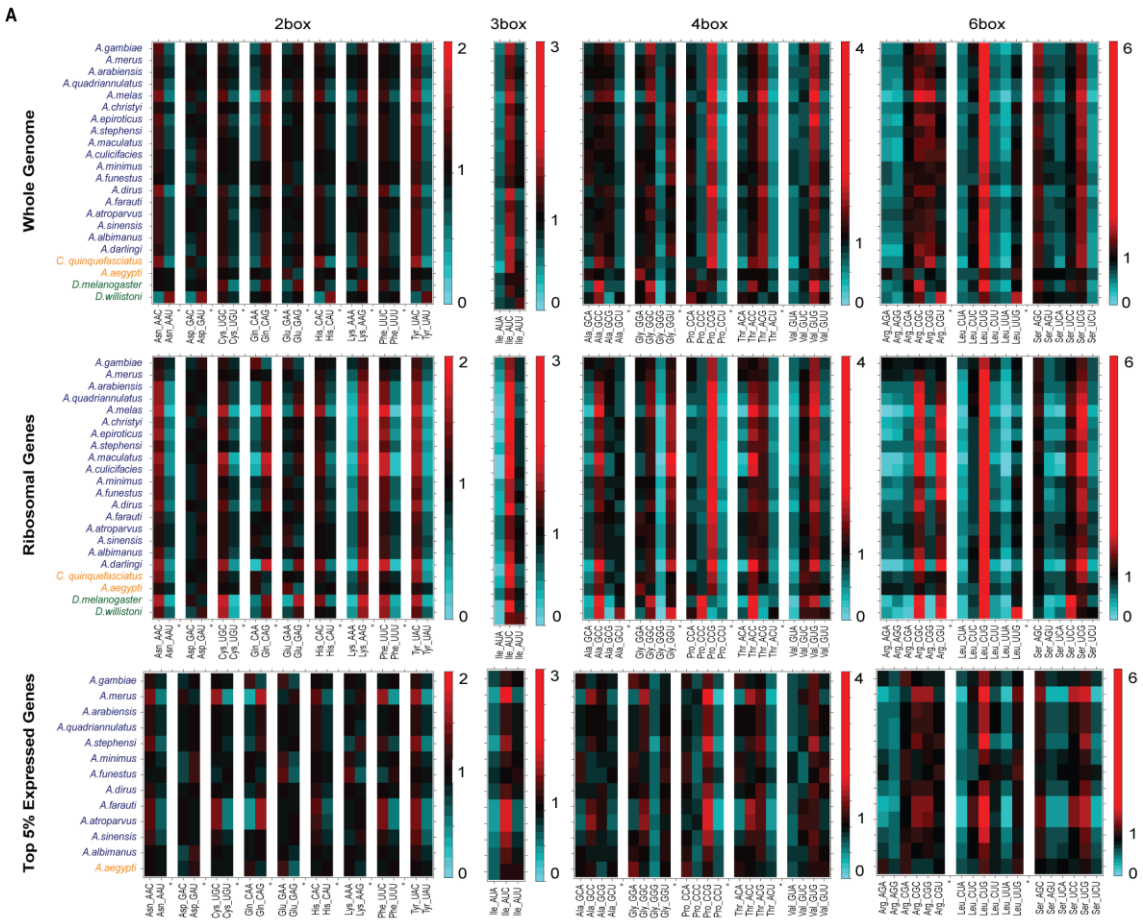
Codon bias

To determine the evolutionary forces affecting codon usage bias across the anophelines and to compare them to other insects, we used a subset of orthologous protein-coding genes, as delineated by the OrthoDB methodology (*66*). Average whole genome relative synonymous codon usage (RSCU) values were computed upon these subsets using the General Codon Usage Analysis (GCUA) software (*92*). Given that highly expressed genes are generally acknowledged to be enriched in "optimal" or "preferred" codons (*93, 94*), we additionally computed RSCU values upon two distinct

subsets of proteins expected to be enriched in "preferred" codons. The first subset consisted exclusively on ribosomal proteins -commonly used as a proxy of highly expressed proteins-, which were retrieved from (http://cegg.unige.ch/orthodbmoz2) OrthoDBmoz2 using the set of manually curated annotations of *Anopheles gambiae* ribosomal proteins as query (*95*). The second subset consisted of highly expressed proteins predicted from RNASeq datasets. BAM alignments were processed using HTSeq (*96*) to obtain read counts per gene, which were then normalized by its corresponding gene length. The top 5% ranked genes (counts/bp) were finally selected as a second subset of highly expressed genes. tRNA gene predictions were annotated using tRNA-scanSE using default eukaryote-specific parameters, and converted into relative gene frequencies (RGF) for correlation analysis purposes.

The degeneracy of the genetic code arises from the fact that there are 64 different codons which encode only for 20 amino acids. As a result of both mutational biases and natural selection for translational optimization, synonymous codons are not equally used within species and across species. Fast-growing organisms such as *Escherichia coli* or *Saccharomyces cerevisiae* are thought to have optimal codons that reflect the composition of their respective genomic tRNA pool (*97*), whereas other organisms such as *Homo sapiens* and *Helicobacter pylori* show low codon usage optimization (*98*), and their codon usage biases are thought to be mainly determined by mutational biases. Between these cases, we find species such as *Drosophila melanogaster* or *Caenorhabditis elegans* which show intermediate levels of codon usage optimization.

The analysis of codon usage bias on the full set of orthologous protein-coding genes of anophelines, including *Culicidae* and *Drosophilidae* species as outgroups shows that all anophelines have consistently the subsets of enriched codons across all amino acids (Figure S5). In contrast, this is not true for *Drosophilidae*, as previously reported (*99*). Moreover, the frequency of favored codons (i.e. codons that are used preferentially in genes that have a strong codon usage bias) increases with increasing expression level. Examining whether these subsets of favored codons could be explained due to their genomic tRNA pools, we find that there is a high correlation between the most abundant tRNA isoacceptor (in terms of tRNA gene copy number) and the subset of enriched codons, suggesting co-evolvability of the translational machinery, rather than just mutational evolutionary forces (Figure S5).



B

AA	Gene number[a]	RGF[a]	tRNA anticodon	Codon	RSCU[1,a]	RSCU[2,a]
Ala	15	**2.62**	**AGC**	GCT	0.94	0.88
	5	0.79	CGC	GCG	0.88	1.04
	0	0.03	GGC	**GCC**	**1.52**	**1.14**
	3	0.28	TGC	GCA	0.66	0.95
Arg	9	**2.70**	**ACG**	CGT	1.70	1.04
	0	0.00	CCG	CGG	0.68	1.23
	2	0.50	CCT	AGG	0.55	0.63
	0	0.02	GCG	**CGC**	**1.83**	**1.25**
	7	2.06	TCG	CGA	0.68	1.13
	2	0.71	TCT	AGA	0.56	0.73
Asn	0	0.00	ATT	AAT	0.60	0.81
	10	**2.00**	**GTT**	**AAC**	**1.40**	**1.19**
Asp	0	0.04	ATC	**GAT**	**1.06**	**1.07**
	17	**1.96**	**GTC**	GAC	0.94	0.93
Cys	0	0.00	ACA	TGT	0.71	0.85
	5	**2.00**	**GCA**	**TGC**	**1.29**	**1.15**
Gln	9	**1.33**	**CTG**	**CAG**	**1.24**	**1.15**
	5	0.67	TTG	CAA	0.76	0.85
Glu	14	**1.13**	**CTC**	GAG	**1.08**	0.92
	13	0.87	TTC	**GAA**	0.92	**1.08**
Gly	0	0.00	ACC	GGT	**1.46**	1.09
	1	0.14	CCC	GGG	0.31	0.63
	12	**2.40**	**GCC**	**GGC**	1.27	**1.23**
	7	1.46	TCC	GGA	0.96	1.05
His	0	0.00	ATG	CAT	0.78	0.85
	16	**2.00**	**GTG**	**CAC**	**1.22**	**1.15**
Ile	11	**2.50**	**AAT**	ATT	0.88	0.92
	0	0.08	GAT	**ATC**	**1.81**	**1.45**
	2	0.43	TAT	ATA	0.32	0.63
Leu	4	1.18	AAG	CTT	0.64	0.93
	4	0.92	CAA	TTG	0.95	1.20
	8	**2.01**	**CAG**	**CTG**	**2.77**	**1.73**
	0	0.00	GAG	CTC	0.91	1.01
	4	1.03	TAA	TTA	0.33	0.60
	3	0.86	TAG	CTA	0.40	0.53
Lys	16	**1.42**	**CTT**	AAG	**1.52**	0.98
	7	0.58	TTT	**AAA**	0.48	**1.02**
Phe	0	0.02	AAA	TTT	0.61	0.89
	10	**1.98**	**GAA**	**TTC**	**1.39**	**1.11**
Pro	5	1.04	AGG	CCT	0.47	0.60
	8	**1.67**	**CGG**	**CCG**	**1.82**	**1.56**
	0	0.00	GGG	CCC	0.86	0.78
	6	1.28	TGG	CCA	0.85	1.07
Ser	0	0.00	ACT	AGT	0.55	0.70
	5	1.64	AGA	TCT	0.66	0.66
	6	**1.86**	**CGA**	**TCG**	**1.74**	**1.45**
	5	1.72	GCT	AGC	1.13	1.25
	0	0.05	GGA	TCC	1.30	1.19
	2	0.73	TGA	TCA	0.62	0.76
Thr	8	**1.91**	**AGT**	ACT	0.58	0.63
	7	1.49	CGT	**ACG**	1.18	**1.24**
	0	0.02	GGT	ACC	**1.53**	1.19
	2	0.58	TGT	ACA	0.71	0.94
Tyr	0	0.00	ATA	TAT	0.52	0.74
	18	**2.00**	**GTA**	**TAC**	**1.48**	**1.26**
Val	14	**1.92**	**AAC**	GTT	0.79	1.01
	19	1.56	CAC	**GTG**	**1.61**	**1.30**
	0	0.01	GAC	GTC	1.12	0.92
	4	0.50	TAC	GTA	0.48	0.76

[1] RSCU values from ribosomal proteins
[2] RSCU values from top 5% ranked proteins using normalized RNASeq counts
[a] Average tRNA gene number, RGF and RSCU values of *Anophelines* are shown

Figure S5. Codon usage.

A. Relative synonymous codon usage bias (RSCU) of *Anophelinae* (blue), *Culicinae* (orange) and *Drosophilidae* (green) species. Species have not been hierarchically clustered, but instead the phylogenetic order has been maintained to facilitate comparison across subsets. Amino acids have been subdivided according to its number of

isodecoding codons into 2-, 3-, 4- and 6-box, respectively. RSCU values below 1 (cyan) indicate under-representation of a given codon, whereas values over 1 (red) indicate over-representation. **B**. Correlation between tRNA gene content and codon usage bias in anophelines. For each amino acid, the "optimal" tRNA and codon are highlighted in bold. The matching between preferred codon and optimal tRNA have been colored differently depending on its pairing abilities: i) green, if the preferred codon can be decoded by the optimal tRNA according to Watson-Crick pairing (including I:C pair), ii) yellow, if it can be decoded using G:U wobble, and iii) orange, if classical rules cannot explain the pairing.

Repetitive elements

Program-assisted manual annotation: Transposable element discovery and classification were performed on the scaffold sequences of different species using analysis pipelines for long terminal repeat (LTR)-retrotransposons, non-LTR-retrotransposons, short interspersed repetitive elements (SINEs), DNA transposons, and miniature inverted-repeat transposable elements (MITEs), followed by manual inspection. All manually-annotated TE sequences are available at TEfam (tefam.biochem.vt.edu). In addition to 10 of the 16 new assemblies (see Table S7 for the list of species annotated) reported in this study, we also re-annotated TEs in the *Anopheles gambiae* PEST genome (*4*) to provide a comparison and reference. Our annotation increased the manually annotated TEs from 158 to 392 in the PEST assembly.

DNA transposon sequences were identified in anopheline genomes by a combination of *de novo*, structural and similarity based searches. Each genome was examined using 1) the program RepeatModeler (*100*) using RepeatMasker libraries (*69*), 2) genome BLAST searches using mosquito transposable element sequences deposited in TEfam and RepBase, 3) a homology-based transposable element discovery pipeline developed by José Riberio (unpublished). The results of these searches were combined and used to map potential transposable element sequences in each genome using the program RepeatMasker. Every mapped transposable element sequence was extended up to 3000 bps on either end and the extended sequences were examined for the presence of terminal inverted repeats (TIR) and target site duplications (TSD) that are characteristic of nearly all full length DNA elements. Data from all searches were summarized and manually curated, with particular attention to identifying complete element sequences (i.e. containing expected TSD, TIR, and complete transposase open reading frame).

RepeatModeler output that are less than 500 bp were used as query to run BLAST against the source genome assembly. Top 20 matches were retrieved together with their flanking sequences (500 bp on either side). These sequences were aligned using Clustal to allow inspection of the TE boundaries. These steps were automated using a program called TEalign. These alignments were used to assess whether a candidate MITE has the requisite terminal inverted repeats (TIRs) and to classify them according to their target

site duplications. After obtaining the initial list of MITEs using methods described above, multiple rounds of self-blast were performed to remove redundancy using a cut-off of overall 80% identity. In other words, two MITEs are considered different if they are at least 20% divergent. SINEs were discovered in a similar manner. Instead of terminal inverted repeats, sequences that show similarity to tRNA or other polymerase III promoters were targeted. SINEs discovered in this study all had target site duplications and tandem repeats at their 3' ends.

LTR retrotransposons were identified in each one the genome assemblies using a structure and homology-based approach. The canonical sequences of LTR retrotransposons from several insect genomes (mainly *An. gambiae* PEST) were retrieved from Repbase (*101*) and TEfam (tefam.biochem.vt.edu). TBLASTN was used to search for sequence homologous to the pol region of representative LTR retrotransposons in each of the anopheline genomes. Those hits showing at least 30% amino acid identity over at least 80% of the length of the query sequence were subjected to further analyses to identify both LTRs of each element by means of BLAST2 sequences. This strategy allowed the identification of canonical sequences representing complete copies that are putatively active and/or consensus sequences of each LTR retrotransposon element in the corresponding anopheline genome.

To discover non-LTR retrotransposons, we utilized an abridged version of TESEEKER, described in detail in (*102*). TESEEKER is an automated homology-based approach for the identification of transposable elements (TEs). In addition to the library of published, representative TEs included with TESEEKER (organized by clade), we performed searches using putative TEs generated by REPEATMODELER. For each genome, TESEEKER was run using both the built-in TESEEKER library as well as the elements identified and classified by RepeatModeler. Simply, TESEEKER performs TBLASTN searches using the representative libraries, and the TBLASTN results are then combined if they are within 50 bp in the genome. Next, elements are extracted from the genome and assembled with the CAP3 assembler (*103*). Finally, we process the CAP3 results, using the highest quality contigs and singlets if there are no contigs. These results are then classified using an in-house classifier based on alignment to known non-LTRs and a maximum likelihood tree. Additionally, all representative ORFs were aligned using CLUSTALX and manually examined. We removed the redundancy at 80% identity cut-off, leaving only a single longest representative of each identified family in a clade.

Integration with RepeatModeler output and analysis of copy number and genome occupancy: Manually annotated TE libraries were also used to compare with the RepeatModeler output to remove redundancy and to correct mis-classification by RepeatModeler (BLASTN, e-30). The remaining RepeatModeler output sequences were compared to the non-redundant protein database (BLASTX, e-5) to remove repetitive protein families and to verify RepeatModeler classifications. The remaining verified RepeatModeler output was combined with the above-mentioned manually annotated TEs

of the species to make the TE library for the species. This combined TE library was used to run RepeatMasker (default setting) to estimate TE copy number and genome occupancy. For RepeatMasker runs of species in the *An. gambiae* complex, the 392 manually annotated TEs from the PEST strain were added in addition to TEs from each species. Copy number and genome occupancy of TEs were calculated according to RepeatMasker output.

Multiple families of LTR and non-LTR retroelements and DNA-mediated transposable elements (TEs) are all found in all genomes (Table S7; TEfam.biochem.vt.edu). These transposable elements and other interspersed repeats represent varying fractions of the *de novo* assemblies, ranging from 1.98% in *An. albimanus* to 11.43% in *An. merus*. These numbers are likely underestimates because short repetitive sequences may be excluded from the assemblies and the amounts of N/X runs, which could consist of TEs and other interspersed repeats, range from a few to 75 Mbp in these assemblies (Table S7). Such variations necessitate caution when performing cross-species comparisons. Here we highlight a few evolutionary insights gleaned from our comparative analysis. For comparative analysis of TEs in the *An. gambiae* species complex, we re-annotated TEs in the *An. gambiae* PEST genome to provide a reference. Our analysis increased the manually annotated TEs from 158 to 392 in the *An. gambiae* PEST assembly. There are no obvious examples where a TE is unique to one or a subset of species within the *An. gambiae* complex, suggesting that all currently annotated TEs existed in the common ancestor of the *An. gambiae* complex. On the other hand, 136 of the 392 *An. gambiae* TEs showed no similarities to any sequence in the *An. christyi* genome at an e-value cutoff of 1e-5. The tRNA-related SINEs in the *An. gambiae* complex showed clear homology with the SINEs in *An. christyi*, with the best match showing 82% identity over the entire 200 bp. These particular families of SINEs are only found in the *An. gambiae* species complex and *An. christyi*, supporting the phylogeny shown in Figure 1, main text. Despite variations in assembly contiguity, our analyses show that there is a great level of plasticity in TE content among the 16 species. *An. albimanus* and *An. christyi*, which belong to two divergent subgenera, have the two smallest genomes among the 16 species. They also showed the lowest TE contents, 1.98% and 2.81%, respectively.

Table S7. Repeat content of anopheline genomes.

Detailed report of repeat content of 11 anophelines. TE: transposable element; LTR: long terminal repeat; LINE: long interspersed nuclear element; SINE: short interspersed nuclear element; MITE: miniature inverted-repeat transposable element.

	***A. gambiae* PEST**			***A. arabiensis***			***A. quadriannulatus***		
TE	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**
LTR	*18901*	*10965554*	*3.94*	*10836*	*3177001*	*1.29*	*8477*	*2176328*	*0.77*
LTR/Pao_Bel	7848	4442083	1.60	5364	1400822	0.57	3073	827628	0.29
LTR/Ty1_copia	2881	1117498	0.40	1881	432256	0.18	2182	433572	0.15
LTR/Ty3_gypsy	6227	3915620	1.41	3573	1324794	0.54	3167	876721	0.31
LTR/RM [1]	1945	1490353	0.54	18	19129	0.01	55	38407	0.01
Non-LTR retrotransposons	*23867*	*9694642*	*3.48*	*16750*	*4230681*	*1.72*	*16237*	*3829302*	*1.35*
LINE/CR1	6883	3946450	1.42	4175	1251953	0.51	4083	1077645	0.38
LINE/I	640	282451	0.10	369	121250	0.05	482	145868	0.05
LINE/Jockey	5732	1324921	0.48	4255	807885	0.33	4230	730967	0.26
LINE/L1	408	230079	0.08	471	104460	0.04	219	80613	0.03
LINE/L2	630	188908	0.07	476	80198	0.03	448	71752	0.03
LINE/Loner	531	208715	0.08	412	69122	0.03	319	65233	0.02
LINE/Outcast	414	267346	0.10	254	136232	0.06	196	101803	0.04
LINE/R1	1806	663174	0.24	1166	248824	0.10	1146	227195	0.08
LINE/R4	60	21692	0.01	55	13463	0.01	53	13419	0.00
LINE/RTE	5083	2165866	0.78	3569	1015697	0.41	4082	1085750	0.38
LINE/Undetermined1	146	8078	0.00	885	209945	0.09	85	5823	0.00
LINE/RM [1]	1534	386962	0.14	663	171652	0.07	894	223234	0.08
SINE*/tSINE*	*7789*	*1196281*	*0.43*	*8426*	*1248921*	*0.51*	*8673*	*1287118*	*0.45*
DNA transposon	*15442*	*3361583*	*1.21*	*10629*	*1809877*	*0.73*	*7695*	*1107665*	*0.39*
DNA/ITmD37E	65	31121	0.01	30	13604	0.01	45	15308	0.01
DNA/P	91	47759	0.02	65	34237	0.01	29	13392	0.00
DNA/PIF	278	101567	0.04	116	27389	0.01	98	16332	0.01
DNA/Tc1	5786	1364649	0.49	5161	1024133	0.42	2893	436996	0.15
DNA/Transib	22	15252	0.01	14	7879	0.00	9	6899	0.00
DNA/gambol	5172	835343	0.30	3857	499112	0.20	2668	349862	0.12
DNA/hAT	150	65062	0.02	102	36485	0.01	151	31511	0.01
DNA/mariner	2193	272776	0.10	1202	139041	0.06	1569	163452	0.06
DNA/piggyBac	148	53704	0.02	82	27997	0.01	117	37131	0.01
DNA/RM [1]	1537	574350	0.21	0	0	0.00	116	36782	0.01
MITEs	*36858*	*8790090*	*3.16*	*24587*	*5379520*	*2.18*	*25367*	*5276102*	*1.86*
MITEs/m3bp	9911	2282290	0.82	6149	1344216	0.55	6781	1387346	0.49
MITEs/m4bp	1312	304724	0.11	704	141046	0.06	857	144085	0.05
MITEs/m8bp	4848	1293871	0.47	3388	893056	0.36	3287	854379	0.30
MITEs/mTA	20419	4839042	1.74	14087	2962487	1.20	14248	2861199	1.01
MITEs/otherMITEs	368	70163	0.03	259	38715	0.02	194	29093	0.01
RC*/Helitron_RM* [1, 2]	*287*	*135675*	*0.05*	*0*	*0*	*0.00*	*128*	*36602*	*0.01*
Unclassified interspersed repeats	85649	15317962	5.51	50861	7280888	2.95	50740	8115240	2.86
Total interspersed repeats			**17.78**			**9.38**			**7.69**
Satellite	0	0	0.00	0	0	0.00	0	0	0.00
Simple_repeat	145023	7232917	2.60	127152	5249410	2.13	127689	5080257	1.79
Low_complexity	13345	640906	0.23	11963	572415	0.23	11792	565130	0.20
Grand Total			**20.61**			**11.74**			**9.68**
Genome size		278237304			246567867			283828998	
Genome size excluding N/X runs [3]		263133019			211454830			208982808	

	A. melas			*A. merus*			*A. christyi*		
TE	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**
LTR	*6941*	*1496802*	*0.66*	*23654*	*6435680*	*2.56*	*375*	*144982*	*0.08*
LTR/Pao_Bel	2421	547744	0.24	9461	2434411	0.97	103	20313	0.01
LTR/Ty1_copia	1507	276306	0.12	2837	708477	0.28	0	0	0.00
LTR/Ty3_gypsy	3013	672752	0.30	11137	3176730	1.26	22	11341	0.01
LTR/RM [1]	0	0	0.00	219	116062	0.05	250	113328	0.07
Non-LTR retrotransposons	*14554*	*2975543*	*1.31*	*18964*	*5450633*	*2.16*	*3544*	*657713*	*0.38*
LINE/CR1	3395	816646	0.36	4925	1692193	0.67	65	16683	0.01
LINE/I	396	100670	0.04	407	165134	0.07	0	0	0.00
LINE/Jockey	4404	617817	0.27	4335	842856	0.33	0	0	0.00
LINE/L1	163	47368	0.02	273	131668	0.05	0	0	0.00
LINE/L2	392	47454	0.02	582	140054	0.06	0	0	0.00
LINE/Loner	413	56988	0.03	464	122269	0.05	37	18884	0.01
LINE/Outcast	181	102729	0.05	306	194111	0.08	0	0	0.00
LINE/R1	1121	184021	0.08	1367	380037	0.15	22	3905	0.00
LINE/R4	29	10056	0.00	62	20539	0.01	0	0	0.00
LINE/RTE	3137	776202	0.34	4591	1361282	0.54	156	24333	0.01
LINE/Undetermined1	121	6787	0.00	189	26623	0.01	0	0	0.00
LINE/RM [1]	802	208805	0.09	1463	373867	0.15	3264	593908	0.34
SINE*/tSINE*	*7491*	*1064732*	*0.47*	*3699*	*530934*	*0.21*	*5458*	*670088*	*0.39*
DNA transposon	*7270*	*953780*	*0.42*	*10832*	*2267035*	*0.90*	*779*	*181193*	*0.10*
DNA/ITmD37E	23	5775	0.00	28	11469	0.00	0	0	0.00
DNA/P	34	16676	0.01	69	40529	0.02	0	0	0.00
DNA/PIF	87	18929	0.01	159	43186	0.02	0	0	0.00
DNA/Tc1	3022	395417	0.17	2775	597393	0.24	199	48309	0.03
DNA/Transib	8	5270	0.00	19	8706	0.00	120	32068	0.02
DNA/gambol	2655	319441	0.14	4043	610951	0.24	11	4976	0.00
DNA/hAT	84	24661	0.01	142	50217	0.02	403	85382	0.05
DNA/mariner	1169	117413	0.05	1307	121746	0.05	0	0	0.00
DNA/piggyBac	34	12374	0.01	111	47530	0.02	4	1431	0.00
DNA/RM [1]	154	37824	0.02	2179	735308	0.29	42	9027	0.01
MITEs	*22704*	*4313227*	*1.90*	*26256*	*5692927*	*2.26*	*321*	*51074*	*0.03*
MITEs/m3bp	5549	1056394	0.46	7536	1579887	0.63	81	11973	0.01
MITEs/m4bp	738	124867	0.05	791	172466	0.07	0	0	0.00
MITEs/m8bp	3199	723782	0.32	3376	881646	0.35	50	5757	0.00
MITEs/mTA	13057	2382457	1.05	14366	3027116	1.20	190	33344	0.02
MITEs/otherMITEs	161	25727	0.01	187	31812	0.01	0	0	0.00
RC*/Helitron_RM* [1, 2]	*103*	*19851*	*0.01*	*270*	*85860*	*0.03*	*0*	*0*	*0.00*
Unclassified interspersed repeats	34601	5752647	2.53	40651	8328080	3.31	22624	3151753	1.83
Total interspersed repeats			**7.29**			**11.43**			**2.81**
Satellite	2849	797370	0.35	0	0	0.00	0	0	0.00
Simple_repeat	135709	4849322	2.13	130215	4923418	1.96	81444	2818168	1.63
Low_complexity	11744	561232	0.25	11660	548321	0.22	9469	426451	0.25
Grand Total			**10.02**			**13.61**			**4.69**
Genome size		227407517			251805912			172658580	
Genome size excluding N/X runs [3]		206735484			218202275			169987814	

	A. epiroticus			*A. dirus*			*A. stephensi SDA*		
TE	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**
LTR	*11283*	*2721923*	*1.22*	*5160*	*1413341*	*0.65*	*1306*	*469479*	*0.21*
LTR/Pao_Bel	6647	1459734	0.65	1377	497365	0.23	201	104602	0.05
LTR/Ty1_copia	506	161876	0.07	492	126921	0.06	634	87300	0.04
LTR/Ty3_gypsy	3751	867315	0.39	3019	684050	0.32	412	244927	0.11
LTR/RM [1]	379	232998	0.10	272	105005	0.05	59	32650	0.01
Non-LTR retrotransposons	*11168*	*2994025*	*1.34*	*9577*	*2313613*	*1.07*	*16251*	*3433844*	*1.52*

LINE/CR1	866	262998	0.12	4728	1119028	0.52	728	241112	0.11
LINE/I	77	44726	0.02	64	40354	0.02	25	11289	0.01
LINE/Jockey	722	97138	0.04	110	52409	0.02	127	57782	0.03
LINE/L1	116	41574	0.02	120	44613	0.02	0	0	0.00
LINE/L2	0	0	0.00	61	15267	0.01	0	0	0.00
LINE/Loner	40	14384	0.01	0	0	0.00	564	135330	0.06
LINE/Outcast	203	55917	0.03	296	47383	0.02	25	17234	0.01
LINE/R1	179	69819	0.03	277	129683	0.06	20	5493	0.00
LINE/R4	0	0	0.00	0	0	0.00	0	0	0.00
LINE/RTE	4687	1361336	0.61	652	139221	0.06	0	0	0.00
LINE/Undetermined1	361	108328	0.05	395	109377	0.05	8052	1725369	0.77
LINE/RM [1]	3917	937805	0.42	2874	616278	0.28	6710	1240235	0.55
***SINE**/tSINE*	*0*	*0*	*0.00*	*1624*	*178310*	*0.08*	*26091*	*3014633*	*1.34*
DNA transposon	*2377*	*719337*	*0.32*	*2096*	*562196*	*0.26*	*1512*	*297989*	*0.13*
DNA/ITmD37E	0	0	0.00	0	0	0.00	0	0	0.00
DNA/P	41	31756	0.01	0	0	0.00	0	0	0.00
DNA/PIF	0	0	0.00	0	0	0.00	0	0	0.00
DNA/Tc1	1696	485424	0.22	891	256746	0.12	1018	218493	0.10
DNA/Transib	157	63819	0.03	100	27129	0.01	0	0	0.00
DNA/gambol	0	0	0.00	0	0	0.00	0	0	0.00
DNA/hAT	12	8333	0.00	110	25857	0.01	283	35288	0.02
DNA/mariner	0	0	0.00	0	0	0.00	0	0	0.00
DNA/piggyBac	103	26759	0.01	269	58498	0.03	29	11291	0.01
DNA/RM [1]	368	103246	0.05	726	193966	0.09	182	32917	0.01
MITEs	*2529*	*512169*	*0.23*	*4240*	*779970*	*0.36*	*2336*	*469751*	*0.21*
MITEs/m3bp	744	161953	0.07	1867	368753	0.17	0	0	0.00
MITEs/m4bp	104	24118	0.01	129	23614	0.01	0	0	0.00
MITEs/m8bp	623	116461	0.05	284	51076	0.02	1714	336043	0.15
MITEs/mTA	463	111663	0.05	1572	284890	0.13	398	83455	0.04
MITEs/otherMITEs	595	97974	0.04	388	51637	0.02	224	50253	0.02
***RC**/Helitron_RM [1, 2]*	*0*	*0*	*0.00*	*0*	*0*	*0.00*	*0*	*0*	*0.00*
Unclassified interspersed repeats	50730	7074011	3.17	40469	5760362	2.66	29603	3672345	1.63
Total interspersed repeats			**6.27**			**5.09**			**5.04**
Satellite	0	0	0.00	0	0	0.00	0	0	0.00
Simple_repeat	82639	3062822	1.37	102594	3834816	1.77	94813	3349612	1.49
Low_complexity	10678	487737	0.22	9056	420104	0.19	8844	424337	0.19
Grand Total			**7.86**			**7.06**			**6.71**
Genome size		223486714			216307690			225369006	
Genome size excluding N/X runs [3]		202637741			197750986			196199663	

	A. funestus			***A. albimanus***		
TE	**Copy**	**Base**	**%Genome**	**Copy**	**Base**	**%Genome**
LTR	*4719*	*1135298*	*0.50*	*654*	*249636*	*0.15*
LTR/Pao_Bel	1328	273717	0.12	49	25233	0.01
LTR/Ty1_copia	2129	301137	0.13	15	5250	0.00
LTR/Ty3_gypsy	786	322932	0.14	536	190147	0.11
LTR/RM [1]	476	237512	0.11	54	29006	0.02
Non-LTR retrotransposons	*6017*	*1448931*	*0.64*	*1782*	*433177*	*0.25*
LINE/CR1	344	115583	0.05	171	57458	0.03
LINE/I	35	16452	0.01	4	5190	0.00
LINE/Jockey	150	68066	0.03	29	14089	0.01
LINE/L1	92	41658	0.02	0	0	0.00
LINE/L2	77	25370	0.01	0	0	0.00
LINE/Loner	87	21900	0.01	0	0	0.00
LINE/Outcast	118	70983	0.03	0	0	0.00
LINE/R1	148	67273	0.03	404	142913	0.08
LINE/R4	0	0	0.00	0	0	0.00

LINE/RTE	0	0	0.00	4	3545	0.00
LINE/Undetermined1	367	95219	0.04	6	2866	0.00
LINE/RM [1]	4599	926427	0.41	1164	207116	0.12
***SINE**/tSINE*	*5946*	*806511*	*0.36*	*0*	*0*	*0.00*
DNA transposon	*1051*	*319500*	*0.14*	*2419*	*678974*	*0.40*
DNA/ITmD37E	0	0	0.00	0	0	0.00
DNA/P	0	0	0.00	0	0	0.00
DNA/PIF	0	0	0.00	0	0	0.00
DNA/Tc1	430	131208	0.06	1653	471323	0.28
DNA/Transib	189	56176	0.02	0	0	0.00
DNA/gambol	0	0	0.00	0	0	0.00
DNA/hAT	0	0	0.00	97	24518	0.01
DNA/mariner	0	0	0.00	0	0	0.00
DNA/piggyBac	0	0	0.00	0	0	0.00
DNA/RM [1]	432	132116	0.06	669	183133	0.11
MITEs	*3074*	*580773*	*0.26*	*667*	*175704*	*0.10*
MITEs/m3bp	358	65366	0.03	667	175704	0.10
MITEs/m4bp	780	134706	0.06	0	0	0.00
MITEs/m8bp	576	92097	0.04	0	0	0.00
MITEs/mTA	1360	288604	0.13	0	0	0.00
MITEs/otherMITEs	0	0	0.00	0	0	0.00
***RC**/Helitron_RM [1, 2]*	*0*	*0*	*0.00*	*0*	*0*	*0.00*
Unclassified interspersed repeats	35791	4790212	2.13	14726	1831115	1.07
Total interspersed repeats			**4.03**			**1.98**
Satellite	0	0	0.00	0	0	0.00
Simple_repeat	66641	2510985	1.11	131808	4504508	2.64
Low_complexity	11475	521321	0.23	9087	428474	0.25
Grand Total			**5.38**			**4.87**
Genome size		225223604			170508315	
Genome size excluding N/X runs [3]		190030892			163551887	

1. RM refers to RepeatModeler outputs that were verified to belong to a particular subclass by BLASTX against the nr database. Detailed classifications (e.g., Tc1, CR1, etc.) were not retained as it requires phylogenetic analysis for further classification.
2. RC: rolling circle/helitrons
3. Note the large variation in the amounts of N/X runs between different assemblies.

Whole genome alignments

Genome assemblies of 21 available anopheline mosquitoes (Table S8) were retrieved from VectorBase, www.vectorbase.org. In addition to the species sequenced as part of the *Anopheles* 16 genomes project (*8*), these included assemblies of *An. gambiae* PEST (*4*) and Pimperena S (*61*), *An. coluzzii* (*61*) (formally *An. gambiae* M molecular form), *An. darling* (*5*), and *An. stephensi* INDIAN (*7*). Before computing the multiple whole genome alignments, repetitive regions of the 21 input genome assemblies were first masked so as to reduce the total number of potential genomic anchors formed by the many matches that occur among regions of repetitive DNA. Assemblies were analyzed using RepeatModeler (*100*) to build libraries of repetitive elements that were then

combined and compared with known repeats from *An. gambiae* (from VectorBase). The combined library made up of repeats from all species was filtered to remove matches to known protein-coding repetitive sequences and then each genome assembly was subsequently masked with RepeatMasker (*69*).

Table S8. Whole genome alignments

Anopheline multiple whole genome alignment statistics describe proportions of gapped, masked, and aligned basepairs (bp). Species: AgamP3, *An. gambiae* (PEST); AgamS1, *An. gambiae* (S); AaraD1, *An. arabiensis*; AmerM1, *An. merus*; AmelC1, *An. melas*; AquaS1, *An. quadriannulatus*; AcolM1, *An. coluzzii*; AchrA1, *An. christyi*; AepiE1, *An. epiroticus*; AfunF1, *An. funestus*; AminM1, *An. minimus*; AculA1, *An. culicifacies*; AsteS1, *An. stephensi* (SDA-500); AstI2, *An. stephensi* (INDIAN); AdirW1, *An. dirus*; AfarF1, *An. farauti*; AmacM1, *An. maculatus*; AatrE1, *An. atroparvus*; AsinS1, *An. sinensis*; AalbS1, *An. albimanus*; AdarC2, *An. darling*.

Assembly	Assembly All (bp)	Aligned All (bp)	% All Aligned	% AgamP3 All Aligned	Gaps (bp)	Masked (bp)	Assembly Non-N‡ (bp)	Aligned Non-N‡ (bp)	% Non-N‡ Aligned	% AgamP3 Non-N‡ Aligned
AgamP3	273,093,681	195,683,074	71.65	100.00	20,654,948	55,247,274	197,191,459	195,359,863	99.07	100.00
AgamS1	236,403,076	185,965,249	78.66	66.44	8,362,861	41,205,745	186,834,470	185,664,708	99.37	91.89
AaraD1	246,567,867	179,887,665	72.96	63.96	35,124,750	30,887,257	180,555,860	179,544,097	99.44	88.49
AmerM1	251,805,912	179,702,979	71.37	63.49	33,613,976	36,451,174	181,740,762	179,389,053	98.71	87.85
AmelC1	227,407,517	178,395,667	78.45	63.24	20,677,584	26,733,654	179,996,279	177,981,897	98.88	87.47
AquaS1	283,828,998	178,008,446	62.72	63.13	74,862,329	30,062,727	178,903,942	177,638,185	99.29	87.33
AcolM1	224,455,335	175,516,911	78.20	62.82	14,926,268	33,114,661	176,414,406	175,287,690	99.36	86.93
AchrA1	172,658,580	138,674,671	80.32	46.17	2,671,395	11,802,245	158,184,940	138,478,155	87.54	63.89
AepiE1	223,486,714	142,926,676	63.95	46.09	20,854,535	21,157,887	181,474,292	142,730,753	78.65	63.79
AfunF1	225,223,604	155,914,914	69.23	39.39	35,208,164	15,767,229	174,248,211	155,738,237	89.38	54.51
AminM1	201,793,324	155,697,810	77.16	38.83	15,386,515	15,442,626	170,964,183	155,569,885	91.00	53.75
AculA1	202,998,806	153,945,405	75.84	37.11	15,840,025	15,543,095	171,615,686	153,808,375	89.62	51.37
AsteS1	225,369,006	175,811,641	78.01	36.89	29,185,349	18,986,918	177,196,739	175,505,101	99.05	51.06
AstI2	221,324,304	180,587,832	81.59	36.58	11,843,490	23,971,602	185,509,212	180,286,501	97.18	50.64
AdirW1	216,307,690	141,771,877	65.54	28.17	18,566,118	18,682,619	179,058,953	141,588,944	79.07	38.99
AfarF1	180,984,331	139,990,231	77.35	25.94	5,030,300	7,654,557	168,299,474	139,824,097	83.08	35.91
AmacM1	141,894,015	105,372,681	74.26	24.27	10,062,842	10,826,477	121,004,696	105,256,995	86.99	33.59
AatrE1	224,290,125	131,720,909	58.73	16.33	35,337,565	7,938,921	181,013,639	131,630,661	72.72	22.61
AsinS1	241,390,279	131,764,047	54.59	14.53	49,229,361	8,626,430	183,534,488	131,681,368	71.75	20.12
AalbS1	170,508,315	113,989,318	66.85	9.54	6,961,065	10,389,034	153,158,216	113,790,410	74.30	13.21
AdarC2	134,715,017	110,551,713	82.06	9.00	26,568	8,764,900	125,923,549	110,349,784	87.63	12.45

‡ Non-N: non-gap and non-masked basepairs.

A similar whole genome alignment strategy was employed to that used for other multi-species whole genome alignments such as for 12 *Drosophila* (*104*) and 29 mammal (*105*) genomes. Multiple whole genome alignments of the 21 available anopheline assemblies were built using the MULTIZ feature of the Threaded-Blockset Aligner suite of tools (*106*). The progressive alignment approach of MULTIZ requires an input dendrogram of the expected relationships between the species so that the closest pairs are aligned first followed by progressively stepping along the phylogeny to the most distant clades. The dendrogram was derived from the 21-species maximum likelihood phylogeny that was estimated using RAxML (*91*) from the concatenated protein sequences of 2,593

Genewise (*107*) gene predictions using Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs) from OrthoDB (*66*), and rooted with predictions from the genomes of *Aedes aegypti* (*62*) and *Culex quinquefasciatus* (*63*).

The first step of the MULTIZ approach consists of running all-against-all pairwise LASTZ alignments, this is followed by a projection to ensure that the reference species is "single-coverage", i.e. in any pairwise alignment, regions of the reference species may only be present once. Following the ROAST (reference dependent multiple alignment tool) alignment strategy, subsequent projection steps are then performed as guided by the species dendrogram to progressively combine the pairwise alignments, and then the multiple alignments, until they encompass the complete phylogeny of all 21 assemblies. Multiple whole genome alignment files (MAF format) were generated with each of the assemblies as the 'reference species'. Statistics of alignment coverage (Table S8) and density (Figure S6) were computed using custom Perl scripts.

Online browsing of the indexed anopheline MAF files is available through a dedicated Codon Alignment Viewer portal: www.broadinstitute.org/compbio1/cav.php. Any region from any of the 21 reference species can be viewed in the forward or reverse strand direction by adding the following minimum set of qualifiers to the base URL: alnset=<Assembly Name>, interval=<chromosome or scaffold prefixed with 'chr'>:<start position>-<end position>, and strand=<+ or ->. Additional optional qualifiers allow further visualization control e.g. to hide gaps in the reference genome, to increase or decrease the number of codons displayed per line, to compute the likely ancestral sequence, to simply output fasta-formatted alignment data, etc..

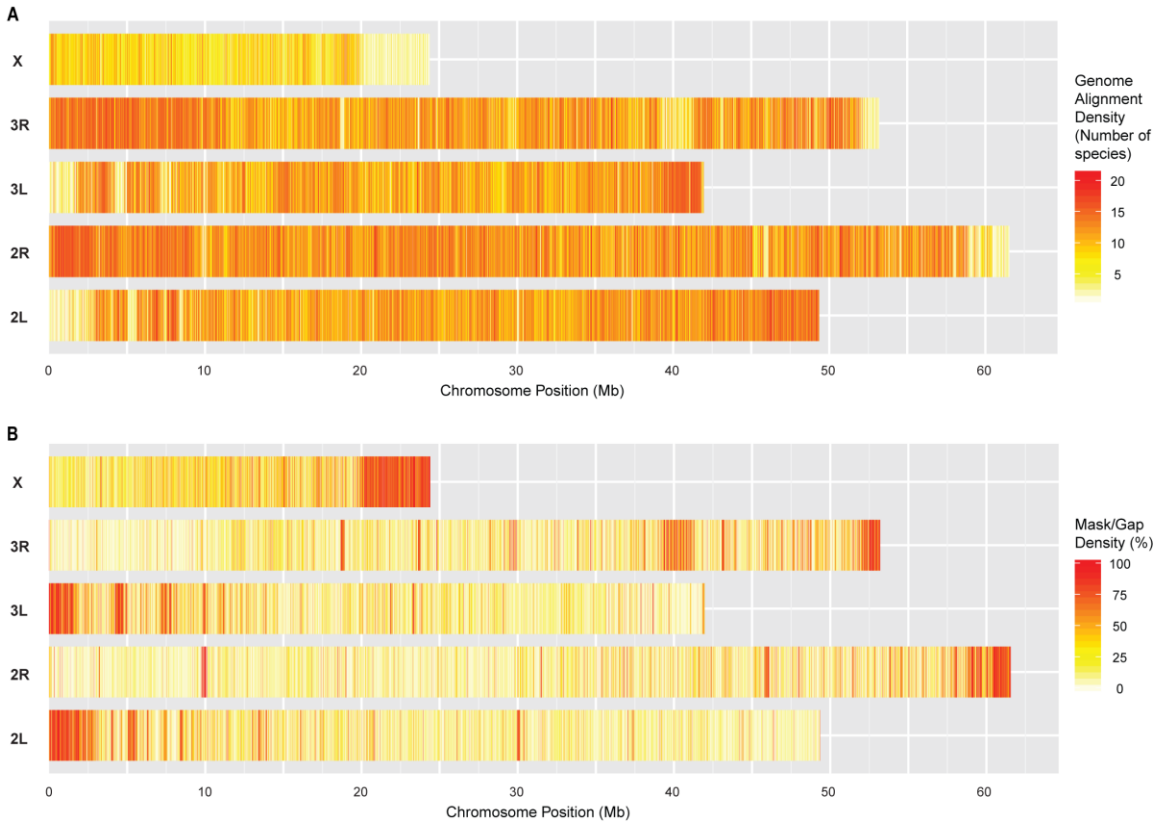


Figure S6. Whole genome alignments.

