MIT Libraries | DSpace@MIT

MIT Open Access Articles

Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Neafsey, D. E., R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, J. E. Allen, J. Amon, et al. "Highly Evolvable Malaria Vectors: The Genomes of 16 Anopheles Mosquitoes." Science 347, no. 6217 (November 27, 2014): 1258522–1258522.

As Published: http://dx.doi.org/10.1126/science.1258522

Publisher: American Association for the Advancement of Science (AAAS)

Persistent URL: http://hdl.handle.net/1721.1/100767

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2016 January 02.

Published in final edited form as:

Science. 2015 January 2; 347(6217): 1258522. doi:10.1126/science.1258522.

"Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes"

A full list of authors and affiliations appears at the end of the article.

Abstract

Variation in vectorial capacity for human malaria among *Anopheles* mosquito species is determined by many factors, including behavior, immunity, and life history. To investigate the genomic basis of vectorial capacity and explore new avenues for vector control, we sequenced the genomes of 16 anopheline mosquito species from diverse locations spanning ~100 million years of evolution. Comparative analyses show faster rates of gene gain and loss, elevated gene shuffling on the X chromosome, and more intron losses, relative to *Drosophila*. Some determinants of vectorial capacity, such as chemosensory genes, do not show elevated turnover, but instead diversify through protein-sequence changes. This dynamism of anopheline genes and genomes may contribute to their flexible capacity to take advantage of new ecological niches, including adapting to humans as primary hosts.

Introduction

Malaria is a complex disease, mediated by obligate eukaryotic parasites with a life cycle requiring adaption to both vertebrate hosts and mosquito vectors. These relationships create a rich co-evolutionary triangle. Just as *Plasmodium* parasites have adapted to their diverse hosts and vectors, infection by *Plasmodium* parasites has reciprocally induced adaptive evolutionary responses in humans and other vertebrates (1), and has also influenced mosquito evolution (2). Human malaria is transmitted only by mosquitoes in the genus *Anopheles*, but not all species within the genus, or even all members of each vector species, are efficient malaria vectors. This suggests an underlying genetic/genomic plasticity that results in variation of key traits determining vectorial capacity within the genus.

In all, five species of *Plasmodium* have adapted to infect humans, and are transmitted by approximately 60 of the 450 known species of anopheline mosquitoes (3). Sequencing the genome of *Anopheles gambiae*, the most important malaria vector in sub-Saharan Africa, has offered numerous insights into how that species became highly specialized to live among and feed upon humans, and how susceptibility to mosquito control strategies is determined (4). Until very recently (5–7), similar genomic resources have not existed for

†Corresponding author. neafsey@broadinstitute.org (D.E.N.); nbesansk@nd.edu (N.J.B.).
*These authors contributed equally to this work.

Supplementary Materials
Materials, Methods, and Information
Figs. S1 to S25
Tables S1 to S36

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

other anophelines, limiting comparisons to individual genes or sets of genomic markers with no genome-wide data to investigate attributes associated with vectorial capacity across the genus.

Thus, we sequenced and assembled the genomes and transcriptomes of 16 anophelines from Africa, Asia, Europe, and Latin America. We chose these 16 species to represent a range of evolutionary distances from *An. gambiae*, a variety of geographic locations and ecological conditions, and varying degrees of vectorial capacity (8) (Fig. 1A, B). For example, *Anopheles quadriannulatus*, while extremely closely related to *An. gambiae*, feeds preferentially on bovines rather than humans, limiting its potential to transmit human malaria. *Anopheles merus*, *Anopheles melas*, *Anopheles farauti*, and *Anopheles albimanus* females can lay eggs in salty or brackish water, instead of the freshwater sites required by other species. With a focus on species most closely related to *An. gambiae* (9), the sampled anophelines span the three main subgenera that shared a common ancestor approximately 100 million years (MYr) ago (10).

Materials and methods summary

Genomic DNA and whole-body RNA were obtained from laboratory colonies and wild-caught specimens (tables S1-S2), with samples for nine species procured from newly established isofemale colonies to reduce heterozygosity. Illumina sequencing libraries spanning a range of insert sizes were constructed, with ~100-fold paired-end 101 base pair (bp) coverage generated for small (180 bp) and medium (1.5 kb) insert libraries and lower coverage for large (38 kb) insert libraries (table S3). DNA template for the small and medium input libraries was sourced from single female mosquitoes from each species to further reduce heterozygosity. High molecular weight DNA template for each large insert library was derived from pooled DNA obtained from several hundred mosquitoes. ALLPATHS-LG (11) genome assemblies were produced using the ‘haploidify’ option to reduce haplotype assemblies caused by high heterozygosity. Assembly quality reflected DNA template quality and homozygosity, with a mean scaffold N50 of 3.6 Mb, ranging to 18.1 Mb for *An. albimanus* (table S4). Despite variation in contiguity, the assemblies were remarkably complete and searches for arthropod-wide single-copy orthologs generally revealed few missing genes (fig. S1) (12).