A. The density of the *An. gambiae* multiple whole genome alignment in 2Kb windows along chromosomes 2, 3, and X. Genome alignment density ranges from 1 (not aligned to any other assembly) to 21 (aligned to all other assemblies). **B**. Mask/Gap density (gaps in the assembly or masked regions, denoted by 'N') in 2Kb windows along chromosomes 2, 3, and X of *An. gambiae*. Density ranges from 0% (no Ns) to 100% (all Ns). Regions with the lowest alignment density shown in panel A correspond to regions with the highest density of Ns in panel B.

Chromosomal evolution

The chromosomal evolution analyses broadly aim to reconstruct the ancestral anopheline genome and to understand the patterns of large-scale rearrangements and gene movements in the anopheline genomes. We used different approaches to perform ancestral genome reconstruction including ANcestral GEnomeS (ANGES) and PATHGROUPS. Having alternative pipelines to address the same research problem is a main strength of our analysis of the ancestral genome architectures. For example, while the PATHGROUPS algorithm offers a rapid heuristic solution to ancestral genome reconstruction, ANGES is a local parsimony approach inspired by techniques of physical

mapping and genome assembly. The choice to use these two methods to investigate ancestral genomes was motivated by the nature of the analyzed data that are represented by highly fragmented assemblies of distant genomes. The ANGES approach offers a counterpoint to the global parsimony approach by PATHGROUPS that might fail to distinguish between true evolutionary rearrangements and artifactual rearrangements due to genome fragmentation. Because the assembly fragmentation is a common problem for many NGS genome projects, we attempted to improve genome assemblies for anopheline species using the Breakpoint Graphs algorithm. The Breakpoint Graph approach offers a hybrid point of view (between local and global parsimony approaches) as adjacencies conserved within only few assemblies can be useful to create connected components in the breakpoint graph of multiple genomes.

MOZGOB and PATHGROUPS

A mosquito synteny browser: Initial synteny within the *Anopheles gambiae* complex and the broader 16 *Anopheles* genome cluster was first viewed with MOZGOB, which is a web-based mosquito genome order browser based on the yeast YGOB platform (*108*). Input to MOZGOB was the VectorBase gene set for the *Anopheles gambiae* PEST reference, predictions for first ten genomes of the cluster (see earlier), the reference gene sets from the other mosquitoes in VectorBase (*Aedes aegypti*, *Culex quinquefasciatus*), and 11 *Drosophila* genomes from FlyBase that included *Drosophila melanogaster*. Initial one-to-one orthologs within the order *Diptera* (true flies) were made by mutual best BLAST hits (e <= 1e-20) from the three reference mosquito genomes to *D. melanogaster*. Next, initial anopheline predictions were made by mutual best hits to *An. gambiae* PEST. One advantage of the underlying YGOB framework is the software can use local synteny to "rescue" syntenic relationships that either were not assigned via the initially ortholog predictions or that were not predicted in all genomes. As an example, Figure S7 shows a visualization of microsynteny in the browser after such local correction was made. Full details are in (*108*). We also implemented an alternative approach to visualize microsynteny using a dotplot (Figure S7).

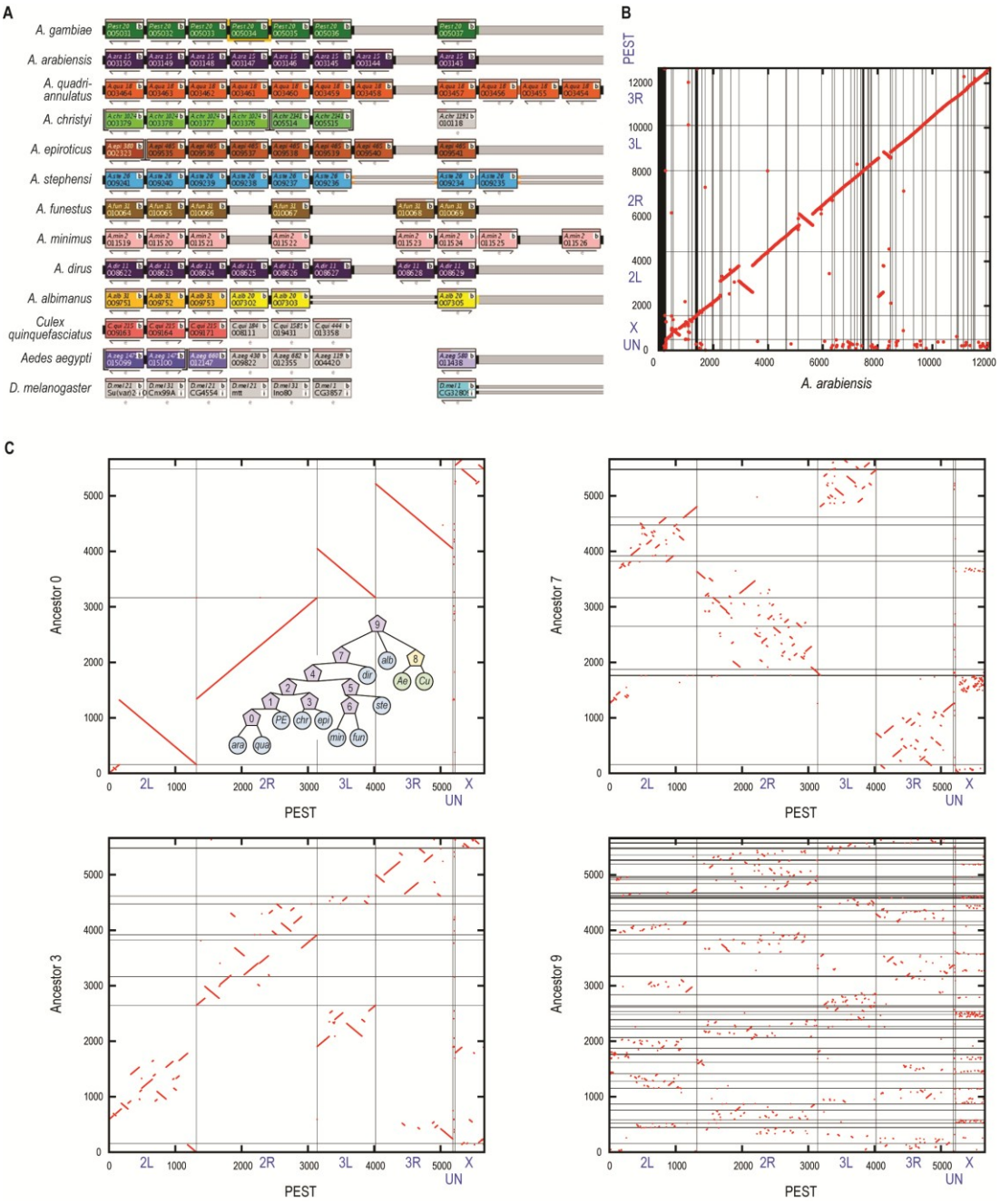


Figure S7. Gene synteny.

A. MOZGOB screenshot depicting a relatively syntenic region across multiple anopheline mosquitoes, where genes are represented by blocks and tracks are species. **B**. Dotplot view of *An. gambiae* PEST genes versus *An. arabiensis* genes after order and orientation based on PEST, with vertical and horizontal lines representing chromosome arm/scaffold breaks. X and Y axes units are number of genes. **C**. Dotplot views (as in panel B of *An. gambiae* PEST genes versus ancestors 0, 3, 7 and 9 with ancestors defined in inset: phylogenetic tree used for PATHGROUPS ancestor reconstruction, pentagons represent ancestors while circles represent extant species - *ara*, *An. arabiensis*;

qua, *An. quadriannulatus*; *PE*, *An. gambiae* PEST; *chr*, *An. christyi*; *epi*, *An. epiroticus*; *min*, *An. minimus*; *fun*, *An. funestus*; *ste*, *An. stephensi*; *dir*, *An. dirus*; *alb*, *An. albimanus*; *Ae*, *Aedes aegypti*; *Cu*, *Culex quinquefasciatus*.

Conservation of synteny within anophelines: Another significant advantage of MOZGOB was that it facilitated ancestral genome reconstruction. Specifically, we wrote custom Python scripts to extract orthologs from MOZGOB's underlying data structure for running PATHGROUPs (*109*) using a phylogenetic tree of the anophelines. Since PATHGROUPs requires orthologs to be present in all of the genomes, a total of 5,652 orthologous genes from MOZGOB were used in this reconstruction. Results using ortholog predictions from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2, (*66*)) conveyed the same global picture, as expected.

To more easily visualize conservation of synteny we generated dotplots that compared predicted ancestral gene order relative to the *Anopheles gambiae* reference. Such dotplots more clearly convey macrosynteny by providing a means to order and orient some of the more fragmented complex genomes relative to PEST, while visualizing any major inversions or shuffling present in the respective assemblies. Ordering and orienting use a majority vote mechanism to assign which strand and what is the most likely location relative to PEST for this specific scaffold/genes. Given the short divergence time of the *Anopheles gambiae* complex, these genomes were highly syntenic outside of known inversions (*9*). As ancestors moved further out from PEST, they expectedly become less syntenic with PEST, and also more fragmented due to heavier influence of less contiguous reference assemblies. Ultimately, comparison of PEST with the top-most ancestor shows only very local synteny with no clear chromosome arm conservation for genes. This result was confirmed using other ancestral reconstruction methods, namely the 16 genomes presented here are too fragmented for accurate ancestral reconstruction. However, the extent of synteny as visualized by the dotplots provides a glimpse of how genes tend to stay within linkage groups/arms, as reported by mapping studies in anopheline (see Ancestral karyotype analyses, physical mapping) and in culicine (*63*) mosquitoes, as well as how diverged the genus is as a whole.

Ancestral karyotype analyses, physical mapping

Chromosome-based genome assemblies for *An. stephensi*, *An. funestus*, *An. atroparvus*, and *An. albimanus*: Our analysis considered portions of anopheline genomes that had been physically mapped to chromosomes. We used previously published physical mapping data for *An. stephensi* (*7, 110*), *An. funestus* (*111*), and *An. albimanus* (*112*). We also performed physical mapping of *An. atroparvus* and additional mapping of *An. albimanus* by fluorescence in situ hybridization (FISH) of PCR probes designed based on the genome sequences. Sequences for chromosomally mapped scaffolds of the four anopheline species were obtained from VectorBase (www.vectorbase.org).

The percent of each genome that was physically mapped varied by species and depended on the size of scaffolds in the assembly. *An. stephensi* had ~41% of the sequenced genome mapped (92.83 Mb of 225 Mb) (Table S9). *An. funestus* had the lowest mapped portion at only about 35% or 78.95 Mb of 225 Mb of the total genome (Table S10). *An. atroparvus* had a slightly higher percentage of the genome mapped with 88.82 Mb of 224 (~40%) (Table S11). Despite having only 17 markers on the physical map of *An. albimanus* (*112*), the mapped portion of *An. albimanus* represented 128.98 Mb of 170.50 Mb (~76%) (Table S12). We mapped scaffold KB672404 to the X chromosome by FISH, as its position was predicted to be adjacent to scaffold KB672457 by a bioinformatics approach of gluing scaffolds (see Improving genome assemblies with the Breakpoint Graphs algorithm). Mapped scaffolds for a single arm were concatenated into a "pseudo-chromosome" that is considered representative of a part of the euchromatin of that chromosomal arm for a particular species.

Table S9. *An. stephensi* chromosome-based genome assembly.

Mapping data to anchor scaffolds to chromosomes for *An. stephensi*.

Chromosome	Positions		AsteS1 Scaffolds
X	1	A	KB664732
X	1	B	KB664732
X	1	C	KB664732
X	2	A	KB665121
X	3	B	KB664910
X	4	B	KB664821
X	4	B	KB664567
X	5	C	KB664677
X	6	A	KB664415
2R	7	A	KB664733
2R	7	B	KB664651
2R	7	B	KB664943
2R	8	A	KB664843
2R	9	C	KB664522
2R	9	A	KB665099
2R	9	A	KB664576
2R	9	C	KB664527
2R	9	C	KB665299
2R	9	D	KB664954
2R	10	A	KB664954
2R	10	A	KB664423
2R	10	C	KB664484
2R	10	D	KB664666
2R	10	D	KB664378
2R	10	D	KB664532
2R	10	D	KB664550
2R	11	A	KB665265
2R	12	BC	KB664323
2R	13	B	KB664460
2R	13	BC	KB664999
2R	14	AB	KB664524
2R	14	B	KB665088
2R	15	A	KB664466
2R	15	AB	KB664623
2R	16	C	KB664966
2R	17	A	KB664899
2R	17	A	KB664334
2R	17	B	KB664389
2R	17	B	KB664517
2R	17	C	KB665376

Chromosome	Positions		AsteS1 Scaffolds
2R	18	A	KB664679
2R	18	B	KB664526
2R	18	B	KB664428
2R	18	C	KB664618
2R	19	A	KB664622
2R	19	BC	KB665065
2R	19	E	KB664495
2L	20	B	KB664766
2L	20	AB	KB664480
2L	21	B	KB664513
2L	22	B	KB664888
2L	22	A	KB664988
2L	24	BC	KB665354
2L	25	B	KB665287
2L	27	AB	KB664855
2L	28	C	KB664462
2L	28	C	KB664473
2L	28	C	KB664565
3R	29	AB	KB664421
3R	29	BC	KB664288
3R	29	DE	KB664437
3R	30	A	KB664559
3R	30	AC	KB664621
3R	31	A	KB664633
3R	31	A	KB664543
3R	31	AB	KB664644
3R	31	C	KB664467
3R	32	A	KB664429
3R	32	C	KB664544
3R	32	C	KB664810
3R	33	C	KB665343
3R	33	C	KB664955
3R	34	AB	KB664549
3R	35	B	KB664457
3R	36	A	KB664708
3R	36	A	KB664401
3R	36	AB	KB665038
3R	36	B	KB665177
3R	37	B	KB664461
3R	37	D	KB664444
3R	37	D	KB664514
3L	38	E	KB664442
3L	39	AB	KB664799
3L	40	CD	KB664510
3L	40	C	KB664422
3L	40	A	KB664459
3L	40	A	KB664482
3L	40	A	KB664545
3L	41	B	KB664583
3L	42	AB	KB664433
3L	43	C	KB664692
3L	43	C	KB664832
3L	43	C	KB665021
3L	43	AC	KB664477
3L	44	A	KB664610
3L	45	C	KB664844
3L	45	A	KB664721
3L	46	D	KB665365
3L	46	C	KB664289
3L	46	B	KB664977
3L	46	AB	KB664599

Table S10. *An. funestus* chromosome-based genome assembly.

Mapping data to anchor scaffolds to chromosomes for *An. funestus*.

Chromosome	Positions		AFunF1 Scaffolds
X	1	B	KB668322
X	1	C	KB668245
X	1	C	KB669125
X	2	B	KB668367
X	3	A	KB668668
X	3	D	KB669003
X	3	D	KB668522
X	5	CD	KB668688
X	5	D	KB668760
X	6		KB669536
2R	7	A	KB668728
2R	7	B	KB668825
2R	9	A	KB668221
2R	9	B	KB669169
2R	10	B	KB668737
2R	10	C	KB668737
2R	11	B	KB668845
2R	12	A	KB668467
2R	12	B	KB668822
2R	12	B	KB669369
2R	12	C	KB668793
2R	12	D	KB668672
2R	12	E	KB668785
2R	13	A	KB668706
2R	13	B	KB668715
2R	13	C	KB668766
2R	13	D	KB668679
2R	14	B	KB668478
2R	14	C	KB669525
2R	14	D	KB668835
2R	15	B	KB668947
2R	15	C	KB669358
2R	15	C	KB668911
2R	15	E	KB669547
2R	16	A	KB668837
2R	16	B	KB669192
2R	16	C	KB668748
2R	18	A	KB668234
2R	18	C	KB668734
2R	18	C	KB668836
2R	19	C	KB669114
2L	28	C	KB668881
2L	28	A	KB668725
2L	27	D	KB668751
2L	27	D	KB668992
2L	27	C	KB668795
2L	27	B	KB669136
2L	26	CD	KB668693
2L	26	C	KB668681
2L	26	A	KB669047
2L	24	C	KB669280
2L	24	A	KB669314
2L	23	A	KB668882
2L	22	D	KB669502
2L	22	B	KB668692
2L	20	D	KB669070
2L	20	C	KB669092
2L	20	B	KB669281
3R	36	F	KB668744
3R	36	E	KB669436
3R	36	D	KB669391
3R	35	F	KB668456
3R	35	F	KB668880
3R	35	F	KB668816

Chromosome	Positions		AFunF1 Scaffolds
3R	35	F	KB668731
3R	35	C	KB668790
3R	35	B	KB668853
3R	35	B	KB668671
3R	35	A	KB668848
3R	34	A	KB668746
3R	33	D	KB668723
3R	33	D	KB669347
3R	33	C	KB668661
3R	33	A	KB668789
3R	32	B	KB668644
3R	31	D	KB668695
3R	30	C	KB668808
3R	30	C	KB668792
3R	29	D	KB669458
3R	29	C	KB668683
3R	29	C	KB669236
3R	29	B	KB668378
3L	38	C	KB668859
3L	38	C	KB668823
3L	39	A	KB669325
3L	39	A	KB668578
3L	39	A	KB668717
3L	39	B	KB668659
3L	40	A	KB669014
3L	40	B	KB668918
3L	40	B	KB668868
3L	44	D	KB668444
3L	44	A	KB668830
3L	44	A	KB668754
3L	43	B	KB668500
3L	43	A	KB668422
3L	42	B	KB668925
3L	41	B	KB669025
3L	41	A	KB669603
3L	41	A	KB668849
3L	40	C	KB668784
3L	46	B	KB668814
3L	46	B	KB668682
3L	46	D	KB668252

Table S11. *An. atroparvus* chromosome-based genome assembly.

Mapping data to anchor scaffolds to chromosomes for *An. atroparvus*.

Chromosome	Positions		AatrE1 Scaffolds
X	4	A	KI421898
X	4	B	KI421898
2R	6	B	KI421882
2R	9	B	KI421882
2L	16	C	KI421886
2L	17	B	KI421886
2L	19	A	KI421884
2L	21	B	KI421884
3R	24	A	KI421883
3R	25	B	KI421883
3R	26	B	KI421885
3R	28	B	KI421885
3L	38	B	KI421887
3L	39	A	KI421887

Table S12. *An. albimanus* chromosome-based genome assembly.

Mapping data to anchor scaffolds to chromosomes for *An. albimanus*.

Chromosome	Positions		AalbS1 Scaffolds
X	1	AB	KB672457
X	3	C	KB672404
2R	8	B	KB672446
2R	11	A	KB672397
2R	11	C	KB672320
2R	13	C	KB672435
2R	14	B	KB672435
2L	25	A	KB672397
2L	24	B	KB672397
2L	17	B	KB672413
3R	30	A	KB672424
3R	28	B	KB672424
3R	27	A	KB672424
3L	40	A	KB672286
3L	42	C	KB672286
3L	43	B	KB672286

Synteny preservation and translocations at the whole chromosome arm level in the evolution of anophelines: Malaria mosquitoes have a 5-arm chromosomal complement. In *An. gambiae*, arms are denoted as chromosomal elements 1 (X), 2+3 (2R+2L), 4+5 (3R+3L) 4,5. The correspondence of other species chromosomal arms is as follows: *An. funestus* 1, 2+4, 3+5; *An. stephensi* 1, 2+5, 3+4; *An. atroparvus* 1, 4+3, 2+5; *An. albimanus* 1, 2+4, 5+3. Therefore, *An. stephensi* and *An. atroparvus* have the same arm association. Also, *An. funestus* and *An. albimanus* have the same arm association. Our mapping data indicate that the anopheline genomes have conserved gene membership on chromosome arms with no pericentric inversions or partial arm translocations. The synteny relationships among anopheline species and *D. melanogaster* (*113*) including divergence times (*10, 15*) are summarized in Table S13. We noticed a major difference between anophelines and *Drosophila* in the pattern of autosomal arm evolution. *Drosophila* species have either all acrocentric or metacentric/submetacentric autosomes consisting of fusions of the chromosomal elements (*114*). In contrast, chromosome arms in anophelines always belong to metacentric/submetacentric autosomes, but they reshuffle multiple times via translocations during ~100 million years (MY) of evolution.

Table S13. Chromosomal arm correspondences.

Syntenic relationships among anopheline species with chromosomal mapping data and *Drosophila melanogaster*.

Species	Element 1	Element 2	Element 3	Element 4	Element 5	Divergence time from *An. gambiae* (MY)
An. gambiae	X	2R	2L	3R	3L	
An. stephensi	X	2R	3L	3R	2L	30.4
An. funestus	X	2R	3R	2L	3L	30.4
An. atroparvus	X	3R	2L	2R	3L	58
An. albimanus	X	2R	3L	2L	3R	100
D. melanogaster	X	3R	3L	2L	2R	250

Patterns of chromosome rearrangements in anophelines: Each of the species, *An. funestus*, *An. stephensi*, *An. atroparvus*, and *An. albimanus*, was individually compared to *An. gambiae*. Single-copy orthologs from *An. gambiae* and each of the other anopheline species were identified using OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2) as delineated by the OrthoDB (*66*) methodology. The gene IDs of *An. gambiae* were retrieved based on GFF3 annotation from VectorBase. The positions of these genes in *An. gambiae* chromosomes were individually compared with their positions on chromosomes in each of the other species. The comparative positions of genes within mapped scaffolds based on orthology relationships were plotted using the R program genoPlotR (*115*). Orientation of scaffolds on chromosomes was obtained from physical mapping data. Scaffolds mapped with only one probe were given the default orientation. The comparative position of genes in each pair of species was used to determine the number of conserved synteny blocks between *An. gambiae* and each of the other species. To determine the number of conserved synteny blocks, two parameters were considered: the orientation and the order of orthologous genes. To be part of the same conserved synteny block, a group of two or more genes must have the same orientation and order in both species. Using these criteria conserved synteny blocks were numbered between *An. gambiae* and in each of the other species. *Anopheles gambiae* was assigned the default gene order of 1, 2, 3, 4...n along the chromosome. Other species were considered rearranged compared to *An. gambiae*. This convention means that the numbering of conserved synteny blocks was the same in both species but the order was rearranged in An. *funestus*, *An. stephensi*, *An. atroparvus*, and *An. albimanus* to reflect the shuffling of conserved synteny blocks that occurred over evolutionary time between species. We estimated the number of chromosomal inversions between the two species using the Genome Rearrangements in Mouse and Man (GRIMM) program (*116*).

An. stephensi – An. gambiae: Based on the comparative positions of 3,908 single-copy orthologs we identified 253 conserved synteny blocks and 130 inversions between *An. stephensi* and *An. gambiae*. The greatest density of inversions between *An. gambiae* and *An. stephensi* is on the sex chromosome or chromosomal element 1 (3.89 inversions/Mb). Of the autosomes, element 2 was the most rearranged with an inversion density of 1.48 inversions/Mb. Elements 3 and 5 had similar densities of inversions with 1.08 and 1.09 inversions/Mb, respectively. Element 4 was far less rearranged than the other autosomal arms with an inversion density of only 0.84 inversions/Mb (Table S14). The most recent estimate of divergence time between *An. stephensi* and *An. gambiae* is 30.4 MY (*15*). Using this divergence time and assuming that each inversion has two breaks, we calculated the number of breaks/Mb/MY. For the *An. gambiae-An. stephensi* pair the rates of evolution were 0.128 for element 1 and 0.049, 0.036, 0.028, 0.036 for elements 2, 5, 4, and 3, respectively. Sex chromosome evolution in *An. stephensi* is approximately 3.47 times faster than the average autosomal rate of evolution.

Table S14. Rates of chromosomal evolution in anophelines.

Synteny blocks per chromosomal arm with counts of inversions to infer rates of chromosomal shuffling in anophelines. Mb: megabasepairs; MY: million years; GRIMM: Genome Rearrangements In Man and Mouse.

An. stephensi – An. gambiae

Element	Size of mapped scaffolds (Mb)	# Conserved synteny blocks	# Inversions (GRIMM)	Inversions/Mb	Breaks/Mb/MY
1	7.97	47	31	3.890	0.128
2	31.83	89	47	1.477	0.049
5	22.11	65	24	1.086	0.036
4	22.60	37	19	0.841	0.028
3	8.33	15	9	1.081	0.036
Total	92.83	253	130	1.675	0.055

An. funestus – An. gambiae

Element	Size of mapped scaffolds (Mb)	# Conserved synteny blocks	# Inversions (GRIMM)	Inversions/Mb	Breaks/Mb/MY
1	6.81	31	25	3.670	0.121
2	25.62	88	50	1.952	0.064
4	15.20	57	30	1.974	0.065
3	18.78	35	15	0.799	0.026
5	12.55	42	25	1.993	0.066
Total	78.96	253	145	2.077	0.068

An. atroparvus – An. gambiae

Element	Size of mapped scaffolds (Mb)	# Conserved synteny blocks	# Inversions (GRIMM)	# Inversions/Mb	Breaks/Mb/MY
1	4.03	31	29	7.203	0.124
4	28.69	106	60	2.091	0.036
3	25.46	104	52	2.043	0.035
2	20.24	59	29	1.433	0.025
5	10.41	40	25	2.403	0.041
Total	88.82	340	195	3.034	0.052

An. albimanus – An. gambiae

Element	Size of mapped scaffolds (Mb)	# Conserved synteny blocks	# Inversions (GRIMM)	Inversions/Mb	Breaks/Mb/MY
1	9.93	134	129	12.996	0.130
2	36.23	234	154	4.251	0.043
3	30.18	228	121	4.009	0.040
4	34.59	194	118	3.412	0.034
5	18.07	113	67	3.708	0.037
Total	128.99	903	589	5.675	0.057

An. funestus – An. gambiae: We identified a total of 253 conserved synteny blocks between *An. gambiae* and *An. funestus* using 3,727 single-copy orthologs. Distribution of conserved synteny blocks among the 5 chromosomal elements was non-uniform with 31, 88, 57, 35, and 42 blocks for chromosomal elements 1, 2, 4, 3, and 5, respectively (Table S14). The program GRIMM estimated a total of 145 chromosomal inversions between *An. gambiae* and *An. funestus*. Chromosomal element 2 displayed the greatest number of inversions equal to 50. The remaining elements 1, 4, 3, and 5 had 25, 30, 15, and 25 inversions, respectively. It is important to note, that anopheline chromosomal arms are of unequal length and the density of chromosomal inversions is more informative than the numbers of inversions per arm. The sex chromosome, or chromosomal element 1, is represented by only 6.8 Mb of the mapped *An. funestus*

genome, and yet this element exhibits an inversion density of 3.67 inversions/Mb. This is approximately 1.5 times greater than the highest density of inversions found on autosomal elements. Among autosomes, elements 2, 4, and 5 displayed similar inversion densities with 1.95, 1.97 and 1.99 inversions/Mb, respectively. The inversion density of chromosomal element 3 was only 0.79 inversions/Mb or less than half of that of other autosomes. To calculate rates of evolution we considered the number of breaks/Mb/MY. *Anopheles funestus* is estimated to have diverged from *An. gambiae* 30.4 MY ago (*15*). The rate of breaks/Mb/MY for the *An. funestus* X was 0.120. By comparison, autosomal rates of evolution ranged from 0.026 (element 3) to 0.066 (element 5). Elements 2 and 4 had rates of evolution of 0.064 and 0.065, respectively. The average rate of evolution across autosomes was 0.055, which is less than half the rate of evolution of the sex chromosome. Interestingly, the *An. funestus* autosomal rate of evolution is higher than those of *An. stephensi*.

An. atroparvus – An. gambiae: We mapped about 40% of the total genome assembly to chromosomes of *An. atroparvus* by FISH. 3,837 single-copy orthologs were used to identify 340 conserved synteny blocks and GRIMM estimated 195 inversions discriminating *An. atroparvus* and *An. gambiae*. *An. atroparvus* is thought to have diverged from *An. gambiae* 58 MY ago, and so a greater number of inversions could be expected. This divergence time was estimated for *An. quadrimaculatus*, which is closely related to *An. atroparvus* (*10*). Despite the difference in divergence times, sex chromosome versus autosome rearrangement trend remains consistent. In *An. atroparvus* the density of inversions for chromosomal element 1 is 3 times greater than that for the most rearranged autosome (7.2 inversions/Mb for element 1 versus 2.4 inversions/Mb for element 5). Chromosomal elements 3 and 4 exhibited similar densities of inversions with 2.04 and 2.09 inversions/Mb, respectively. Element 2 was less rearranged with only 1.43 inversions/Mb. The rates of evolution were 0.124, 0.036, 0.035, 0.025, and 0.041 breaks/Mb/MY for elements 1, 4, 3, 2, and 5, respectively (Table S14). The difference in the rate of evolution between the sex chromosome and autosomes is more than 3.65 times, which is more pronounced than the difference seen in *An. stephensi*.

An. albimanus – An. gambiae: Changes in gene order between *An. gambiae* and *An. albimanus* were reconstructed using 6,364 single copy orthologs. The analysis resulted in 903 conserved synteny blocks and an estimated number of 589 chromosomal inversions. Densities of inversion varied from 12.99 to 3.41 of inversions/Mb in elements 1 and 3, respectively. Elements 2, 5, and 4 had 4.25, 4.01, and 3.71 inversions/Mb, respectively. (Table S14). Divergence time between *An. gambiae* and *An. albimanus* is 100 MY, which was calculated using a closely-related species to *An. albimanus*, *An. darlingi* (*10*). Similarly to the other anopheline species, the highest rates of evolution are on the sex chromosome. Element 1 had a rate of evolution of 0.130 breaks/Mb/MY. This is the highest rate of evolution in the species that we determined. However, it is interesting to note that the difference in sex chromosome versus autosomal rates of

evolution is not as pronounced in this species as in some of the others (0.130 for element 1 versus 0.038 autosomal average). This difference of 3.38 times is slightly less than for *An. atroparvus* and *An. stephensi*.

We used two independent groups t-test (STATISTICA 10.0, StatSoft Inc. 2014) to compare evolutionary rates between the X chromosomes and autosomes in *An. stephensi*, *An. funestus*, *An. atroparvus*, and *An. albimanus*. We found that the rate of evolution between the X chromosomes and autosomes is significantly different, $t(18)=12.527$, $p < 0.0000001$, with means of 0.126 and 0.041, respectively.

Rates of chromosome rearrangements in anophelines in comparison with *Drosophila*: We obtained the number of inversions between *D. melanogaster* and 8 other *Drosophila* species from the published Dataset S1 (*117*). This study used the same definition of conserved synteny blocks as in our study: more than one gene in the same order and orientation. The numbers of inversions were calculated using Multiple Genome Rearrangement (MRG) program (*118*), which uses the same algorithm as GRIMM. To calculate the number of breaks/Mb/MY for these species we used the length of the mapped portion of the genome assembly to each chromosomal arm of a particular species and the divergence time from *D. melanogaster* (*114, 119*). We calculated the rates of rearrangements separately for the X chromosome and for the total mapped genome.

Drosophila melanogaster diverged from *An. gambiae* ~260 MY ago (*113*), and a comparison of rates of evolution in these two genera can provide a glimpse of evolutionary trends that encapsulates a large part of the phylogenetic tree of Diptera. The availability of genomic data for several *Drosophila* species has permitted multispecies comparison of evolutionary rates between genera *Anopheles* and *Drosophila*. Rates of rearrangements were calculated separately for the X chromosome (Table S15) and for the total mapped genome (Table S16). We considered Muller's element A in *Drosophila* separately regardless of whether it was connected to another element (like in *D. pseudoobscura* and *D. willistoni*) or if it was free (like in most other species).

Table S15. Rates of X chromosome evolution in *Drosophila*.

Rates of X chromosome shuffling in fruit flies in terms of inversions relative to *Drosophila melanogaster*. Mb: megabasepairs; MY: million years.

Species	# of inversions with *D. melanogaster*	Size of mapped scaffolds (Mb)	Inversions/Mb	Divergence, MY	Breaks/Mb/MY
D. erecta	3	21.3	0.141	12.6	0.011
D. yakuba	6	21.8	0.275	12.6	0.022
D. ananassae	159	31.5	5.048	44.2	0.114
D. pseudoobscura	198	20.3	9.754	54.9	0.178
D. willistoni	381	27.9	13.656	62.2	0.220
D. virilis	298	30.5	9.770	62.9	0.155
D. mojavensis	300	32	9.375	62.9	0.149
D. grimshawi	330	26.4	12.500	62.9	0.199

Table S16. Genome rearrangements in *Drosophila*.

Rates of the genome rearrangement in the total mapped genome of fruit flies in terms of inversions relative to *Drosophila melanogaster*. Mb: megabasepairs; MY: million years.

	# of inversions with *D. melanogaster*	Size of mapped scaffolds (Mb)	Inversions/Mb	Divergence, MY	Breaks/Mb/MY
D. erecta	20	124.5	0.161	12.6	0.013
D. yakuba	35	138.2	0.253	12.6	0.020
D. ananassae	507	130.5	3.885	44.2	0.088
D. pseudoobscura	790	129	6.124	54.9	0.112
D. willistoni	1624	153.1	10.607	62.2	0.171
D. virilis	1295	148.7	8.709	62.9	0.138
D. mojavensis	1317	152.7	8.625	62.9	0.137
D. grimshawi	1355	135.4	10.007	62.9	0.159

Overall rates of chromosome evolution in anophelines fall within the rates of rearrangement in *Drosophila*, with rates being higher in the majority of *Drosophila* species, except in *D. erecta* and *D. yakuba*. Two independent groups t-test (STATISTICA 10.0, StatSoft Inc. 2014) was used to compare genome-wide rates of rearrangements between the group of four anopheline species (*An. stephensi*, *An. funestus*, *An. atroparvus*, *An. albimanus*) and the group of eight *Drosophila* species (*D. erecta*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, *D. virilis*, *D. mojavensis*, *D. grimshawi*). We found that the rates of chromosomal evolution are not significantly different, $t(10)=1.842$, $p=0.095$, with means of 0.047 and 0.105, respectively. Similarly, we found that the rates of X chromosome evolution are not significantly different between the anopheline and *Drosophila* species, $t(10)=0.132$, $p=0.898$, with means of 0.126 and 0.131, respectively. However, our data reveals some interesting discrepancies in rates of evolution between the two groups of dipterans when we considered the ratio of the X chromosome evolution rate to the total rates of rearrangement (Figure S8). We found that the ratio of the rates of evolution of sex chromosome to all chromosomes is significantly higher in anophelines than *Drosophila*, $t(10)=7.299$, $p=0.000026$, with means of 2.726 and 1.200, respectively. We also used two independent groups t-test to compare evolutionary rates between the X chromosomes and total rearrangement rates in fruit flies and in mosquitoes. We found that the rates of evolution between the X chromosomes and all chromosomes are not significantly different in *Drosophila* species, $t(14)=-0.756$, $p=0.462$, with means of 0.131 and 0.105, respectively. In contrast, the rates of evolution between the X chromosomes and all chromosomes were significantly different in anopheline species, $t(6)=-15.247$, $p=0.000005$, with means of 0.126 and 0.048, respectively. The fast rate of X chromosome rearrangements is in sharp contrast with the paucity of polymorphic inversions on the X in anopheline species. It is possible that heterozygote X chromosome inversions are underdominant in females. We explored whether this result could be sensitive to genome

assembly fragmentation by conducting a simple analysis in which rearrangements were noted only when gene adjacencies from pairs of genomes both supported a rearrangement. This finding could be indicative of a greater role of the X chromosome rearrangements in speciation of malaria mosquitoes. Previous studies indicated that the X chromosome has a disproportionately large effect on male and female hybrid sterility and inviability in *An. gambiae* and *An. arabiensis* (*120, 121*). Future genomic studies are necessary to dissect the role of X chromosome inversions in speciation of malaria vectors.

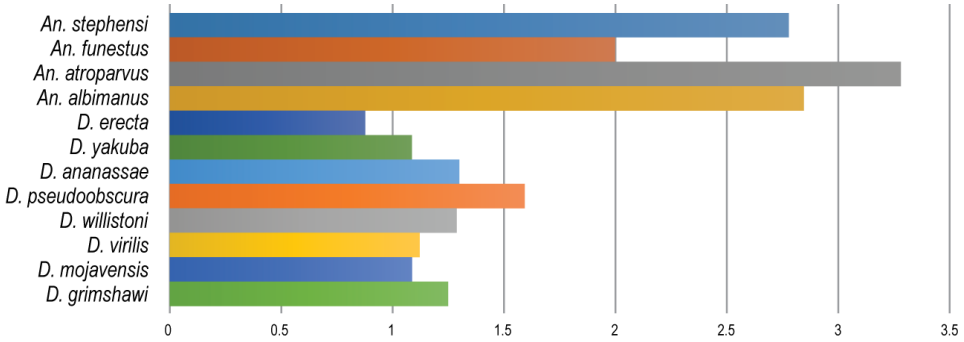


Figure S8. X versus autosome rearrangement rates.

The ratio of X chromosome evolutionary rate to the total rate of rearrangement in anophelines and drosophilids.

Gene movements among the anophelines

For the gene movement analysis, we used the six species that have the physical mapping results available for anopheline species (see Ancestral karyotype analyses, physical mapping): *An. gambiae* PEST, *An. funestus*, *An. stephensi*, *An. atroparvus*, *An. albimanus*, and for *Aedes aegypti* (*122*). The gene families across these species were extracted from the Culicidae level orthologous groups from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2). There were a total of 18,559 gene families, covering 71,777 genes across the six species. Among them, a total of 40,276 genes were physically mapped to a chromosome, and thus could be analyzed for gene movements in this study: 11,822 *An. gambiae*; 4,720 *An. funestus*; 5,413 *An. stephensi*; 5,292 *An. atroparvus*; 8,332 *An. albimanus*; and 4,697 *Aedes aegypti.*

Using the chromosome elements 1 to 5 as representations of each arm, we encoded the distribution of a gene family as a vector u = (u1, u2, … , u5) with each element specifying the number of gene homologs on each of the muller elements 1 to 5. Then, we performed parsimonious reconstruction on the genomic distribution of gene families following the methods described in (*123*). To test for the overrepresentation of movements, we counted the movements between arms and compared them against the expectation. When there was more than one parsimonious ancestral state, those cases

were excluded from the count (Including the ambiguous counts did not change the results). The expected proportions of movements were calculated using the formula similar to (*124*). But instead of the length of the mapped regions on chromosome arms, we used the total number of genes mapped on each of the arms as a proxy for the length.

We identified a total of 132 unambiguous gene movements between chromosome arms. The gene movements identified are listed in Table S17 and Table S18 shows the overall pattern of gene movements between chromosome arms. *Tests for significance of movements out of the X*: We tested the null hypothesis that the proportion of genes moving out of the X chromosome (59/132) is less than or equal to the expected proportion (0.116). One sample proportion test gave a p-value < 2.2e-16, showing a very significant over-representation of genes moving out of the X-chromosome. The 95 percent confidence interval for the true proportion of genes moving out of the X chromosome was (0.38, 1.00). *Limitations of the method and data*: Due to large proportions of the scaffolds not being assigned to chromosomes, there were many genes in the gene families that were missing chromosome information. Because we ignored the genes that have missing information, we may have identified movements on the wrong branch, or identified multiple independent movements when the real movement happened only once earlier in the ancestor. Although this may lead to overestimation of gene movements, we believe that it does not bias the detection of movements in any particular direction.

Table S17. Catalog of gene movements.

Cataloged gene movements (translocations) between chromosome arms.

family ID	branch	from	to
ODBMOZ00470	AFUN	5	3
ODBMOZ00574	AALB	1	4
ODBMOZ00670	AALB	1	5
ODBMOZ00670	AATE	1	3
ODBMOZ00889	AGAP	5	1
ODBMOZ00889	AGAP	5	3
ODBMOZ00889	AGAP	5	4
ODBMOZ01035	AGAP	3	4
ODBMOZ01087	AFUN	3	5
ODBMOZ01180	AATE	4	3
ODBMOZ01313	Cellia	1	5
ODBMOZ01361	AALB	4	1
ODBMOZ01565	AATE	5	2
ODBMOZ01565	AGAP	5	4
ODBMOZ01809	AALB	4	2
ODBMOZ01809	AGAP	4	2
ODBMOZ02277	AALB	1	2
ODBMOZ02279	AGAP	4	5
ODBMOZ02475	AALB	3	2
ODBMOZ02475	AGAP	3	2
ODBMOZ02725	AGAP	2	4
ODBMOZ03093	Cellia	4	5
ODBMOZ03201	AGAP	1	2
ODBMOZ03204	Anopheles	4	2
ODBMOZ03493	AATE	1	4
ODBMOZ04181	AATE	3	5
ODBMOZ05558	ASTE	4	5
ODBMOZ06144	ASTE	3	2

family ID	branch	from	to
ODBMOZ06308	ASTE	1	3
ODBMOZ06392	AALB	1	5
ODBMOZ06392	AGAP	1	5
ODBMOZ06632	AATE	1	3
ODBMOZ07065	AGAP	4	2
ODBMOZ07694	AALB	4	2
ODBMOZ07745	AATE	3	2
ODBMOZ07853	AFUN	3	2
ODBMOZ07998	Cellia	1	4
ODBMOZ08131	AALB	1	4
ODBMOZ08699	AGAP	1	3
ODBMOZ08907	Cellia	1	2
ODBMOZ09455	ASTEFUN	5	1
ODBMOZ09572	AFUN	5	2
ODBMOZ09666	AGAP	4	2
ODBMOZ09741	AFUN	3	2
ODBMOZ10076	AATE	1	4
ODBMOZ10478	AALB	1	4
ODBMOZ11259	AALB	1	3
ODBMOZ11370	Cellia	3	2
ODBMOZ11567	AGAP	3	1
ODBMOZ11569	AALB	1	5
ODBMOZ11970	AALB	2	3
ODBMOZ12434	AATE	1	3
ODBMOZ12434	AGAP	1	4
ODBMOZ12557	Cellia	5	3
ODBMOZ12612	AGAP	1	3
ODBMOZ12618	Cellia	3	2
ODBMOZ13159	AGAP	1	4
ODBMOZ13749	AGAP	4	1
ODBMOZ14044	AALB	1	2
ODBMOZ14337	Cellia	3	1
ODBMOZ14405	AGAP	1	3
ODBMOZ14431	AALB	1	3
ODBMOZ14431	Cellia	1	2
ODBMOZ14824	ASTE	2	1
ODBMOZ14830	AGAP	2	1
ODBMOZ14906	AGAP	1	5
ODBMOZ15028	AALB	3	5
ODBMOZ15702	AGAP	2	1
ODBMOZ15751	AGAP	1	5
ODBMOZ15980	AGAP	3	2
ODBMOZ16103	AALB	1	5
ODBMOZ16452	AALB	5	2
ODBMOZ16507	AGAP	1	2
ODBMOZ16905	Anopheles	3	4
ODBMOZ17507	AALB	4	2
ODBMOZ17596	AALB	5	4
ODBMOZ17619	Cellia	5	1
ODBMOZ17685	AFUN	3	5
ODBMOZ18317	AALB	1	5
ODBMOZ18768	Cellia	5	2
ODBMOZ19000	Cellia	4	1
ODBMOZ19096	AALB	1	2
ODBMOZ19096	Cellia	1	3
ODBMOZ19113	AALB	1	4
ODBMOZ19113	Cellia	1	2
ODBMOZ19340	ASTE	3	2
ODBMOZ19618	AATE	3	2
ODBMOZ19737	Cellia	3	4
ODBMOZ19761	Anopheles	1	4
ODBMOZ19904	AALB	3	2
ODBMOZ20091	AALB	1	4
ODBMOZ20144	AGAP	1	2
ODBMOZ20295	AATE	1	2
ODBMOZ20356	Cellia	4	3
ODBMOZ20385	AGAP	3	2
ODBMOZ20410	AALB	1	4
ODBMOZ20410	AATE	1	2

family ID	branch	from	to
ODBMOZ20555	AALB	3	1
ODBMOZ20782	Cellia	1	5
ODBMOZ20900	AGAP	1	5
ODBMOZ21315	AATE	2	3
ODBMOZ21636	AALB	1	4
ODBMOZ21759	AGAP	3	4
ODBMOZ22150	AGAP	4	3
ODBMOZ22505	AALB	1	5
ODBMOZ23449	AGAP	5	1
ODBMOZ23846	AATE	4	5
ODBMOZ23859	AGAP	1	3
ODBMOZ23883	AALB	2	3
ODBMOZ23883	AGAP	2	3
ODBMOZ23883	AGAP	2	5
ODBMOZ23929	AALB	1	3
ODBMOZ23929	AGAP	1	4
ODBMOZ24120	AALB	1	2
ODBMOZ24120	AGAP	1	2
ODBMOZ24545	AALB	1	5
ODBMOZ24620	AALB	1	2
ODBMOZ24731	Anopheles	1	2
ODBMOZ25115	AATE	1	2
ODBMOZ25314	AGAP	2	5
ODBMOZ25564	AALB	1	4
ODBMOZ26018	AALB	1	2
ODBMOZ26103	AGAP	2	1
ODBMOZ26478	AGAP	1	5
ODBMOZ26721	AALB	5	3
ODBMOZ26721	AGAP	5	2
ODBMOZ26721	AGAP	5	3
ODBMOZ27191	AFUN	3	4
ODBMOZ27281	AFUN	5	1
ODBMOZ27640	AALB	1	4
ODBMOZ30741	AATE	5	2
ODBMOZ30741	AATE	5	4

Table S18. Movements of genes between chromosome arms.

Summary of movements of genes between chromosome arms excluding ambiguous calls.

From / To	1	2	3	4	5	Grand Total
1		17	12	16	14	**59**
2	4		4	1	2	**11**
3	3	13		5	4	**25**
4	3	7	3		4	**17**
5	5	6	5	4		**20**
Grand Total	**15**	**43**	**24**	**26**	**24**	**132**

ANGES: reconstructing ancestral genomes

We reconstructed Contiguous Ancestral Regions (CARs) for several *An. gambiae* complex and anopheline ancestral genomes using the software ANGES (*125*), that assembles CARs from conserved gene adjacencies and intervals using local parsimony algorithms initially developed for computing physical maps (*126*).

Input data: gene families and species tree.

The input data were composed of gene families of orthologous groups delineated at the Dipteran level from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2) and corresponding GFF files for the following 11 species, selected according to the quality of their assemblies: *An. gambiae*, *An. arabiensis*, *An. quadriannulatus*, *An. merus*, *An. stephensi*, *An. minimus*, *An. funestus*, *An. dirus*, *An. farauti*, *An. atroparvus*, *An. albimanus*. The input files were processed to extract gene families with a unique occurrence in each considered genome (one-to-one orthologous families); to resolve the issue caused by overlapping genes, genes were ordered along chromosomes/scaffolds/supercontigs according to their middle point. This resulted in a set of 5,343 gene families. The considered species tree is the same as the one used for the CAFE gene family analysis (see gene families section), restricted to the 11 considered species. In order to confirm the topology of the *An. gambiae* complex subtree the subset of 446 one-to-one orthologous gene families whose occurrence in *An. gambiae* is on the X chromosome was considered, as X chromosomes are known to be less subject to introgression (*9*).

For each internal node of the considered phylogeny but the root, ANGES was applied using the following options: markers were doubled to account for the orientation of genes, conserved syntenies composed of oriented gene adjacencies and strong common intervals were detected according to the Dollo parsimony, locally conserved syntenies were assembled into CARs using the greedy heuristic.

The analysis of the X chromosome showed a very clear syntenic signal suggesting the chosen topology of the *An. gambiae* complex: the hypothesis of *An. gambiae* and *An. arabiensis* forming a clade leads to a much higher number of syntenic conflicts in assembling adjacencies and intervals into CARs (20) compared to the hypothesis of *An. arabiensis* and *An. quadriannulatus* forming a clade (4), confirming a new species phylogeny (*9*). Moreover, for all other ancestral nodes, the number of discarded adjacencies and intervals to clear syntenic conflicts is upper-bounded by 10, even for deep ancestors. However, the increasing fragmentation of deep ancestral genomes, even when limited to the X chromosome dataset, illustrates the loss of syntenic signal.

Besides *An. gambiae*, and to a lesser extent *An. albimanus* and *An. minimus*, most other genomes were fragmented in a large number of chromosomal segments (scaffolds or supercontigs), thus potentially preventing the detection of conserved syntenies. Nevertheless, the availability of the three relatively well assembled genomes and their location in the anopheline phylogeny allowed the reconstruction of relatively well defined sets of CARs (Figure S9). In all cases, less than 1% of detected local conserved syntenies needed to be discarded during the assembly phase to respect the constraint of linear CARs. Moreover, in most cases, a few number of CARs covers most gene families, and a large majority of genes that do not belong to one-to-one orthologous families

belong to conserved adjacencies present in CARs, suggesting that a few CARs capture the evolution of most of the anopheline genomes, except for the ancestor of all anophelines and *An. albimanus* (node A6 in Figure S9), where gene order seems to have been much less conserved. At the level of all ancestors, most CARs were limited to genes from a single *An. gambiae* arm, showing that the current fragmentation level does not allow for reconstruction of the chromosomal arm structure of ancestral genomes.

Double-Cut-and-Join genome rearrangement distance.

To estimate the genome rearrangement distances along branches of the tree, the Double-Cut-and-Join distance (*127*) was computed, using the UniMog server (*128*) with the rDCJ option, for the X chromosome dataset, composed of gene families whose occurrence in *An. gambiae* is located on the X chromosome.

The results (Figure S9) show the impact of assembly fragmentation on the observed rearrangement distance, suggesting that many of the observed rearrangement on terminal branches are in fact artifacts due to assembly breaks. Distances between pairs of internal nodes are very close to the sum of the number of X chromosome CARs, suggesting here again that most of the observed rearrangements represent well conserved gene order in X chromosomes. Finally, the comparison of the distances between X chromosomes of ancestor A6 with *An. atroparvus* and *An. albimanus* suggests that this ancestor is much closer to *An. atroparvus* than to *An. albimanus*.

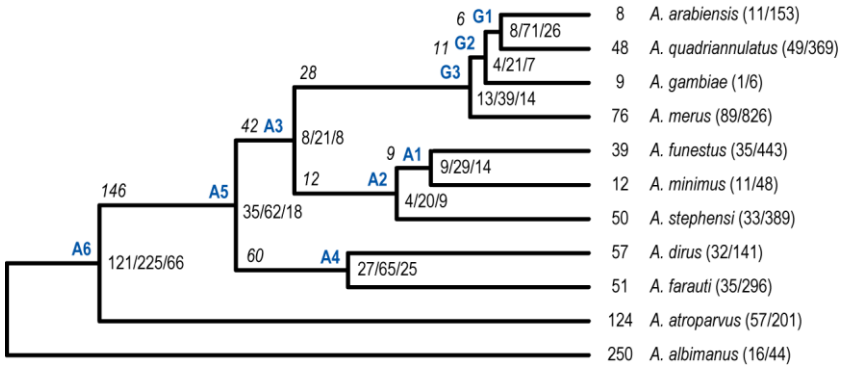


Figure S9. *Anopheles* Contiguous Ancestral Regions (CARs).

Numbers associated to extant genomes indicate the number of chromosomal fragments containing one-to-one orthologous families in the X chromosome dataset (left) and in the whole-genome dataset (right). Labels associated to ancestral nodes indicate the number of X chromosome CARs (left), the number of whole genome CARs (middle) and the number of whole genome CARs that cover 90% of one-to-one orthologous families (right). Numbers associated to branches represent the DCJ distance between the two sets of X chromosome CARs for the genomes at the extremities of the branch.

The data used for this analysis also substantiate the observation of higher rate of rearrangements on the X than the autosomes. For each branch of the tree, we considered the pair of species that define it (pair = (ancestral genome, extant genome) or (ancestral genome, ancestral genome) and looked at the two sets of gene adjacencies that define

both genomes. We counted how many adjacencies were found in one genome but not in the other but could not be ruled out due to assembly/reconstruction fragmentation. This last condition could be easily identified when the adjacencies were both localized to extremities of scaffolds (for extant genomes) or CARs (for ancestral genomes). We call such adjacencies "potentially conserved". We defined the distance between the two genomes as the number of adjacencies present only in one of both and NOT potentially conserved. This is a useful proxy for rearrangement distance, and this metric tends to minimize the distance and erase effects of fragmentation by effectively converting potentially conserved adjacencies to conserved status. Calculating the X:autosomal rearrangement rate ratios on the branches of the tree using this approach, we found a ratio of 2 or greater on 11 of 17 branches of sufficient length to yield reliable rates. Thus, this analysis indicates that the higher rate of X rearrangement relative to autosomes is conservative with respect to genome assembly fragmentation.

Improving genome assemblies with the Breakpoint Graphs algorithm

We considered the following 6 anopheline genomes: *An gambiae* (PEST strain), *An, quadriannulatus*, *An. arabiensis*, *An. merus*, *An. melas*, *An. dirus* and *An. albimanus*. Since all these genomes (except *An. gambiae*) are relatively fragmented and are represented by a large number of scaffolds, we posed a problem of decreasing genome fragmentation and designed an algorithm for scaffold assembly, which is based on gene order and genome rearrangement analysis (*129*). Our algorithm is integrated with the Multiple Genome Rearrangements and Ancestors (MGRA) framework (*130*), which further increases the quality of the resulting assembly.

We extracted gene families from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2). Since our algorithm expects each gene to appear in a single copy in every genome, we filtered the gene families and obtained 6,837 gene families that are uniquely represented in every genome. We remark that this filtration eliminated all genes from some scaffolds and thus we exclude such scaffolds from assembly. Table S19 gives the scaffold statistics for anopheline genomes before and after filtering. We determined the gene order and orientation using the GFF3 annotation from VectorBase (*74*), where each gene is represented by a sequence of coding exons of various lengths. Namely, we defined the coordinate of a gene within a genomic fragment (i.e., chromosome or scaffold) as the mean coordinate of all its coding exons start/end coordinates. We further used these coordinates to represent the fragments as ordered sequences of genes and posed the scaffolds assembly problem as the reconstruction of the global gene order (along genome chromosomes) from the gene sub-orders defined by the scaffolds.

Table S19. Gene scaffolds for breakpoint graph analysis.

Statistics on the number of non-empty scaffolds before and after gene family filtering for improving genome assemblies for anopheline species using breakpoint graphs.

Species	# scaffolds before filtering	# scaffolds after filtering
An. gambiae	6	6
An. arabiensis	340	95
An. quadriannulatus	647	306
An. merus	1078	816
An. dirus	302	124
An. albimanus	57	39

We view gene sub-orders defined by scaffolds as the result of both evolutionary events (i.e., genome rearrangements) and technological fragmentation in the genomes. We notice that technological fragmentation can be modeled by artificial "fissions" that break genomic chromosomes into scaffolds. Scaffold assembly can therefore be reduced to the search for "fusions" that revert technological "fissions" and glue scaffolds back into chromosomes. This observation inspired us to employ the genome rearrangement analysis techniques for scaffolding purposes. Our scaffold assembly algorithm relies on the "standard" tool for rearrangement analysis, called break-point graph (*130*). While traditionally the breakpoint graph is constructed for complete genomes, it can similarly be constructed for fragmented genomes whose scaffolds are treated as "chromosomes". We demonstrated that the breakpoint graph of multiple genomes possesses an important property that its connected components are robust with respect to genome fragmentation and mostly retain information about the complete genomes, even when the breakpoint graph is constructed on their scaffolds. One therefore can utilize connected components of the breakpoint graph for the scaffold assembly of fragmented genomes (*129*). Our scaffold assembly algorithm was further integrated with the MGRA rearrangement analysis framework, which improved the algorithm's sensitivity and resulted in more comprehensive scaffold assembly.

Table S20 reports the number of scaffold assemblies for anopheline genomes obtained by our scaffold assembly algorithm. We compared our assembly results to anopheline genome mapping studies. One study performed analysis of *An. gambiae* and *An. arabiensis* genomes from the same source, where *An. gambiae* represents a complete genome, while *An. arabiensis* is more fragmented. The relationships between these genes and their order on scaffolds were visualized in genoPlotR (*115*) and further compared to the cytogenetic (*4*) and physical (*131*) maps. The *An. gambiae* genome assembly was used as a reference for scaffolding in *An. arabiensis*. Among 10 assemblies in *An. arabiensis* genome identified by our algorithm, the comparison study was able to identify and confirm 6. For example, our algorithm suggested an assembly of the scaffolds KB704562 and KB704374 as well as scaffolds KB704518 and KB704685 in *An. arabiensis* genome, which were confirmed by the comparison study with the gene

reference-based mapping. In addition, scaffolds KB672404 and KB672457 were physically mapped to the polytene X chromosome of *An. albimanus* confirming their adjacency (See Ancestral karyotype analyses, physical mapping). All gluing chains of scaffolds for five anopheline genomes are listed in Table S21.

Table S20. Scaffold assemblies from breakpoint graph analysis.

Statistics on the number of reported scaffold assemblies, both with and without integration with Multiple Genome Rearrangements and Ancestors (MGRA).

Species	Without MGRA	Integrated with MGRA
An. gambiae	0	0
An. arabiensis	6	10
An. quadriannulatus	75	91
An. merus	466	550
An. dirus	30	45
An. albimanus	6	10

Table S21. Inferred scaffold chains.

Inferred scaffold chains (glued ends in order) for five anopheline genomes, with strand indicated by (+), forward, and (-), reverse.

Species	Inferred Scaffold Chains
An. quadriannulatus	KB665644 (+) <==> KB667442 (-) <==> KB667527 (-) <==> KB666955 (-) <==> KB666176 (-)
	KB665666 (+) <==> KB668011 (+)
	KB667555 (-) <==> KB667922 (-)
	KB667600 (+) <==> KB668012 (+)
	KB666254 (+) <==> KB667533 (-) <==> KB665588 (-)
	KB665555 (-) <==> KB666165 (+) <==> KB666522 (+)
	KB667343 (-) <==> KB665400 (-) <==> KB667489 (+) <==> KB667221 (+)
	KB667833 (+) <==> KB667733 (+) <==> KB667877 (+) <==> KB667678 (-)
	KB665488 (-) <==> KB666376 (+)
	KB665810 (+) <==> KB665844 (+) <==> KB666855 (+)
	KB665677 (+) <==> KB666399 (+) <==> KB666465 (-)
	KB667608 (-) <==> KB667544 (-)
	KB666143 (+) <==> KB666365 (-)
	KB666210 (-) <==> KB665411 (+) <==> KB667310 (+)
	KB667778 (+) <==> KB666777 (+)
	KB668111 (+) <==> KB666866 (-) <==> KB667999 (-) <==> KB667433 (+) <==> KB665510 (+)
	KB666666 (-) <==> KB667844 (-) <==> KB667642 (+)
	KB665610 (+) <==> KB666943 (-) <==> KB667543 (+)
	KB668110 (-) <==> KB665511 (-) <==> KB667445 (-)
	KB667010 (-) <==> KB667866 (+) <==> KB667911 (+) <==> KB667811 (-)
	KB666332 (-) <==> KB665444 (+)
	KB665877 (-) <==> KB665422 (+)
	KB668033 (-) <==> KB667722 (+)
	KB667598 (-) <==> KB667210 (-)
	KB667717 (+) <==> KB666788 (-)
	KB668166 (-) <==> KB667900 (-)
	KB667478 (-) <==> KB666066 (-) <==> KB665433 (+)
	KB665633 (+) <==> KB665732 (+) <==> KB667888 (+)
	KB666721 (-) <==> KB667875 (+)
	KB667644 (+) <==> KB668188 (-)
	KB667567 (+) <==> KB666021 (+)
	KB667655 (+) <==> KB666398 (+)
	KB665544 (+) <==> KB666177 (-) <==> KB666732 (+)
	KB666188 (-) <==> KB666610 (+) <==> KB665899 (+)
	KB666677 (+) <==> KB668044 (+)
	KB665710 (+) <==> KB667789 (+)

	KB666509 (-) <==> KB667176 (+) KB667689 (-) <==> KB667667 (-) KB665921 (-) <==> KB665398 (+) KB665466 (-) <==> KB666010 (-) KB666810 (+) <==> KB666354 (+) KB667977 (+) <==> KB667711 (-) KB665744 (+) <==> KB665799 (+) KB667522 (-) <==> KB666088 (+) KB667456 (+) <==> KB668210 (+) KB667065 (+) <==> KB665655 (+) <==> KB665955 (-) KB666065 (-) <==> KB667611 (-) KB666843 (+) <==> KB667556 (+) KB665477 (+) <==> KB666132 (+)
An. arabiensis	KB704685 (-) <==> KB704518 (+) KB705228 (+) <==> KB704740 (+) KB704895 (+) <==> KB704496 (+) KB705151 (+) <==> KB704540 (+) KB704562 (+) <==> KB704374 (+) KB704418 (-) <==> KB704348 (-)
An. merus	KI439120 (+) <==> KI439674 (+) KI439128 (-) <==> KI439034 (-) <==> KI439234 (+) <==> KI439299 (+) <==> KI439372 (-) <==> KI439134 (+) <==> KI439098 (+) <==> KI439413 (-) <==> KI439305 (+) <==> KI439467 (+) <==> KI439338 (-) KI439770 (-) <==> KI439370 (+) <==> KI439563 (+) <==> KI439528 (-) KI439274 (+) <==> KI439140 (-) <==> KI439279 (+) <==> KI438989 (+) <==> KI439015 (+) <==> KI439165 (+) <==> KI439024 (+) <==> KI439561 (+) KI439085 (-) <==> KI438985 (-) <==> KI438990 (+) <==> AXCQ01006373 (+) <==> AXCQ01006807 (+) <==> AXCQ01005936 (+) KI439067 (-) <==> KI439869 (-) <==> KI439124 (-) <==> KI438974 (-) <==> KI439544 (-) KI439316 (-) <==> KI439463 (-) KI439611 (+) <==> KI439111 (-) <==> KI439373 (+) <==> AXCQ01007282 (-) <==> KI439110 (+) AXCQ01007281 (-) <==> KI439253 (+) <==> KI439200 (-) <==> KI439233 (+) <==> KI439351 (-) KI439355 (-) <==> KI439255 (-) KI439432 (+) <==> KI439947 (-) <==> KI439648 (-) <==> AXCQ01011266 (+) <==> KI439839 (+) KI439303 (-) <==> KI439004 (+) <==> AXCQ01006890 (-) <==> KI439367 (+) <==> KI438997 (-) <==> KI439682 (-) <==> KI439330 (+) AXCQ01007707 (+) <==> KI439683 (-) AXCQ01010989 (-) <==> KI439401 (-) KI438981 (-) <==> KI439106 (-) <==> KI439087 (-) <==> KI439041 (-) <==> AXCQ01007467 (-) <==> KI439627 (-) KI439404 (-) <==> KI439282 (-) KI439396 (+) <==> KI439055 (-) <==> KI439938 (+) <==> KI439347 (+) AXCQ01007603 (+) <==> KI438998 (-) <==> KI438991 (-) <==> AXCQ01006738 (+) <==> KI438982 (+) <==> KI439247 (-) <==> KI439332 (-) <==> KI439029 (+) <==> KI439075 (+) <==> KI439417 (+) <==> KI439395 (+) KI439091 (+) <==> KI439159 (-) KI439585 (+) <==> KI439118 (+) <==> KI439062 (+) <==> KI439205 (+) <==> KI439531 (+) KI439037 (+) <==> KI439387 (+) AXCQ01006110 (+) <==> KI439781 (+) KI439471 (-) <==> KI439715 (+) <==> KI439065 (+) KI439190 (-) <==> KI439431 (-) <==> KI439377 (-) <==> KI439545 (+) <==> KI439354 (-) <==> KI439605 (-) KI440294 (+) <==> KI440080 (-) KI439295 (-) <==> KI439517 (-) <==> KI439442 (-) <==> KI439101 (-) <==> KI439202 (+) <==> KI439686 (+) <==> KI439181 (-) KI439012 (+) <==> KI439177 (+) <==> KI439076 (+) KI439039 (+) <==> KI439215 (-) <==> KI439093 (-) <==> KI439430 (+) KI439050 (-) <==> KI439591 (-) <==> KI439156 (+) KI439494 (+) <==> KI439705 (+) <==> KI439855 (-) <==> KI439837 (-) KI439710 (+) <==> KI439604 (+) <==> KI439261 (-) KI439105 (+) <==> KI439196 (+) KI439126 (-) <==> KI438979 (+) <==> KI439153 (+) <==> KI439235 (+) <==> KI439245 (+) KI439713 (-) <==> AXCQ01009552 (-) <==> AXCQ01009212 (+) KI439403 (+) <==> KI439480 (+) KI439056 (-) <==> KI439084 (+) KI439263 (-) <==> KI439412 (-) <==> KI439040 (-) <==> KI439524 (+) <==> KI439549 (-) <==> KI439244 (+) <==> KI439514 (-) <==> KI439331 (-) KI439016 (-) <==> KI439304 (+) KI439660 (+) <==> AXCQ01005798 (+) AXCQ01007662 (+) <==> KI439341 (-) <==> KI439122 (+) <==> KI439639 (+) <==> KI439288 (-) <==> KI439078 (+) <==> KI439382 (+) KI439620 (+) <==> AXCQ01005496 (+) KI439046 (+) <==> KI439379 (-) KI438978 (-) <==> KI439237 (-) <==> KI439097 (-) <==> KI438977 (-)

KI439185 (-) <==> KI439150 (-) <==> KI439166 (-)
KI439366 (+) <==> KI440196 (+) <==> KI439043 (-) <==> KI439139 (-)
KI439314 (-) <==> KI439309 (+)
KI439703 (+) <==> KI439252 (+)
KI439262 (+) <==> KI439053 (+)
KI439113 (+) <==> KI439521 (+) <==> KI439176 (+) <==> KI439021 (-) <==> AXCQ01003868 (+)
KI439302 (+) <==> KI439392 (-) <==> KI439007 (-) <==> KI439066 (-) <==> KI439152 (+) <==> KI439428 (-) <==> KI439081 (-)
AXCQ01006016 (+) <==> AXCQ01007691 (-) <==> AXCQ01008494 (-) <==> KI439137 (-) <==> KI439018 (+) <==> KI439684 (+) <==> KI438994 (-)
KI439475 (-) <==> KI439568 (-)
KI439135 (+) <==> KI439538 (+) <==> KI439116 (+) <==> AXCQ01006997 (+) <==> KI439507 (+) <==> KI439558 (-) <==> KI439221 (-) <==> KI439700 (+) <==> AXCQ01008077 (+)
KI439092 (+) <==> KI439508 (+) <==> KI439163 (-) <==> KI439458 (+) <==> AXCQ01005045 (-) <==> KI439451 (-) <==> KI438986 (-)
KI439182 (-) <==> KI439509 (-) <==> KI439204 (-) <==> KI439623 (-) <==> KI439192 (-)
KI439499 (-) <==> KI439535 (+) <==> KI439358 (-)
KI439588 (+) <==> KI439164 (+)
KI439019 (-) <==> KI439601 (-) <==> KI439290 (-) <==> KI439191 (+) <==> KI439058 (-) <==> KI439465 (-) <==> KI439184 (-)
KI439436 (+) <==> KI439049 (+) <==> AXCQ01006814 (-) <==> KI439147 (-)
KI439297 (+) <==> KI439408 (+) <==> KI439363 (+)
KI439575 (+) <==> KI439525 (-)
AXCQ01007398 (+) <==> KI439115 (+) <==> KI439376 (+) <==> KI439014 (-) <==> KI439173 (+)
KI439144 (-) <==> KI439589 (-) <==> KI439393 (-)
KI439167 (+) <==> KI439281 (+) <==> KI439438 (+) <==> KI439194 (-) <==> KI438996 (-)
KI439423 (-) <==> AXCQ01007369 (-)
KI439615 (-) <==> KI439294 (-) <==> KI439231 (-)
KI439321 (+) <==> KI439609 (-) <==> KI439406 (+) <==> KI439188 (+) <==> KI439488 (+) <==> KI439189 (-) <==> KI439335 (-)
KI439260 (-) <==> KI439522 (-) <==> KI439380 (-) <==> AXCQ01006360 (+) <==> KI439677 (+)
KI439328 (-) <==> KI439044 (+) <==> KI439251 (+) <==> KI439313 (+) <==> KI439170 (-) <==> KI439308 (-) <==> AXCQ01007310 (-) <==> AXCQ01005256 (+) <==> KI439161 (-)
KI439273 (+) <==> KI439219 (+) <==> KI439756 (+) <==> KI439095 (+)
KI439258 (+) <==> KI439008 (-)
KI439178 (+) <==> KI439209 (-)
KI439676 (+) <==> KI439138 (+) <==> KI439154 (+) <==> KI439675 (+) <==> KI439399 (+) <==> KI439462 (+) <==> KI439214 (+) <==> KI438999 (+) <==> KI439352 (+)
AXCQ01008174 (-) <==> KI439339 (+) <==> KI439929 (+) <==> KI439526 (-) <==> AXCQ01010979 (-) <==> KI439434 (+) <==> KI439141 (+) <==> KI439961 (-)
KI439211 (-) <==> KI439285 (+) <==> KI439203 (+) <==> KI439155 (+) <==> KI439083 (+)
KI439239 (-) <==> KI439348 (+)
KI439389 (-) <==> KI439419 (-)
KI439069 (-) <==> KI439688 (-)
KI439548 (+) <==> KI439171 (-) <==> KI439473 (-) <==> KI439074 (+)
KI439550 (+) <==> KI440150 (-)
KI439665 (-) <==> KI439942 (+)
AXCQ01006959 (-) <==> KI439013 (+) <==> KI439148 (+) <==> KI439421 (+) <==> KI439045 (+) <==> KI438973 (+) <==> KI439057 (+) <==> KI439278 (+) <==> KI439006 (+) <==> AXCQ01007472 (+) <==> KI439547 (+) <==> KI439276 (+)
KI439357 (-) <==> AXCQ01006848 (-) <==> KI439653 (-)
KI439038 (+) <==> KI439661 (-)
KI439217 (+) <==> KI439425 (-)
KI439257 (+) <==> KI439284 (+)
KI439378 (-) <==> KI439248 (+) <==> KI439342 (-) <==> KI439198 (-)
KI439275 (+) <==> KI439264 (+) <==> AXCQ01006451 (-) <==> KI439108 (+) <==> KI439746 (+)
KI439238 (+) <==> AXCQ01005666 (-)
KI439022 (+) <==> KI439329 (-) <==> KI439457 (+) <==> KI439673 (-) <==> KI439320 (-) <==> KI439070 (-) <==> KI439424 (+)
KI439172 (+) <==> KI439595 (-) <==> KI439546 (+) <==> KI440216 (+) <==> KI439391 (-) <==> KI439449 (-) <==> KI439259 (-)
KI439796 (-) <==> KI439174 (+) <==> KI438984 (+) <==> KI439270 (-)
KI439291 (-) <==> KI439009 (+)
AXCQ01005866 (-) <==> KI439446 (-) <==> AXCQ01006689 (-) <==> KI439256 (+)
KI439286 (-) <==> AXCQ01006590 (-) <==> KI439515 (-)
KI439090 (+) <==> KI439236 (-)
KI439079 (-) <==> KI439530 (-) <==> KI439453 (-) <==> KI439169 (+)
KI439621 (-) <==> KI439068 (-) <==> KI439361 (-) <==> KI439375 (-)
KI439519 (-) <==> KI438995 (-) <==> AXCQ01009491 (+) <==> KI439503 (-) <==> KI439073 (-)
KI439003 (-) <==> AXCQ01008896 (+)

KI438987 (-) <==> KI438980 (+)
KI439759 (+) <==> KI439643 (-)
KI439349 (+) <==> KI439570 (+)
KI439860 (+) <==> KI439323 (+) <==> KI439435 (+)
KI439344 (+) <==> KI439809 (+)
KI439452 (-) <==> KI439483 (-) <==> KI439311 (+)
KI439315 (+) <==> KI439280 (+)
KI439359 (-) <==> KI439240 (+) <==> AXCQ01005844 (-)
KI439017 (+) <==> KI439618 (+)
KI440257 (-) <==> AXCQ01010930 (+)
KI439761 (-) <==> KI439112 (-)
AXCQ01007424 (-) <==> KI439243 (+) <==> KI439445 (+)
KI439131 (-) <==> KI439107 (-) <==> KI439133 (-) <==> KI439082 (-) <==> KI439582 (-) <==> KI439104 (+) <==>
KI439334 (+) <==> KI439345 (+)
KI439512 (+) <==> KI439157 (-) <==> KI439032 (+) <==> KI439482 (-) <==> KI439096 (+) <==> KI439077 (+) <==>
KI439293 (-)
KI440079 (+) <==> KI439662 (-)
AXCQ01004778 (-) <==> KI439385 (-)
AXCQ01004895 (-) <==> KI439489 (+) <==> KI439574 (+) <==> KI439511 (+) <==> KI439175 (+) <==> KI439485 (+)
<==> KI438992 (-) <==> KI439030 (-) <==> KI439000 (-) <==> KI439130 (+) <==> KI439411 (+) <==> KI438988 (+)
KI439916 (+) <==> KI439967 (-)
KI439529 (+) <==> KI439292 (+)
KI439119 (-) <==> KI439088 (-) <==> KI439186 (-) <==> KI439268 (+)
KI439114 (+) <==> AXCQ01006452 (+)
KI439513 (-) <==> KI439470 (+) <==> KI439229 (-) <==> KI439222 (-) <==> KI439010 (-) <==> KI439216 (-) <==>
AXCQ01007602 (-) <==> KI439064 (+) <==> KI438975 (-) <==> KI439533 (-)
KI439498 (-) <==> KI439109 (-)
AXCQ01010479 (-) <==> AXCQ01009122 (-)
KI439094 (+) <==> KI439265 (+)
KI439207 (-) <==> KI439322 (+) <==> KI439283 (-)
KI439028 (+) <==> KI439518 (-) <==> KI439491 (-)
KI439125 (+) <==> KI439497 (-) <==> KI439633 (-) <==> AXCQ01005899 (-)
KI439414 (-) <==> KI439054 (-) <==> AXCQ01005900 (-)
KI439036 (+) <==> KI439072 (-) <==> KI439657 (-)
KI439052 (-) <==> KI439804 (-)
KI439760 (-) <==> KI439863 (+) <==> KI439450 (-)
KI439593 (-) <==> KI439365 (-)
KI439023 (-) <==> KI439779 (+)
KI439356 (+) <==> AXCQ01006567 (-) <==> KI439129 (+)
KI439272 (-) <==> KI439086 (+)
KI439298 (+) <==> KI439629 (+) <==> KI439310 (+)
KI439136 (-) <==> KI439468 (-) <==> KI439149 (+) <==> KI439386 (-)
KI439089 (-) <==> KI439187 (-)
KI439307 (-) <==> KI439195 (+) <==> AXCQ01005219 (+) <==> AXCQ01010210 (+) <==> KI439346 (+)
KI439566 (+) <==> KI439061 (+) <==> KI439213 (-)
KI440010 (+) <==> KI439460 (-) <==> KI439454 (+) <==> KI439360 (-)
KI439394 (+) <==> KI439031 (+) <==> KI439246 (+) <==> KI439254 (+) <==> KI439647 (+) <==> KI439193 (+)
KI439033 (+) <==> KI439123 (-)
AXCQ01006731 (-) <==> KI439362 (-) <==> KI439668 (-) <==> KI439484 (-)
AXCQ01006958 (-) <==> AXCQ01005946 (+) <==> KI439059 (-)
KI439448 (+) <==> KI439479 (+)
KI439685 (+) <==> KI439383 (+)
AXCQ01006181 (+) <==> KI438993 (+) <==> KI439461 (+)
KI439599 (+) <==> AXCQ01010898 (+)
AXCQ01006176 (-) <==> KI439655 (-) <==> KI439437 (+) <==> KI439327 (+) <==> KI439474 (+) <==>
AXCQ01010344 (-) <==> KI439748 (-) <==> KI439337 (+) <==> KI439520 (+)
KI439266 (+) <==> KI439011 (+) <==> KI439277 (+) <==> KI439142 (-)
KI439201 (+) <==> KI439168 (+) <==> KI439001 (-) <==> KI438976 (+) <==> AXCQ01007397 (+) <==> KI439487 (-)
<==> KI439642 (+) <==> KI439422 (-) <==> KI439060 (+) <==> KI439319 (+) <==> KI439132 (+)
KI439670 (-) <==> KI439407 (+) <==> KI439369 (-) <==> AXCQ01008112 (+)
AXCQ01011450 (+) <==> KI439696 (+) <==> KI439340 (-)
KI439343 (-) <==> KI439602 (-)
KI439301 (+) <==> KI439612 (-)
KI439146 (+) <==> AXCQ01006704 (-) <==> KI439218 (+) <==> AXCQ01007355 (-) <==> KI439296 (-)
KI439047 (-) <==> KI439592 (+) <==> KI439542 (+)
KI439466 (+) <==> KI439230 (-) <==> KI439005 (-)
KI439350 (-) <==> KI439701 (-)
KI439409 (-) <==> KI439427 (-) <==> AXCQ01005945 (-)
KI440119 (+) <==> KI439100 (+)
KI439151 (+) <==> KI439269 (+)

	KI439666 (-) <==> KI439250 (-) <==> KI439210 (-) KI439048 (+) <==> KI439486 (-) <==> KI439318 (-)
An. dirus	KB672868 (+) <==> KB672813 (-) <==> KB672979 (-) KB673202 (-) <==> KB673257 (-) KB673668 (-) <==> KB673557 (-) KB672525 (+) <==> KB673490 (+) KB673090 (-) <==> KB672957 (-) <==> KB673091 (+) KB672891 (+) <==> KB672968 (+) KB673059 (+) <==> KB673346 (+) KB672935 (-) <==> KB672946 (-) KB672802 (+) <==> KB672867 (-) <==> KB673534 (-) KB672835 (-) <==> KB672602 (-) <==> KB672791 (-) <==> KB672902 (-) <==> KB672490 (-) <==> KB673335 (-) <==> KB672896 (+) KB672945 (-) <==> KB672614 (+) KB673103 (+) <==> KB673224 (-) <==> KB673368 (+) KB672841 (+) <==> KB673007 (-) <==> KB672941 (-) <==> KB673075 (-) <==> KB673424 (-) <==> KB672758 (-) KB673057 (+) <==> KB673645 (-) <==> KB673135 (+) KB672848 (-) <==> KB673324 (+) KB673046 (-) <==> KB672491 (-)
An. albimanus	KB672423 (-) <==> KB672417 (-) KB672404 (-) <==> KB672457 (+) KB672287 (+) <==> KB672364 (+) <==> KB672405 (-) KB672409 (-) <==> KB672353 (+) KB672411 (-) <==> KB672408 (+)

Summary of chromosomal evolution analyses

Analysis of synteny showed that the anopheline genomes have conserved gene membership on chromosome arms. Unlike *Drosophila*, arms reshuffle between chromosomes multiple times across the anopheline phylogeny via translocations and do not undergo fission or fusions. The anopheline X chromosome exhibits a significantly higher rate of rearrangement compared to autosomes despite the paucity of polymorphic inversions on the X. The ratio of the X chromosome to the total rearrangement rates is significantly higher in anophelines than in *Drosophila*. The X chromosome also has a significant excess of gene movement to other chromosomes, further underscoring its structurally dynamic nature. The ANGES comparison of the distances between X chromosomes of the anopheline ancestor with *An. atroparvus* and *An. albimanus* suggests that this ancestor is much closer to *An. atroparvus* than to *An. albimanus*. Both ANGES and PATHGROUPS analyses demonstrated a very clear syntenic signal indicating that *An. arabiensis* and *An. quadriannulatus* rather than *An. gambiae* and *An. arabiensis* form a sister clade. The PATHGROUPS comparison of *An. gambiae* with the top-most ancestor showed only very local synteny with no clear chromosome arm conservation for gene order. Both ANGES and PATHGROUPS analyses were impacted by the issue of assembly fragmentation resulting in observing rearrangement on terminal branches, which are in fact artifacts due to assembly breaks. To address the issue of high fragmentation of genome assemblies obtained by NGS, we developed a new Breakpoint Graphs algorithm to improve assemblies. We provide gluing chains of scaffolds for five anopheline genomes including *An. quadriannulatus*, *An. merus*, *An. arabiensis*, *An. dirus* and *An. albimanus*. Physical genome mapping data for *An. arabiensis* and *An. albimanus* confirmed proposed scaffold adjacencies.

Gene families

Gene families are orthologous groups of genes that perform similar functions. The number of genes in a gene family can indicate some aspect of organismal function. Therefore identifying gene families that are rapidly evolving can provide insights into important differences between mosquitoes with different ecologies, behaviors, and physiological traits. We used CAFE v3.0 (*18*), which performs maximum likelihood reconstruction of ancestral states of gene families, to calculate the rate of gene gain and loss (λ) for all 18 anopheline species. This not only gives us the rate of gene family evolution across the phylogeny, but also enables us to identify families that are rapidly evolving along specific lineages.

Input data: ultrametric tree (divergence times) and gene families. To generate an ultrametric tree, peptide alignments of 3,899 single-copy orthologs from 21 *Culicidae* species were used to generate a maximum likelihood tree with RAxML (*91*). After removing the non-anopheline outgroups, the tree was smoothed using the r8s program, which uses a semiparametric, penalized likelihood approach to estimate substitution rates and divergence times across a phylogeny, given at least one calibration time (*132*). We calibrated the times on the tree based on the divergence between *An. gambiae* and *An. darlingi* at 100 million years (*5*) (Figure S10). In order to classify genes from all species into gene families, we obtained peptides for the 18 anopheline species from VectorBase and performed an all-vs-all BLAST (*133*) search on these data. The resulting e-values from the search were used as the main clustering criterion for the MCL program to group peptides into gene families (*134*). This resulted in 23,335 clusters. Of these, all single-peptide clusters were filtered out, leaving 11,636 gene families. With the ultrametric phylogeny and gene family data as input, we estimated gene gain and loss rates (λ) with CAFE v3.0 (*18*).

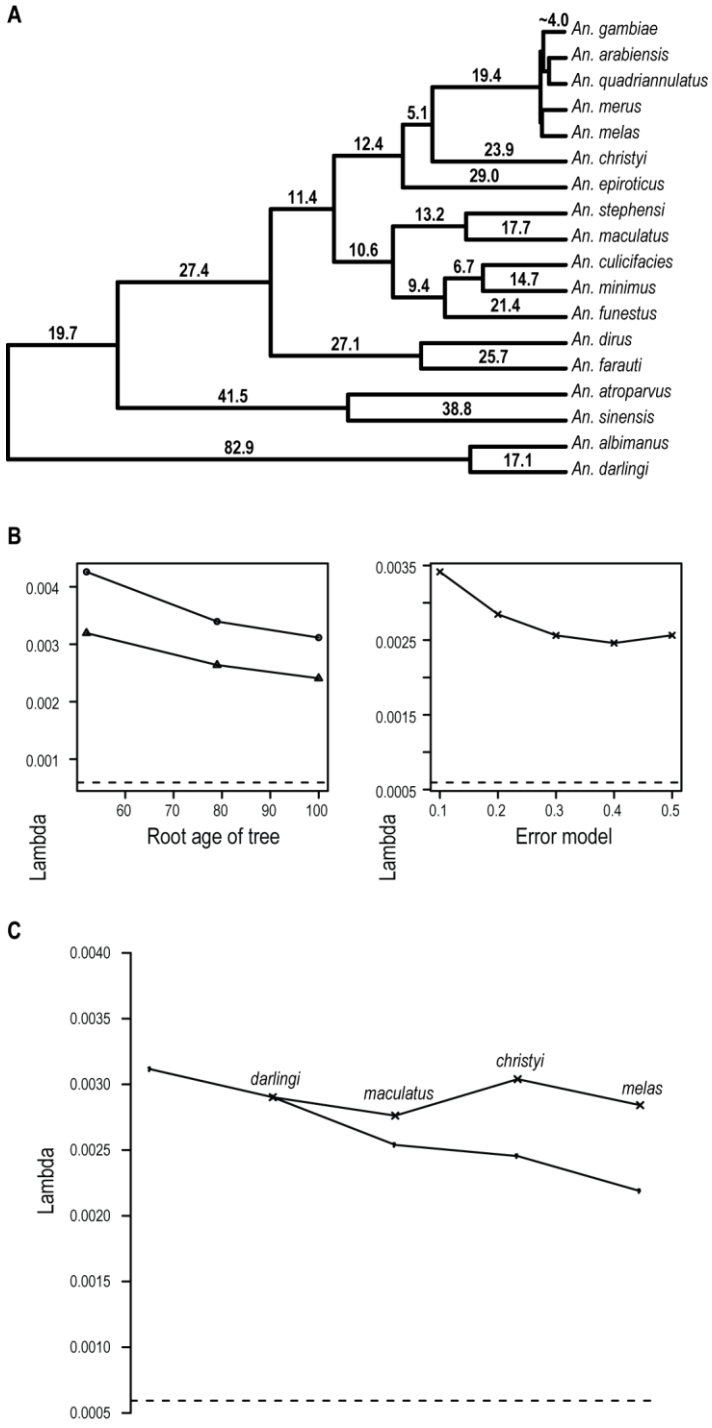


Figure S10. CAFE gene family dynamics.

A. Ultrametric tree with branch lengths in millions of years for the 18 anophelines. Divergence times were estimated using r8s and calibrated on the 100 million years divergence between *An. gambiae* and *An. darlingi*. **B**. Diagnostic tests to ensure the robustness of the anopheline rate of gene gain and loss (λ) with respect to the *Drosophila* rate. Root age of tree: by varying the length of the anopheline tree (circles) or removing the gambiae species complex and varying the length of the tree (triangles), λ never approaches that of the *Drosophila* rate (dashed line). Error model: using realistic estimates of assembly and annotation error with the anopheline data, λ never approaches that of the

Drosophila rate (dashed line). **C**. Diagnostic tests excluding four species with potentially lower-quality annotations, independently removing one species at a time (crosses) or cumulatively removing species from left to right (dots). Again, λ never approaches that of the *Drosophila* rate (dashed line).

CAFE: Computational Analysis of (gene) Family Evolution. With the ultrametric phylogeny and gene family data as input, we estimated gene gain and loss rates (λ) with CAFE v3.0 (*18*). This version of CAFE is able to estimate the amount of assembly and annotation error (ε) present in the input data using a distribution across the observed gene family counts. CAFE is then able to correct for this error and obtain a more accurate estimate of λ. We find an ε of about 0.14, which implies that 14% of gene families have observed counts that are not equal to their true counts. After correcting for this error rate, we find λ = 0.006. This rate is approximately five times higher than the rate measured across 12 *Drosophila* species (*18*) (Table S22).

Table S22. CAFE gene gain/loss rates and error estimates.

Assembly/Annotation error estimation and gene gain/loss rates in 18 anopheline species and 12 *Drosophila* species.

	λ (No Error Model)	**ε** (Estimated)	**λ** (Error Model = **ε**)
Anopheles	0.00602	0.14307	0.00312
Drosophila	0.00121	0.04102	0.00059

To ensure the robustness of the anopheline gene gain and loss rate with respect to the much lower *Drosophila* rate, we ran three diagnostic tests. First, to make sure the difference in rates was not simply due to the larger phylogeny available for the anopheline data, we pruned the tree to make it more similar to that of *Drosophila*, and ran CAFE using six variations on the anopheline phylogeny. First, we used the full phylogeny (18 species, root age 100 mya). Next we removed the species *An. darlingi* and *An. albimanus* which results in a tree with 16 species and a root age of 79 mya. And finally we removed the species *An. darlingi*, *An. albimanus*, *An. sinensis*, and *An. atroparvus* resulting in a tree with 14 species and a root age of 52 mya. Lambda values of these three phylogenies (Figure S10, circles) were compared with the full *Drosophila* phylogeny of 12 species with a root age of 64 mya (Figure S10, dotted line). We also performed three trials in which we removed the *gambiae* species complex (*An. gambiae*, *An. quadriannulatus*, *An. arabiensis*, *An. merus*, and *An. melas*) in addition to the species mentioned above. This results in lambda values from trees containing 13, 11, and 9 species with root ages of 100, 79, and 52 mya, respectively (Figure S10, triangles). Using these varying phylogenies the rate of gene gain and loss for anophelines never approaches that of the *Drosophila* rate. To ensure that CAFE was not underestimating the error in the anopheline data (and therefore overestimating λ) we re-ran CAFE while varying the values of the error model distribution. This tells us if it is possible to measure

a rate as low as that in *Drosophila* while correcting for realistic amounts of error. Figure S10 shows that regardless of the amount of error used in the anopheline data, the rate still never gets as low as the *Drosophila* rate. Finally, we re-ran the analyses excluding four species for which assessments of assemblies and gene sets suggested lower quality annotations. Figure S10 shows that while removing species with potentially lower quality annotations does reduce the estimated rate of gene gain and loss, it still never gets as low as the Drosophila rate. These tests give us confidence in the elevated rate of gene gain and loss observed across the anophelines.