Genome annotation with MAKER (13) supported with RNAseq transcriptomes (produced from pooled male and female larvae, pupae, and adults; table S5), and comprehensive non-coding RNA gene prediction (fig. S2), yielded relatively complete gene sets (fig. S3), with between 10,738 and 16,149 protein-coding genes identified for each species. Gene count was generally commensurate with assembly contiguity (table S6). Some of this variation in total gene counts may be attributed to the challenges of gene annotations with variable levels of assembly contiguity and supporting RNAseq data. To estimate the prevalence of erroneous gene model fusions and/or fragmentations, we compared the new gene annotations to *An. gambiae* gene models and found an average of 3.3% and 9.7% potentially fused and fragmented gene models, respectively. For analyses described below that may be sensitive to variation in gene model accuracy or gene set completeness, we have conducted sensitivity analyses to rule out confounding results from these factors (12).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Rapidly evolving genes and genomes

Orthology delineation identified lineage-restricted and species-specific genes, as well as ancient genes found across insect taxa, of which universal single-copy orthologs were employed to estimate the molecular species phylogeny (Fig. 1B, 1C, fig. S4). Analysis of codon frequencies in these orthologs revealed that anophelines, unlike drosophilids, exhibit relatively uniform codon usage preferences (fig. S5).

Polytene chromosomes have provided a glimpse into anopheline chromosome evolution (14). Our genome-sequence-based view confirmed the cytological observations, and offers many new insights. At the base pair level, ~90% of the non-gapped and non-masked *An. gambiae* genome (*i.e.*, excluding transposable elements, as detailed in table S7) is alignable to the most closely related species, while only ~13% aligns to the most distant (Fig. 1D, fig. S6, table S8), with reduced alignability in centromeres and on the X chromosome (Fig. 1D). At chromosomal levels, mapping data anchored 35–76% of the *Anopheles stephensi*, *Anopheles funestus*, *Anopheles atroparvus*, and *An. albimanus* genome assemblies to chromosomal arms (tables S9-S12). Analysis of genes in anchored regions showed that synteny at the whole-arm level is highly conserved, despite several whole-arm translocations (Fig. 2A, table S13). In contrast, small-scale rearrangements disrupt gene colinearity within arms over time, leading to extensive shuffling of gene order over a timescale of 29 MYA or more (10, 15) (Fig. 2B, fig. S7). As in *Drosophila*, rearrangement rates are higher on the X chromosome than on autosomes (Fig. 2C, tables S14-S16). However, the difference is significantly more pronounced in *Anopheles*, where X chromosome rearrangements are 2.7-fold more frequent than autosomal rearrangements; in *Drosophila*, the corresponding ratio is only 1.2 (*t*-test, $t_{10} = 7.3$, $P < 1\times10^{-5}$) (fig. S8). The X chromosome is also notable for a significant degree of observed gene movement to other chromosomes relative to *Drosophila* (one sample proportion test, $P < 2.2\times10^{-16}$; Fig. 2D, tables S17-S18), as was previously noted for *Anopheles* relative to *Aedes* (16), further underscoring its distinctive evolutionary profile in *Anopheles* compared to other dipteran genera.

Such dynamic gene shuffling and movement may be facilitated by the multiple families of DNA transposons and LTR and non-LTR retroelements found in all genomes (table S7), as well as a weaker dosage compensation phenotype in *Anopheles* compared to *Drosophila* (17). Despite such shuffling, comparing genomic locations of orthologs can be successfully employed to reconstruct ancestral chromosomal arrangements (fig. S9) and to confidently improve assembly contiguity (tables S19-S21).

Copy number variation in homologous gene families also reveals striking evolutionary dynamism. Analysis of 11,636 gene families with CAFE 3 (18) indicates a rate of gene gain/loss at least five-fold higher than that observed for 12 *Drosophila* genomes (19). Overall, these *Anopheles* genomes exhibit a rate of gain or loss/gene/million years of 3.12×10^{-3} compared to 5.90×10^{-4} for *Drosophila*, suggesting substantially higher gene turnover within anophelines relative to fruit flies. This five-fold greater gain/loss rate in anophelines holds true under models that account for uncertainly in gene family sizes at the tips of the species tree due to annotation/assembly errors, and is not sensitive to inclusion or exclusion of taxa affecting the root age of the tree, nor to the exclusion of taxa with the poorest assemblies

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and gene sets (fig. S10, tables S22-S23). Examples include expansions of cuticular proteins in *Anopheles arabiensis* and neurotransmitter-gated ion channels in *An. albimanus* (table S24).

The evolutionary dynamism of *Anopheles* genes extends to their architecture. Comparisons of single-copy orthologs at deeper phylogenetic depths showed losses of introns at the root of the true fly order Diptera, and revealed continued losses as the group diversified into the lineages leading to fruit flies and mosquitoes. However, anopheline orthologs have sustained greater intron loss than drosophilids, leading to a relative paucity of introns in the genes of extant anophelines (fig. S11, table S25). Comparative analysis also revealed that gene fusion and fission played a substantial role in the evolution of mosquito genes, with apparent rearrangements affecting an average of 10.1% of all genes in the genomes of the 10 species with the most contiguous assemblies (fig. S12). Furthermore, gene boundaries can be flexible; whole genome alignments identified 325 candidates for stop-codon readthrough (fig. S13, table S26).