With respect to gene family evolution across specific lineages, Table S23 summarizes these results. Column 1 denotes the extant species of the lineage. Columns 2 through 6 show the number of families and genes that were gained, lost, or did not change along that lineage. Column 7 reports the average number of expansions per gene family. A negative average expansion indicates that families in this species are, on average, contracting in size. Columns 8 through 10 show the number of families that show significant changes in gene family size (p-value < 0.01) and these are families that are rapidly evolving. Column 11 shows the number of families lost in that lineage with respect to the nearest sister species. Gene families in most species are, on average, contracting in size (Table S23, column 7). *An. melas* has the most expansions per family (0.158058) while *An. christyi* has the most contractions (-0.323335). Again *An. melas* has the most rapidly evolving families (381) with expansions vastly outnumbering contractions at 362 to 19, respectively. It should be noted that *An. melas* has a rather fragmented assembly so the rates for this species may be erroneously inflated. We have also annotated rapidly-evolving families using the most common InterPro domain per family for each species. Some interesting examples are shown in Table S24.

Table S23. CAFE gene family results.

Summary of CAFE results for individual lineages across the anopheline phylogeny.

Species	Expan-sions	Genes Gained	No Change	Contr-actions	Genes Lost	Avg. Expan-sion	Sig. Expan-sions	Sig. Contr-actions	Total Sig. Changes	Families lost
An. albimanus	283	342	10469	883	998	-0.056382	15	34	49	250
An. arabiensis	233	325	10852	550	606	-0.024151	90	29	119	115
An. atroparvus	435	585	10080	1120	1184	-0.051483	15	7	22	296
An. christyi	362	399	7403	3870	4161	-0.323335	5	65	70	897
An. culicifacies	1170	1582	9717	748	827	0.064890	111	27	138	407
An. darlingi	541	665	9342	1752	1897	-0.105887	20	68	88	1180
An. dirus	353	524	9709	1573	1687	-0.099957	23	11	34	576
An. epiroticus	374	543	8300	2961	3045	-0.215041	19	7	26	165
An. farauti	535	723	10060	1040	1147	-0.036442	26	20	46	355
An. funestus	364	602	9708	1563	1662	-0.091104	41	23	64	468
An. gambiae	733	1324	9372	1530	1680	-0.030597	210	75	285	710
An. maculatus	2135	3530	7869	1631	1768	0.151440	242	52	294	1306
An. melas	2073	2907	8523	1039	1068	0.158058	362	19	381	385
An. merus	288	381	10489	858	915	-0.045896	104	23	127	240
An. minimus	234	271	9933	1468	1626	-0.116459	10	36	46	164
An. quadriannulatus	271	304	10817	547	596	-0.025097	75	29	104	123
An. sinensis	1052	1542	9540	1043	1059	0.041513	64	2	66	278
An. stephensi	287	360	10336	1012	1146	-0.067555	27	33	60	345

Table S24. Dynamic gene families.

Interesting examples of rapidly evolving gene families from CAFE analysis.

CAFE ID	Species	Type of change	InterPro Annotation (ID)
36	*An. albimanus*	Expansion	Neurotransmitter-gated ion-channel (IPR006201)
36	*An. darlingi*	Contraction	Neurotransmitter-gated ion-channel (IPR006201)
68	*An. arabiensis*	Expansion	Insect cuticle protein (IPR000618)
131	*An. christyi*	Contraction	Insect pheromone/odorant binding protein PhBP (IPR006625)
189	*An. christyi*	Contraction	Gustatory receptor (IPR009318)
76	*An. gambiae*	Expansion	Heat shock protein 70 family (IPR013126)
134	*An. gambiae*	Expansion	Olfactory receptor, Drosophila (IPR004117)
470	*An. gambiae*	Expansion	Insulin family (IPR022352)
191	*An. gambiae*	Contraction	Frizzled domain (IPR020067)
30	*An. melas*	Expansion	ABC transporter type 1, transmembrane domain (IPR011527)

Introns

To examine the evolutionary histories of intron gains and losses across the anopheline phylogeny and compare them to those in other insects, orthologous protein-coding genes were selected from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2) as delineated by the OrthoDB (*66*) methodology. A total of 32 species were selected for the analysis: 4 outgroup insects, *Pediculus humanus*, *Apis mellifera*, *Tribolium castaneum*, and *Danaus plexippus*; 12 *Drosophila*, *D. grimshawi*, *D. mojavensis*, *D. virilis*, *D. willistoni*, *D. persimilis*, *D. pseudoobscura*, *D. ananassae*, *D. erecta*, *D. yakuba*, *D. melanogaster*, *D. sechellia*, and *D. simulans*; 2 culicine mosquitoes, *Aedes aegypti*, and *Culex quinquefasciatus*; and 14 anophelines, *An. darlingi*, *An. albimanus*, *An. sinensis*, *An. atroparvus*, *An. farauti*, *An. dirus*, *An. funestus*, *An. minimus*, *An. culicifacies*, *An. maculatus*, *An. stephensi* (SDA-500), *An. epiroticus*, *An. christyi*, and *An. gambiae* (PEST).

The phylogeny of the selected species was determined from the concatenated protein sequence alignments (using MUSCLE (*76*) followed by alignment trimming with trimAl (*90*)) of 1,085 relaxed single-copy orthologs (a maximum of 3 paralogs allowed in no more than 5 species, longest protein selected) across all 43 species included in OrthoDBmoz2. RAxML (*91*) with the PROTGAMMAJTT model was used with the resulting 720,022 amino acid columns (with 526,151 distinct alignment patterns) to estimate the maximum likelihood species phylogeny. A total of 5,871 orthologous groups were selected for intron analyses according to the following criteria: single-copy orthologs in at least 25 species, missing from no more than 7 species, and duplications in no more than 7 species. Where duplications were present, the longest protein was

selected for analysis. Orthologous protein sequences were aligned using MUSCLE (*76*) and formatted to include intron position and length data from General Feature Format files using custom Perl scripts. The formatted alignments were loaded into MALIN (*135*) for maximum likelihood analyses of intron evolutionary histories.

A total of 58,823 informative intron sites from the ortholog alignments were selected using MALIN (*135*) requiring a minimum of 5 non-gap positions on both sides of the intron site, and maximum of 7 ambiguous sties (missing alignment data for no more than 7 species). Intron gain and loss rates were optimized using default MALIN (*135*) parameters and estimates of intron presence, gains, and losses across the phylogeny were computed using the posterior probabilities method with 10 bootstrap analyses.

Examining the evolutionary histories of intron gains and losses across the anopheline phylogeny and comparing them to those in other insects reveals more intron losses in anophelines compared to drosophilids. The analysis recovers the known dramatic losses of introns in the dipteran ancestor since the divergence from Lepidoptera, and furthermore shows continued high losses in the *Drosophilidae* and the *Culicidae* ancestors. Although total intron gain and loss events along the *An. gambiae* and *Dr. melanogaster* lineages since their last common ancestor are similar, anopheline orthologs have lost relatively more introns, leading to the lower numbers of introns observed in the genes of extant species (Figure S11 and Table S25).

Intron loss through retrotransposition: orthology analysis revealed 922 groups of paralogs, where in the same species at least one multi-exon and one single-exon paralog of similar protein length are present. Of such groups, 185 groups were found occurring in at both fly and mosquito species, suggesting an ancient origin. Furthermore, 189 groups were found exclusively in the fly lineage (present in two or more species); whereas 161 groups were present exclusively in the mosquito lineage. Finally, species-specific groups were identified, of which 159 were found in flies and 228 groups were found in mosquitoes.

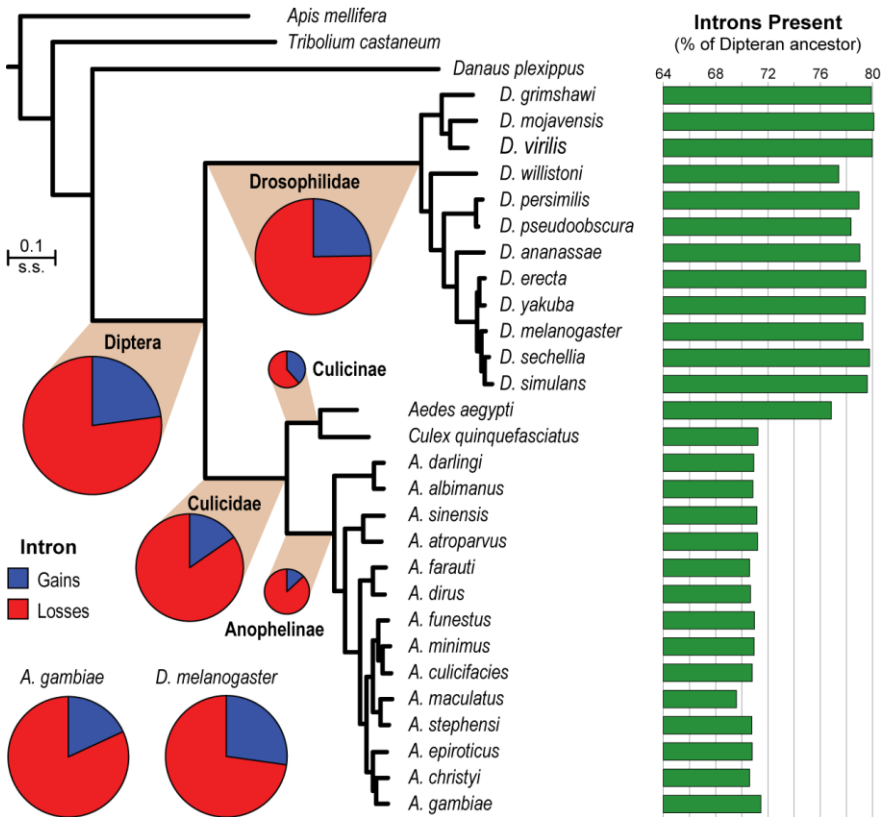


Figure S11. Intron evolution.

Intron evolution in mosquitoes and fruit flies. The proportions of estimated intron gains (blue) and losses (red) from orthologous genes are shown in the pie charts for major branches of the maximum likelihood species phylogeny. The dramatic losses of introns in the dipteran ancestor since the divergence from Lepidoptera (represented by *Danaus plexippus*) continue in the *Drosophilidae* and the *Culicidae* ancestors. The *An. gambiae* and *D. melanogaster* pie charts (bottom left) show the total intron gain and loss events along the mosquito and fruit fly lineages since their common ancestor. The areas of all pie charts are proportional to the total number of estimated gain and loss events. Although the anophelines and drosophilids show similar total numbers of events, anopheline genes have lost relatively more introns, leading to the lower numbers of introns present in the genes of extant species (green bars, right).

Table S25. Intron evolution.

Intron presence, gain, and loss estimates in mosquitoes and fruit flies. The proportions of estimated introns present, gained, and lost from orthologous genes for each ancestral node and extant species are indicated as a percentage of the total estimated intron count in the dipteran ancestor. SD: standard deviation computed from 10 bootstrap samples.

Species/Node	Present	SD	Gain	SD	Loss	SD
A. gambiae	71.45	0.48	1.55	0.12	0.67	0.06
A. gambiae + A. christyi	70.57	0.44	0.03	0.01	0.06	0.01
A. christyi	70.60	0.46	0.41	0.05	0.38	0.05
Pyretophorus	70.60	0.45	0.01	0.01	0.12	0.01
A. epiroticus	70.80	0.46	0.60	0.07	0.40	0.06
Pyretophorus + Neocellia + Myzomyia	70.71	0.46	0.07	0.03	0.01	0.01
A. stephensi	70.78	0.47	0.40	0.08	0.11	0.03
Neocellia	70.48	0.45	0.07	0.03	0.26	0.05
A. maculatus	69.59	0.37	0.49	0.03	1.39	0.18
Neocellia + Myzomyia	70.68	0.46	0.05	0.01	0.08	0.02
A. culicifacies	70.80	0.42	0.65	0.07	0.50	0.03

Species/Node	Present	SD	Gain	SD	Loss	SD
A. culicifacies + A. minimus	70.65	0.44	0.01	0.01	0.07	0.02
A. minimus	70.95	0.44	0.42	0.06	0.12	0.02
Myzomyia	70.71	0.46	0.05	0.01	0.02	0.02
A. funestus	70.98	0.46	0.44	0.05	0.17	0.04
Cellia	70.64	0.46	0.13	0.03	0.28	0.03
A. dirus	70.66	0.41	0.57	0.03	0.21	0.04
Neomyzomyia	70.31	0.42	0.08	0.02	0.41	0.05
A. farauti	70.60	0.39	0.59	0.05	0.29	0.04
Cellia + Anopheles	70.79	0.45	0.04	0.03	0.06	0.03
A. atroparvus	71.20	0.43	0.76	0.08	0.33	0.07
Anopheles	70.78	0.43	0.08	0.03	0.10	0.03
A. sinensis	71.16	0.42	1.01	0.09	0.61	0.07
Anophelinae	70.80	0.42	0.82	0.09	5.42	0.25
A. albimanus	70.84	0.43	0.89	0.07	0.32	0.05
Nyssorhynchus	70.27	0.43	0.33	0.05	0.86	0.06
A. darlingi	70.91	0.33	1.03	0.13	0.39	0.06
Culicidae	75.40	0.38	5.46	0.40	30.05	0.59
Culex quinquefasciatus	71.24	0.53	3.99	0.18	7.20	0.37
Culicinae	74.44	0.36	1.60	0.18	2.56	0.18
Aedes aegypti	76.85	0.39	3.53	0.17	1.12	0.16
Diptera	100.00	1.07	13.46	0.48	45.32	0.77
D. simulans	79.60	0.46	0.97	0.07	0.76	0.07
D. simulans + D. sechellia	79.40	0.43	0.07	0.03	0.05	0.03
D. sechellia	79.77	0.41	0.70	0.06	0.34	0.04
D. simulans + D. sechellia + D. melanogaster	79.38	0.42	0.03	0.02	0.09	0.03
D. melanogaster	79.27	0.46	0.16	0.03	0.26	0.04
Melanogaster Subgroup	79.43	0.39	0.44	0.07	0.29	0.05
D. yakuba	79.44	0.41	0.15	0.04	0.19	0.03
D. yakuba + D. erecta	79.47	0.40	0.07	0.02	0.03	0.02
D. erecta	79.49	0.41	0.14	0.03	0.13	0.05
Melanogaster Group	79.28	0.45	0.54	0.04	0.26	0.03
D. ananassae	79.03	0.48	0.86	0.10	1.11	0.12
Melanogaster + Obscura	78.99	0.41	0.30	0.05	0.44	0.03
D. pseudoobscura	78.36	0.48	0.37	0.06	0.27	0.06
D. pseudoobscura + D. persimilis	78.25	0.50	0.43	0.08	1.17	0.07
D. persimilis	78.98	0.49	1.31	0.12	0.57	0.03
Sophophora	79.13	0.43	0.68	0.15	0.66	0.09
D. willistoni	77.42	0.41	1.57	0.11	3.27	0.09
Drosophila	79.10	0.34	10.23	0.48	31.13	0.65
D. virilis	79.99	0.29	0.41	0.06	0.21	0.06
D. virilis + D. mojavensis	79.78	0.32	0.11	0.05	0.09	0.02
D. mojavensis	80.11	0.34	0.63	0.06	0.31	0.03
D. virilis + D. mojavensis + D. grimshawi	79.76	0.32	1.32	0.11	0.66	0.11
D. grimshawi	79.91	0.30	0.50	0.06	0.36	0.06

Gene fusions and fissions

A total of 21 species were selected for this analysis: 19 anophelines, *An. gambiae* (PEST), *An. merus*, *An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. christyi*, *An. epiroticus*, *An. stephensi* (INDIAN), *An. stephensi* (SDA-500), *An. maculatus*, *An. culicifacies*, *An. minimus*, *An. funestus*, *An. dirus*, *An. farauti*, *An. atroparvus*, *An. sinensis*, *An. albimanus*, *An. darlingi*; and 2 culicine mosquitoes, *Culex quinquefasciatus* and *Aedes aegypti*. The longest protein per gene was selected and all-against-all sequence alignments were performed with SWIPE (*136*) using default parameters. The SWIPE hits were filtered to retain hits with e-value ≤ 1e-5, percent identity ≥ 60%, and either aligned query length or aligned subject length ≥ 50 amino acids. Fused and fragmented gene candidates were then identified through non-transitive hits in each pairwise species comparison: a potential fusion/fragmentation event was identified when a single gene in

species A mapped to two or more genes in species B if the genes in species B did not map to each other (or vice versa).

To distinguish between likely annotation errors and possible true fusion or fission events the set of fused and fragmented gene candidates were partitioned according to their genomic locations. Candidates with a fragment at the end of a scaffold or with any two fragments that were immediate neighbors on the same scaffold were considered likely annotation errors, while the remaining candidates were considered possible true gene fusion or fission events.

To compare against fusion and fission rates in flies, we selected a total of 12 fly species: *Drosophila melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*, and ran the same analysis.

Estimates of annotation errors through analysis of gene fusions and fissions:

The identification of gene fusion or fission events that have created novel gene architectures can be achieved by comparing the alignments of homologous proteins to identify cases where two or more genes align contiguously along the length of a single homolog (*137*). However, automated gene annotation pipelines may confound the search for true fusion or fission events by erroneously fusing (merging) or fragmenting (splitting) gene models, especially when annotating tandem arrays of homologs or when assemblies are relatively fragmented. Quantification of potential fusion or fission events can therefore highlight the relative levels of potential annotation errors that lead to incorrectly merged or split gene models.

On average, across the anopheline genomes, 24.9% [34.9%] of genes per genome were detected as fused [fragmented] candidates, compared to 19.5% [33.2%] of genes in *An. gambiae* (Figure S12A). Note that any fused/fragmented candidate is based on a pairwise comparison where gene model(s) in species A compared to gene model(s) in species B suggest a conflict that can be resolved either by multiple gene models merging (in one of the two species) or a gene model splitting (in one of the two species). Thus, a single fusion or fission event contributes **both** fused and fragmented candidates.

To identify genomes of potentially poor quality, we compared the levels of fusion and fission across individual genomes to detect "outlier" genomes, that is, those with either an unusual number of fusions or fissions (Figure S12B). Unsurprisingly, these outlier genomes tended to (i) have highly fragmented assemblies, (ii) have a higher than usual percentage of fragmented genes, and (iii) have a lower than usual percentage of fused genes.

Next, we estimated the percentage of genes per genome that are detected as fused [fragmented] candidates and could be attributable to poor genome assembly or gene annotation. On average, across the anopheline genomes, 19.2% [25.3%] of genes per

genome could be erroneously fused [fragmented] candidates, compared to 13.0% [20.9%] in *An. gambiae*.

Finally, for a more direct estimate, we treated the *An. gambiae* genome annotations as "correct" and compared genomes against only *An. gambiae*. Using this error estimate, across the anopheline genomes, the average percentage of erroneously fused [fragmented] gene models is 3.3% [9.7%]. Fusions, i.e. the gene was associated with multiple genes in *An. gambiae*, and the genes in *An. gambiae* are neighbors on the same scaffold: mean, 3.3%; median, 4.1%, and std. dev., 1.7%. Fissions, i.e. multiple genes were associated with a single gene in *An. gambiae*, and (a) the non-*An. gambiae* genes are neighbors on the same scaffold: mean, 5.6; median, 5.7%, and std. dev., 2.4% or (b) at least one non-*An. gambiae* gene is at the end of its scaffold: mean, 4.4; median, 0.7%, and std. dev., 8.3%.

Searches for true gene fusions and fissions:

After accounting for possible errors due to fragmented assemblies and/or erroneous gene annotations, on average, across the anopheline genomes, 5.7% [9.6%] of genes per genome were detected as fused [fragmented] candidates. For comparison, the same analysis of 12 fly genomes estimated 2.8% [4.7%] of genes per genome were as fused [fragmented] candidates.

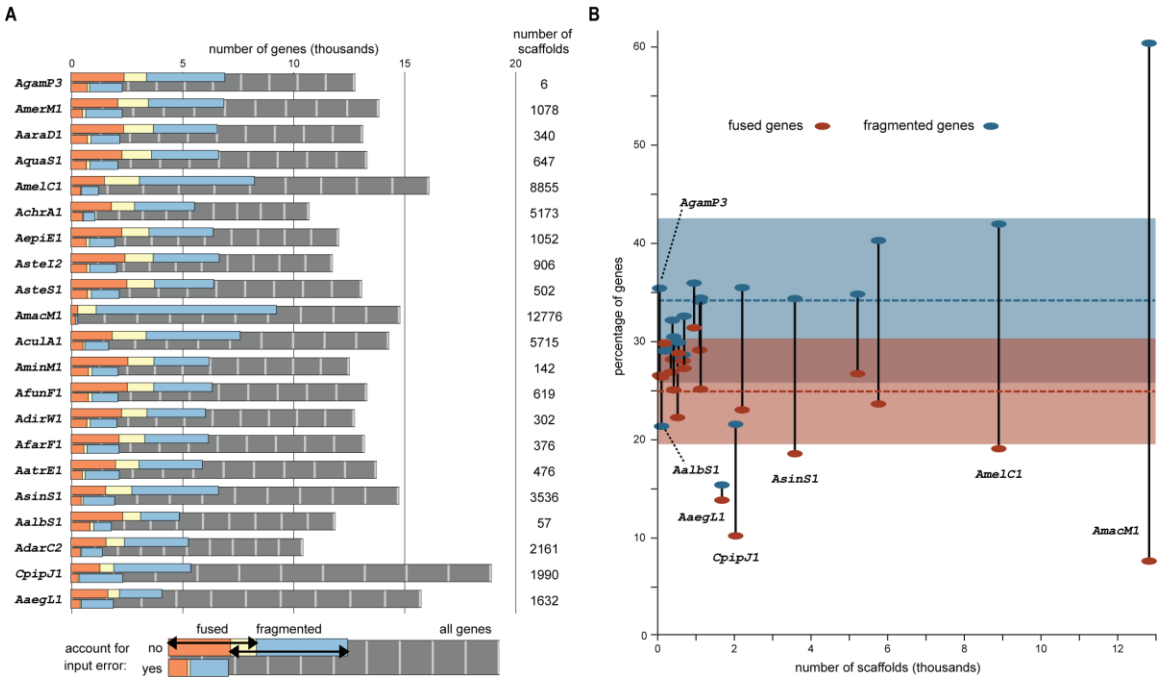


Figure S12. Gene fission and fusion.

A. For each genome, the total number of genes, the number of candidate fused and fragmented genes, and the numbers of scaffolds are shown. **B**. The percentage of genes per genome that are candidate fused or fragmented genes is plotted against the number of scaffolds. Horizontal lines and bars depict the mean ± one standard deviation of

the percentage across all anopheline genomes. *An. gambiae* (PEST) and species whose number of fused genes or number of fragmented genes are outliers are annotated. Species: AgamP3, *An. gambiae* (PEST); AmerM1, *An. merus*; AaraD1, *An. arabiensis*; AquaS1, *An. quadriannulatus*; AmelC1, *An. melas*; AchrA1, *An. christyi*; AepiE1, *An. epiroticus*; AsteI2, *An. stephensi* (INDIAN); AsteS1, *An. stephensi* (SDA-500); AmacM1, *An. maculatus*; AculA1, *An. culicifacies*; AminM1, *An. minimus*; AfunF1, *An. funestus*; AdirW1, *An. dirus*; AfarF1, *An. farauti*; AatrE1, *An. atroparvus*; AsinS1, *An. sinensis*; AalbS1, *An. albimanus*; AdarC2, *An. darlingi*; CpipJ1, *Culex quinquefasciatus*; AaegL1, *Aedes aegypti*.

Stop-codon readthrough

We used the 21-way whole genome alignments to find evolutionary evidence of functional translational stop-codon readthrough in hundreds of *Anopheles* genes. Specifically, we found 325 annotated *An. gambiae* PEST transcripts ("readthrough candidates") for which the region between the annotated stop-codon and the subsequent in-frame stop-codon showed an evolutionary signature specific to protein-coding regions, as measured by PhyloCSF (*138*), and for which readthrough was a more likely explanation than alternative splicing, dicistronic translation, or a recent nonsense substitution (Figure S13). A similar approach had been used previously to find 283 readthrough candidates in *Drosophila melanogaster*, many of which have been experimentally verified (*139-142*). No species other than *Drosophila* has been found to have such an extensive catalogue of stop-codon readthrough genes, as similar searches in *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and Human had found only a handful of readthrough candidates in those species, and our findings offer the first confirmation of a previous prediction based on k-mer statistics that insect and crustacean species have hundreds of readthrough genes but other metazoans have considerably fewer (*140*).

Figure S13. Stop-codon readthrough.

Alignment of the readthrough region of AGAP006444-RA, one of the *An. gambiae* readthrough candidates, color coded by CodAlignView. The high concentration of synonymous substitutions (light green) and conservative amino acid changes (dark green), and lack of radical amino acid changes (red) and frame shifted regions (orange) in the 17

codons between the annotated stop-codon and the next in-frame stop-codon is characteristic of protein-coding regions. The region's evolutionary coding potential, as measured by PhyloCSF (*138*), is 1,600 times more likely to occur in a coding region than a non-coding region, implying that it has been functional at the amino acid level in much of the anopheline tree. The perfectly conserved TGA-C stop-codon context, which occurs in roughly one third of the readthrough candidates, is known to promote inefficient termination.

Our *An. gambiae* stop-codon readthrough candidates (Table S26) display many of the unusual properties previously found to distinguish the *Drosophila* readthrough candidates, as a class, from other *Drosophila* transcripts (*140*), demonstrating that these properties are not unique to *Drosophila*. Among the readthrough candidates the distribution of the four-base context consisting of the stop-codon and subsequent base, which is known to affect termination efficiency, is almost the reverse of the distribution among other transcripts, with TGA-C, the least common among other transcripts, present in about one third of the readthrough candidates. Substitutions between stop-codons are extremely rare among the readthrough candidates: among transcripts for which a majority of the anopheline tree has an aligned stop-codon, over 97% of the readthrough candidates have the same stop-codon in all species that have an aligned stop-codon versus 30% among non-readthrough candidates, perhaps because the three stop-codons encode different amino acids when read through or modulate the readthrough rate. We applied RNAz (*79*) to find that there is a predicted conserved RNA structure in the 100nt regions 3' of the stop-codon of 31 (10%) of the readthrough candidates compared to <1% of other transcripts. Such a structure has been found to trigger readthrough in the *Drosophila hdc* gene (*143*).

Table S26. Stop-codon readthrough candidates.

List of *An. gambiae* stop-codon readthrough candidates with transcript IDs, genomic locations, lengths, and presence/absence of predicted RNA structures.

Transcript	RT intervals	strand	RT length (codons)	Stop Codon Context	RNA structure
AGAP001773-RC	chr2R:9941060-9941098	-	13	TAA-C	no
AGAP001773-RA	chr2R:9784181-9784219	-	13	TAA-C	no
AGAP001774-RA	chr2R:10055589-10055627	+	13	TAA-C	no
AGAP012179-RA	chr3L:38353459-38353635	-	59	TAA-T	no
AGAP009931-RA	chr3R:45438147-45438236	+	30	TAG-G	no
AGAP012186-RA	chr3L:38528175-38528249	+	25	TGA-A	no
AGAP004823-RA	chr2L:3915084-3915260	-	59	TGA-C	no
AGAP007008-RA	chr2L:40792437-40792493	-	19	TAA-C	no
AGAP009274-RA	chr3R:30516752-30516784	+	11	TGA-G	no
AGAP009138-RA	chr3R:27106050-27106151	+	34	TGA-C	no
AGAP000018-RA	chrX:251811-251849	+	13	TGA-C	no
AGAP000190-RA	chrX:3158052-3158675	+	208	TGA-A	yes
AGAP004331-RA	chr2R:54678854-54679381	+	176	TGA-C	no
AGAP003428-RA	chr2R:37671309-37672910	+	534	TGA-A	no
AGAP002793-RA	chr2R:27628493-27630691	+	733	TGA-C	no
AGAP007739-RA	chr3R:169997-170089	-	31	TGA-G	no
AGAP001232-RA	chr2R:1747346-1747453	+	36	TAG-C	no
AGAP013495-RA	chrX:17555750-17555926	-	59	TGA-G	no
AGAP006942-RA	chr2L:40187649-40187747	+	33	TAG-A	no

AGAP000019-RA	chrX:288874-289038	+	55	TGA-C	no
AGAP010201-RA	chr3R:50614016-50614348	-	111	TGA-T	no
AGAP006346-RA	chr2L:30128728-30128757	-	10	TAA-C	no
AGAP006496-RA	chr2L:33023261-33023290	-	10	TGA-G	no
AGAP002566-RA	chr2R:22970562-22970567	+	2	TAG-C	no
AGAP008574-RA	chr3R:13114868-13115767	+	300	TGA-C	yes
AGAP005980-RA	chr2L:24447517-24447639	+	41	TGA-C	yes
AGAP003140-RA	chr2R:33179679-33179864	+	62	TAG-T	no
AGAP004989-RA	chr2L:7555596-7555634	-	13	TAA-T	no
AGAP013461-RA	chr2R:57866592-57866612	+	7	TAA-G	no
AGAP006382-RA	chr2L:31014469-31014492	+	8	TGA-G	yes
AGAP011956-RA	chr3L:35532242-35532316	-	25	TAA-T	no
AGAP000123-RA	chrX:2014166-2014255	+	30	TGA-C	no
AGAP004767-RA	chr2L:3193876-3194034	-	53	TGA-T	no
AGAP028130-RA	chr3L:35742536-35742856	-	107	TGA-C	no
AGAP005229-RA	chr2L:12549161-12549283	-	41	TAG-C	no
AGAP001536-RA	chr2R:6034372-6034392	+	7	TGA-C	no
AGAP006800-RA	chr2L:38659388-38659471	+	28	TAG-G	no
AGAP003349-RB	chr2R:36361975-36362133	-	53	TGA-G	no
AGAP005034-RA	chr2L:8643962-8644042	-	27	TGA-T	no
AGAP005127-RA	chr2L:10332744-10332854	-	37	TAG-G	no
AGAP009158-RA	chr3R:27712443-27712553	-	37	TAG-C	no
AGAP000314-RA	chrX:5511709-5514036	-	776	TGA-C	no
AGAP008803-RA	chr3R:18057920-18058549	-	210	TAG-C	no
AGAP002233-RA	chr2R:17959130-17959162	+	11	TGA-G	no
AGAP000962-RA	chrX:18426303-18426677	-	125	TAA-C	no
AGAP010184-RA	chr3R:49983850-49983894	-	15	TGA-G	yes
AGAP004143-RA	chr2R:50706905-50706916	-	4	TGA-G	no
AGAP002115-RA	chr2R:15784352-15784369	-	6	TAG-G	yes
AGAP005356-RA	chr2L:14607017-14607028	+	4	TGA-T	yes
AGAP003798-RA	chr2R:43442832-43442879	-	16	TAG-G	yes
AGAP000351-RB	chrX:6188277-6188297	-	7	TGA-T	no
AGAP005898-RA	chr2L:23430018-23430041	+	8	TGA-C	no
AGAP002544-RA	chr2R:22660169-22660204	+	12	TGA-C	no
AGAP002924-RB	chr2R:29636881-29636898	-	6	TGA-G	no
AGAP005633-RA	chr2L:18044209-18044226	+	6	TGA-G	no
AGAP011985-RA	chr3L:35881111-35881137	-	9	TGA-C	no
AGAP008077-RA	chr3R:5111185-5111214	-	10	TGA-G	no
AGAP009952-RA	chr3R:46166464-46166757	+	98	TAA-A	no
AGAP004664-RA	chr2R:60410736-60411038	-	101	TGA-C	no
AGAP006537-RA	chr2L:34024345-34024404	+	20	TGA-C	no
AGAP002886-RA	chr2R:28811748-28811801	+	18	TAA-C	no
AGAP000457-RA	chrX:7930983-7931087	-	35	TGA-T	no
AGAP027996-RA	chr3R:21544689-21544862	-	58	TGA-C	no
AGAP005408-RA	chr2L:15189310-15189324	-	5	TAG-C	no
AGAP008656-RA	chr3R:14277282-14278340	-	353	TGA-T	no
AGAP004990-RA	chr2L:7586855-7586872	+	6	TGA-G	no
AGAP005011-RA	chr2L:8136323-8136340	+	6	TGA-C	no
AGAP001671-RA	chr2R:7740963-7740983	-	7	TGA-C	no
AGAP000446-RA	chrX:7843486-7843671	+	62	TAG-T	no
AGAP010769-RA	chr3L:10104006-10104050	+	15	TGA-C	no
AGAP008588-RA	chr3R:13330756-13330839	-	28	TAG-C	no
AGAP005729-RA	chr2L:19791460-19791507	-	16	TGA-G	no
AGAP000502-RA	chrX:8847274-8847777	+	168	TAG-C	no
AGAP004613-RA	chr2R:58410911-58410919	-	3	TGA-C	no
AGAP012085-RA	chr3L:37482772-37482900	-	43	TGA-C	no
AGAP003318-RA	chr2R:35928047-35928355	+	103	TAG-G	no
AGAP005878-RA	chr2L:23058812-23059411	-	200	TAG-C	no
AGAP010355-RA	chr3L:2111321-2111365	+	15	TAA-T	no
AGAP010474-RA	chr3L:4170963-4170968	+	2	TAG-C	no
AGAP003457-RA	chr2R:37886773-37886802	-	10	TGA-T	no
AGAP010357-RA	chr3L:2139147-2139203	+	19	TGA-T	no
AGAP002794-RA	chr2R:27630692-27630781	+	30	TAG-G	no
AGAP006884-RA	chr2L:39348566-39348640	-	25	TAG-C	no
AGAP002942-RA	chr2R:29985541-29985558	+	6	TGA-T	no
AGAP011883-RA	chr3L:34626090-34626140	+	17	TAG-C	no
AGAP010910-RA	chr3L:12896168-12896257	-	30	TAA-C	no
AGAP005088-RA	chr2L:9877087-9877629	+	181	TGA-T	no
AGAP005641-RA	chr2L:18131953-18132153	-	67	TAG-G	no
AGAP007373-RA	chr2L:46136701-46136712	-	4	TGA-C	no

AGAP004504-RA	chr2R:57149060-57149104	+	15	TGA-T	no
AGAP012987-RA	chr2R:36943257-36943694	-	146	TGA-T	no
AGAP007992-RA	chr3R:3850496-3850537	-	14	TGA-C	no
AGAP006653-RA	chr2L:35709219-35709263	+	15	TGA-T	no
AGAP007333-RA	chr2L:45633849-45633893	+	15	TAG-C	no
AGAP002719-RA	chr2R:26227734-26227784	+	17	TAG-C	no
AGAP004229-RA	chr2R:52465842-52465895	+	18	TGA-C	no
AGAP001310-RA	chr2R:2840583-2840855	-	91	TGA-T	no
AGAP000063-RA	chrX:980032-980088	-	19	TAA-T	no
AGAP009950-RA	chr3R:46104981-46105031	+	17	TAA-C	no
AGAP011254-RA	chr3L:20103910-20103924	-	5	TGA-C	no
AGAP008312-RA	chr3R:8862100-8862222	-	41	TGA-G	no
AGAP004273-RC	chr2R:53538235-53538291	-	19	TAA-A	no
AGAP005245-RK	chr2L:12980093-12980200	+	36	TGA-C	no
AGAP001773-RB	chr2R:9856446-9856484	-	13	TAA-C	no
AGAP006536-RA	chr2L:34016198-34016263	+	22	TAA-C	no
AGAP006943-RA	chr2L:40190483-40190503	-	7	TAG-C	no
AGAP003651-RA	chr2R:41005365-41005520	-	52	TGA-C	no
AGAP004453-RA	chr2R:56483560-56483697	-	46	TGA-G	no
AGAP011115-RA	chr3L:17227495-17227566	+	24	TGA-G	yes
AGAP010460-RA	chr3L:3929149-3929175	-	9	TAA-T	no
AGAP010390-RA	chr3L:2482826-2483236	-	137	TGA-C	no
AGAP008533-RA	chr3R:11963025-11963132	-	36	TGA-C	no
AGAP000532-RA	chrX:9504470-9504553	+	28	TGA-T	no
AGAP013174-RA	chr2R:20643044-20643124	-	27	TGA-T	no
AGAP007847-RA	chr3R:1975559-1975612	-	18	TGA-C	no
AGAP009177-RA	chr3R:28152886-28152939	-	18	TGA-G	no
AGAP004311-RA	chr2R:54375941-54376051	+	37	TGA-C	no
AGAP008518-RA	chr3R:11748110-11748154	+	15	TGA-C	no
AGAP009446-RA	chr3R:33539325-33539495	-	57	TGA-T	no
AGAP010442-RA	chr3L:3726304-3726438	+	45	TGA-C	no
AGAP010585-RA	chr3L:6544883-6544918	+	12	TGA-T	no
AGAP006590-RA	chr2L:34294403-34294519	-	39	TAG-C	yes
AGAP002578-RA	chr2R:23163838-23163894	-	19	TAG-C	no
AGAP005567-RA	chr2L:17316196-17316225	+	10	TAG-C	no
AGAP006762-RA	chr2L:38011942-38011986	+	15	TGA-C	no
AGAP013455-RA	chr2R:42781056-42781136	+	27	TAG-C	no
AGAP001535-RA	chr2R:5955494-5955550	+	19	TGA-G	no
AGAP004982-RA	chr2L:7516006-7516056	-	17	TGA-G	no
AGAP009286-RA	chr3R:30818488-30818616	-	43	TGA-C	no
AGAP001225-RA	chr2R:1642915-1642947	-	11	TGA-T	no
AGAP001337-RA	chr2R:3311393-3311416	+	8	TGA-C	no
AGAP011745-RA	chr3L:32840087-32840158	-	24	TGA-T	no
AGAP009180-RA	chr3R:28180139-28180180	+	14	TGA-C	no
AGAP001867-RA	chr2R:11698495-11698521	+	9	TGA-G	no
AGAP005099-RA	chr2L:10113494-10113511	+	6	TGA-C	no
AGAP007106-RA	chr2L:42800422-42800487	-	22	TGA-T	no
AGAP004178-RA	chr2R:51246246-51246269	+	8	TGA-G	no
AGAP013018-RA	chr2R:21598347-21598625	-	93	TGA-C	no
AGAP007807-RA	chr3R:1276294-1276347	+	18	TGA-G	no
AGAP009255-RA	chr3R:29969907-29970029	+	41	TGA-G	no
AGAP011055-RA	chr3L:15794843-15794878	-	12	TGA-G	no
AGAP000185-RA	chrX:3032207-3032329	-	41	TGA-C	no
AGAP003797-RA	chr2R:43418854-43418913	+	20	TGA-C	no
AGAP003997-RA	chr2R:47730437-47730625	+	63	TGA-C	no
AGAP010206-RA	chr3R:50702199-50702258	-	20	TAA-C	no
AGAP006059-RA	chr2L:25801746-25801799	-	18	TGA-C	yes
AGAP006475-RA	chr2L:32681036-32681089	-	18	TGA-T	no
AGAP010752-RA	chr3L:9824704-9824754	+	17	TGA-G	no
AGAP008989-RA	chr3R:22774005-22774037	+	11	TGA-C	no
AGAP005440-RA	chr2L:15477310-15477390	-	27	TAG-G	no
AGAP001367-RA	chr2R:3631556-3631645	+	30	TAG-A	no
AGAP000874-RA	chrX:16435202-16435219	+	6	TGA-A	no
AGAP001592-RA	chr2R:6527222-6527323	-	34	TAG-C	no
AGAP010388-RA	chr3L:2456283-2456303	-	7	TAG-C	no
AGAP002702-RA	chr2R:25774788-25774826	+	13	TGA-C	no
AGAP002824-RA	chr2R:28036209-28036226	+	6	TGA-C	no
AGAP001683-RA	chr2R:8247379-8247459	-	27	TGA-C	no
AGAP007720-RA	chr2L:49268946-49268993	-	16	TGA-T	no
AGAP003821-RA	chr2R:43768864-43768902	+	13	TGA-G	no

AGAP012118-RA	chr3L:37832620-37832679	+	20	TGA-C	no
AGAP007865-RA	chr3R:2366328-2366375	+	16	TGA-C	yes
AGAP000080-RA	chrX:1308428-1308484	-	19	TGA-T	no
AGAP002671-RA	chr2R:25287525-25287599	+	25	TGA-T	no
AGAP004270-RA	chr2R:53509137-53509169	-	11	TAG-C	no
AGAP005258-RA	chr2L:13331898-13331993	+	32	TGA-A	no
AGAP000672-RA	chrX:11856370-11856459	-	30	TGA-T	no
AGAP012018-RA	chr3L:36404211-36404249	+	13	TAG-G	no
AGAP007108-RA	chr2L:42829316-42829432	+	39	TGA-G	no
AGAP003593-RA	chr2R:40372811-40372840	-	10	TAG-C	no
AGAP003814-RA	chr2R:43727759-43727785	+	9	TGA-C	no
AGAP009123-RA	chr3R:26217302-26217370	-	23	TGA-C	no
AGAP001255-RA	chr2R:2052545-2052607	+	21	TAG-A	no
AGAP004120-RA	chr2R:50320359-50320397	-	13	TGA-C	no
AGAP011694-RA	chr3L:31484371-31484400	+	10	TGA-C	no
AGAP005599-RA	chr2L:17739676-17739729	+	18	TAG-A	no
AGAP001762-RA	chr2R:9493076-9493096	+	7	TAG-C	no
AGAP011916-RA	chr3L:34979635-34979676	-	14	TAG-C	yes
AGAP008252-RA	chr3R:7902562-7902576	-	5	TGA-T	no
AGAP000078-RA	chrX:1264682-1264726	-	15	TGA-C	no
AGAP006606-RA	chr2L:34687987-34688055	-	23	TGA-C	no
AGAP007046-RA	chr2L:41624643-41624687	-	15	TGA-T	yes
AGAP001947-RA	chr2R:12726312-12727415	-	368	TGA-T	no
AGAP008980-RA	chr3R:22547540-22547689	-	50	TAG-C	no
AGAP001413-RA	chr2R:4558488-4558523	+	12	TAG-C	yes
AGAP010067-RA	chr3R:48569145-48569279	-	45	TGA-C	yes
AGAP000606-RA	chrX:11101429-11101548	+	40	TGA-T	no
AGAP003892-RA	chr2R:45624525-45624569	+	15	TAG-G	yes
AGAP013022-RA	chrX:14833564-14833662	-	33	TAG-C	no
AGAP001961-RA	chr2R:12909643-12909768	+	42	TAG-C	no
AGAP005716-RA	chr2L:19567192-19567224	+	11	TGA-C	no
AGAP005074-RA	chr2L:9616140-9616298	+	53	TGA-T	no
AGAP002296-RA	chr2R:18702388-18703200	-	271	TAA-C	no
AGAP000932-RA	chrX:17601897-17602019	-	41	TAG-T	no
AGAP006454-RA	chr2L:32305731-32305862	-	44	TAA-G	no
AGAP002913-RA	chr2R:29228405-29228611	-	69	TAG-C	no
AGAP006173-RA	chr2L:27653072-27653122	-	17	TGA-T	no
AGAP013145-RB	chr2R:8660687-8660944	+	86	TGA-C	no
AGAP003059-RA	chr2R:31793043-31794464	+	474	TAA-C	no
AGAP002993-RA	chr2R:30668705-30668926	-	74	TGA-C	yes
AGAP007302-RA	chr2L:45100319-45100594	-	92	TGA-C	no
AGAP003658-RA	chr2R:41282916-41282984	+	23	TGA-G	no
AGAP002672-RA	chr2R:25293460-25293762	-	101	TAG-C	no
AGAP011378-RA	chr3L:22704946-22705914	+	323	TAG-C	no
AGAP002709-RA	chr2R:26001206-26001232	-	9	TAG-C	no
AGAP011935-RA	chr3L:35264366-35264497	-	44	TGA-G	no
AGAP002951-RA	chr2R:30068042-30068095	-	18	TAG-A	yes
AGAP005393-RA	chr2L:14909287-14909379	+	31	TAA-G	no
AGAP011988-RF	chr3L:35903071-35903190 +chr3L:35905437-35905475	-	53	TGA-C	no
AGAP004434-RA	chr2R:55991372-55991863	-	164	TGA-T	no
AGAP027982-RA	chr3L:1333593-1333667	-	25	TGA-C	no
AGAP004401-RA	chr2R:55677241-55677363	-	41	TAA-C	no
AGAP006528-RA	chr2L:33920822-33921007	+	62	TGA-T	yes
AGAP007585-RA	chr2L:47839470-47839529	-	20	TGA-T	no
AGAP011097-RA	chr3L:16798569-16798640	-	24	TGA-C	yes
AGAP003849-RA	chr2R:44403149-44403226	-	26	TAG-C	no
AGAP005737-RA	chr2L:20024721-20025203	+	161	TAG-C	no
AGAP007765-RA	chr3R:498376-498480	-	35	TAA-G	no
AGAP003572-RA	chr2R:39982953-39983264	+	104	TGA-C	no
AGAP006242-RA	chr2L:28588454-28588546	+	31	TGA-C	no
AGAP008379-RA	chr3R:10123654-10123701	-	16	TAG-C	no
AGAP003202-RA	chr2R:33806716-33806763	-	16	TAG-C	no
AGAP003201-RA	chr2R:33788769-33788816	-	16	TAG-C	no
AGAP000058-RA	chrX:932546-932590	-	15	TAG-G	no
AGAP002610-RA	chr2R:24057081-24057143	+	21	TGA-C	no
AGAP004106-RA	chr2R:49935653-49935787	+	45	TAG-C	no
AGAP000456-RA	chrX:7926878-7926913	+	12	TAG-G	no
AGAP004494-RA	chr2R:57042215-57042265	+	17	TAG-G	no
AGAP008814-RA	chr3R:18284293-18284325	+	11	TAG-C	no

AGAP006133-RA	chr2L:27049091-27049135	-	15	TAG-T	yes
AGAP007646-RA	chr2L:48665958-48666050	-	31	TAA-T	yes
AGAP004403-RA	chr2R:55712143-55712274	-	44	TAA-C	no
AGAP010776-RA	chr3L:10179669-10179755	-	29	TAG-C	no
AGAP008221-RA	chr3R:7044275-7044307	+	11	TGA-A	no
AGAP007712-RA	chr2L:49190952-49190978	+	9	TGA-T	no
AGAP005139-RA	chr2L:10463982-10464002	-	7	TGA-C	no
AGAP028090-RA	chr3R:22315062-22315112	-	17	TGA-C	no
AGAP006444-RA	chr2L:32185812-32185865	+	18	TGA-C	yes
AGAP000801-RA	chrX:14618183-14618284	-	34	TAG-G	no
AGAP001806-RA	chr2R:10784790-10784825	+	12	TGA-C	no
AGAP012026-RA	chr3L:36647338-36647424	+	29	TAG-C	no
AGAP010599-RA	chr3L:7068611-7068655	+	15	TAG-C	no
AGAP007657-RA	chr2L:48814615-48814650	+	12	TGA-G	yes
AGAP012115-RA	chr3L:37774109-37774150	-	14	TGA-C	no
AGAP005916-RA	chr2L:23649311-23649403	-	31	TGA-T	no
AGAP002991-RA	chr2R:30637456-30637491	+	12	TGA-T	yes
AGAP003709-RA	chr2R:42280985-42281032	+	16	TGA-C	no
AGAP006474-RA	chr2L:32674673-32674695 +chr2L:32674763-32674853	+	38	TGA-C	no
AGAP012090-RA	chr3L:37578054-37578055 +chr3L:37578140-37578296	+	53	TAA-T	no
AGAP006403-RA	chr2L:31292789-31292809	+	7	TGA-G	no
AGAP002666-RA	chr2R:25260341-25260358	-	6	TGA-C	no
AGAP008811-RA	chr3R:18133763-18133960	-	66	TGA-T	no
AGAP004119-RA	chr2R:50309018-50309095	+	26	TGA-C	no
AGAP007367-RA	chr2L:46108937-46108984	+	16	TGA-G	no
AGAP009888-RA	chr3R:44753879-44753917	+	13	TGA-A	no
AGAP005684-RA	chr2L:18745920-18745949	+	10	TGA-C	no
AGAP007984-RA	chr3R:3670908-3670931	+	8	TGA-C	no
AGAP007754-RA	chr3R:273312-273323	-	4	TGA-C	no
AGAP002344-RA	chr2R:20488300-20488338	-	13	TGA-T	no
AGAP008977-RA	chr3R:22451967-22451987	+	7	TGA-T	no
AGAP001175-RA	chr2R:722285-722368	+	28	TGA-C	no
AGAP012325-RA	chr3L:40221503-40221550	+	16	TGA-T	no
AGAP005245-RJ	chr2L:12955042-12955503	+	154	TGA-G	no
AGAP000070-RA	chrX:1099833-1100126	-	98	TGA-C	no
AGAP005831-RA	chr2L:22279174-22279212	-	13	TGA-G	no
AGAP003631-RA	chr2R:40725832-40725963	-	44	TGA-A	no
AGAP009970-RA	chr3R:46921073-46921138	+	22	TAG-G	no
AGAP009677-RA	chr3R:38314792-38314899	+	36	TAG-C	no
AGAP000250-RA	chrX:4599585-4599626	+	14	TAG-C	no
AGAP010260-RA	chr3R:51755482-51755583	+	34	TGA-G	no
AGAP009852-RA	chr3R:44347912-44347926	-	5	TGA-A	no
AGAP000045-RA	chrX:735719-736168	+	150	TGA-C	no
AGAP007522-RA	chr2L:47143842-47144129	+	96	TGA-C	no
AGAP007485-RA	chr2L:46858480-46858530	+	17	TGA-T	no
AGAP012372-RA	chr3L:41151031-41151045	-	5	TGA-T	yes
AGAP008228-RA	chr3R:7176499-7176531	-	11	TGA-C	no
AGAP007846-RA	chr3R:1866207-1866227	-	7	TGA-G	no
AGAP012748-RA	chrUNKN:26692377-26692421	+	15	TAG-G	yes
AGAP011563-RA	chr3L:27773188-27773265	-	26	TGA-T	no
AGAP003614-RA	chr2R:40515332-40515337	-	2	TGA-C	no
AGAP009925-RA	chr3R:45390740-45390862	+	41	TGA-T	no
AGAP000758-RA	chrX:13768303-13768359	-	19	TGA-C	no
AGAP002366-RA	chr2R:20677673-20677708	+	12	TGA-A	no
AGAP003479-RA	chr2R:38407148-38407168	+	7	TGA-C	no
AGAP007512-RA	chr2L:47106986-47107009	+	8	TGA-G	no
AGAP013747-RA	chr3R:4021843-4021890	-	16	TGA-C	yes
AGAP002927-RA	chr2R:29692374-29692694	-	107	TGA-C	no
AGAP002030-RA	chr2R:14096583-14096924	+	114	TAG-T	no
AGAP009333-RA	chr3R:31516548-31516880	-	111	TGA-G	no
AGAP002567-RA	chr2R:22996997-22997011	+	5	TAG-C	no
AGAP012288-RA	chr3L:39796269-39796412	-	48	TGA-C	no
AGAP008166-RA	chr3R:6296518-6296583	-	22	TGA-T	no
AGAP004522-RA	chr2R:57348661-57348729	+	23	TGA-C	no
AGAP003312-RA	chr2R:35775171-35775458	-	96	TAG-T	no
AGAP006776-RA	chr2L:38134171-38134185	-	5	TGA-C	no
AGAP013210-RA	chrX:8255651-8255689	+	13	TGA-G	no
AGAP002591-RA	chr2R:23656657-23656725	+	23	TGA-G	no

AGAP002896-RA	chr2R:28997033-28997200	-	56	TAG-G	no
AGAP005901-RB	chr2L:23482905-23484056	+	384	TGA-C	no
AGAP012237-RA	chr3L:39209810-39209947	-	46	TGA-G	no
AGAP008241-RA	chr3R:7590725-7591279	+	185	TAG-C	no
AGAP012976-RA	chrX:9600985-9601194	-	70	TGA-C	no
AGAP000045-RB	chrX:727796-728263	+	156	TGA-C	no
AGAP011271-RA	chr3L:20419140-20419304	+	55	TGA-C	no
AGAP008928-RA	chr3R:21099032-21099106	-	25	TGA-A	no
AGAP001606-RA	chr2R:6687806-6687889	+	28	TGA-T	no
AGAP000038-RA	chrX:481577-481591	-	5	TGA-C	no
AGAP007686-RA	chr2L:49055349-49055381	-	11	TGA-C	no
AGAP007548-RA	chr2L:47479920-47479988	-	23	TAA-C	no
AGAP010138-RA	chr3R:49345247-49345279	+	11	TAG-G	yes
AGAP011996-RA	chr3L:36028887-36029093	+	69	TGA-C	no
AGAP007623-RA	chr2L:48362787-48362804	-	6	TGA-T	no
AGAP002974-RA	chr2R:30478829-30478909	+	27	TAG-C	yes
AGAP004659-RA	chr2R:59732409-59732438	-	10	TAG-G	no
AGAP008826-RA	chr3R:18934239-18934280	-	14	TGA-C	no
AGAP007080-RA	chr2L:42229998-42230021	-	8	TGA-T	no
AGAP011133-RA	chr3L:17694780-17694812	+	11	TGA-G	no
AGAP005063-RA	chr2L:9167112-9167153	-	14	TGA-C	no
AGAP004707-RA	chr2L:2431618-2431671	+	18	TGA-G	no
AGAP010867-RA	chr3L:12159045-12159086	-	14	TGA-C	no
AGAP009932-RA	chr3R:45442251-45442322	-	24	TAG-C	no
AGAP012565-RA	chrUNKN:16625393-16625887	+	165	TGA-C	no
AGAP005327-RA	chr2L:14115470-14115487	+	6	TGA-C	no
AGAP009664-RA	chr3R:37779693-37779743	-	17	TGA-C	no
AGAP007886-RA	chr3R:2472714-2472737	-	8	TGA-G	no
AGAP005116-RA	chr2L:10240160-10240381	+	74	TGA-C	no

Selection

Selective constraints on gene sequence evolution were estimated using the dN/dS statistic calculated for orthologous group multiple sequence alignments spanning all of the sequenced genomes or just those species belonging to the *An. gambiae* species complex. Protein sequence multiple alignments were generated first using MUSCLE (*76*), and then used to inform CDS alignments with the codon-aware PAL2NAL alignment program (*144*). Ambiguous regions were removed from the CDS alignments using TrimAl (*90*), with the gap tolerance parameter set to 0.8 to exclude alignment columns with missing data in 20% or more of the sequences in the orthologous group. Sequences were removed from the CDS alignments if more than 40% of their length was gap characters following multiple alignment. Phylogenetic trees were constructed for each orthologous group using RAxML (*91*) with a GTR+Gamma model of evolution. PAML v4.7 (*20*) was used to calculate dN/dS ratios for each aligned orthologous group using the corresponding phylogenetic trees (codeml model=0, NSsites=0, ncatG=1).

Gene functional and evolutionary traits

Gene functional annotations and their orthologous group evolutionary properties from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2, see orthology section for details) were interrogated to investigate relationships among gene functional properties and gene evolutionary traits across the anophelines. *An. gambiae* genes were categorized into

functional classes according to their Gene Ontology (GO) and InterPro domain annotations: using GO-slim terms for biological processes and molecular functions, and selecting the InterPro domains with the highest numbers of annotated *An. gambiae* genes. From VectorBase, a total of 8,727 *An. gambiae* genes were annotated with a total of 2,247 GO terms (71,065 gene-term pairs). In order to make the GO annotations richer, the VectorBase annotations were supplemented with GO terms from *Drosophila melanogaster* orthologs: only FlyBase GO terms for orthologs from Dipteran OrthoDBmoz2 orthologous groups with single-copy orthologs in at least 30 of 33 species (a total of 4,652 orthologous groups). This increased the total annotations to 9,432 *An. gambiae* genes with 5,529 GO terms (94,264 gene-term pairs).

Anopheline orthologous group evolutionary traits examined included: a measure of amino acid conservation/divergence - evolutionary rate, a measure of selective pressure - dN/dS, a measure of gene duplicability - mean gene copy-number, and a measure of ortholog universality - number of species. Evolutionary rates are computed for each orthologous group as the average of inter-species identities normalized to the average identity of all inter-species best reciprocal hits, computed from pairwise Smith-Waterman alignments of protein sequences. The dN/dS ratios - the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) - were computed with PAML (*20*) on filtered alignments of orthologs; described in detail in the selection section above. The mean number of genes per orthologous group is calculated simply as the total number of genes divided by the total number of species present - a value of 1 indicates no duplications and values of >1 quantify the relative number of paralogs. The total number of species is a simple count and ranges from 2 species (lineage-restricted or multiple losses) to the full set of 19 anophelines (universally maintained). Similar measures of evolutionary traits have been successfully used to explore gene evolutionary dynamics across arthropods, vertebrates, and fungi (*145*).

It should be noted that assignment of GO terms is usually biased towards slower-evolving, well-conserved genes as assignments are often made through homology/orthology to genes with annotated terms from *Drosophila melanogaster*. The histograms in Figure S14 indicate how the subset of annotated orthologous groups is generally skewed to less dynamic values of the evolutionary traits analyzed; this effect is less pronounced for the InterPro annotations (Figure S14). Categories with the fastest evolutionary rates and highest dN/dS ratios include ‘neurological system process’, ‘DNA metabolic process’, ‘nuclease activity’, ‘peptidase activity’, ‘signal transducer activity’, ‘7TM chemoreceptor’, and ‘olfactory receptor, *Drosophila*’. Categories with multi-copy orthologous groups (high duplicability) include ‘protein complex assembly’, ‘chromosome organization’, ‘helicase activity’, ‘peptidase activity’, ‘insect cuticle protein’, and ‘fibrinogen, alpha/beta/gamma chain, C-terminal globular domain’. Categories of the most lineage-specific orthologous groups (or with the most losses)

include ‘cofactor metabolic process’, ‘protein folding’, ‘peptidase activity’, ‘structural molecule activity’, ‘fibrinogen, alpha/beta/gamma chain, C-terminal globular domain’, and ‘insect cuticle protein’.

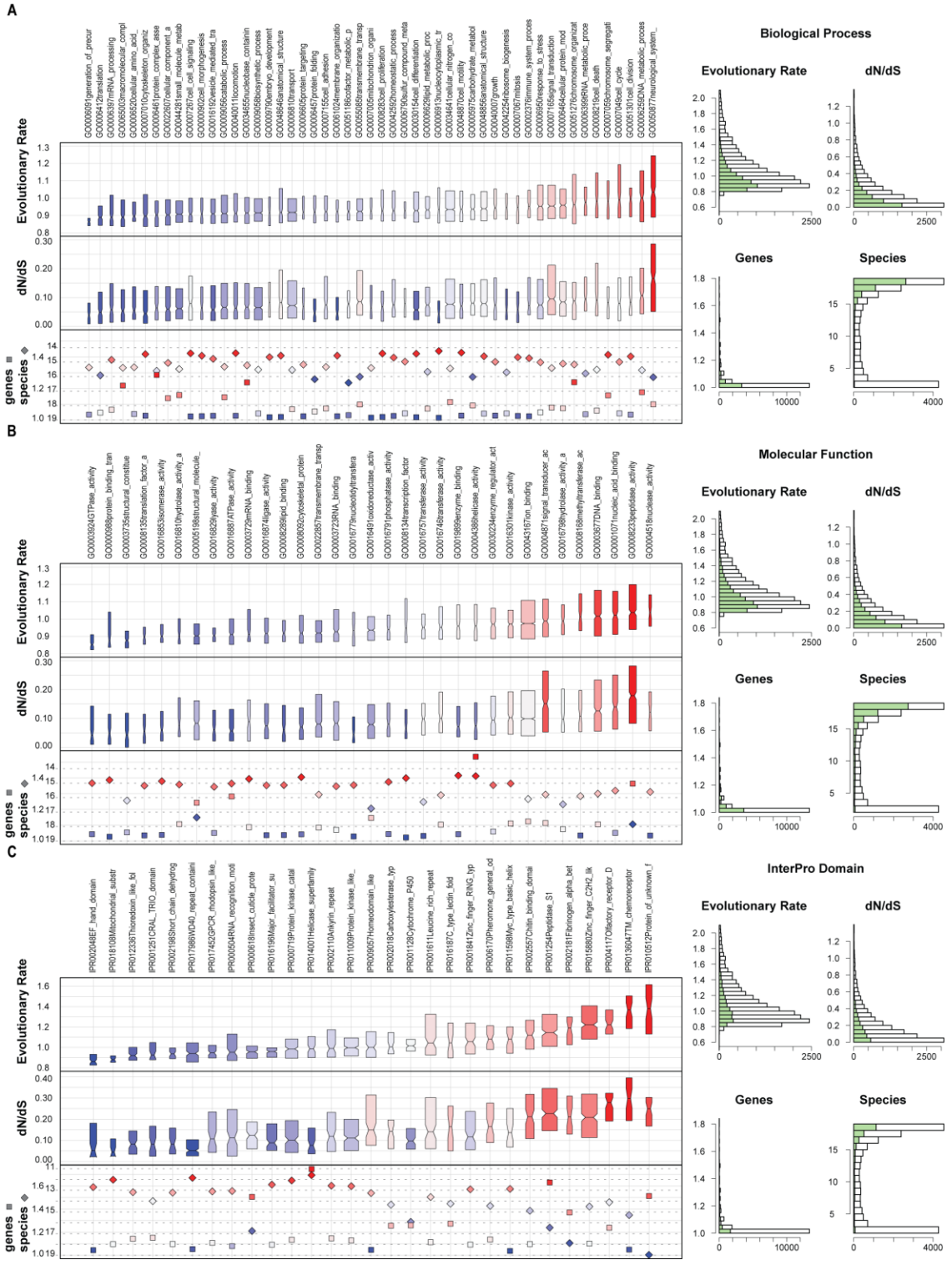


Figure S14. Evolutionary and functional traits.

Evolutionary traits of genes categorized by Gene Ontology (GO-slim) **A**. biological processes, **B**. molecular functions and **C**. the most abundant InterPro domains. Categories are sorted by evolutionary rate from the most conservative (left) to the most dynamic (right) and colored from the highest values (red) to the median value (gray) to the lowest values (blue). Notched boxes show medians of orthologous group values with the limits of the upper and lower quartiles, and box widths are proportional to the number of orthologous groups in each category. Histograms show value distributions for all orthologous groups and for orthologous groups with genes with associated biological process GO terms or InterPro domains (green).

Although the fastest evolving genes are generally less likely to have any functional annotations (as discussed above), comparing the top enriched functional categories in the slowest and fastest subsets of genes can complement the GO-slim and InterPro analyses described above. Orthologous groups with evolutionary rates and dN/dS ratios less than the 20th percentile or greater than the 80th percentile were selected to represent the lowest and highest gene sets, respectively (Figure S15).

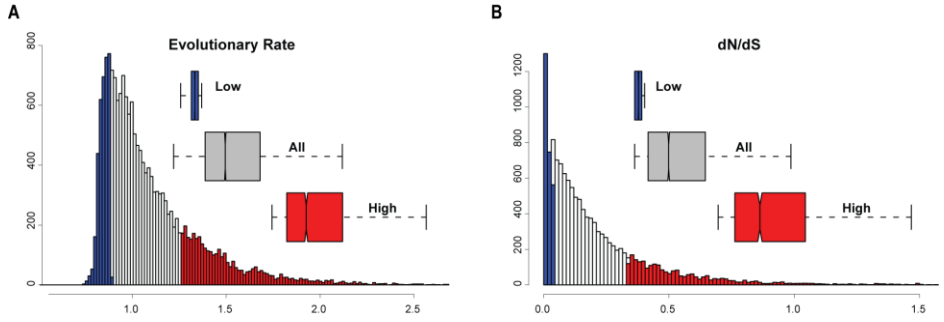


Figure S15. Evolutionary rate and dN/dS ratio distributions.

A. Distribution of orthologous group evolutionary rates (measured in terms of protein sequence divergence) highlighting those less than the 20th percentile or greater than the 80th percentile. **B**. Distribution of orthologous group dN/dS ratios highlighting those less than the 20th percentile or greater than the 80th percentile.

Enrichment tests on GO Biological Processes and Molecular Functions were performed using Bioconductor's GOstats hypergeometric test (*146*) and with the topGO (*147*) implementations of the classic Fisher and the weighted fisher tests (Tables S27 and S28). The background gene sets in each case were all *An. gambiae* genes that were classified into any orthologous group and were annotated with Biological Process or Molecular Function GO-terms. The results were combined using a conservative strategy: terms must appear significant with a p-value <0.05 for all three enrichment tests, and there must be more than five genes in the test set. Genes with the fastest evolutionary rates were enriched for processes and functions including 'proteolysis', 'serine-type endopeptidase activity', 'detection of chemical stimulus involved in sensory perception

of smell', 'olfactory receptor activity', 'odorant binding', 'sensory perception of taste', 'chitin metabolic process', 'chitin binding', 'syncytial blastoderm mitotic cell cycle', 'peptidyl-lysine methylation', 'G-protein coupled receptor signaling pathway', 'G-protein coupled receptor activity', and 'metal ion binding'. Similar terms were found for genes with the highest dN/dS ratios, as well as 'potassium ion transmembrane transport'. Genes with the slowest evolutionary rates and the lowest dN/dS ratios were enriched for essential house-keeping processes and functions including 'protein transport', 'nucleosome assembly', 'translation', 'electron transport chain', 'protein complex assembly', 'glycolysis', 'tricarboxylic acid cycle', 'nucleoside-triphosphatase activity', and 'neurogenesis'.

Table S27. Functional enrichments of genes with low/high evolutionary rates.

Orthologous groups with evolutionary rates < 20th percentile or > 80th percentile represent the lowest and highest sets, respectively. Tables show best results of Hypergeometric (HyperG) enrichment tests on Gene Ontology Biological Processes and Molecular Functions, where they are also identified by both the classic Fisher (Fisher), and weighted Fisher (FisherW) tests. #, number in test set, Tot., number in total (background).

	Top 20% Evo.Rate Biological Process	HyperG	Fisher	FisherW	#	Tot.
GO:0006508	proteolysis	4.64E-24	4.60E-24	4.40E-30	128	690
GO:0050911	detection of chemical stimulus involved ...	6.07E-19	6.10E-19	6.10E-19	35	81
GO:0050909	sensory perception of taste	1.05E-18	1.00E-18	1.00E-18	31	64
GO:0006030	chitin metabolic process	4.76E-08	4.80E-08	3.90E-08	29	128
GO:0035186	syncytial blastoderm mitotic cell cycle	1.18E-05	3.60E-07	1.20E-05	6	9
GO:0018022	peptidyl-lysine methylation	2.79E-05	2.80E-05	1.70E-03	7	14
GO:0007186	G-protein coupled receptor signaling pat...	5.86E-04	5.90E-04	1.00E-03	35	263
GO:0006355	regulation of transcription, DNA-depende...	6.65E-03	1.24E-03	8.80E-05	49	463
GO:0006351	transcription, DNA-dependent	9.62E-03	1.00E-04	1.20E-03	11	69
GO:0006260	DNA replication	1.25E-02	2.88E-03	1.28E-02	12	78
GO:0006325	chromatin organization	1.66E-02	1.66E-02	3.30E-02	19	150
	Top 20% Evo.Rate Molecular Function	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0004252	serine-type endopeptidase activity	5.70E-31	<1e-30	<1e-30	111	333
GO:0003676	nucleic acid binding	3.05E-26	8.20E-27	<1e-30	169	751
GO:0008270	zinc ion binding	2.27E-23	2.30E-23	2.30E-23	149	631
GO:0046872	metal ion binding	2.51E-17	2.50E-17	2.00E-12	251	1483
GO:0004984	olfactory receptor activity	9.75E-15	9.70E-15	9.70E-15	35	79
GO:0005549	odorant binding	2.00E-10	2.00E-10	2.00E-10	42	142
GO:0008061	chitin binding	4.72E-06	4.70E-06	4.70E-06	27	103
GO:0003677	DNA binding	8.97E-05	9.00E-05	1.50E-04	95	625
GO:0004930	G-protein coupled receptor activity	2.21E-03	2.21E-03	1.93E-03	35	204
GO:0003682	chromatin binding	1.00E-02	1.01E-02	4.77E-02	15	75
GO:0003700	sequence-specific DNA binding transcript...	3.01E-02	3.01E-02	2.18E-03	46	333
	Bottom 20% Evo.Rate Biological Process	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0015031	protein transport	1.04E-10	1.80E-11	5.90E-14	46	75
GO:0006334	nucleosome assembly	2.02E-09	2.00E-09	5.20E-09	33	49
GO:0006412	translation	2.08E-09	7.00E-12	4.10E-13	117	276
GO:0022900	electron transport chain	9.11E-09	9.10E-09	1.79E-02	28	40
GO:0000022	mitotic spindle elongation	9.42E-09	9.40E-09	9.40E-09	33	51
GO:0006461	protein complex assembly	1.50E-08	3.10E-09	2.10E-09	83	183
GO:0006096	glycolysis	8.53E-08	8.50E-08	8.50E-08	14	15
GO:0043161	proteasomal ubiquitin-dependent protein ...	8.86E-08	1.90E-08	5.80E-09	23	32
GO:0000398	mRNA splicing, via spliceosome	1.30E-07	2.10E-08	1.50E-05	64	136

GO:0007264	small GTPase mediated signal transductio...	1.32E-07	2.20E-06	3.80E-09	75	167
GO:0006099	tricarboxylic acid cycle	2.11E-06	2.10E-06	2.10E-06	19	27
	Bottom 20% Evo.Rate Molecular Function	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0005525	GTP binding	6.18E-27	6.20E-27	6.20E-27	95	144
GO:0003924	GTPase activity	1.33E-22	1.30E-22	1.30E-22	80	122
GO:0032549	ribonucleoside binding	7.23E-21	7.20E-21	3.81E-02	311	822
GO:0000166	nucleotide binding	5.77E-19	1.80E-20	4.64E-02	385	1103
GO:0003735	structural constituent of ribosome	9.74E-17	9.70E-17	9.70E-17	83	150
GO:0004298	threonine-type endopeptidase activity	1.16E-10	1.20E-10	1.20E-10	16	16
GO:0005509	calcium ion binding	4.42E-09	4.40E-09	4.40E-09	84	197
GO:0003779	actin binding	1.68E-08	1.70E-08	3.20E-09	48	95
GO:0046982	protein heterodimerization activity	1.16E-07	1.20E-07	1.20E-07	46	94
GO:0001104	RNA polymerase II transcription cofactor...	1.50E-07	1.50E-07	6.00E-07	20	28
GO:0017111	nucleoside-triphosphatase activity	3.61E-07	2.60E-24	1.47E-02	102	281

Table S28. Functional enrichments of genes with low/high dN/dS ratios.

Orthologous groups with dN/dS ratios < 20th percentile or > 80th percentile were selected to represent the lowest and highest sets, respectively. Tables show best results of Hypergeometric (HyperG) enrichment tests on Gene Ontology Biological Processes and Molecular Functions, where they are also identified by both the classic Fisher (Fisher), and weighted Fisher (FisherW) tests. #, number in test set, Tot., number in total (background).

	Top 20% dN/dS Biological Process	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0006508	proteolysis	9.30E-11	9.30E-11	2.60E-15	102	690
GO:0050909	sensory perception of taste	8.13E-09	8.10E-09	2.50E-08	21	64
GO:0006030	chitin metabolic process	5.23E-05	5.20E-05	7.10E-05	24	128
GO:0006355	regulation of transcription, DNA-depende...	6.62E-03	6.62E-03	4.70E-08	63	587
GO:0071805	potassium ion transmembrane transport	2.35E-02	2.35E-02	2.35E-02	6	29
	Top 20% dN/dS Molecular Function	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0004252	serine-type endopeptidase activity	3.42E-19	3.40E-19	3.40E-19	89	333
GO:0008061	chitin binding	4.06E-04	4.10E-04	4.10E-04	22	103
GO:0043565	sequence-specific DNA binding	1.25E-03	1.25E-03	5.50E-04	40	247
GO:0005549	odorant binding	3.20E-03	3.20E-03	4.89E-03	25	142
GO:0003676	nucleic acid binding	4.91E-03	4.91E-03	2.20E-06	164	1376
GO:0003700	sequence-specific DNA binding transcript...	5.08E-03	5.08E-03	2.00E-04	48	333
	Bottom 20% dN/dS Biological Process	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0022008	neurogenesis	1.26E-11	1.30E-11	2.30E-11	268	774
GO:0016192	vesicle-mediated transport	2.83E-07	2.80E-08	2.61E-02	112	301
GO:0007018	microtubule-based movement	3.10E-07	4.50E-08	8.43E-03	35	65
GO:0006457	protein folding	1.54E-06	1.50E-06	5.40E-06	41	85
GO:0000022	mitotic spindle elongation	2.91E-06	2.90E-06	2.90E-06	28	51
GO:0000281	mitotic cytokinesis	1.16E-05	1.20E-05	1.20E-05	17	26
GO:0007298	border follicle cell migration	1.36E-05	1.40E-05	1.40E-05	33	68
GO:0006367	transcription initiation from RNA polyme...	1.42E-05	1.30E-06	2.60E-05	21	36
GO:0043161	proteasomal ubiquitin-dependent protein ...	1.87E-05	2.40E-06	2.40E-08	22	39
GO:0000398	mRNA splicing, via spliceosome	5.13E-05	1.10E-05	1.53E-03	54	136
GO:0006886	intracellular protein transport	5.80E-05	5.80E-05	9.90E-04	69	185
	Bottom 20% dN/dS Molecular Function	**HyperG**	**Fisher**	**FisherW**	**#**	**Tot.**
GO:0032549	ribonucleoside binding	6.07E-16	6.10E-16	3.21E-02	280	822
GO:0005525	GTP binding	1.00E-09	1.00E-09	1.00E-09	65	144
GO:0003899	DNA-directed RNA polymerase activity	4.19E-09	4.20E-09	4.80E-09	22	30
GO:0005524	ATP binding	1.28E-08	1.30E-08	1.30E-08	212	675
GO:0003924	GTPase activity	2.06E-08	2.10E-08	2.10E-08	55	122
GO:0004298	threonine-type endopeptidase activity	5.98E-08	6.00E-08	6.00E-08	14	16

GO:0003735	**structural constituent of ribosome**	8.51E-07	8.50E-07	8.50E-07	60	150
GO:0051082	**unfolded protein binding**	9.64E-07	9.60E-07	9.60E-07	26	47
GO:0001104	**RNA polymerase II transcription cofactor...**	2.43E-06	2.40E-06	1.90E-06	18	28
GO:0005544	**calcium-dependent phospholipid binding**	4.62E-05	4.60E-05	4.60E-05	10	13
GO:0003779	**actin binding**	2.00E-04	2.00E-04	1.10E-04	37	95

InterPro domain annotations across all 20 mosquito species with annotations available from VectorBase BioMart (http://biomart.vectorbase.org/biomart/martview) were analyzed to identify the protein domains with highest variation in gene counts across the anophelines. A crude measure that highlights such variations in copy-number was computed as the standard deviation divided by the mean of the anopheline gene counts matching a particular InterPro domain. Results were filtered to focus on abundant domains by requiring more than 200 genes in total and more than 5 genes in each species. Among the most highly variable are genes with fibrinogen, protein kinase, potassium channel, C-type lectin, ribonuclease H-like, proteinase inhibitor, glutathione S-transferase, and zinc-finger domains (Table S29).

Table S29. Most variable InterPro domains counts.

InterPro domains with the most variable counts across anophelines. The top 40 InterPro domains in terms of their variability in gene counts across the anophelines. CNV, copy-number variation = standard deviation / mean of anopheline gene counts with matching InterPro domains; TOT, total gene counts with matching InterPro domains across all the mosquitoes.

Domain Identifier	Domain Description	CNV	TOT	*A. gambiae*	*A. merus*	*A. arabiensis*	*A. quadriannulatus*	*A. melas*	*A. christyi*	*A. epiroticus*	*A. stephensi*	*A. maculatus*	*A. culicifacies*	*A. minimus*	*A. funestus*	*A. dirus*	*A. farauti*	*A. atroparvus*	*A. sinensis*	*A. albimanus*	*A. darlingi*	*Culex quinquefasciatus*	*Aedes aegypti*
IPR002181	Fibrinogen, alpha/beta/gamma chain, C-terminal globular domain	0.681	776	50	39	41	15	34	27	23	21	20	29	26	15	29	33	55	123	46	23	93	34
IPR020837	Fibrinogen, conserved site	0.583	321	31	15	20	6	13	13	13	12	8	13	17	7	14	10	25	43	18	8	21	14
IPR010512	Protein of unknown function DUF1091	0.528	469	66	22	22	21	22	17	25	21	18	25	34	34	22	23	13	18	14	7	34	11
IPR008266	Tyrosine-protein kinase, active site	0.497	532	34	32	15	14	38	13	16	18	25	35	12	15	18	33	33	57	11	36	38	39
IPR020635	Tyrosine-protein kinase, catalytic domain	0.461	420	28	27	13	11	31	10	10	13	24	30	12	9	14	28	28	33	9	28	30	32
IPR000033	LDLR class B repeat	0.389	221	9	9	8	12	12	9	9	10	16	10	10	9	8	9	15	26	9	10	10	11
IPR003091	Potassium channel	0.387	253	9	7	18	16	7	15	18	18	8	9	20	18	18	9	7	7	19	14	7	9
IPR018378	C-type lectin, conserved site	0.327	224	15	9	9	8	7	8	6	12	7	11	14	8	15	13	6	6	11	7	34	18
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic	0.323	878	50	46	37	31	60	25	34	33	46	49	37	30	32	46	51	81	28	52	54	56

	domain																						
IPR012337	Ribonuclease H-like domain	0.303	830	33	30	47	49	35	42	51	46	33	31	52	66	54	29	26	28	32	23	81	42
IPR001878	Zinc finger, CCHC-type	0.302	508	14	16	20	24	16	13	30	29	19	22	28	37	28	17	20	21	20	15	80	39
IPR003146	Proteinase inhibitor, carboxypeptidase propeptide	0.301	234	16	14	8	8	13	8	8	9	13	15	7	7	8	15	13	12	7	13	21	19
IPR002035	von Willebrand factor, type A	0.292	282	17	17	10	11	20	9	9	10	18	18	10	10	10	18	16	15	12	10	21	21
IPR004045	Glutathione S-transferase, N-terminal	0.290	484	33	26	22	22	25	16	17	19	25	27	20	20	8	29	34	28	15	29	39	30
IPR002017	Spectrin repeat	0.286	216	8	10	10	9	13	10	8	10	22	12	10	10	10	9	10	12	10	11	13	9
IPR006578	MADF domain	0.283	321	19	17	23	22	14	11	15	16	11	14	17	20	20	7	15	8	17	15	17	23
IPR018957	Zinc finger, C3HC4 RING-type	0.283	210	13	12	7	8	14	6	8	10	8	14	10	9	7	16	14	13	9	11	11	10
IPR002041	Ran GTPase	0.281	406	29	14	14	12	13	15	20	27	13	23	23	26	25	26	25	27	24	19	5	26
IPR000863	Sulfotransferase domain	0.273	242	18	14	8	9	17	6	10	9	11	12	10	9	10	11	13	13	9	13	19	21
IPR022755	Zinc finger, double-stranded RNA binding	0.270	233	16	17	11	11	18	9	9	10	9	12	11	12	7	13	9	10	8	11	15	15
IPR009072	Histone-fold	0.267	699	54	27	27	25	27	25	25	25	17	32	24	27	29	28	22	25	28	27	123	82
IPR000734	Lipase	0.266	575	27	30	19	20	32	19	20	22	37	38	17	22	20	34	34	32	21	29	53	49
IPR011038	Calycin-like	0.266	247	15	16	12	10	13	6	12	10	9	12	9	8	9	16	15	17	10	13	21	14
IPR020479	Homeodomain, metazoa	0.263	540	34	23	29	30	23	21	27	30	27	30	31	30	32	28	29	25	6	14	33	38
IPR000618	Insect cuticle protein	0.260	2211	170	86	102	99	116	44	97	81	73	111	92	94	79	116	123	108	94	92	186	248
IPR002893	Zinc finger, MYND-type	0.258	518	37	32	32	30	30	19	21	20	16	19	21	22	25	23	23	15	19	22	69	23
IPR004117	Olfactory receptor, Drosophila	0.258	1078	79	45	54	54	52	31	52	43	43	47	49	52	51	45	40	42	42	18	145	94
IPR011527	ABC transporter type 1, transmembrane domain	0.255	443	23	20	19	18	34	19	20	19	36	22	16	21	16	20	20	26	17	21	27	29
IPR017972	Cytochrome P450, conserved site	0.252	1558	102	82	60	62	85	48	33	57	65	81	60	59	54	74	87	79	51	80	173	166
IPR005135	Endonuclease/exonuclease/phosphatase	0.249	298	11	11	14	12	11	11	15	18	14	11	13	14	19	11	24	16	16	13	31	13
IPR009020	Proteinase inhibitor, propeptide	0.247	299	19	17	11	11	16	10	11	12	15	18	10	10	12	19	17	18	11	17	23	22
IPR013106	Immunoglobulin V-set domain	0.246	646	43	36	31	20	34	26	37	35	13	38	39	34	37	35	42	31	37	20	29	29
IPR001314	Peptidase S1A, chymotrypsin-type	0.245	3513	257	192	120	129	189	116	121	155	118	181	118	151	158	180	185	215	137	154	328	309
IPR001140	ABC transporter, transmembrane domain	0.242	383	20	17	16	16	30	17	17	15	29	20	14	17	14	18	18	21	15	19	24	26
IPR006589	Glycosyl hydrolase, family 13, subfamily, catalytic domain	0.241	320	18	19	12	13	22	9	10	13	17	17	12	12	13	18	16	18	12	13	30	26
IPR000834	Peptidase M14, carboxypeptidase A	0.241	417	23	23	18	15	28	16	16	16	26	25	15	15	15	26	24	23	14	20	30	29
IPR001128	Cytochrome P450	0.240	1866	113	93	68	70	105	58	56	64	101	104	67	67	64	91	103	102	61	93	207	179
IPR025110	AMP-binding enzyme C-terminal domain	0.239	352	22	17	12	12	21	11	12	13	19	19	14	13	15	21	18	20	11	17	27	38
IPR008930	Terpenoid cyclases/protein prenyltransf. alpha-alpha toroid	0.235	255	18	14	14	16	13	9	11	9	12	13	16	8	15	11	12	18	11	10	13	12
IPR004046	Glutathione S-transferase, C-terminal	0.232	415	28	23	20	20	24	17	14	18	19	22	19	18	10	19	28	21	14	23	34	24

Reproduction

Male accessory gland cluster comparative analysis: A dedicated analysis workflow was designed to perform the identification of putative orthologs of *An. gambiae* male accessory gland (MAG) genes localized to the 3R chromosome in 10 anopheline species (*An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. merus*, *An. stephensi*, *An. funestus*, *An. minimus*, *An. dirus*, *An. atroparvus* and *An. albimanus*) (Figure S16). The global alignments were performed by means of the Needleman-Wunsch algorithm (*148*); additional, local alignments (when needed) were executed with the Smith-Waterman algorithm (*149*). All the analyses were implemented with the Bioinformatics toolbox (Version 4.1 - R2012a) in the Matlab programming environment (MathWorks). Multiple alignment parameters were tested including different substitution matrices, gap penalty and gap extension. The best results were obtained via BLOSUM 62 scoring matrix, gap penalty of 10 and extension 1. Each run resulted in a list of candidate proteins in the 10 selected genomes, which were subjected to reciprocal best-hit analysis against the *An. gambiae* genome. The resulting putative orthologs were further tested for DNA homology to *An. gambiae* using the region comparison (trans Blat net) tool available at VectorBase.

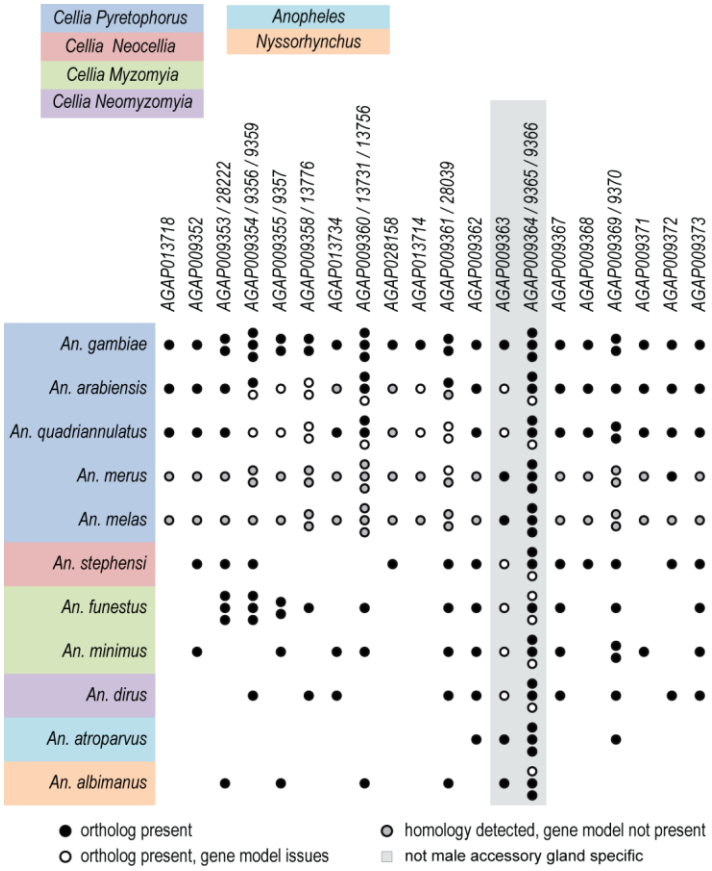


Figure S16. Male accessory gland genes.

Comparative analysis in 10 anopheline species of male accessory gland genes clustered on chromosome arm 3R in An. gambiae. The figure shows the putative orthologs of *An. gambiae* male accessory gland (MAG) genes present in the different anophelines. Paralogous *An. gambiae* genes are in the same column.

Ex vivo enzymatic activity of transglutaminase from male accessory glands: For each species (*An. gambiae*, *An. arabiensis*, *An. stephensi*, *An. dirus*, *An. atroparvus*, *An. albimanus*) 20-25 male accessory gland (MAG) pairs were dissected from 4-day old virgin males, homogenized in either TGase ‘‘+’’ buffer (50 mM Tris pH 7.6, 1 mM DTT, 5 mM CaCl2) or TGase ‘‘-“ buffer (TGase ‘‘+’’ buffer with 250 mM EDTA and 0.3 mM dGTP) and freeze/thawed on dry ice three times before the addition of 5 mM monodansylcadaverine (MDC). Samples were incubated at 37 °C for 60 min, vortexed briefly, and spun down 10 min at 13,000 rpm. Proteins in the supernatant were separated by SDS-PAGE and visualized via the fluorescent property of cross-linked MDC under UV illumination. Gels were visually inspected and activity quantified relative to *An. gambiae* (Table S30).

Table S30. Transglutaminase activity from male accessory glands.

Activity of transglutaminase from male accessory glands (MAGs) *ex vivo*. Monodansylcadaverine (MDC) crosslinking activity of transglutaminase (TG) enzyme present in 6 anopheline species exhibiting variable male mating plug phenotypes. TG activity was visually inspected via the fluorescent property of cross-linked MDC under UV illumination and quantified relative to *An. gambiae* (+++).

Species	Plug phenotype	TG activity
An. gambiae	Full	+++
An. arabiensis	Full	+++
An. stephensi	Full	++
An. dirus	Intermediate	+
An. atroparvus	Intermediate	++
An. albimanus	No plug	n/a

The male accessory glands (MAGs) of many insect species produce and secrete a number of seminal proteins, which are essential for male fertility and upon sexual transfer act as direct or indirect regulators of female reproductive biology (*150, 151*). It has been shown that these proteins are rapidly evolving in *Drosophila melanogaster* (*152*) and other insect species (*151, 153*). In *An. gambiae*, MAG proteins are transferred to the female as a mating plug, which is formed through the action of a transglutaminase enzyme (TG3 – AGAP09099) on the major plug protein Plugin (AGAP009368). In *An. gambiae*, MAG genes are predominantly organized in a large cluster on chromosome arm

3R (division 33D: 31,743,101 - 31,827,569), known as the "fertility island" (*21, 22*). This fertility cluster encompasses 27 genes exclusively expressed in the MAGs, including Plugin, and contains several paralogs, indicative of extensive gene duplication in this region. This is in contrast to the organization of MAG genes in *Drosophila*, where seminal genes are predominantly dispersed throughout the genome. The co-localization of MAG genes suggests a potential co-regulation that is likely to be important for male reproductive biology. However, the extent of conservation and/or co-localization of MAG genes across anopheline species was completely unknown to date. Our comparative genomic analyses identified putative orthologs of *An. gambiae* male accessory gland (MAG) genes localized to the 3R chromosome in 10 anopheline species (*An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. merus*, *An. stephensi*, *An. funestus*, *An. minimus*, *An. dirus*, *An. atroparvus* and *An. albimanus*) (Figure S16). MAG genes were highly conserved within the *Gambiae* complex; however conservation was mostly lost outside of the complex. Four genes present in this region but expressed in both males and females on the other hand were well conserved across species, suggesting rapid evolution of genes encoding male seminal components compared to other genes. These results confirm previous findings in *Drosophila* species, where high levels of gene gain/loss and evidence for rapid evolution were reported in male reproductive genes compared to the rest of the genome (*152*).

We performed mating experiments in eight anopheline species to determine whether males of species other than *An. gambiae* transfer a mating plug. While a fully formed plug was found in *An. arabiensis*, *An. stephensi* and *An. funestus*, no plug was identified in *An. albimanus* and the remaining species showed plug-like structures with intermediate levels of coagulation. Interestingly, plug phenotype was correlated with transglutaminase activity in the MAGs, as assessed using an *ex vivo* assay on MAGs dissected from virgin males in six species (Table S30). *An. gambiae* and *An. arabiensis*, both exhibiting fully formed plugs, showed considerable transglutaminase activity, whereas species with intermediate coagulation phenotypes generally had lower enzyme activity. No TG activity was measured in the plug-less species *An. albimanus*, which suggests that although TG3 is present in the genome (Figure 4A, main text), it is not active in the MAGs.

Moreover, conservation of the major plug protein Plugin was lost outside the *An. gambiae* complex, with only *An. stephensi* showing an ortholog (Figure S16), suggesting an alternate plug substrate or mechanism of plug formation in the other plug-forming species. The transfer of the mating plug in *An. gambiae* has multiple key reproductive functions, as it ensures correct sperm storage by yet unknown mechanisms (*26*) and it delivers high levels of the steroid hormone 20-hydroxyecdysone that trigger an increase in female fecundity (*27*) and fertility (*28*). Comparative functional analyses will allow identification of the mechanisms that plugless anophelines utilize towards the same reproductive processes.

Sex biased gene expression

RNA extraction was carried out on sexed samples derived from wild type or transgenic G3 strains of *An. gambiae*. Samples comprised wild type 3-4 day old adult male and females from a mixed cage and L1 instar, L2-L3 instar and L3-L4 instar pools of sexed larvae that were obtained based on the inheritance of an X-chromosome-linked 3xP3::dsRED transgene in a G3 background. RNA was prepared for sequencing using the Illumina mRNA-Seq Sample Preparation kit. All samples were paired-end sequenced at either 100bp or 85bp read lengths (Raw data are submitted as part of the *Anopheles*-Y-chromosome consortium).

Paired reads were processed and aligned against the reference genome of *An. gambiae* (AgamP3.7) using RNA-STAR (*154*). Reads were filtered for those that map uniquely, allowing only one alignment per read, and were quantified using HTSeq (*96*). Counts were normalized between sexes of the same time point using the effective library size of concordantly mapping read pairs. Biological replicates were collected for larval timepoints and we observed good correlations between replicates. Adult data are derived from a single replicate. The sum of reads of all genes in all conditions from L1 to adults was calculated separately for male and female samples. In addition, data from a published array experiment ("MozAtlas") assaying whole males and whole females (blood-fed) were also used to assess sex-bias. Data files containing quantile-normalized gene expression values were downloaded from www.tissue-atlas.org (*155*).

Estimates of male and female expression for both the RNAseq and the MozAtlas datasets described above were calculated as described below. The estimated dN and dS across the genus (described in selection section) were determined for each annotated *An. gambiae* gene (AGAP_id). In some cases, multiple AGAPs fell in a single orthogroup, and in these cases, the PAML calculated dN and dS was assigned to all AGAPs in that orthogroup. We then filtered the dataset to include only genes for which there was an alignment of at least 200 bp across a minimum of 6 taxa in the genus. This resulted in a dataset of 8,310 genes which passed quality control and had dN/dS and expression in two distinct expression datasets (7,518 on the autosomes and 792 on the X chromosome). We categorized genes by differing degrees of sex-bias using the metric "female expression/(male expression+female expression)". This metric was calculated for both datasets. If both datasets had values > 0.8, this was considered "Strong Female-Bias" (N=27); above 0.6 (but < 0.8) = "Weak Female-Bias" (N=73); below 0.2 = "Strong Male-Bias" (N=50); and below 0.4 (but > 0.2) = "Weak Male-Bias" (N=233). Anything that fell in the range of 0.4-0.6 for both datasets was considered "Unbiased". Data from MozAtlas arrays assaying particular tissues were also used to assess tissue-bias using the following metric: "specific tissue/(sum of all tissues)" > 0.5 = "Tissue-biased".

We estimated dN/dS for orthologous gene classes categorized by these differing degrees of sex bias (Figure S17) and discover that both categories of female-biased genes

are faster than male-biased genes (Strong Female vs Strong Male Wilcoxon rank sum test p=0.0005; Weak Female vs Weak Male Wilcoxon rank sum test p= 0.013). Thus, the trends are consistent regardless of where the cut-off of "bias" is determined. In the main manuscript, we report the results for only the strong sex-biased classes.

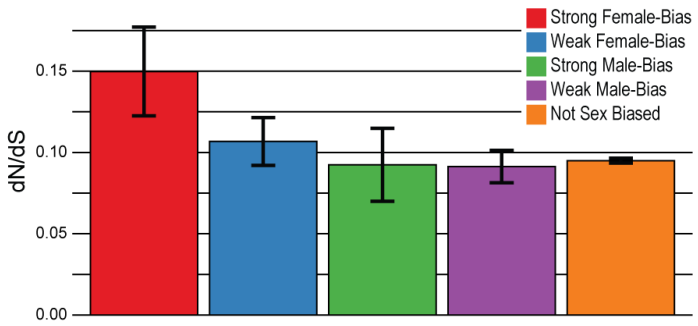


Figure S17. Sex biased genes.

dN/dS ratios for orthologous gene classes categorized by differing degrees of sex bias, where sex biased gene expression is determined using the metric female expression/(male expression+female expression).

While the RNAseq dataset includes larval sex biased expression data, because of our requirement that both MozAtlas data and RNAseq data support the assigned sex-bias category, all of these genes are expressed in the adult stages (as MozAtlas only examined adults). To explore whether there were any trends of particular tissue-biases among the most sex-biased genes, we used the MozAtlas data to further categorize genes. If the sum of the expression data across all of the tissues suggested that more than 75% of the total expression came from a particular tissue, this was considered tissue-biased. Thus the individual genes can be classified according to their tissue bias (we included both strong and weak biased genes as they show the same trends). Among the 27 strong female-biased genes, Midgut&Abdomen&MT (a measure summarizing bias in midgut, abdominal carcass, and malpighian tubules, each examined individually in the MozAtlas dataset but considered together for our purposes) and salivary gland biased genes are distinctly faster than the few male-biased genes highly-biased in the same tissues. Additionally, these tend to be strong female-biased genes while the male biased ones are weak. The more rapid evolution of female-biased genes, highly expressed in gut and salivary glands in comparison to male-biased genes that are highly expressed in similar tissues, lends support to the notion that the extremely distinct life histories of males and females in anophelines might drive different patterns of sex-biased gene evolution than are seen in the vast majority of other organisms (*156*). However, male anophelines are not exempt from the patterns observed in other organisms. Genes expressed in the accessory glands of males display the highest rates of evolution of any other male tissue examined and none of the 50 strong male-biased genes are present on the X chromosome,

which is consistent with previous findings that the X chromosome is not an environment conducive to male-biased expression.

Cuticular proteins

We investigated the extent and time scale of sequence-cluster homogenization within the anophelines by estimating gene trees for the co-orthologous regions of all project species. Trees were constructed in MEGA5 (*157*) using maximum likelihood and the Tamura-Nei substitution model. To investigate expression of concertedly evolving CPLCW and CPLCG genes, DIG-labeled in situ probes targeting conserved sequence of all *An. gambiae* CPLCW genes and all "group A" CPLCG genes identified in (*32*) were hybridized to larval sections following the methods of (*158*).

Cuticular proteins (CPs) account for about 2% of known protein-coding sequences in mosquitoes, but the functional significance of this molecular diversity is not understood (*29*). CP expression (*159*) and cuticular thickening (*160*) have been associated with insecticide resistance in anophelines, however, and the serosal cuticle is critical for embryonic desiccation tolerance (*161*). CPs are also abundant in the selectively maintained 2La inversion polymorphism (*162*). For their sheer number as well as developmental and ecological importance, CPs have been a focus of manual annotation in *An. gambiae* (see (*29*) for a review of CP families). Comparisons among Diptera have revealed numerous amplifications of CP genes undergoing concerted evolution (*30-33*), which were termed "sequence clusters". Estimating gene trees for the co-orthologous regions allowed us to investigate the extent and time scale of sequence-cluster homogenization within the genus *Anopheles* (Figure S18).

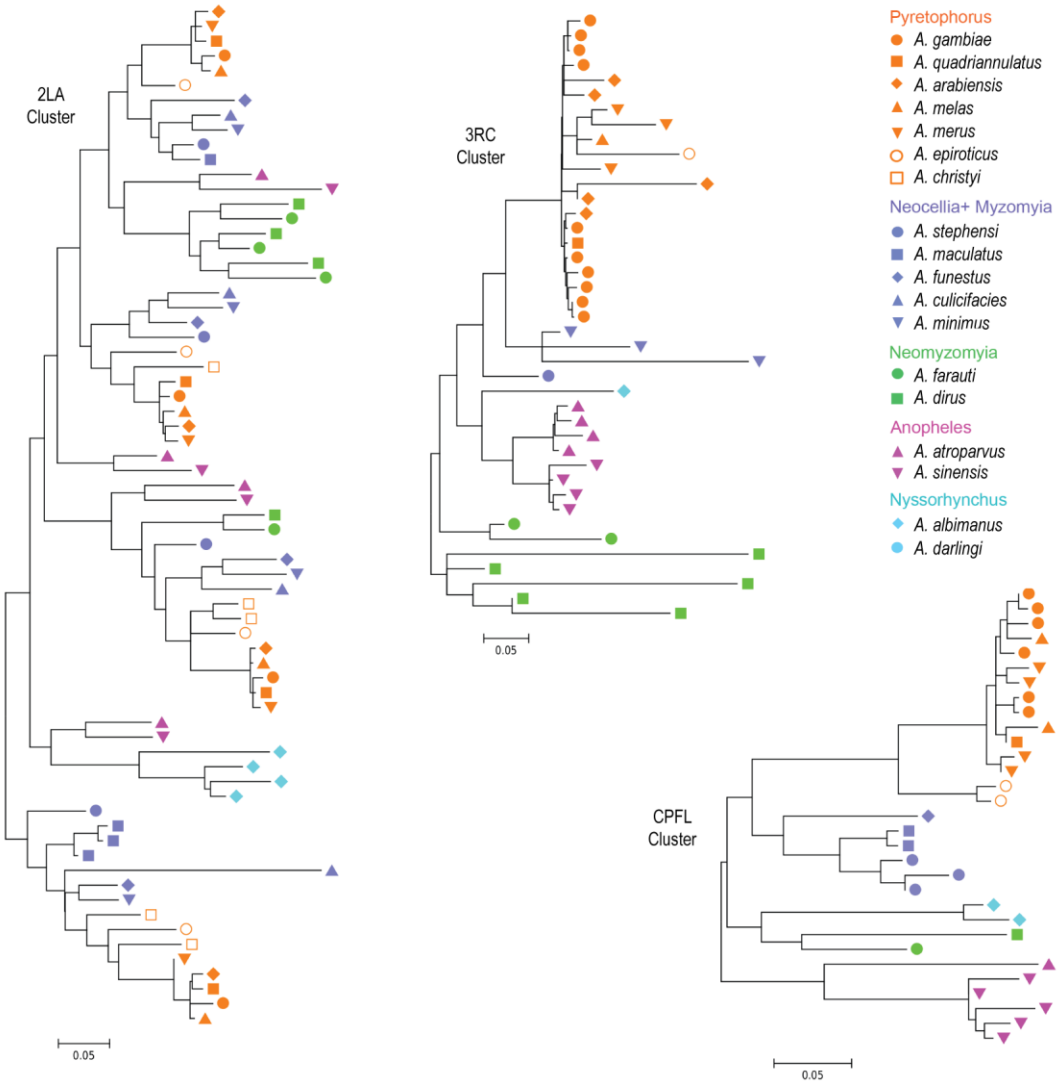


Figure S18. Cuticular proteins.

Representative gene trees of cuticle protein "sequence clusters", sets of genes with high sequence similarity that have been inferred to be concertedly evolving based on comparisons among dipterans (i.e., divergence times >100 million years). Trees were constructed in MEGA5 (*157*) using maximum likelihood and the Tamura-Nei substitution model. No bootstrap values or branch statistics are reported, as only the most likely topology was of interest and not statistical support for individual nodes.

We then identified the largest single cluster of paralogous genes within each species, as well as the largest clusters grouping by taxon at deeper nodes within the organismal phylogeny. We found sequence clusters identified in *An. gambiae* to be consistently associated with paralog clustering throughout the genus, but most extensively in a subset of clusters (Table S31). These include the "3RB" and "3RC" sequence clusters of CPR genes (defined in (*30*)), the CPLCG "group A" and CPLCW

sequence clusters found elsewhere on 3R (detailed in (*32*)), and six tandemly arrayed genes on 3L designated CPFL2 through CPFL7 (*163*). For these five sequence clusters, complete clustering by organismal lineage was observed for most deep nodes as well as for many individual species outside the shallow *Pyretophorus* clade. The gene trees imply that active clusters appear reciprocally monophyletic on the order of 20 million years post-divergence, based on current molecular-clock estimates.

Table S31. Co-orthologous cuticular gene sequence clusters.

Number of CP genes from each anopheline species considered co-orthologous to *An. gambiae* sequence clusters. The total number of genes identified is in the left column for each sequence cluster, and the number of genes constituting the largest single-taxon phylogenetic cluster is given in the right column. Dots are given where no single-taxon cluster is possible. Deeper nodes of the anopheline phylogeny are given below individual species. Tot. = Total.

	2LA		2LB+2LC		2RA		2RB		3RB		3RC		CPFL		CPLCG-groupA		CPLCW		CPLCP	
	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster	Tot.	Max cluster
A. gambiae	4	1	15	2	6	3	8	1	8	2	10	5	6	2	12	9	9	3	14	1
A. arabiensis	4	1	12	1	3	1	5	2	4	1	5	2	0	.	2	1	3	1	5	1
A. quadri-annulatus	4	1	19	3	5	1	4	1	0	.	1	.	1	.	0	.	0	.	4	1
A. merus	4	1	25	2	4	1	3	1	1	1	3	2	4	2	6	1	2	1	9	1
A. melas	4	1	14	2	5	1	3	1	2	1	1	.	2	1	4	1	0	.	8	1
A. christyi	5	2	4	1	4	2	1	.	2	1	0	.	0	.	1	.	2	1	5	3
A. epiroticus	4	1	11	3	5	3	6	3	0	.	1	.	**2**	**2**	**3**	**3**	**2**	**2**	5	3
A. stephensi	4	1	21	4	4	2	4	1	0	.	1	.	**3**	**3**	**4**	**4**	1	.	5	1
A. maculatus	4	3	2	1	1	.	6	2	0	.	0	.	**2**	**2**	1	.	1	.	7	2
A. culicifacies	4	1	6	3	2	2	4	2	0	.	0	.	0	.	**2**	**2**	2	1	5	2
A. funestus	4	1	8	3	2	1	5	2	**2**	**2**	0	.	1	.	**2**	**2**	1	.	2	1
A. minimus	4	1	9	2	5	2	4	2	**3**	**3**	**3**	**3**	0	.	1	.	0	.	5	2
A. dirus	4	1	14	4	3	1	4	1	**7**	**7**	**5**	**5**	1	.	3	2	2	1	5	2
A. farauti	4	1	11	3	3	1	5	2	0	.	**2**	**2**	1	.	4	2	**3**	**3**	7	3
A. sinensis	4	1	11	6	1	.	4	2	2	1	**4**	**4**	**5**	**5**	**2**	**2**	0	.	4	1
A. atroparvus	4	1	12	6	5	4	5	2	**2**	**2**	**4**	**4**	1	.	**4**	**4**	**2**	**2**	10	5
A. albimanus/ Nyssorhynchus	**4**	**4**	8	5	4	3	4	3	**2**	**2**	1	.	**2**	**2**	**2**	**2**	**2**	**2**	6	3
Pyretophorus	29	3	100	12	32	6	30	11	**17**	**17**	**21**	**21**	**15**	**15**	**28**	**28**	**18**	**18**	50	13
Neocellia + Myzomyia	20	6	46	11	9	3	23	5	**5**	**5**	4	3	**6**	**6**	**10**	**10**	**5**	**5**	24	5
Neomyzomyia	8	6	25	6	6	4	9	3	**7**	**7**	**7**	**7**	**2**	**2**	**7**	**7**	**5**	**5**	12	4
Anopheles	8	4	23	6	6	4	9	4	4	3	**8**	**8**	**6**	**6**	**6**	**6**	**2**	**2**	14	6

For most sequence clusters, the number of genes identified in the newly sequenced genomes was substantially lower than that identified in *An. gambiae* (Table S31). Since CP gene sequences of the PEST strain are broadly concordant with the M (now *An. coluzzii*) and S genome sequences (R.S.C., unpublished observations) and with molecular studies of the G3 strain ((*29*) and references therein), the large difference is not a peculiarity of the PEST assembly. More likely it is an artifact of short-read sequencing technology. Duplication and gene conversion, both hallmarks of CP sequence clusters (*31, 33*), are phenomena well known to complicate genome assembly, particularly for graph-based methods, and dot plots indicate that chromosomal regions containing

sequence clusters are often interrupted by scaffold gaps or ends. The apparent dropout of sequence-cluster genes in the current assemblies should bias downward the apparent rate of concerted evolution because, by implication, highly similar paralog sets are not being completely recovered. A genuine and massive increase in CP gene number peculiar to *An. gambiae* remains a formal possibility to be refuted, but not a parsimonious one given that the sequence clusters with greatest evidence of concerted evolution also show the greatest disparity in gene number (Table S31) and because change in gene number outside of sequence clusters is modest. For example, of 82 CPR genes identified in *An. albimanus* (the earliest diverging lineage) that are not co-orthologous with *An. gambiae* sequence clusters, 79 have a single best-aligning match in *An. gambiae* with on average 80.0% identity and 93.4% alignment coverage. Two have matches in *An. gambiae* that lack simple reciprocity and one is novel. The emerging pattern of anopheline CP evolution is thus one of relative stasis for a majority of single-copy orthologs, juxtaposed with consistent concerted evolution of a subset of genes, a pattern dissimilar to the gain-and-loss (birth-and-death) process that frequently characterizes large gene families (*164*).

While complete CPLCW genes were rarely recovered in the new assemblies, in *An. gambiae* all such genes occur in a head-to-head arrangement with a CPLCG "group A" sequence cluster gene (but note that some of the latter genes are not paired with CPLCWs). The non-coding and intergenic sequence between the respective start codons is short and conserved within and among species (Figure S19).

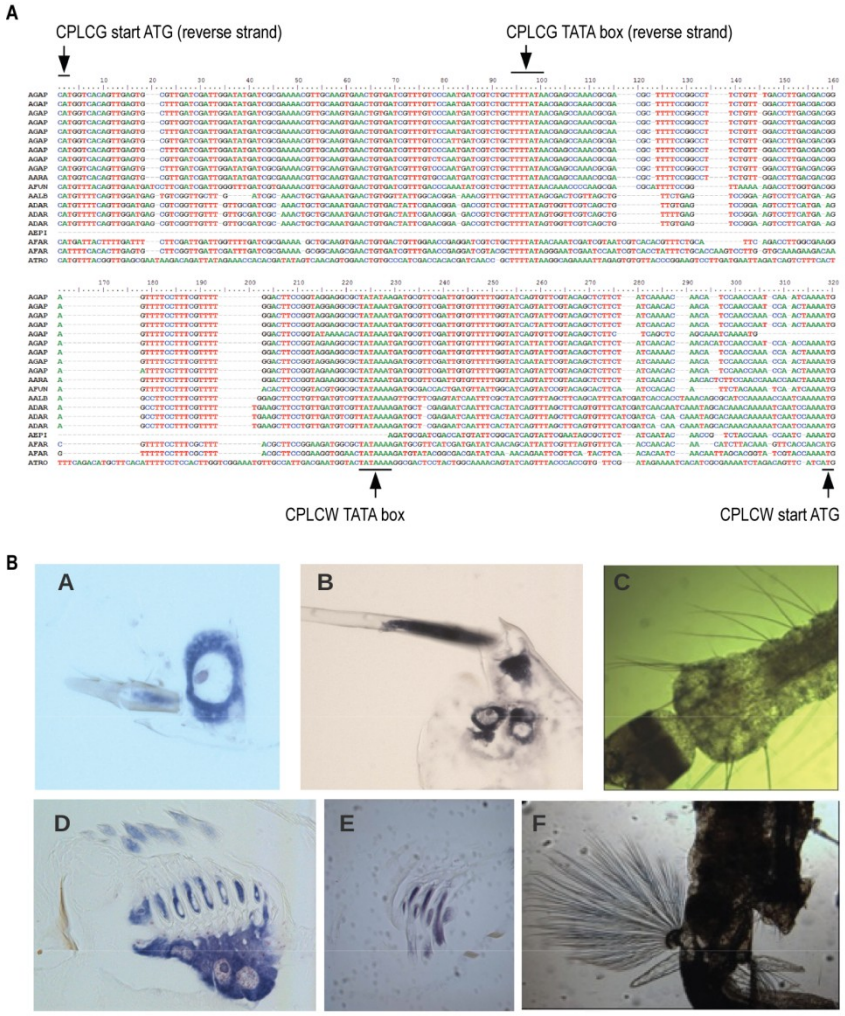


Figure S19. CPLCG and CPLCW genes.

A. Nucleotide sequence alignment showing non-coding sequence between adjacent CPLCG "groupA" and CPLCW gene pairs. These genes are arranged head-to-head, each gene has a distinct TATA box that is in close proximity to the other, an arrangement that differs from bidirectional promoters. The "conjoined" promoters are strongly conserved within the *An. gambiae* complex and more loosely conserved at greater divergence times. Species codes follow VectorBase prefixes (but ATRO, *An. atroparvus*). **B**. Overlapping expression of concertedly evolving CPLCG and CPLCW genes. DIG-labeled in situ probes targeting conserved sequence of all *An. gambiae* CPLCW genes and all "group A" CPLCG genes identified in (*32*) were hybridized to larval sections. Identical hybridization patterns were detected for the two probes, localizing to abdominal and thoracic bristles (A and B for CPLCG group A and CPLCW genes, respectively) and the grid supporting the ventral brush (D and E for CPLCG group A and CPLCW genes, respectively). Examples of intact bristles (C) and the ventral brush (F) are also shown.

The CPLCG and CPLCW sequence clusters are temporally co-expressed (*32*) and have identical in situ hybridization signal (Figure S19), suggesting they are regulated and function in tandem akin to a bacterial operon or eukaryotic bidirectional promoter, although each member of a pair has an unambiguous core promoter. Hybridization was

observed in relatively few cells of developing larvae, specifically those supporting growth of thoracic and abdominal bristles and the ventral brush. Even so, whole-body estimates of the expression of these genes are comparable to or higher than many other CPs that have broader distribution (*32, 158, 165*), implying a prodigious per-nucleus output of message. Since monomers from different co-expressed loci are expected to co-aggregate into fibrils during cuticle self-assembly, a high demand for transcript coupled with the requirement that monomers from different genes be functionally interchangeable could well provide the selective basis for concerted evolution of these two conjoined sequence clusters. The large majority of CPCLG, CPLCW, and CPLCP genes share a simple gene architecture that is not limited to concertedly evolving sets. This shared architecture includes an archetypal TATAA promoter sequence and a short first exon containing a UTR of approximately 10-100 bases and ending with the codons ATGAAG. Short introns of less than 100 bases and a terminal exon encoding all of the mature peptide are also characteristic. To assess whether the 5' pattern is in fact specific to cuticular protein genes, we identified all *An. gambiae* genes with upstream and first-exon sequence matching the pattern: TATA[T or A]A followed by 40-150 bases (but containing no ATG after the first 25 bases) followed by the codons ATGAAG followed by an intron. Of the 63 matches, 54 were CPLCG, CPLCW, or CPLCP genes and 4 were CPRs. The remaining five all encoded proteins with signal peptides and low-complexity sequence composition typical of CPs (*29*). Four of these (AGAP008448, AGAP008450, AGAP028001, AGAP028180) are interposed within the larger array of CPLCG and CPLCW genes on 3R but do not contain the conserved domains of those families nor have evident orthologs in *Aedes* or *Culex*. Thus, despite its simplicity, this 5' sequence pattern is very strongly and specifically associated with CPs and its union with novel terminal exons appears to have underlain *de novo* gene formation in this region.

Chemosensation

To identify chemosensory genes in the 16 anopheline species, TBLASTN searches v.2.28 (with 1e-5 as e-value cutoff) were conducted against each genome assembly using protein sequences of annotated OR/GR/IR genes in *An. gambiae*, *Ae. aegypti*, *Cu. quinquefasciatus*, and other insects as queries. For all putative chemosensory gene coding regions, gene models were predicted using GeneWise v2.2.0 (*107*) and curated manually to ensure the quality of annotation. The process was repeated with newly identified chemosensory genes as queries until no additional genes were found. For each of the OR, GR, and IR families, a preliminary phylogenetic tree containing all family members was first created to divide genes into subfamilies, each representing a single gene in the most recent common ancestor of all sequenced anopheline species. Phylogenetic trees were then built for each subfamily individually. In all phylogenetic analyses, multiple sequence alignments were generated using MAFFT v7.130 (*166*),

poorly aligned regions in the alignments were filtered using TrimAl v1.4 (with "automated1" option) (*90*), and Maximum-likelihood trees were reconstructed using RAxML v8.0.12 (with "PROTGAMMAAUTO" option and 100 bootstrap replicates) (*91*). Two approaches were used to estimate the number of chemosensory gene gain and loss events: first, subfamily trees were collapsed at the bootstrap support level of 80% and then reconciled with the species phylogeny using Notung v2.6 (*167*); second, the number of chemosensory genes in all species and subfamilies were used as input for CAFE v3.0 (*18*) to calculate rates of gene gain and loss and gene copy numbers at ancestral nodes. Fitmodel v20140407 was used to detect positive selection in chemosensory gene subfamilies.

The diverse vectorial capacities of anopheline mosquitoes are closely associated with their different host preferences. For instance, despite of its competence for malaria parasite, *An. quadriannulatus* is considered a non-vector member of the *An. gambiae* complex mainly due to its zoophagic feeding behavior (*168*). The host-seeking and other chemosensory behaviors of anopheline mosquitoes are largely mediated by several major types of chemoreceptors, including odorant receptors (ORs), gustatory receptors (GRs), and variant ionotropic glutamate receptors (IRs). Given the rapid gain-and-loss (birth-and-death) of chemosensory genes observed in many other insects, it is natural to expect that the different host preferences of anopheline mosquitoes could be also attributed to substantial variations of chemosensory gene copy numbers among these species.

Interestingly, we found that the overall size of chemosensory gene repertoire has been relatively conserved across the anopheline genomes (Figure S20). Our phylogenetic analyses estimated that the most recent common ancestor of the 18 anopheline mosquitoes had about 60 genes in each of the OR and GR families, which is very close to the number of OR and GR genes in most anopheline genomes. The estimated gain-and-loss rates of OR and GR genes (λ (no error model) = 0.00241 for ORs and 0.00066 for GRs, respectively) are both much lower than the overall level of anopheline gene families (λ (no error model) = 0.00602). Similarly, we found almost the same number of IRs (about 20) in all anopheline genomes that belong to several conserved IR subfamilies whose members are expressed in the antennae of *An. gambiae* and other insects. Despite the overall conservation in chemosensory gene number, we also noticed a few examples of rapid gene gain-and-loss on specific lineages. For example, there is a net gain of at least 10 ORs in the common ancestor of the *An. gambiae* complex, leading to higher numbers of ORs in the 5 species in the complex compared to other anopheline mosquitoes.

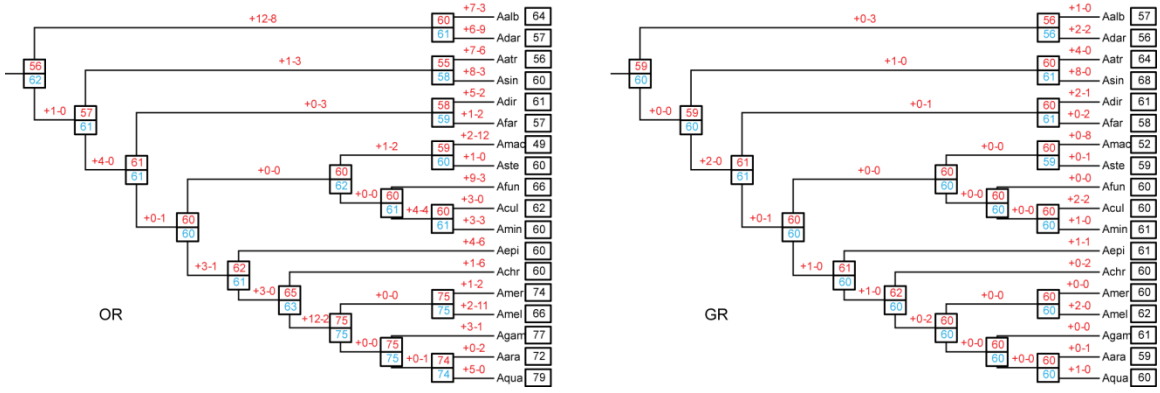


Figure S20. Odorant receptors and gustatory receptors.

Evolutionary histories of gains and losses of odorant receptors (ORs) and gustatory receptors (GRs) across the anophelines. Numbers on each node and branch are estimated ancestral copy numbers and gene duplication/loss events. Red: Notung estimation. Blue: CAFE3 estimation. Species: Aalb, *An. albimanus*; Adar, *An. darlingi*; Aatr, *An. atroparvus*; Asin, *An. sinensis*; Adir, *An. dirus*; Afar, *An. farauti*; Amam, *An. maculatus*; Aste, *An. stephensi*; Afun, *An. funestus*; Acul, *An. culicifacies*; Amin, *An. minimus*; Aepi, *An. epiroticus*; Achr, *An. christyi*, Amer, *An. merus*; Amel, *An. melas*; Agam, *An. gambiae*; Aara, *An. arabiensis*; Aqua, *An. quadriannulatus*.

The data indicate that the chemosensory gene repertoire has been relatively stable throughout the evolution of anopheline mosquitoes. It is likely that the majority of these chemosensory genes are required to carry out several critical behaviors, most notably host preference insofar as blood-feeding, during the anopheline lifecycle. Host-preference differences among the anopheline species can be caused by a combination of functional divergence and transcriptional modulation of orthologous genes, which is supported by previous study of antennal transcriptomes in *An. gambiae* before and after blood-feeding (*35*), and cross-species comparison between *An. gambiae* and *An. quadriannulatus* (*36*). Furthermore, we found evidence of positive selection in 19 out of 53 OR subfamilies and 17 out of 59 GR subfamilies, suggestive of potential functional divergence. Additional data on chemosensory genes expression in the 18 anopheline species would further help to elucidate the genomics basis for the diverse host-preferences.

Neuropeptide hormones/receptors

Optimal alignments of each protein matrix were made using default setting in OPAL (*169*) as implemented in Mesquite (*170*). PartitionFinderProtein (*171*) was the used to find the best model of protein evolution for each gene. Maximum likelihood (ML) trees were constructed using RAxML-HPC2 on XSEDE v7.2.7 with optimized parameters on the CIPRES Science Gateway (*172, 173*). For each single gene matrix, 500 search replicates were conducted to find the maximum likelihood trees, and 500 non-parametric bootstrap replicates were used to calculate support values for groups of interest.

Peptide hormones: Peptide hormones regulate numerous processes in mosquitoes and other eukaryotes (*37*). These small peptides are synthesized, processed and released from nervous and endocrine systems in the mosquito and elicit their effects through binding appropriate receptors at distant tissues. They also are typically synthesized as a preprohormone consisting of a signal peptide, the small, biologically active peptide hormone, and additional amino acids that are proteolytically removed during post-translational processing (*38*). Because of this it is likely that the mature hormone responsible for its physiological effects would be highly conserved, whereas the non-essential structural amino acids flanking the mature peptide may be more variable. The 16 anopheline genomes offer a unique opportunity to compare the evolution of these peptides in a closely related group with a long evolutionary history.

In total, 39 peptide hormones were identified from each of the 16 anopheline genomes. These peptides were subsequently analyzed to determine their phylogenetic relatedness and compared with the predicted phylogenetic tree of the anopheline genomes. Of the 39 peptide hormones, six varied dramatically from the predicted tree (adipokinetic hormone 2 (AKH2), allatostatin (AST2), bursicon β (Bursβ), diuretic hormone 31 (DH31), ion transport peptide (ITP) and RYamide (RYa); Figure S21). Interestingly, four of these hormones, AKH2, AST2, Burs β and DH31 represent hormones that share a core function with another peptide hormone. Furthermore, as expected mature peptides tended to be conserved, whereas amino acids in non-retained regions of the prohormone tended to be more variable. We did not observe any consistent rearrangement of species among the six atypical phylogenetic trees. However, the five species in the *Anopheles gambiae* complex remained grouped together except in the cases of DH31 and ITP.

It is also interesting to note that the head peptide (HP) hormone was not identified in any of the anopheline genomes. In the mosquito *Aedes aegypti* HP is responsible for inhibiting the host seeking behavior of the mosquito following a blood meal (*39*). It is thought to be a paralog of short neuropeptide F in *Aedes aegypti* and its absence from any of the anopheline genomes suggests that the duplication event leading to HP occurred after anopheline mosquitoes diverged from culicine mosquitoes.

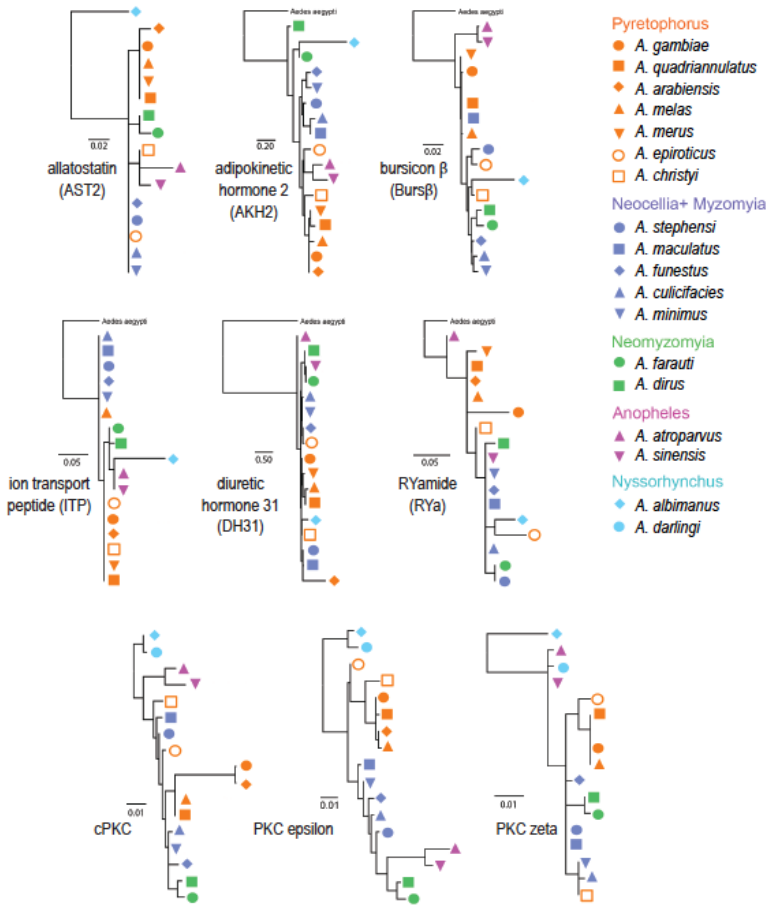


Figure S21. Neuropeptides.

Maximum likelihood trees for six peptide hormones and three PKCs with atypical phylogenies.

Insulin-like peptides, insulin signaling, Tor signaling: The insulin/insulin growth factor 1 signaling (IIS) cascade and its sister pathway, the target of rapamycin (TOR) signaling cascade, regulate diverse physiological functions in anopheline mosquitoes including innate immunity, reproduction, metabolism and lifespan (*174*). The IIS cascade is activated upon the binding of one of five insulin-like peptides (ILP1-5) encoded within all 16 anopheline genomes to the single insulin receptor ortholog. Four of five anopheline ILPs arise from an apparent radiation that resulted in all four peptides being expressed within a 6,221-20,982 bp fragment of the anopheline genomes (Figure S22). Remarkably, this genomic synteny of ILPs is strongly conserved across all 16 anopheline genomes indicating that this radiation occurred over ~100 million years ago during which time this grouping has remained intact. The fifth anopheline ILP is a more distantly related member of the insulin family and is not encoded in the single locus. Also of note, an ortholog for insulin growth factor 1 (IGF1) was not identified in any of the

anopheline genomes even though IGF orthologs have been identified in other dipterans, including the fruit fly *Drosophila melanogaster* and the culicine mosquito *Aedes aegypti* (*41, 42*).

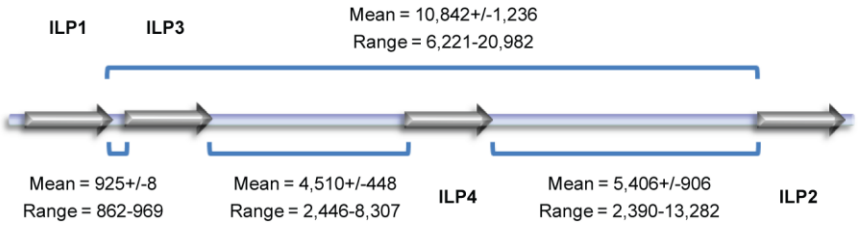


Figure S22. Conserved synteny of insulin-like peptides.

Conserved genomic synteny of four insulin-like peptides in all 16 anopheline genomes. Four of the five insulin-like peptides (ILPs) identified in the anopheline genomes are tightly linked. Organization of the ILP genes are detailed in the cartoon. Mean distances in base-pairs (± SE) between the genes are shown along with the range of distances among the 16 genomes.

Following activation of the insulin receptor by the ILPs, the IIS cascade diverges into two distinct branches, the phosphatidylinositol 3-kinase (PI3K)/Akt branch and the mitogen-activated protein kinase (MAPK) branch. Signaling proteins in these two branches are highly conserved from insects to vertebrates and are encoded with multiple representatives in the anopheline genomes. As expected, nearly all of these signaling proteins have a high degree of conservation with their *An. gambiae* and *Aedes aegypti* orthologs. Phylogenetic analysis of the PI3K/Akt and upstream MAPK signaling molecules strongly support the predicted phylogenies of the 16 anopheline species. In addition to activation by IIS, the MAPKs are activated by various other molecules including PKCs and TGFβ receptors (*175*). The genomes of all 16 anopheline species encode six PKCs that are characterized by a conserved kinase domain coupled to a series of differentially activated regulatory domains (*176*). PKCs can respond to multiple inputs with spatial and temporal specificity and can control the behavior of scaffolded protein complexes by influencing their assembly/disassembly and their subcellular localization (*177*). PKCs play a significant role in the innate immune responses of plants, insects and mammals (*178*). Specifically in mammals, a "signalosome" of PKCε, PKCδ, JNK and p38 MAPK can transduce information between mitochondria and other cellular compartments to modulate not only mitochondrial energy homeostasis but also host immunity (*179*). The best described pathway for activation of ERK is via the Ras-Raf1-MEK signaling cascade. However, PKCζ can also activate LPS-dependent macrophage MEK-ERK signaling in independently of Ras-Raf1 via enhanced ceramide synthesis or through PI3K activation. In *Drosophila*, PKCs regulate the responsiveness of photoreceptors and odorant receptors (*180, 181*), both of which are critical components of mosquito host seeking behavior as well. Phylogenetic analysis revealed that relationships

among the anopheline genes encoding Ras, MEK (MAP2K1), cPKC, PKCε, PKCζ, and JNK (MAPK8) significantly differed from the predicted anopheline species tree (Figure S21), suggesting selection associated with variation in innate immune response function via these related pathway proteins may also influence host-seeking behavior in anopheline species.

Transporters

We identified at least partial sequences orthologous to the seven known *An. gambiae* aquaporins, (AGAP010325, AGAP010326, AGAP008842, AGAP008766, AGAP008767, AGAP008843 and AGAP010878), in all 16 genome sequences, with the exception of AGAP008843 in *An. maculatus*. Rather than a deletion of the gene, it appears that the absence of this gene may be due to a gap in the sequencing coverage.

dN/dS estimates for each ortholog compared to *An. gambiae* coding sequences show all aquaporins to be under purifying selection, the average dN/dS ratio across all genes being 0.098 (AGAP010325 not included for technical reasons), with a minimum of 0.014 and maximum of 0.336. The highest dN/dS ratios were for the *An. melas* orthologs to AGAP010326 (AQP4) (dN/dS of 0.973) and AGAP008767 (*Big Brain* in *Drosophila*) (dN/dS of 0.56). No other dN/dS ratio was above 0.2. The percent identity was still above 95% for these two *An. melas* orthologs. It may be of interest that *An. melas* is described as primarily a brackish water breeder in the context of these results. Sequence homology was quite conserved across the 7 genes in the 16 species, the average identity across all was 87%. The lowest average identity across all 7 AQP genes was *An. albimanus*, at 78%. There was a trend observed where AGAP008842 (*AQP1*) consistently had the highest percent identity results with its orthologs outside of the gambiae complex. Identity was consistently 5% or more higher between AGAP008842 and its orthologs outside of the *gambiae* complex. Inside the *gambiae* complex, all genes had 98% or higher identity except in the case noted above for *An. melas AQP4*. The set of orthologs for AGAP00842 also had the lowest dN/dS on average. These results indicate that *AQP1* may be under special selection pressure to preserve its structure and/or function. *Aquaporin 1* has been shown to be important in diuresis following blood feeding, water homeostasis and humidity response and reproduction, due to expression in the ovaries (*182-184*). The aquaporins have not undergone large-scale changes, but have changed slowly under purifying selection.

Epigenetic modifiers

A *D. melanogaster* epigenetic regulatory gene set was assembled based on those genes associated with Gene Ontology (GO) terms acetyltransferase, ACG/Chrac-Complex, beta-heterochromatin, chromatin remodeling, heterochromatin, histone acetylation, histone deacetylation, histone methylation, histone demethylation, histone ubiquitination, histone deubiquitination, histone phosphorylation, Ino80 complex,

intercalary heterochromatin, Nu4A, nuclear centromeric heterochromatin, nuclear heterochromatin, NuRD complex, RSF complex, Set-N chromatin protein, telomeric heterochromatin and DNA methylation, and identified in (*185*), (*186*), (*187*), (*188*), (*189*) and (*44*). TBLASTN using *D. melanogaster* open reading frames was performed against *An. gambiae* assembly AgamP3 from VectorBase (*74*) (www.vectorbase.org), followed by reciprocal best BLAST to define orthologous genes. OrthoDB (*66*) and eggNOG (*190*) databases were used to further validate orthologous genes between *D. melanogaster* and *An. gambiae*. Sequences of putative retrogenes closely related to the *D. melanogaster effete* gene were identified using TBLASTN and were identifiable only in *An. albimanus*, *An. darlingi*, *An. sinensis*, *An. atroparvus*, *An. farauti*, *An. dirus* (2 copies), *An. funestus*, and *An. minimus*, among the species analyzed. Alignments were performed with putative open reading frames using Geneious software (version 5.53 created by Biomatters. Available from www.genious.com).

Drosophila melanogaster has served as an excellent model organism for genetic and functional genomic analysis of epigenetic regulatory genes and their control over chromatin structure, telomere remodeling and transcriptional control (*44, 45, 191, 192*). Given this extensive characterization of the epigenetic regulatory gene set in *D. melanogaster*, we identified orthologous epigenetic regulatory genes in *Anopheles gambiae*. Based on a set of 215 epigenetic regulatory genes in *D. melanogaster*, we identified 169 putative orthologous genes in *An. gambiae* (Table S32). The positive identification in *An. gambiae* of orthologs for over 75 percent of epigenetic regulatory genes found in *D. melanogaster* supports the premise that similar mechanisms of epigenetic control function in the genus *Anopheles* and the species *D. melanogaster*.

Table S32. Epigenetic regulatory genes.

Orthologous epigenetic regulatory gene set members in *Drosophila melanogaster* and *Anopheles gambiae*.

Epigenetic Regulatory Gene Class	Genes in *Drosophila melanogaster*	Orthologous Genes in *Anopheles gambiae*
Acetylation/Deacetylation	33	29
Methylation/Demethylation	43	39
Chromatin-Associated Complexes	28	25
Chromatin States/Remodeling	42	30
Ubiquitylation/Deubiquitylation	7	7
Phosphorylation/Dephosphorylation	7	5
SET-N Proteins	40	23
Miscellaneous Genes	15	11

Expansion and diversification of gene families is attributed most frequently to sequence divergence following gene duplication that results from unequal crossing over or the breakage and nonhomologous end-joining of chromosomal segments. However,

recent studies have implicated retrotransposition as a source for the genesis of functional paralogs that lead to increased gene diversity and behavioral differences in *Drosophila* (*193-195*). Among the epigenetic regulatory genes we have analyzed, the ubiquitin-conjugating enzyme E2D (orthologous to effete in *D. melanogaster*) has undergone duplication via retrotransposition (Figure 5A, main text). Orthologs of this retrogene are found in a subset of anopheline species. The presence of retrogenes in multiple subgenera among the *Anophelinae* may be consistent with the hypothesis that the initial E2D retrotransposition occurred only once after divergence of the subfamilies *Anophelinae* and *Culicinae*. If this were the case, the retrogene must have been lost within the series *Pyretophorous* and *Neocellia*, and within a subset of the series *Myzomyia*. Alternatively, the retrotransposition may have occurred independently within two or more subgenera within the subfamily *Anophelinae*. The presence of two E2D retrogenes within *An. dirus* implies that there has been either a second retrotransposition event or a conventional duplication of the E2D retrogene within this species. The inference that the retrogene persists as a functional ortholog under selective pressure is supported by the preservation of the full-length E2D open reading frame in all eight species in which it is found, with substantial sequence conservation. The identification of this apparently functional retrogene is consistent with the hypothesis that expansion of gene families through the genesis of functional retrogenes contributes to genetic diversity and phenotypic differences among rapidly divergent anopheline species.

Salivary proteins

Manual reannotation of salivary genes was done with the tool Artemis (*196*), and also using the VectorBase website (*74*). The GARD, FEL and MEME programs (*197, 198*) were run locally from programs available at the HYPHY website (*197*), and were based on orthologous (based on reciprocal blastp match and same derived blast score) alignments of *Cellia* sequences when four or more species were obtained. These alignments were treated first by the HYPHY GARD analysis for breakpoint recombination determination, the output of which was used as input to the FEL or MEME programs. Genes were classified based on known anopheline salivary proteins, on their COG (*199*) matches and on a 300 key word vocabulary, their e-value, coverage and their order of appearance on annotated matches of the SwissProt and Refseq databases, as described previously (*200*).

Saliva of blood sucking arthropods contains a complex cocktail of pharmacologically active compounds that disarms their host's platelet aggregation and blood clotting, while promoting vasodilation and immunomodulation. Salivary gland transcriptomes of *An. gambiae*¸ *An. stephensi*, *An. funestus* and *An. darlingi* have been previously described, indicating that anopheline salivary cocktails consist of near 100 polypeptides, many of which are unique to mosquitoes or to anophelines, or even to

individual species. About one half of these salivary-coding genes result from gene duplications and are found in tandem repeated families across the mosquito genome (*201*).

Saliva helps adult females to feed blood by impairing host hemostasis and affecting inflammation and immunity. The salivary potion of mosquitoes is complex and in *An. gambiae* it is estimated to contain the product of at least 75 genes, most being expressed solely in the adult female salivary glands. Perhaps due to the host immune pressure, these genes appear to be at an accelerated pace of evolution, with a significant component of gene "birth and death" (*164*). Indeed a scenario of high plasticity of salivary genes emerges following annotation of the existing 18 anopheline genomes, revealing a relatively large number of gene gains/losses involving both individual salivary genes and multi-gene families (Figure S23).

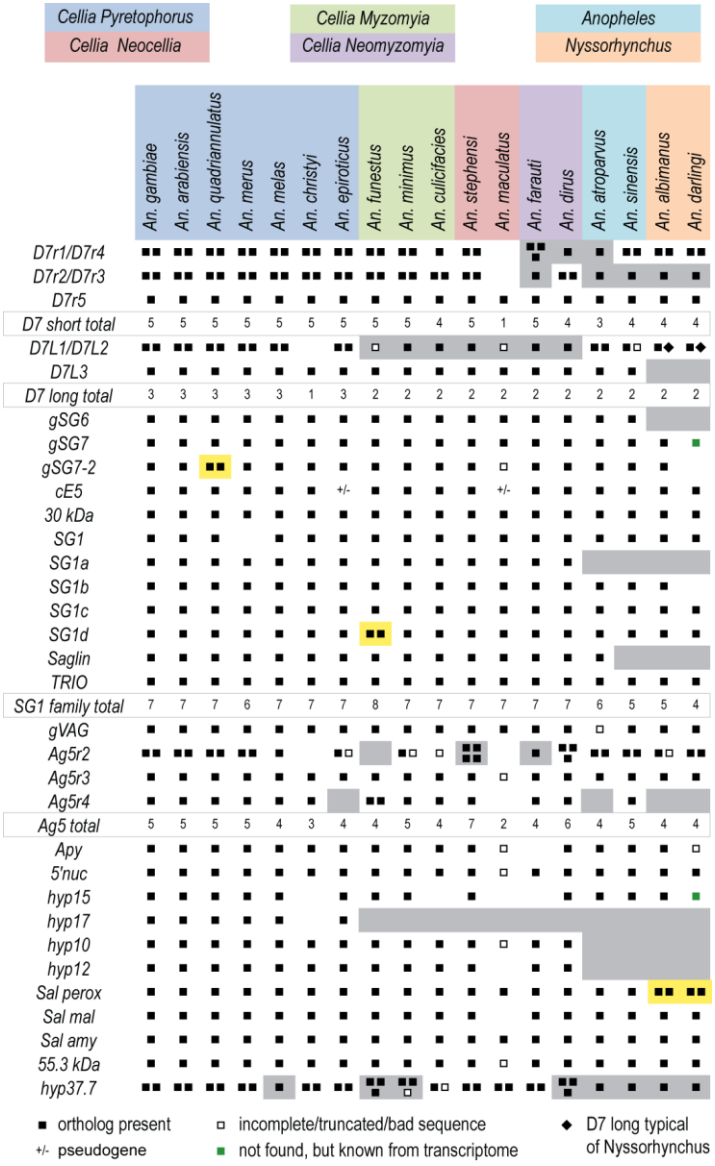


Figure S23. Salivary gland genes.

Gene gain/losses among a subset of salivary gland expressed genes in adult mosquitoes. Likely gene gain/loss events are highlighted in grey, and in yellow for sporadic cases. Empty cells indicate that the gene is apparently absent but there is no clear evidence of gene loss.

A limited set of salivary genes were previously studied and observed to carry positive selection signatures (*46*). With the availability of additional genomes we used ortholog alignments of the Cellia subgenus to analyze ~ 1,000 genes from seven different classes to provide for an independent evaluation of the unique evolutionary characteristics of salivary coding genes using the HYPHY package (*197*). Both fixed effect likelihood (FEL) analysis as well as a mixed effects model of evolution (MEME) analysis showed that salivary gland genes are among those having the highest rate of positively selected codons among seven gene classes, further corroborating the estimates that salivary gland genes are at an accelerated pace of evolution.

Gene gain/loss among anophelines: Based on a previous *An. gambiae* catalogue (*202*) , 36 salivary proteins were selected and used to search orthologs in the genome of the anopheline species analyzed here. Events of salivary gene gain/loss were previously reported within the family *Culicidae*: in fact the saliva of *Anophelinae* carries several subfamily-specific proteins, presumably originated after the separation from *Culicinae*, around 120-150 mya (that is the case for SG6, SG7/SG7-2, cE5/anophelin, hyp10/hyp12, hyp15/hyp17 among those included in Figure S23; (*201*)). Mechanisms as different as horizontal transfer, gene duplication and "gene scrambling" may have contributed to their appearance, and certainly played more general roles in salivary gene evolution. The comparative analysis within the subfamily *Anophelinae* points up the highly dynamic nature of salivary genes, revealing events of gene gain/loss involving both single genes and multi-gene families at all taxonomic levels (Figure S23). Several examples of salivary gene gain/loss are highlighted in Figure S23 and some representative cases are shortly discussed below:

(1) A few salivary genes, i.e. *SG6*, *D7L3* and the *SG1* family member *saglin*, were not found in the *Nyssorhynchus* species *An. albimanus* and *An. darlingi* but were present in Old World anophelines, suggesting they were possibly "invented" after the separation from New World anophelines, around 80-100 mya. *Saglin* is known to play a role in salivary gland invasion by *Plasmodium* sporozoites (*203*) and its absence from the main malaria vectors in Central and South America suggests redundancy of salivary gland receptor function and the potential involvement of other SG1 family proteins. The *An. gambiae* SG6 protein has been used as marker of human exposure to Afrotropical malaria vectors (*204*) and the relatively high conservation (>50% identity among anophelines) supports its possible exploitation as more generalized indicator of human exposure to Old World malaria vectors.

(2) The SG1a and hyp10/hyp12 were not found in species of the *Anopheles* and *Nyssorhynchus* subgenera suggesting they represent events of gene gain followed by gene duplication that happened sometime in the lineage leading to *Cellia*.

(3) The hyp15/hyp17 family is a more specific case, with one of the two family members (*hyp15*) widely spread among anophelines and the other (*hyp17*) restricted to *Pyretophorus* species, most likely as a result of a gene duplication.

(4) Multi-gene families, especially D7 and Ag5 families, underwent more complex events of expansion/contraction. The D7 family encodes proteins with anti-hemostatic and anti-inflammatory functions which bind biogenic amines and inhibit activation of kallikrein (*205, 206*). *An. gambiae* and the other members of the complex have a well conserved D7 cluster including five short and three long *D7* genes. Based on relative divergence and to simplify classification some members of the family (D7r1/D7r4, D7r2/D7r3 and D7L1/D7L2) have been paired in Figure S23. Multifaceted examples of gene gain/loss are found in the D7 family. *Cellia* species of the *Myzomyia*, *Neocellia* and *Neomyzomyia* series they all carry one member of the D7L1/D7L2 pair whereas both are found in all other anophelines. The D7L3 gene seems to be lost in the *Nyssorhynchus* species. As far as the D7 short cassette is concerned, different events of gene gain/loss involving the D7r1/D7r4 and D7r2/D7r3 pairs were found in *Neomyzomyia* and in members of the *Anopheles* and *Nyssorhynchus* subgenus. A similar situation was found in the Ag5 family where the variation in copy number (from one to four) of *Ag5r2* is especially striking.

(5) Also a smaller family as the hyp37.7 went through dynamic expansion/contraction in anophelines. Only one gene could be traced in the two species of the *Anopheles* subgenus *An. atroparvus* and *An. sinensis* whereas all other species appear to carry at least two genes with clear cases of loss in *An. melas*, *An. albimanus* and *An. darlingi*.

(6) Finally, some isolated duplications, regarding single genes/species were also detected. That is the case for *SG7-2* in *An. quadriannulatus*, *Sg1d* in *An. funestus* or the salivary peroxidase in *An. albimanus* and *An. darlingi*.

Evolutionary signatures among different gene classes reveal high degree of positive selection in salivary gland genes: The diversity and divergence of mosquito salivary proteins suggest their genes might be at an accelerated pace of evolution, and indeed signatures of positive selection were found in a limited set of salivary gland genes from *An. gambiae* (*46*). The completion of additional anopheline genomes allows the use of orthologous protein alignments from various species for powerful evolutionary analysis, such as those on the HYPHY/FEL/MEME package (*197, 198*). We used ortholog alignments of the *Cellia* subgenus to analyze 1,055 genes from seven different classes (more were not used due to excessive computer time involved) to provide for an independent evaluation of the unique evolutionary characteristics of salivary coding genes. Fixed effect likelihood (FEL) analysis (Figure S24) showed salivary gland genes

to have the highest rate of positively selected codons among the functional classes analyzed, while the mixed effects model of evolution (MEME) (*197, 198*) (Figure S24) showed that salivary gland genes had a similar number of positively selected codons as the immune genes, further corroborating the estimates that salivary gland genes are at an accelerated pace of evolution.

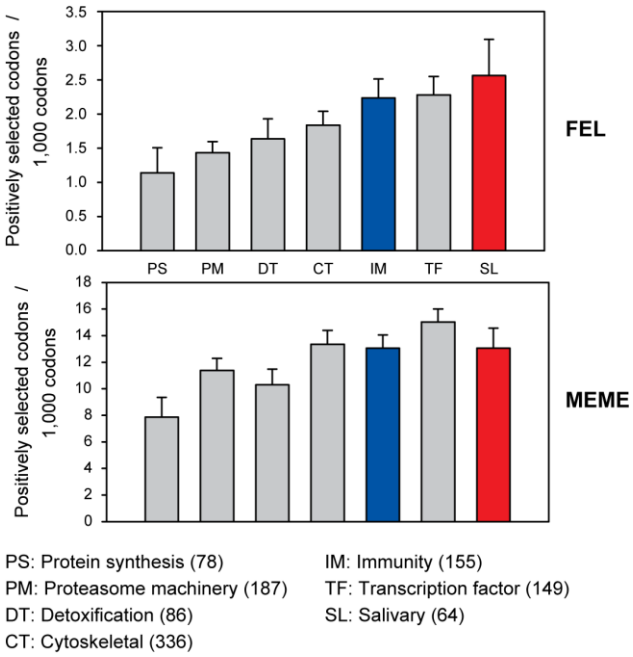


Figure S24. Salivary gene sequence analysis.

Fixed effect likelihood (FEL, top) and mixed effects model of evolution (MEME, bottom) analysis of 1,056 anopheline genes of the *Cellia* subgenus according to different functional classes.

Insecticide resistance

Representatives of the *Pyretophorus* group - *An. epiroticus* and four members of the *gambiae* complex (*An. gambiae*, *An. arabiensis*, *An. merus*, *An. quadriannulatus*) together with *An. stephensi* (*Neocellia*), *An. funestus* (*Myzomyia*), *An. sinensis* (*Anopheles*) and *An. albimanus* (*Nyssorhynchus*) were studied. CYP and GST genes were identified in the genomes, using a tBLASTn approach (no e-value cut off) with default options and PEST CYP and GST protein sequences as queries. Input files included all known glutathione-S transferase sequences and splice variants from PEST, and all known cytochromes P450 from PEST supplemented with additional, or completed P450 sequences from the annotation of Ranson et al. (*207*). Prior to BLAST searches we checked for additional incorrectly labeled P450 and GSTs with IPR identifiers IPR004045 (Glutathione S-transferase, N-terminal), IPR010987 (Glutathione S-transferase, C-terminal) and IPR001128 (Cytochrome P450). Resulting high-scoring

segment pairs (HSPs) were clustered based on their coordinates on the genomes using in-house Perl scripts. Subsequently, HSP clusters were used to annotate CYP and GST genes in the genomes aided by automated gene prediction models conducted using Augustus (*71*) with default settings. Gene models were further verified by performing reciprocal BLASTn searches against the non-redundant nucleotide database of NCBI. Pseudogenes (gene models with one or two frameshifts, late start codon or premature stop-codon compared to the queries) and gene fragments were separated from putative full length CYP and GST coding sequences. Full length CYP (name: [species]CYPxxx) and GST (name: [species]GSTxxx) genes were translated into protein sequences.

Malaria vector control of anophelines relies upon insecticides applied to bednets or the internal walls of buildings where mosquitoes rest. The availability of only four insecticidal classes which share just two target-sites has been a major factor in the development of resistance in vector mosquitoes. Resistance to all four classes of public health approved insecticides is widespread - especially in the sub-Saharan African vectors of the *Anopheles gambiae* group (*208*) and in *An. funestus* but is also present in other vector species (e.g. *An. epiroticus*, *An. stephensi*, *An. dirus* and *An. sinensis*). The best characterized insecticide resistance mutations are in target-sites which have arisen multiple times (*209*) and confer selective advantage to mosquitoes resulting in a rapid increase to high frequency (*210, 211*). Metabolic resistance arising as a result of allelic variation or over-expression of enzymes involved in detoxification or sequestration of insecticides has been detected in multiple populations of those species with pre-existing genomic resources to facilitate the development and deployment of microarrays. Enzymes of the cytochromes P450 and glutathione-S transferase classes are often found up-regulated in insecticide resistant populations (*212, 213*). In addition, there is evidence that non-synonymous mutations in particular members of these classes can facilitate resistance (*214, 215*). A previous study using a primer-walking approach to sequence through the insect specific glutathione-S transferase epsilon cluster detected gene gain, loss, pseudogenization and positive selection on particular branches (*216*). Here we have developed a new analysis pipeline, necessarily heavily dependent upon manual curation, to extend and develop this study to examine the evolution of the glutathione-S transferases and cytochromes P450 throughout the genus *Anopheles* which is crucial for application in insecticide resistance studies.

Both P450s and GSTs are composed of closely related gene families often found clustered within the genome (e.g. (*207, 216*)). Such organization has led to difficulties with automated gene identification with manual annotation necessary for these gene classes in all species, including *An. gambiae* where a number of cytochrome P450 models were revised (see Table S33 for an example of the detailed revisions required in two species - *An. stephensi* and *An. quadriannulatus* - though these figures are generalizable across the genus).

Table S33. Cytochrome P450 and GST gene annotations.

Differences between automated (MAKER) annotations of cytochrome P450 and GST families in *An. quadriannulatus* (VectorBase AQUA gene models) and *An. stephensi* (VectorBase ASTE gene models). Only 28 cytochrome P450 automated P450 models in *An. stephensi* were deemed correct whilst 35 models were revised, resulting in 65 manually annotated genes. One additional P450 was annotated where no gene model was predicted.

Species	Curation Status	P450	GST
An. quadriannulatus	Untouched VectorBase AQUA annotations	26	9
	Modified VectorBase AQUA annotations	35 (68)	12 (18)
	No VectorBase AQUA annotations	0	1
An. stephensi	Untouched VectorBase ASTE annotations	28	5
	Modified VectorBase ASTE annotations	35 (65)	13 (22)
	No VectorBase ASTE annotations	1	0

In total we manually annotated 1,048 complete gene models (776 CYP450s and 272 GSTs plus additional partial annotations and pseudogenes - see Table S34). Overall, gene number is highly conserved with only limited examples of lineage specific duplication or losses. Orthologs of those genes repeatedly implicated in insecticide resistance through up-regulation or allelic variation in insecticide resistant populations (e.g. Cyp6m2, Cyp6p3 (=Cyp6p9 in *An. funestus*), Gste2, Gste4) are found in all studied species suggestive that potentially all species are pre-adapted to develop resistance. Additional partial gene sequences have been identified, typically at scaffold edges, with some pseudogenes evident. Individual gene gains/losses are observed e.g. the absence of Gste1 seen previously in *An. plumbeus* and *An. darlingi* (*216*). The absence of Gste1 in *An. albimanus* but presence in all other species examined is supportive of gene gain in the Cellia lineage. Cyp18a1, a gene conserved throughout the Insecta (*49*) and essential for ecdysteroid inactivation during molting, but not previously detected in *An. gambiae* PEST (*217*) was also not detected in any member of the *An. gambiae* complex, but was found in all other species examined. This gene is found in 3'-3' orientation with Cyp306a1 in all insect species in which both genes were annotated (*218*). Whilst Cyp306a1 is found in the *An. gambiae* complex no proximal copy of Cyp18a1 was seen in any species. These genes are located centromerically on chromosome 2R in *An. gambiae* and absence of this gene may reflect the difficulty in assembly of this region. However, not only do we not find Cyp18a1 in any member of the *An. gambiae* complex whilst detecting it in all other species examined, but we have searched the available genomic and transcriptomic Sequence Read Archives (SRA) and whilst Cyp18a1 can be detected in SRAs of *An. funestus*, *An. stephensi* and *An. albimanus* we are unable to find any match to this gene in any SRA archive or RNA-Seq archive of members of the *An. gambiae* complex. Through BLAST search of the *An. christyi* sequence we identified both Cyp18a1 and Cyp306a1 indicating that loss of Cyp18a1 is specific to the *An. gambiae* complex.

Whilst the evidence shows Cyp18a1 is absent in the *An. gambiae* lineage, further experimentation is necessary to identify which gene(s) has been co-opted to perform this essential function. Click probes (*219*) based on the ecdysteroid targets of Cyp18a1 may allow identification of this gene which would be an excellent demonstration of how comparative genomics may lead to the identification of potential novel targets for vector control.

Table S34. CYP450s and GSTs in *An. gambiae* and seven other anophelines.

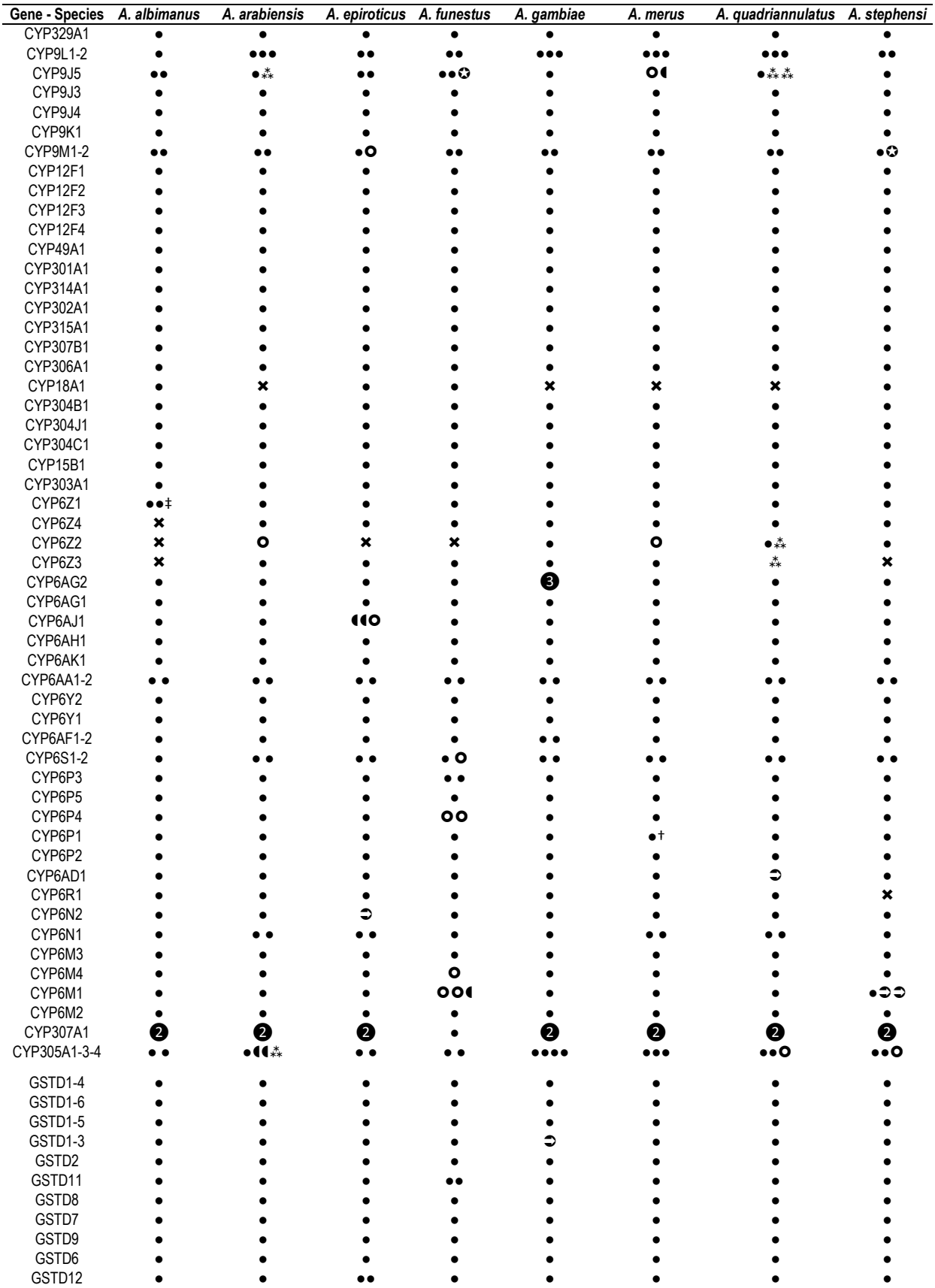
Table symbols:

●	gene annotated	❷	2 alternative transcripts (etc.)
✖	not found	➲	frameshift
●●	duplicated	✪	premature stop-codon
O	sequence incomplete due to internal gap	⁂	fragments of sequence
◖	Sequence incomplete - next to a gap or scaffold start	†	missing exonic region in sequence (deletion)

* CYP4D22 is incorrectly labeled as a P450, manually annotated but removed from subsequent analyses

‡ *An. albimanus* CYP6Z genes were not clearly ascribable to a single CYP6Z (see tree)

Gene - Species	*A. albimanus*	*A. arabiensis*	*A. epiroticus*	*A. funestus*	*A. gambiae*	*A. merus*	*A. quadriannulatus*	*A. stephensi*
CYP4C26	●	●	●	●	●	●	●	●
CYP4C25	●	●	●	●	●	●	●	●
CYP4C37	●	●	●	●	●	●	●	●
CYP4C28	●	●	●	●	●	●	●	●
CYP4C35	●	●	●	◖	●	●	●	●
CYP4C36	✖	●	●	●	●	●	●	●
CYP4C27	●	●	●	●	●	●	●	●
CYP4D22*	➲	●	●	●	●	●	●	●
CYP4D15	●	●	●	●	●	●	●	●
CYP4D16	●	O	●	●	●	●	●	●
CYP4D17	●	O	●	●	●	●	●	●
CYP4H19-24	●●	●●	●O	OO	●●	●●	●●	●●
CYP4H14	●	●	●	●	●	●	●	●
CYP4H18	●	●	●	●	●	●	●	●
CYP4H16	✖	●	●	●	●	●	●	●
CYP4H17	●●	●	●	●	●	●	●	●
CYP4H26	●	●	●	●	●	●	●	●
CYP4H15-25-27	➲OO	●(H15) ●(H25) ●(H27) O(H27)	●(H15) ●(H25)	●(4H25)	●(H15)●(H25) ●(H27)	●(H15)●(H27) O(H25)	●(H15) ●(H25) ●(H27) O(H27)	●(4H25)
CYP4K2	●	●	●	●	●	●	●	●
CYP4AR1	◖✪	●	●	●	●	●	●	●
CYP4J5	●	●	●	●	●	●	●	●
CYP4J10	●	●	●	●	●	●	●	◖
CYP4J9	●	●	●	●	●	●	●	●
CYP4AA1	●	●	●	●	●	●	●	●
CYP4G17	●	●	●	●	●	●	●	●
CYP4G16	●	❷	●	●	❹	●➲	❷	●
CYP325F2	●	❷	●	●	❷	❷	❷	●
CYP325F1	✖	●	✖	●	●	●	●	●
CYP325G1	●	●	●	●	●	➲	●	●
CYP325H1	●	●	●	●	●	●	●	●
CYP325J1	●●	●	●	O	✪	●	●	●
CYP325A1-2-3	●	●●◖	●	●O	●●●	◖◖◖	●◖◖	●
CYP325K1	●	●	●	●	●	●	●	●
CYP325E1	●	●	●	●	●	●	●	●
CYP325B1	✖	●	◖	●◖	●	●	●	●
CYP325D1	✖	●	●	●	●	✖	●◖	●
CYP325C1-2-3	OOOO	●●⁂⁂	●◖◖	●	●●●	●⁂⁂	●◖O⁂⁂	●
CYP615P	●	●	●	O	●	●	●	●


Gene - Species
A. albimanus
A. arabiensis
A. epiroticus
A. funestus
A. gambiae
A. merus
A. quadriannulatus
A. stephensi
CYP329A1
CYP9L1-2
CYP9J5
CYP9J3
CYP9J4
CYP9K1
CYP9M1-2
CYP12F1
CYP12F2
CYP12F3
CYP12F4
CYP49A1
CYP301A1
CYP314A1
CYP302A1
CYP315A1
CYP307B1
CYP306A1
CYP18A1
CYP304B1
CYP304J1
CYP304C1
CYP15B1
CYP303A1
CYP6Z1
CYP6Z4
CYP6Z2
CYP6Z3
CYP6AG2
CYP6AG1
CYP6AJ1
CYP6AH1
CYP6AK1
CYP6AA1-2
CYP6Y2
CYP6Y1
CYP6AF1-2
CYP6S1-2
CYP6P3
CYP6P5
CYP6P4
CYP6P1
CYP6P2
CYP6AD1
CYP6R1
CYP6N2
CYP6N1
CYP6M3
CYP6M4
CYP6M1
CYP6M2
CYP307A1
CYP305A1-3-4
GSTD1-4
GSTD1-6
GSTD1-5
GSTD1-3
GSTD2
GSTD11
GSTD8
GSTD7
GSTD9
GSTD6
GSTD12

Gene - Species	*A. albimanus*	*A. arabiensis*	*A. epiroticus*	*A. funestus*	*A. gambiae*	*A. merus*	*A. quadriannulatus*	*A. stephensi*
GSTD3	●	●	●	●	●	●	●	●
GSTD4	✖	●	●	●	●	●	●	✖
GSTD10	●	●	●	●	●	●	●	●
GSTD5	●	●	●	●	●	●	●	●
GSTU3	●	●	●	●	●	●	●	●
GSTU2	●	●	●	●	●	●	●	●
GSTE8/GSTU4	●	●	●	●	●	●	●	●
GSTE6	●	●	●	●	●	●	●	●
GSTE2	●	●	●	●	●	●	●	●
GSTE1	✖	●	●	●	●	●	●	●
GSTE7	●	●	●➲	●	●	●	●	●
GSTE3	●	●	●●	●	●	●	●	●
GSTE4	●	●	●	●	●	●	●	●
GSTE5	●	●	●	●	●	●	●	●
GSTU1	●	●	●	●	●	●	●	●
GSTT2	●	●	●	●	❷	●	●	●
GSTT1	●	●	●	●	●	●	●	●
GSTO1	●	●	●	●	●	●	●	●
GSTZ1-1	●	●	●	●	●	●	●	●
GSTZ1-2	✖	●	●	●	●	●	●	●
GSTZ1-3/4	✖	●	●	●	❷	●	●	●
GSTS1-1	●	●	●	●	●	●	●	●
GSTS1-2	●	●	●	●	●	●	●	●

Immunity

To examine the evolutionary characteristics of immune-related genes across the anopheline phylogeny, orthologs of cataloged immunity genes from previous studies (*50, 51, 220*) were retrieved from OrthoDBmoz2 (http://cegg.unige.ch/orthodbmoz2). These lists of orthologs were supplemented by searches for genes with protein domains characteristic of particular immune-related gene families, e.g. IPR002181, the fibrinogen domain that defines the fibrinogen related proteins (FREPs) or IPR023796, the serpin domain that defines the serine protease inhibitors (SRPNs). Manual curation of the resulting lists revised gene counts where erroneous automated gene predictions had resulted in gene model fusions or fissions, or where no gene had been predicted but a likely ortholog could be identified with sequence searches of the genome assemblies. The examined pathways include the Imd, Jak/Stat, and Toll pathways, expanded to include genes recently identified as pathway members particularly from RNAi screens in *Dr. melanogaster*. The examined gene families include anti-microbial peptides (attacins, cecropins, defensins, gambicins etc.), autophagy-related proteins, caspases and their inhibitors, catalases, CLIP-domain serine proteases (families A-E), C-type lectins, fibrinogen-related proteins, galectins, gram-negative binding proteins, inhibitors of apoptosis, leucine-rich repeat immune proteins, lysozymes, MD2-like proteins, nimrod-related proteins, peptidoglycan recognition proteins, prophenoloxidases, peroxidases (heme, glutathione, and thioredoxin), scavenger receptors (A, B, and C-types), supoeroxide dismutases, spaetzle-like proteins, serine protease inhibitors, small regulatory RNA pathway members, thioester-containing proteins, and toll-like receptors.

Mosquitoes are largely refractory to malaria parasites - only relatively few species are competent vectors - and this resistance is mainly attributable to the mosquito's immune responses to the invading parasite (*221*). Molecular and cell biology research has focused on the roles of genes known to be important in immunity from studies in *Drosophila melanogaster* and others identified as key mosquito agonists and antagonists of parasite development that together make up the canonical mosquito immune repertoire (*50, 51, 220*). Some of the best-studied immune-related factors include thioester-containing proteins, TEPs, (*222-225*), leucine-rich repeat immune proteins, LRIMs, (*42, 226-232*), fibrinogen-related proteins, FREPs, (*233, 234*), clip-domain proteases, CLIPs, (*235-238*), serine protease inhibitors, SRPNs, (*239, 240*), and C-type-lectins, CTLs, (*236, 241*). The results of such functional studies have prompted further investigations into the evolutionary signatures of many of these immune-related genes, often through targeted DNA sequencing from different *An. gambiae* populations as well as other anopheline species (*242-248*). The sequencing of multiple anopheline genomes (*8*) now allows for genome-wide comparative phylogenomics studies to examine the complete catalog of immune-related genes across many millions of years of mosquito evolution.

The framework of insect innate immunity that classifies immune-related genes and families into broad functional categories - recognition, signal transduction, modulation, and effector components - is supported by many years of *Drosophila melanogaster* research and complemented by more recent explorations of non-model organisms, especially in *An. gambiae* (*50, 51, 220*). The canonical immune-related gene repertoire is composed of genes that have been directly implicated in immune responses through experimental research in different insect systems, or they have been indirectly linked to immunity through homology to known immune proteins. Classification according to this framework facilitates the examination of each of the distinct phases and thereby helps to recognize their defining characteristics and how they combine to provide protection from many different challenges. This framework has proven to be robust and informative, and it continues to be used as the basis for dissecting the immune system into inter-related functional components.

The catalog of immune-related genes of the 19 anopheline and 2 culicine mosquitoes, as well as from *Drosophila melanogaster*, encompasses a total of more than 9,000 genes from 28 different gene families or pathways (Table S35). The anophelines average about 400 immunity genes each, with only *An. sinensis* having more identified immunity genes than *An. gambiae*, a difference mainly attributable to the presence of at least 120 FREPs - about 3 times as many as *An. gambiae*.

Table S35. Catalog of immune-related genes and gene families.

Gene counts are shown for each gene family or pathway for each of the 21 mosquitoes and *Drosophila melanogaster*. The gene families and pathways include: AMP, anti-microbial peptides; APHAG, autophagy-related proteins; CASP, caspases; CASPA, caspase inhibitors; CAT, catalases; CLIP, CLIP-domain serine proteases; CTL, C-type lectins; FREP, fibrinogen-related proteins; GALE, galectins; GNBP, gram-negative binding proteins; IAP, inhibitors of apoptosis; IMDPATH, Imd pathway members; JAKSTAT, Jak/Stat pathway members; LRIM, leucine-rich repeat immune proteins; LYS, lysozymes; ML, MD2-like proteins; NIMROD, nimrod-related proteins; PGRP, peptidoglycan recognition proteins; PPO, prophenoloxidases; PRDX, peroxidases; SCR, scavenger receptors; SOD, supoeroxide dismutases; SPZ, spaetzle-like proteins; SRPN, serine protease inhibitors; SRRP, small regulatory RNA pathway members; TEP, thioester-containing proteins; TOLL, toll-like receptors; TOLLPA, Toll pathway members.

Family	*A. gambiae*	*A. merus*	*A. arabiensis*	*A. quadriannulatus*	*A. melas*	*A. christyi*	*A. epiroticus*	*A. stephensi* (Indian)	*A. stephensi* (SDA-500)	*A. maculatus*	*A. culicifacies*	*A. minimus*	*A. funestus*	*A. dirus*	*A. farauti*	*A. atroparvus*	*A. sinensis*	*A. albimanus*	*A. darlingi*	*Culex quinquefasciatus*	*Aedes aegypti*	*Drosophila melanogaster*	Totals
AMP	11	5	9	8	3	4	7	5	11	1	2	8	9	8	5	5	8	9	6	8	18	21	171
APHAG	20	21	20	19	20	19	20	19	19	15	19	20	20	19	20	19	20	17	19	19	21	22	427
CASP	15	7	9	10	6	8	7	9	10	9	9	9	11	8	9	7	7	9	5	15	11	7	197
CASPA	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	2	3	5	33
CAT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	23
CLIP	86	75	61	63	79	57	58	55	61	64	71	53	59	67	75	73	95	58	65	104	92	49	1520
CTL	28	23	20	22	22	21	16	26	25	24	23	26	20	27	25	19	19	15	18	56	45	39	559
FREP	42	35	38	15	33	27	23	26	22	19	30	26	19	27	27	50	120	36	20	84	32	14	765
GALE	10	9	8	8	8	8	9	9	8	6	9	9	8	8	9	9	9	8	9	11	12	6	190
GNBP	7	6	4	5	6	5	5	5	5	7	5	6	5	5	5	5	6	4	5	11	7	3	122
IAP	8	4	6	6	7	4	4	4	4	2	4	4	5	5	5	4	5	4	5	6	5	4	105
IMDPATH	17	17	16	17	17	16	16	16	16	15	17	17	17	17	17	16	15	16	17	18	19	19	368
JAKSTAT	4	3	3	3	3	3	3	4	4	3	4	3	4	3	3	4	3	3	3	5	3	3	74
LRIM	28	21	20	19	25	18	22	23	19	21	21	20	22	25	30	23	25	22	21	31	29	0	485
LYS	8	6	3	3	3	3	3	4	3	4	4	3	3	4	5	5	5	3	3	7	8	13	103
ML	16	14	10	8	14	8	9	10	11	11	12	9	13	11	13	14	13	10	12	19	24	9	270
NIMROD	6	5	4	5	6	4	3	5	4	7	6	4	4	4	6	5	8	5	6	8	5	12	122
PGRP	7	7	7	6	7	5	6	7	6	5	6	7	5	6	7	7	6	6	7	8	8	13	149
PPO	18	17	12	11	16	8	11	9	9	18	16	5	9	8	15	17	17	9	11	22	21	10	289
PRDX	24	25	21	24	30	20	23	21	21	29	27	21	19	19	24	24	26	22	23	20	21	20	504
SCR	21	23	22	23	26	21	25	25	23	28	26	22	21	21	20	21	21	19	22	26	21	24	501
SOD	6	6	5	7	6	5	6	6	6	6	6	6	6	6	6	6	6	5	5	8	10	6	135
SPZ	6	6	6	6	6	2	6	6	6	5	6	6	6	6	6	6	7	6	5	7	9	6	131
SRPN	19	17	13	13	19	12	15	16	17	20	16	13	15	15	17	23	20	19	19	38	23	30	409
SRRP	29	28	21	29	33	26	29	30	30	27	32	32	30	32	32	30	30	29	28	36	36	29	658
TEP	13	11	11	13	11	5	8	6	5	9	12	13	4	16	10	10	22	16	11	10	8	5	229
TOLL	13	12	13	13	12	9	10	8	8	14	9	8	9	8	9	10	10	8	8	10	13	10	224
TOLLPA	19	19	18	19	20	16	19	19	19	18	19	18	18	18	19	18	18	17	16	20	20	22	409
Totals	484	424	383	377	440	336	365	375	374	389	413	370	363	395	421	433	544	377	371	610	525	403	9172

From the complete catalog we focused on genes involved in classical recognition, signal transduction, cascade modulation, effector enzymes, and a selection of other immune-related gene families, to compare their evolutionary traits. The orthologous group evolutionary traits examined include: (1) the evolutionary rate as measured by the average normalized ortholog protein sequence percent identity as computed by OrthoDB (*66*) - from fast-evolving orthologs with high sequence divergence to slow-evolving orthologs with low sequence divergence; (2) the universality as measured simply by the number of species with an ortholog - from universal orthologs present in all species to lineage-specific orthologs present in few species; and (3) the duplicability as measured simply by the average number of genes per species - from single-copy orthologs with only one gene per species to multi-copy orthologs with many gene duplications. Comparing these evolutionary traits across the different categories of immune-related genes reveals considerable variation that clearly distinguishes the different immune phases (Figure S25). Firstly, compared to a background of all orthologous groups with *An. gambiae* genes, those implicated in immunity show, on average, elevated evolutionary rates and more gene duplications, and are slightly more lineage-specific. However, breaking this down into its constituent parts reveals much more detailed patterns: classical recognition genes exhibit relatively low levels of sequence divergence and effector enzymes are in fact highly conserved at the sequence level; signal transducers rarely duplicate, are found universally across all species, and show elevated rates of sequence divergence; cascade modulators also exhibit higher rates of sequence divergence but they are much more lineage-specific and generally have more gene duplications. The leucine-rich repeat immune proteins (LRIMs) and the thioester-containing proteins (TEPs) exhibit high sequence divergence but while the LRIMs are found across most mosquitoes the TEPs are more lineage-specific and have duplicated more.

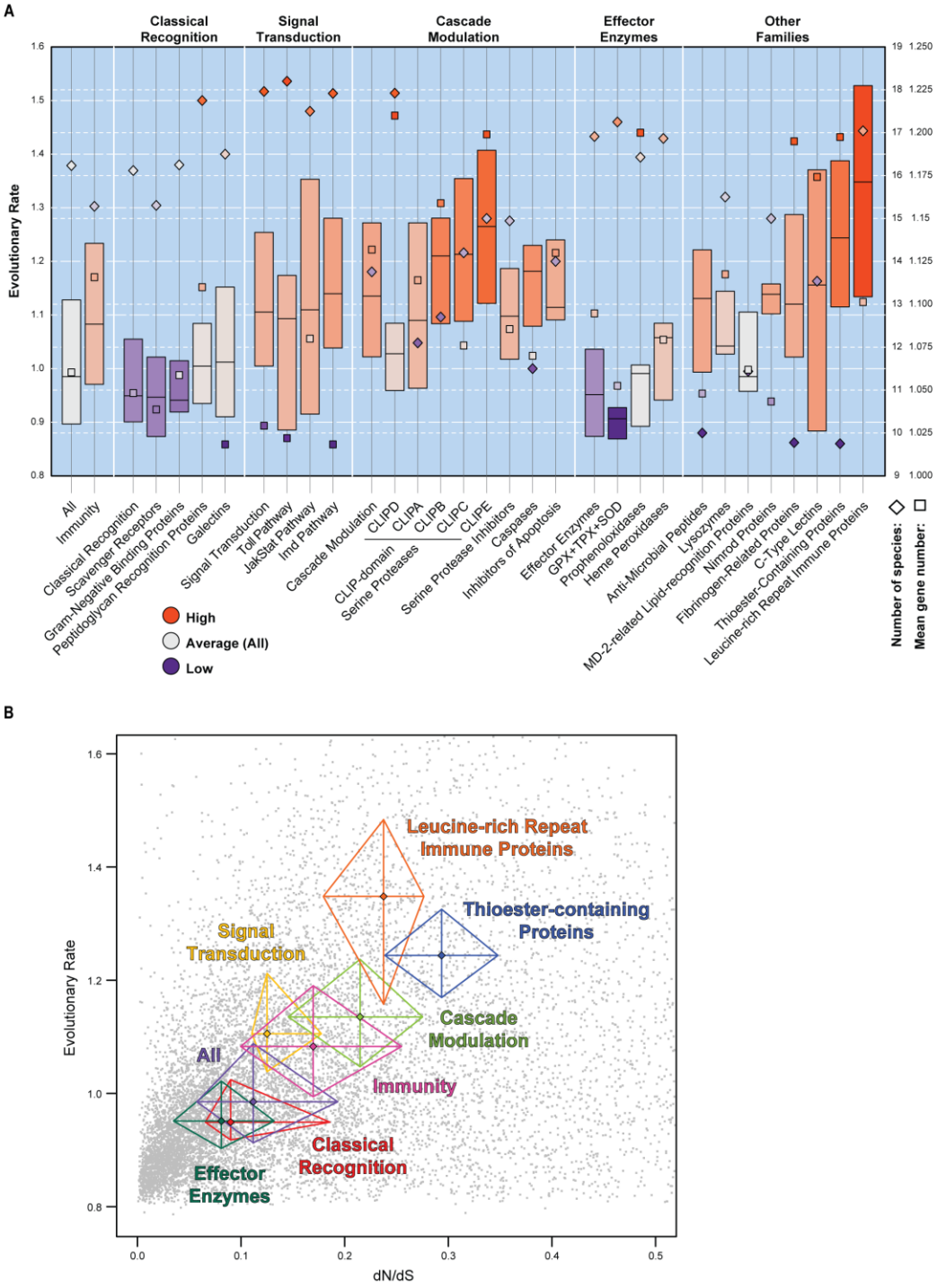


Figure S25. Evolutionary characteristics immune genes.

A. For each category of immune genes, boxplots show medians of orthologous group evolutionary rates with the limits of the upper and lower quartiles, diamonds show the mean number of species, and squares show the mean number of genes per species. Boxplots, diamonds, and squares are colored from dark purple (lower than the average for all orthologous groups) to grey (similar to the average for all orthologous groups) to dark orange (higher than the average for all orthologous groups). **B**. Evolutionary rates versus dN/dS ratios of different categories of immune genes. Colored

diamonds represent the median evolutionary rate and dN/dS ratio values for each category of immune genes, and for all immunity genes and for all genes. The 'kites' show the limits of the 30th and 70th percentile values in each case. Grey points in the background show values for all orthologous groups individually, within the plotted area.

Using the dN/dS ratios computed across the *Pyretophorus* species for all anopheline orthologous groups (see selection section) we compared the levels of selective constraint of the canonical immune phases as well as the LRIMs and the TEPs with their levels of sequence divergence (Figure S25). Although the LRIMs generally show higher evolutionary rates in terms of sequence divergence, the TEPs exhibit higher dN/dS ratios. The immune genes as a whole show higher rates of sequence evolution and higher dN/dS ratios than the background of all orthologous groups, but this trend is driven by the signal transducers and modulators while the effector enzymes and classical recognition genes generally show more conservative patterns of sequence evolution.

An intronless signal transducer and activator of transcription (STAT) gene, called *STAT1* (also called *Ag-STAT* or *AgSTAT-B*) was cloned from *An. gambiae* and found to be activated in response to bacterial challenge (*52*). Sequencing of the *An. gambiae* genome later revealed a second STAT gene, *STAT2* (also called *AgSTAT-A*), a multi-exon gene and the likely progenitor of the *STAT1* retrogene. Genomic searches of the mosquito genomes reveal that while the multi-exon *STAT2* gene is present in all species, the *STAT1* retrogene is present in all *Cellia* species except *An. dirus* and *An. farauti* (Table S36). Thus the retrotransposition event that lead to the generation of *STAT1* occurred after the divergence of *An. dirus* and *An. farauti* and the gene has been maintained in all descendent lineages. Phylogenetic analyses using neighbor-joining tree estimation from the nucleotide alignments guided by protein alignments indicate that *STAT1* genes are much more divergent than *STAT2* genes and an independent retrotransposition event created a retrogene copy in *An. atroparvus* (Figure 5C, main text). The STAT pathway has been shown to mediate late-phase immunity against *Plasmodium* in the *An. gambiae* (*53*) and to control *Plasmodium vivax* load in early stages of *An. aquasalis* infection (*54*). Thus, the generation of a second STAT gene may have had profound consequences on the regulatory networks governing immune responses in this subset of anophelines.

Table S36. Genomic locations of mosquito STAT genes.

Locations of BLAST hits of *An. gambiae* signal transducer and activator of transcription (STAT) genes.

Gene ID	Assembly	Scaffold/Chrom.	Start	End	Comment
		STAT1			
AGAP010423	AgamP3	3L	3092566	3094731	retrogene
None	AgamS1	scf_1106392397099	1088213	1090393	retrogene
None	AgamM1	scf_1925491352	1195389	1197551	retrogene
None	AmerM1	KI439491	562	2739	retrogene
None	AaraD1	KB704574	1155611	1157794	retrogene
None	AquaS1	KB665799	479182	481362	retrogene
None	AmelC1	AXCO01019596	1721	3898	retrogene
ACHR010209	AchrA1	KB682195	3	1298	retrogene
None	AepiE1	KB671674	62198	64408	retrogene
ASTEI09465	AsteI2	scaffold_00093	221615	223681	retrogene
ASTE004392	AsteS1	KB664688	904470	906653	retrogene
AMAM004727	AmacM1	AXCL01013607	1169	3358	retrogene
ACUA024657	AculA1	KI422507	89730	91847	retrogene
AMIN004977	AminM1	KB663788	356161	358329	retrogene
AFUN003506	AfunF1	KB669325	636732	638921	retrogene
		STAT2			
AGAP000099	AgamP3	X	1609322	1620718	
None	AgamS1	scf_1106392397092	1625467	1636794	
ACOM024273	AgamM1	scf_1925489324	6721	9712	
AMEM003143	AmerM1	KI439162	99530	110941	
AARA001072	AaraD1	KB704474	2077131	2089755	
AQUA002469	AquaS1	KB667567	577507	590014	
AMEC022522	AmelC1	KI432123	1019	2418	start
AMEC008521	AmelC1	KI429051	19826	29738	end
ACHR006444	AchrA1	KB696577	4577	7339	
AEPI005639	AepiE1	KB671565	564640	575858	
ASTEI06240	AsteI2	scaffold_00038	807412	818750	
ASTE011642	AsteS1	KB664677	916201	927540	
AMAM013720	AmacM1	AXCL01034287	7	1697	start
AMAM007460	AmacM1	AXCL01020239	1859	2703	end
ACUA012928	AculA1	AXCM01003811	6374	15998	
AMIN005063	AminM1	KB664054	203373	212834	
AFUN001738	AfunF1	KB668688	139085	149837	
ADIR000992	AdirW1	KB672824	3890385	3899718	
AFAF008425	AfarF1	KI421542	1634663	1643481	
AATE011504	AatrE1	KI421895	3048052	3050881	
AATE001182	AatrE1	KI421888	3253244	3255493	retrogene
ASIS010499	AsinS1	AXCK01023743	3778	6049	
AALB006346	AalbS1	KB672457	824162	831456	
ADAC010176	AdarC2	scaffold_1175	10384	13034	
CPIJ016471	CpipJ1	supercont3.767	61608	75273	
CPIJ016470	CpipJ1	supercont3.767	39998	41607	fragment
CPIJ016469	CpipJ1	supercont3.767	33677	37485	fragment
AAEL009692	AaegL1	supercont1.418	987517	1008947	

Computational Tools

The following list provides brief descriptions and associated references or website addresses (URLs) of all the tools employed for the analyses performed as part of the 16 *Anopheles* genomes project.

Name	Description	Reference or URL
ALLPATHS-LG	Genome assembly	(*11*)
ANGES	Ancestral genome reconstruction	(*125*)
Artemis	Genome browser for annotation	(*196*)
Augustus	Gene prediction	(*71*)
BLAST	Sequence homology searches	(*133*)
BUSCOs	Assessing assembly completeness	(*66*)
CAFE	Gene gain/loss rate estimation	(*18*)
CAP3	Sequence assembly	(*103*)
Clustal	Sequence alignment and tree building	(*77*)
Exonerate	Gene prediction	(*65*)
GAEMR	Assembly evaluation	www.broadinstitute.org/software/gaemr
GCUA	General codon usage analysis	(*92*)
Geneious	Evolutionary biology analysis suite	www.genious.com
Genewise	Gene prediction	(*107*)
genoPlotR	Graphics for gene and genome maps	(*115*)
GOstats	Gene Ontology enrichment analyses	(*146*)
GRIMM	Genome rearrangement analysis	(*116*)
HHMMiR	miRNA gene prediction	(*81*)
HTSeq	high-throughput sequencing data processing	(*96*)
HYPHY	Selection analysis suite	(*197*)
MAFFT	Protein sequence alignments	(*166*)
MAKER	Genome annotation	(*13*)
MALIN	Maximum likelihood analysis of intron evolution	(*135*)
MCL	Protein family clustering	(*134*)
MEGA5	Molecular Evolutionary Genetics Analysis	(*157*)
Mesquite	Evolutionary biology analysis suite	(*170*)
MGRA	Genome rearrangement analysis	(*130*)
MiRPara	miRNA gene prediction	(*82*)
MRG	Genome rearrangement analysis	(*118*)
MUSCLE	Sequence alignment	(*76*)
Needleman-Wunsch	Protein sequence alignments	(*148*)
Notung	Gene gain/loss rate estimation	(*167*)
OPAL	Protein sequence alignments	(*169*)
OrthoDB	Orthology delineation	(*66*)
PAL2NAL	Protein to DNA alignment conversion	(*144*)
PAML	Phylogenetic analysis by maximum likelihood	(*20*)
PartitionFinderProtein	Protein sequence evolution analysis	(*171*)
PATHGROUPs	Ancestral genome reconstruction	(*109*)
PhyloCSF	Protein-coding potential analysis	(*138*)
Pilon	Post-assembly improvement	www.broadinstitute.org/software/pilon

r8s	Ultrametric tree analysis	(*132*)
RAxML	Maximum likelihood phylogenies	(*91*)
RECON	*De novo* repeat identification	(*68*)
RepeatMasker	Repeat identification	www.repeatmasker.org
RepeatModeler	Repeat identification	www.repeatmasker.org
RepeatScout	*De novo* repeat identification	(*67*)
RNA-STAR	RNAseq read mapping	(*154*)
RNAz	RNA gene prediction	(*79*)
Smith-Waterman	Protein sequence alignments	(*149*)
SNAP	Gene prediction	(*70*)
SnoReport	snoRNA gene prediction	(*85*)
Snoscan	snoRNA gene prediction	(*84*)
STATISTICA	Statistical analyses	www.statsoft.com/Products/STATISTICA
SWIPE	Protein sequence alignments	(*136*)
TBA-MULTIZ	Whole genome alignment	(*106*)
TESEEKER	Repeat identification	(*102*)
topGO	Gene Ontology enrichment analyses	(*147*)
Tophat and Bowtie	RNAseq read mapping and analysis	(*72*)
TrimAl	Sequence alignment trimming	(*90*)
Trinity	Transcriptome assembly	(*60*)
tRNAScan-SE	tRNA gene prediction	(*78*)
UniMog	Genome rearrangement distance analysis	(*128*)

References

1. D. P. Kwiatkowski, How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* **77**, 171-192 (2005).
2. A. Cohuet, C. Harris, V. Robert, D. Fontenille, Evolutionary forces on Anopheles: what makes a malaria vector? *Trends Parasitol* **26**, 130-136 (2010).
3. S. Manguin, P. Carnevale, J. Mouchet, *Biodiversity of Malaria in the World*. (John Libbey Eurotext, 2008).
4. R. A. Holt *et al.*, The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129-149 (2002).
5. O. Marinotti *et al.*, The genome of Anopheles darlingi, the main neotropical malaria vector. *Nucleic Acids Res* **41**, 7387-7400 (2013).
6. D. Zhou *et al.*, Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. *BMC Genomics* **15**, 42 (2014).
7. X. Jiang *et al.*, Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol* **15**, 459 (2014).
8. D. E. Neafsey *et al.*, The Evolution of the Anopheles 16 Genomes Project. *G3-Genes Genomes Genetics* **3**, 1191-1194 (2013).
9. M. C. Fontaine *et al.*, Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science. in press*.
10. M. Moreno *et al.*, Complete mtDNA genomes of Anopheles darlingi and an approach to anopheline divergence time. *Malar J* **9**, 127 (2010).
11. S. Gnerre *et al.*, High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**, 1513-1518 (2011).
12. Materials and methods are available as supplementary material on *Science* Online.
13. C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
14. M. Coluzzi, A. Sabatini, A. della Torre, M. A. Di Deco, V. Petrarca, A polytene chromosome analysis of the Anopheles gambiae species complex. *Science* **298**, 1415-1418 (2002).
15. M. Kamali *et al.*, Multigene phylogenetics reveals temporal diversification of major african malaria vectors. *PLoS One* **9**, e93580 (2014).
16. M. A. Toups, M. W. Hahn, Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genetics* **186**, 763-766 (2010).
17. D. A. Baker, S. Russell, Role of testis-specific gene expression in sex-chromosome evolution of Anopheles gambiae. *Genetics* **189**, 1117-1120 (2011).
18. M. V. Han, G. W. Thomas, J. Lugo-Martinez, M. W. Hahn, Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **30**, 1987-1997 (2013).
19. M. Hahn, M. Han, S. Han, Gene family evolution across 12 Drosophila genomes. *PLoS Genet* **3**, e197 (2007).
20. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007).
21. T. Dottorini *et al.*, A genome-wide analysis in Anopheles gambiae mosquitoes reveals 46 male accessory gland genes, possible modulators of female behavior. *Proc Natl Acad Sci U S A* **104**, 16215-16220 (2007).
22. F. Baldini, P. Gabrieli, D. W. Rogers, F. Catteruccia, Function and composition of male accessory gland secretions in Anopheles gambiae: a comparison with other insect vectors of infectious diseases. *Pathog Glob Health* **106**, 82-93 (2012).
23. R. Assis, Q. Zhou, D. Bachtrog, Sex-biased transcriptome evolution in Drosophila. *Genome Biol Evol* **4**, 1189-1200 (2012).
24. S. Grath, J. Parsch, Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of Drosophila evolution. *Genome Biol Evol* **4**, 346-359 (2012).
25. J. C. Perry, P. W. Harrison, J. E. Mank, The ontogeny and evolution of sex-biased gene expression in Drosophila melanogaster. *Mol Biol Evol* **31**, 1206-1219 (2014).

26. D. W. Rogers *et al.*, Transglutaminase-mediated semen coagulation controls sperm storage in the malaria mosquito. *PLoS Biol* **7**, e1000272 (2009).
27. F. Baldini *et al.*, The interaction between a sexually transferred steroid hormone and a female protein regulates oogenesis in the malaria mosquito Anopheles gambiae. *PLoS Biol* **11**, e1001695 (2013).
28. W. R. Shaw *et al.*, Mating activates the heme peroxidase HPX15 in the sperm storage organ to ensure fertility in Anopheles gambiae. *Proc Natl Acad Sci U S A* **111**, 5854-5859 (2014).
29. J. H. Willis, Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochem Mol Biol* **40**, 189-204 (2010).
30. R. S. Cornman *et al.*, Annotation and analysis of a large cuticular protein family with the R&R Consensus in Anopheles gambiae. *BMC Genomics* **9**, 22 (2008).
31. R. S. Cornman, J. H. Willis, Extensive gene amplification and concerted evolution within the CPR family of cuticular proteins in mosquitoes. *Insect Biochem Mol Biol* **38**, 661-676 (2008).
32. R. S. Cornman, J. H. Willis, Annotation and analysis of low-complexity protein families of Anopheles gambiae that are associated with cuticle. *Insect Mol Biol* **18**, 607-622 (2009).
33. R. S. Cornman, Molecular evolution of Drosophila cuticular protein genes. *PLoS One* **4**, e8345 (2009).
34. T. Togawa, W. A. Dunn, A. C. Emmons, J. Nagao, J. H. Willis, Developmental expression patterns of cuticular protein genes with the R&R Consensus from Anopheles gambiae. *Insect Biochem Mol Biol* **38**, 508-519 (2008).
35. D. C. Rinker *et al.*, Blood meal-induced changes to antennal transcriptome profiles reveal shifts in odor sensitivities in Anopheles gambiae. *Proc Natl Acad Sci U S A* **110**, 8260-8265 (2013).
36. D. C. Rinker *et al.*, Antennal transcriptome profiles of anopheline mosquitoes reveal human host olfactory specialization in Anopheles gambiae. *BMC Genomics* **14**, 749 (2013).
37. M. Altstein, D. R. Nässel, Neuropeptide signaling in insects. *Adv Exp Med Biol* **692**, 155-165 (2010).
38. J. P. Goetze, I. Hunter, S. K. Lippert, L. Bardram, J. F. Rehfeld, Processing-independent analysis of peptide hormones and prohormones in plasma. *Front Biosci (Landmark Ed)* **17**, 1804-1815 (2012).
39. T. H. Stracker, S. Thompson, G. L. Grossman, M. A. Riehle, M. R. Brown, Characterization of the AeaHP gene and its expression in the mosquito Aedes aegypti (Diptera: Culicidae). *J Med Entomol* **39**, 331-342 (2002).
40. C. D. de Oliveira, W. P. Tadei, F. C. Abdalla, P. F. Paolucci Pimenta, O. Marinotti, Multiple blood meals in Anopheles darlingi (Diptera: Culicidae). *J Vector Ecol* **37**, 351-358 (2012).
41. N. Okamoto *et al.*, A fat body-derived IGF-like peptide regulates postfeeding growth in Drosophila. *Dev Cell* **17**, 885-891 (2009).
42. M. A. Riehle, Y. Fan, C. Cao, M. R. Brown, Molecular characterization of insulin-like peptides in the yellow fever mosquito, Aedes aegypti: expression, cellular localization, and phylogeny. *Peptides* **27**, 2547-2560 (2006).
43. Y. Antonova, A. J. Arik, W. Moore, M. M. Riehle, M. R. Brown, in Insect Endocrinology, L. I. Gilbert, Ed. (Academic Press, 2012), pp. 63-92.
44. A. Swaminathan, A. Gajan, L. A. Pile, Epigenetic regulation of transcription in Drosophila. *Front Biosci (Landmark Ed)* **17**, 909-937 (2012).
45. F. Cipressa *et al.*, Effete, a Drosophila chromatin-associated ubiquitin-conjugating enzyme that affects telomeric and heterochromatic position effect variegation. *Genetics* **195**, 147-158 (2013).
46. B. Arcà *et al.*, Positive selection drives accelerated evolution of mosquito salivary genes associated with blood-feeding. *Insect Mol Biol* **23**, 122-131 (2014).
47. H. Isawa, M. Yuda, Y. Orito, Y. Chinzei, A mosquito salivary protein inhibits activation of the plasma contact system by binding to factor XII and high molecular weight kininogen. *J Biol Chem* **277**, 27651-27658 (2002).
48. E. Calvo *et al.*, Aegyptin, a novel mosquito salivary gland protein, specifically binds to collagen and prevents its interaction with platelet glycoprotein VI, integrin alpha2beta1, and von Willebrand factor. *J Biol Chem* **282**, 26928-26938 (2007).
49. E. Guittard *et al.*, CYP18A1, a key enzyme of Drosophila steroid hormone inactivation, is essential for metamorphosis. *Dev Biol* **349**, 35-45 (2011).
50. R. M. Waterhouse *et al.*, Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* **316**, 1738-1743 (2007).
51. L. C. Bartholomay *et al.*, Pathogenomics of Culex quinquefasciatus and Meta-Analysis of Infection Responses to Diverse Pathogens. *Science* **330**, 88-90 (2010).

52. C. Barillas-Mury, Y. S. Han, D. Seeley, F. C. Kafatos, Anopheles gambiae Ag-STAT, a new insect member of the STAT family, is activated in response to bacterial infection. *EMBO J* **18**, 959-967 (1999).
53. L. Gupta *et al.*, The STAT pathway mediates late-phase immunity against Plasmodium in the mosquito Anopheles gambiae. *Cell Host Microbe* **5**, 498-507 (2009).
54. A. C. Bahia *et al.*, The JAK-STAT pathway controls Plasmodium vivax load in early stages of Anopheles aquasalis infection. *PLoS Negl Trop Dis* **5**, e1317 (2011).
55. M. E. Sinka *et al.*, A global map of dominant malaria vectors. *Parasit Vectors* **5**, 69 (2012).
56. M. E. Sinka *et al.*, The dominant Anopheles vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. *Parasit Vectors* **4**, 89 (2011).
57. M. E. Sinka *et al.*, The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasit Vectors* **3**, 117 (2010).
58. M. E. Sinka *et al.*, The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. *Parasit Vectors* **3**, 72 (2010).
59. L. J. Williams *et al.*, Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res* **22**, 2241-2249 (2012).
60. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).
61. M. K. Lawniczak *et al.*, Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science* **330**, 512-514 (2010).
62. V. Nene *et al.*, Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* **316**, 1718-1723 (2007).
63. P. Arensburger *et al.*, Sequencing of Culex quinquefasciatus Establishes a Platform for Mosquito Comparative Genomics. *Science* **330**, 86-88 (2010).
64. M. Adams *et al.*, The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).
65. G. S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
66. R. M. Waterhouse, F. Tegenfeldt, J. Li, E. M. Zdobnov, E. V. Kriventseva, OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research* **41**, D358-D365 (2013).
67. A. L. Price, N. C. Jones, P. A. Pevzner, De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).
68. Z. Bao, S. R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269-1276 (2002).
69. A. Smit, R. Hubley, P. Green. (1996-2010).
70. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
71. M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465-467 (2005).
72. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
73. U. Consortium, Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42**, 7486 (2014).
74. K. Megy *et al.*, VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res* **40**, D729-734 (2012).
75. J. L. Jakubczak, W. D. Burke, T. H. Eickbush, Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc Natl Acad Sci U S A* **88**, 3295-3299 (1991).
76. R. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
77. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
78. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
79. A. R. Gruber, S. Findeiß, S. Washietl, I. L. Hofacker, P. F. Stadler, RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput*, 69-79 (2010).
80. S. W. Burge *et al.*, Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**, D226-232 (2013).
81. S. Kadri, V. Hinman, P. V. Benos, HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics* **10 Suppl 1**, S35 (2009).

82. Y. Wu, B. Wei, H. Liu, T. Li, S. Rayner, MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics* **12**, 107 (2011).
83. A. Kozomara, S. Griffiths-Jones, miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**, D68-73 (2014).
84. P. Schattner, A. N. Brooks, T. M. Lowe, The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686-689 (2005).
85. J. Hertel, I. L. Hofacker, P. F. Stadler, SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* **24**, 158-164 (2008).
86. D. Tautz, J. M. Hancock, D. A. Webb, C. Tautz, G. A. Dover, Complete sequences of the rRNA genes of Drosophila melanogaster. *Mol Biol Evol* **5**, 366-376 (1988).
87. S. Artavanis-Tsakonas *et al.*, The 5S genes of Drosophila melanogaster. *Cell* **12**, 1057-1067 (1977).
88. R. M. Waterhouse, E. M. Zdobnov, F. Tegenfeldt, J. Li, E. V. Kriventseva, OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Research* **39**, D283-D288 (2011).
89. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Research* **36**, D271-D275 (2008).
90. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
91. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, (2014).
92. J. O. McInerney, GCUA: general codon usage analysis. *Bioinformatics* **14**, 372-373 (1998).
93. E. M. Novoa, L. Ribas de Pouplana, Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet* **28**, 574-581 (2012).
94. S. Pechmann, J. Frydman, Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol* **20**, 237-243 (2013).
95. A. Nakao, M. Yoshihama, N. Kenmochi, RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res* **32**, D168-170 (2004).
96. S. Anders, P. Pyl, W. Huber, HTSeq: Analysing high-throughput sequencing data with Python. *bioRxiv preprint*, (2014).
97. T. Ikemura, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**, 389-409 (1981).
98. M. dos Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036-5044 (2004).
99. S. Vicario, E. N. Moriyama, J. R. Powell, Codon usage in twelve species of Drosophila. *BMC Evol Biol* **7**, 226 (2007).
100. A. Smit, R. Hubley, P. Green. (2008-2010).
101. J. Jurka *et al.*, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467 (2005).
102. R. C. Kennedy, M. F. Unger, S. Christley, F. H. Collins, G. R. Madey, An automated homology-based approach for identifying transposable elements. *BMC Bioinformatics* **12**, 130 (2011).
103. X. Huang, A. Madan, CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-877 (1999).
104. A. Stark *et al.*, Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**, 219-232 (2007).
105. K. Lindblad-Toh *et al.*, A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482 (2011).
106. M. Blanchette *et al.*, Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715 (2004).
107. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res* **14**, 988-995 (2004).
108. K. P. Byrne, K. H. Wolfe, The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**, 1456-1461 (2005).
109. C. Zheng, Pathgroups, a dynamic data structure for genome reconstruction problems. *Bioinformatics* **26**, 1587-1594 (2010).
110. M. V. Sharakhova *et al.*, A physical map for an Asian malaria mosquito, Anopheles stephensi. *Am J Trop Med Hyg* **83**, 1023-1027 (2010).
111. I. V. Sharakhov *et al.*, Inversions and gene order shuffling in Anopheles gambiae and A. funestus. *Science* **298**, 182-185 (2002).

112. A. J. Cornel, F. H. Collins, Maintenance of chromosome arm integrity between two Anopheles mosquito subgenera. *J Hered* **91**, 364-370 (2000).
113. E. M. Zdobnov *et al.*, Comparative genome and proteome analysis of anopheles gambiae and Drosophila melanogaster. *Science* **298**, 149-159 (2002).
114. S. W. Schaeffer *et al.*, Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics* **179**, 1601-1655 (2008).
115. L. Guy, J. R. Kultima, S. G. Andersson, genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334-2335 (2010).
116. G. Tesler, GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492-493 (2002).
117. M. von Grotthuss, M. Ashburner, J. M. Ranz, Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila. *Genome Res* **20**, 1084-1096 (2010).
118. G. Bourque, P. A. Pevzner, Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res* **12**, 26-36 (2002).
119. A. Bhutkar *et al.*, Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. *Genetics* **179**, 1657-1680 (2008).
120. M. Slotman, A. Della Torre, J. R. Powell, Female sterility in hybrids between Anopheles gambiae and A. arabiensis, and the causes of Haldane's rule. *Evolution* **59**, 1016-1026 (2005).
121. M. Slotman, A. Della Torre, J. R. Powell, The genetics of inviability and male sterility in hybrids between Anopheles gambiae and An. arabiensis. *Genetics* **167**, 275-287 (2004).
122. V. A. Timoshevskiy *et al.*, Genomic composition and evolution of Aedes aegypti chromosomes revealed by the analysis of physically mapped supercontigs. *BMC Biol* **12**, 27 (2014).
123. M. V. Han, M. W. Hahn, Inferring the history of interchromosomal gene transposition in Drosophila using n-dimensional parsimony. *Genetics* **190**, 813-825 (2012).
124. E. Betrán, K. Thornton, M. Long, Retroposed new genes out of the X in Drosophila. *Genome Res* **12**, 1854-1859 (2002).
125. B. R. Jones, A. Rajaraman, E. Tannier, C. Chauve, ANGES: reconstructing ANcestral GEnomeS maps. *Bioinformatics* **28**, 2388-2390 (2012).
126. C. Chauve, E. Tannier, A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol* **4**, e1000234 (2008).
127. A. Bergeron, J. Mixtacki, J. Stoye. (2006).
128. R. Hilker, C. Sickinger, C. N. Pedersen, J. Stoye, UniMoG--a unifying framework for genomic distance calculation and sorting based on DCJ. *Bioinformatics* **28**, 2509-2511 (2012).
129. S. Aganezov, N. Sydtnikova, A. Consortium, M. A. Alekseyev, (2014).
130. M. A. Alekseyev, P. A. Pevzner, Breakpoint graphs and ancestral genome reconstructions. *Genome Res* **19**, 943-957 (2009).
131. P. George, M. V. Sharakhova, I. V. Sharakhov, High-resolution cytogenetic map for the African malaria vector Anopheles gambiae. *Insect Mol Biol* **19**, 675-682 (2010).
132. M. J. Sanderson, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-302 (2003).
133. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
134. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).
135. M. Csurös, Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics* **24**, 1538-1539 (2008).
136. T. Rognes, Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics* **12**, 221 (2011).
137. Y. C. Wu, M. D. Rasmussen, M. Kellis, Evolution at the subgene level: domain rearrangements in the Drosophila phylogeny. *Mol Biol Evol* **29**, 689-705 (2012).
138. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-i282 (2011).
139. M. F. Lin *et al.*, Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17**, 1823-1836 (2007).

140. I. Jungreis *et al.*, Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res* **21**, 2096-2113 (2011).
141. J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, J. S. Weissman, Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *Elife* **2**, e01179 (2013).
142. C. S. Chan, I. Jungreis, M. Kellis, Heterologous stop codon readthrough of metazoan readthrough candidates in yeast. *PLoS One* **8**, e59450 (2013).
143. P. Steneberg, C. Samakovlis, A novel stop codon readthrough mechanism produces functional Headcase protein in Drosophila trachea. *EMBO Rep* **2**, 593-597 (2001).
144. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-612 (2006).
145. R. M. Waterhouse, E. M. Zdobnov, E. V. Kriventseva, Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi. *Genome Biology and Evolution* **3**, 75-86 (2011).
146. S. Falcon, R. Gentleman, Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258 (2007).
147. A. Alexa, J. Rahnenfuhrer. (2010).
148. S. B. Needleman, C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).
149. T. Smith, M. Waterman, Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
150. C. Gillott, Male accessory gland secretions: modulators of female reproductive physiology and behavior. *Annu Rev Entomol* **48**, 163-184 (2003).
151. F. W. Avila, L. K. Sirot, B. A. LaFlamme, C. D. Rubinstein, M. F. Wolfner, Insect seminal fluid proteins: identification and function. *Annu Rev Entomol* **56**, 21-40 (2011).
152. A. G. Clark *et al.*, Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203-218 (2007).
153. J. A. Andrés, L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, R. G. Harrison, Molecular evolution of seminal proteins in field crickets. *Mol Biol Evol* **23**, 1574-1584 (2006).
154. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
155. D. A. Baker *et al.*, A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, Anopheles gambiae. *BMC Genomics* **12**, 296 (2011).
156. J. Parsch, H. Ellegren, The evolutionary causes and consequences of sex-biased gene expression. *Nat Rev Genet* **14**, 83-87 (2013).
157. K. Tamura *et al.*, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739 (2011).
158. L. Vannini, T. W. Reed, J. H. Willis, Temporal and spatial expression of cuticular proteins of Anopheles gambiae implicated in insecticide resistance or differentiation of M/S incipient species. *Parasit Vectors* **7**, 24 (2014).
159. J. Vontas *et al.*, Transcriptional analysis of insecticide resistance in Anopheles stephensi using cross-species microarray hybridization. *Insect Mol Biol* **16**, 315-324 (2007).
160. O. Wood, S. Hanrahan, M. Coetzee, L. Koekemoer, B. Brooke, Cuticle thickening associated with pyrethroid resistance in the major malaria vector Anopheles funestus. *Parasit Vectors* **3**, 67 (2010).
161. Y. Goltsev *et al.*, Developmental and evolutionary basis for drought tolerance of the Anopheles gambiae embryo. *Dev Biol* **330**, 462-470 (2009).
162. B. J. White *et al.*, Localization of candidate regions maintaining a common polymorphic inversion (2La) in Anopheles gambiae. *PLoS Genet* **3**, e217 (2007).
163. T. Togawa, W. Augustine Dunn, A. C. Emmons, J. H. Willis, CPF and CPFL, two related gene families encoding cuticular proteins of Anopheles gambiae and other insects. *Insect Biochem Mol Biol* **37**, 675-688 (2007).
164. M. Nei, A. P. Rooney, Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**, 121-152 (2005).
165. J. Qiu, P. E. Hardin, Temporal and spatial expression of an adult cuticle protein gene from Drosophila suggests that its protein product may impart some specialized cuticle function. *Dev Biol* **167**, 416-425 (1995).
166. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002).

167. K. Chen, D. Durand, M. Farach-Colton, NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol* **7**, 429-447 (2000).
168. W. Takken, N. O. Verhulst, Host preferences of blood-feeding mosquitoes. *Annu Rev Entomol* **58**, 433-453 (2013).
169. T. J. Wheeler, J. D. Kececioglu, Multiple alignment by aligning alignments. *Bioinformatics* **23**, i559-568 (2007).
170. W. P. Maddison, D. R. Maddison. (2011).
171. R. Lanfear, B. Calcott, S. Y. Ho, S. Guindon, Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* **29**, 1695-1701 (2012).
172. A. Stamatakis, P. Hoover, J. Rougemont, A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* **57**, 758-771 (2008).
173. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
174. Y. Antonova, A. Arik, W. Moore, M. Riehle, M. Brown, Insulin-like peptides: Structure, signaling, and function. *Insect Endocrinology. Gilbert L.I. (Ed) Academic Press*, 63-92 (2012).
175. A. A. Horton *et al.*, The mitogen-activated protein kinome from Anopheles gambiae: identification, phylogeny and functional characterization of the ERK, JNK and p38 MAP kinases. *BMC Genomics* **12**, 574 (2011).
176. N. Pakpour *et al.*, Protein kinase C-dependent signaling controls the midgut epithelial barrier to malaria parasite infection in anopheline mosquitoes. *PLoS One* **8**, e76535 (2013).
177. C. Rosse *et al.*, PKC and the control of localized signal dynamics. *Nat Rev Mol Cell Biol* **11**, 103-112 (2010).
178. S. L. Tan, P. J. Parker, Emerging and diverse roles of protein kinase C in immune cell signalling. *Biochem J* **376**, 545-552 (2003).
179. J. Gong *et al.*, Two protein kinase C isoforms, δ and ε, regulate energy homeostasis in mitochondria by transmitting opposing signals to the pyruvate dehydrogenase complex. *FASEB J* **26**, 3537-3549 (2012).
180. B. H. Shieh, L. Parker, D. Popescu, Protein kinase C (PKC) isoforms in Drosophila. *J Biochem* **132**, 523-527 (2002).
181. V. Sargsyan *et al.*, Phosphorylation via PKC Regulates the Function of the Drosophila Odorant Co-Receptor. *Front Cell Neurosci* **5**, 5 (2011).
182. K. Liu, H. Tsujimoto, S. J. Cha, P. Agre, J. L. Rasgon, Aquaporin water channel AgAQP1 in the malaria vector mosquito Anopheles gambiae during blood feeding and humidity adaptation. *Proc Natl Acad Sci U S A* **108**, 6062-6066 (2011).
183. H. Tsujimoto, K. Liu, P. J. Linser, P. Agre, J. L. Rasgon, Organ-specific splice variants of aquaporin water channel AgAQP1 in the malaria vector Anopheles gambiae. *PLoS One* **8**, e75888 (2013).
184. L. L. Drake *et al.*, The Aquaporin gene family of the yellow fever mosquito, Aedes aegypti. *PLoS One* **5**, e15578 (2010).
185. G. J. Filion *et al.*, Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-224 (2010).
186. E. L. Greer, Y. Shi, Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet* **13**, 343-357 (2012).
187. J. G. van Bemmel *et al.*, A network model of the molecular organization of chromatin in Drosophila. *Mol Cell* **49**, 759-771 (2013).
188. C. H. Arrowsmith, C. Bountra, P. V. Fish, K. Lee, M. Schapira, Epigenetic protein families: a new frontier for drug discovery. *Nat Rev Drug Discov* **11**, 384-400 (2012).
189. S. R. Schulze, L. L. Wallrath, Gene regulation by chromatin structure: paradigms established in Drosophila melanogaster. *Annu Rev Entomol* **52**, 171-192 (2007).
190. S. Powell *et al.*, eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**, D231-239 (2014).
191. E. Darbo, C. Herrmann, T. Lecuit, D. Thieffry, J. van Helden, Transcriptional and epigenetic signatures of zygotic genome activation during early Drosophila embryogenesis. *BMC Genomics* **14**, 226 (2013).
192. P. Dimitri, C. Pisano, Position effect variegation in Drosophila melanogaster: relationship between suppression effect and the amount of Y chromosome. *Genetics* **122**, 793-800 (1989).

193. E. Betrán, M. Long, Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. *Genetics* **164**, 977-988 (2003).
194. R. Kalamegham, D. Sturgill, E. Siegfried, B. Oliver, Drosophila mojoless, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. *Mol Biol Evol* **24**, 732-742 (2007).
195. W. Wang, F. G. Brunet, E. Nevo, M. Long, Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **99**, 4448-4453 (2002).
196. M. Berriman, K. Rutherford, Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4**, 124-132 (2003).
197. S. L. Pond, S. D. Frost, S. V. Muse, HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679 (2005).
198. B. Murrell *et al.*, Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764 (2012).
199. R. Tatusov *et al.*, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
200. S. Karim, P. Singh, J. M. Ribeiro, A deep insight into the sialotranscriptome of the gulf coast tick, Amblyomma maculatum. *PLoS One* **6**, e28525 (2011).
201. J. M. Ribeiro, B. J. Mans, B. Arcà, An insight into the sialome of blood-feeding Nematocera. *Insect Biochem Mol Biol* **40**, 767-784 (2010).
202. B. Arcà *et al.*, An updated catalogue of salivary gland transcripts in the adult female mosquito, Anopheles gambiae. *J Exp Biol* **208**, 3971-3986 (2005).
203. M. A. Okulate *et al.*, Identification and molecular characterization of a novel protein Saglin as a target of monoclonal antibodies affecting salivary gland infectivity of Plasmodium sporozoites. *Insect Mol Biol* **16**, 711-722 (2007).
204. C. Rizzo *et al.*, Wide cross-reactivity between Anopheles gambiae and Anopheles funestus SG6 salivary proteins supports exploitation of gSG6 as a marker of human exposure to major malaria vectors in tropical Africa. *Malar J* **10**, 206 (2011).
205. E. Calvo, B. J. Mans, J. F. Andersen, J. M. Ribeiro, Function and evolution of a mosquito salivary protein family. *J Biol Chem* **281**, 1935-1942 (2006).
206. H. Isawa *et al.*, Identification and characterization of a new kallikrein-kinin system inhibitor from the salivary glands of the malaria vector mosquito Anopheles stephensi. *Insect Biochem Mol Biol* **37**, 466-477 (2007).
207. H. Ranson *et al.*, Evolution of supergene families associated with insecticide resistance. *Science* **298**, 179-181 (2002).
208. C. V. Edi, B. G. Koudou, C. M. Jones, D. Weetman, H. Ranson, Multiple-insecticide resistance in Anopheles gambiae mosquitoes, Southern Côte d'Ivoire. *Emerg Infect Dis* **18**, 1508-1511 (2012).
209. J. Pinto *et al.*, Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector Anopheles gambiae. *PLoS One* **2**, e1243 (2007).
210. A. Lynd *et al.*, Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in Anopheles gambiae s.s. *Mol Biol Evol* **27**, 1117-1125 (2010).
211. D. K. Mathias *et al.*, Spatial and temporal variation in the kdr allele L1014S in Anopheles gambiae s.s. and phenotypic variability in susceptibility to insecticides in Western Kenya. *Malar J* **10**, 10 (2011).
212. C. V. Edi *et al.*, CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito Anopheles gambiae. *PLoS Genet* **10**, e1004236 (2014).
213. S. N. Mitchell *et al.*, Identification and validation of a gene causing cross-resistance between insecticide classes in Anopheles gambiae from Ghana. *Proc Natl Acad Sci U S A* **109**, 6147-6152 (2012).
214. S. N. Mitchell *et al.*, Metabolic and target-site mechanisms combine to confer strong DDT resistance in Anopheles gambiae. *PLoS One* **9**, e92662 (2014).
215. J. M. Riveron *et al.*, A single mutation in the GSTe2 gene allows tracking of metabolically-based insecticide resistance in a major malaria vector. *Genome Biol* **15**, R27 (2014).
216. C. F. Ayres *et al.*, Comparative genomics of the anopheline glutathione S-transferase epsilon cluster. *PLoS One* **6**, e29237 (2011).
217. R. Feyereisen, Evolution of insect P450. *Biochem Soc Trans* **34**, 1252-1255 (2006).
218. R. Niwa *et al.*, CYP306A1, a cytochrome P450 enzyme, is essential for ecdysteroid biosynthesis in the prothoracic glands of Bombyx and Drosophila. *J Biol Chem* **279**, 35942-35949 (2004).

219. H. M. Ismail *et al.*, Pyrethroid activity-based probes for profiling cytochrome P450 activities associated with insecticide interactions. *Proc Natl Acad Sci U S A* **110**, 19766-19771 (2013).
220. G. K. Christophides *et al.*, Immunity-related genes and gene families in Anopheles gambiae. *Science* **298**, 159-165 (2002).
221. D. Vlachou, F. Kafatos, The complex interplay between mosquito positive and negative regulators of Plasmodium development. *Curr Opin Microbiol* **8**, 415-421 (2005).
222. E. Levashina *et al.*, Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, Anopheles gambiae. *Cell* **104**, 709-718 (2001).
223. S. Blandin *et al.*, Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. *Cell* **116**, 661-670 (2004).
224. S. Blandin, E. Levashina, Thioester-containing proteins and insect immunity. *Mol Immunol* **40**, 903-908 (2004).
225. R. Baxter *et al.*, Structural basis for conserved complement factor-like function in the antimalarial protein TEP1. *Proc Natl Acad Sci U S A* **104**, 11615-11620 (2007).
226. M. Osta, G. Christophides, F. Kafatos, Effects of mosquito genes on Plasmodium development. *Science* **303**, 2030-2032 (2004).
227. M. Riehle *et al.*, Anopheles gambiae APL1 is a family of variable LRR proteins required for Rel1-mediated protection from the malaria parasite, Plasmodium berghei. *PLoS One* **3**, e3672 (2008).
228. M. Fraiture *et al.*, Two mosquito LRR proteins function as complement control factors in the TEP1-mediated killing of Plasmodium. *Cell Host Microbe* **5**, 273-284 (2009).
229. C. Mitri *et al.*, Fine pathogen discrimination within the APL1 gene family protects Anopheles gambiae against human and rodent malaria species. *PLoS Pathog* **5**, e1000576 (2009).
230. M. Povelones, L. M. Upton, K. A. Sala, G. K. Christophides, Structure-function analysis of the Anopheles gambiae LRIM1/APL1C complex and its interaction with complement C3-like protein TEP1. *PLoS Pathog* **7**, e1002023 (2011).
231. R. M. Waterhouse, M. Povelones, G. K. Christophides, Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. *Bmc Genomics* **11**, (2010).
232. M. Povelones, R. M. Waterhouse, F. C. Kafatos, G. K. Christophides, Leucine-Rich Repeat Protein Complex Activates Mosquito Complement in Defense Against Plasmodium Parasites. *Science* **324**, 258-261 (2009).
233. Y. Dong, G. Dimopoulos, Anopheles fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *J Biol Chem* **284**, 9835-9844 (2009).
234. Y. Dong *et al.*, Anopheles gambiae immune responses to human and rodent Plasmodium parasite species. *PLoS Pathog* **2**, e52 (2006).
235. M. Povelones *et al.*, The CLIP-domain serine protease homolog SPCLIP1 regulates complement recruitment to microbial surfaces in the malaria mosquito Anopheles gambiae. *PLoS Pathog* **9**, e1003623 (2013).
236. J. Volz, H. Müller, A. Zdanowicz, F. Kafatos, M. Osta, A genetic module regulates the melanization response of Anopheles to Plasmodium. *Cell Microbiol* **8**, 1392-1405 (2006).
237. J. Volz, M. Osta, F. Kafatos, H. Müller, The roles of two clip domain serine proteases in innate immune responses of the malaria vector Anopheles gambiae. *J Biol Chem* **280**, 40161-40168 (2005).
238. C. Barillas-Mury, CLIP proteases and Plasmodium melanization in Anopheles gambiae. *Trends Parasitol* **23**, 297-299 (2007).
239. E. Abraham *et al.*, An immune-responsive serpin, SRPN6, mediates mosquito defense against malaria parasites. *Proc Natl Acad Sci U S A* **102**, 16327-16332 (2005).
240. K. Michel, A. Budd, S. Pinto, T. Gibson, F. Kafatos, Anopheles gambiae SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. *EMBO Rep* **6**, 891-897 (2005).
241. A. Schnitger, H. Yassine, F. Kafatos, M. Osta, Two C-type lectins cooperate to defend Anopheles gambiae against Gram-negative bacteria. *J Biol Chem* **284**, 17616-17624 (2009).
242. T. Little, N. Cobbe, The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein. *Insect Mol Biol* **14**, 599-605 (2005).
243. A. Parmakelis *et al.*, Anopheles immune genes and amino acid sites evolving under the effect of positive selection. *PLoS One* **5**, e8885 (2010).
244. M. Slotman *et al.*, Patterns of selection in anti-malarial immune genes in malaria vectors: evidence for adaptive evolution in LRIM1 in Anopheles arabiensis. *PLoS One* **2**, e793 (2007).

245. D. Obbard, J. Welch, T. Little, Inferring selection in the Anopheles gambiae species complex: an example from immune-related serine protease inhibitors. *Malar J* **8**, 117 (2009).
246. D. Obbard *et al.*, The evolution of TEP1, an exceptionally polymorphic immunity gene in Anopheles gambiae. *BMC Evol Biol* **8**, 274 (2008).
247. T. Lehmann *et al.*, Molecular evolution of immune genes in the malaria mosquito Anopheles gambiae. *PLoS One* **4**, e4549 (2009).
248. C. Mendes *et al.*, Molecular evolution of the three short PGRPs of the malaria vectors Anopheles gambiae and Anopheles arabiensis in East Africa. *BMC Evol Biol* **10**, 9 (2010).