As molecular evolution of protein-coding sequences is a well-known source of phenotypic change, we compared evolutionary rates among different functional categories of anopheline orthologs. We quantified evolutionary divergence in terms of protein sequence identity of aligned orthologs and the d_N/d_S statistic computed using PAML (12, 20). Among curated sets of genes linked to vectorial capacity or species-specific traits against a background of functional categories defined by Gene Ontology or InterPro annotations, odorant and gustatory receptors show high evolutionary rates and male accessory gland proteins exhibit exceptionally high d_N/d_S ratios (Fig. 3, figs. S14-S15, tables S27-S29). Rapid divergence in functional categories related to malaria transmission and/or mosquito control strategies led us to examine the genomic basis of several facets of anopheline biology in closer detail.

Insights into mosquito biology and vectorial capacity

Mosquito reproductive biology evolves rapidly and presents a compelling target for vector control. This is exemplified by the *An. gambiae* male accessory gland protein (Acp cluster on chromosome 3R (21, 22), where conservation is mostly lost outside the *An. gambiae* species complex (fig. S16). In *Drosophila*, male-biased genes such as Acps tend to evolve faster than loci without male-biased expression (23–25). We looked for a similar pattern in anophelines after assessing each gene for sex-biased expression using microarray and RNAseq datasets for *An. gambiae* (12). In contrast to *Drosophila*, female-biased genes show dramatically faster rates of evolution across the genus than male-biased genes (Wilcoxon rank sum test, $P = 5\times10^{-4}$) (fig. S17).

Differences in reproductive genes among anophelines may provide insight into the origin and function of sex-related traits. During copulation, *An. gambiae* males transfer a gelatinous mating plug, a complex of seminal proteins, lipids, and hormones that are essential for successful sperm storage by females and for reproductive success (26–28). Coagulation of the plug is mediated by a seminal transglutaminase (*TG3*), which is found in anophelines but is absent in other mosquito genera that do not form a mating plug (26). We examined *TG3* and its two paralogs (*TG1* and *TG2*) in the sequenced anophelines, and investigated the rate of evolution of each gene (Fig. 4A). Silent sites were saturated at the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

whole-genus level, making d_S difficult to estimate reliably, but *TG1* (the gene presumed to be ancestral due to broadest taxonomic representation) exhibited the lowest rate of amino acid change (d_N = 0.20), *TG2* exhibited an intermediate rate (d_N = 0.93), and the anopheline-specific *TG3* has evolved even more rapidly (d_N = 1.50), perhaps due to male/male or male/female evolutionary conflict. Interestingly, plug formation appears to be a derived trait within anophelines, as it is not exhibited by *An. albimanus* and intermediate, poorly coagulated plugs were observed in taxa descending from early-branching lineages within the genus (table S30). Functional studies of mating plugs will be necessary to understand what drove the origin and rapid evolution of *TG3*.

Proteins that constitute the mosquito cuticular exoskeleton play important roles in diverse aspects of anopheline biology, including development, ecology, and insecticide resistance, and constitute approximately 2% of all protein-coding genes (29). Comparisons among dipterans have revealed numerous amplifications of cuticular protein (CP) genes undergoing concerted evolution at physically clustered loci (30–33). We investigated the extent and timescale of gene cluster homogenization within anophelines by generating phylogenies of orthologous gene clusters (fig. S18, table S31). Throughout the genus, these gene clusters often group phylogenetically by species rather than by position within tandem arrays, particularly in a subset of clusters. These include the 3RB and 3RC clusters of CP genes (30), the CPLCG group A and CPLCW clusters found elsewhere on 3R (32), and six tandemly arrayed genes on 3L designated *CPFL2* through *CPFL7* (34). CPLCW genes occur in a head-to-head arrangement with CPLCG group A genes, and exhibit highly conserved intergenic sequences (fig. S19). Furthermore, transcript localization studies using *in situ* hybridization revealed identical spatial expression patterns for CLPCW and CPLCG group A gene pairs suggestive of co-regulation (fig. S19). For these five gene clusters, complete grouping by organismal lineage was observed for most deep nodes as well as for many individual species outside the shallow *An. gambiae* species complex (Fig. 4B), consistent with a relatively rapid (less than 20 million years) homogenization of sequences via concerted evolution. 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.

Anophelines identify hosts, oviposition sites, and other environmental cues through specialized chemosensory membrane-bound receptors. We examined three of the major gene families that encode these molecules: the odorant receptors (*OR*s), gustatory receptors (*GR*s), and variant ionotropic glutamate receptors (*IR*s). Given rapid chemosensory gene turnover observed in many other insects, we explored whether varying host preferences of anopheline mosquitoes could be attributed to chemosensory gene gains and losses. Unexpectedly in light of the elevated genome-wide rate of gene turnover, we found that the overall size and content of the chemosensory gene repertoire is relatively conserved across the genus. CAFE 3 (18) analyses estimated that the most recent common ancestor of the anophelines had approximately 60 genes in each of the *OR* and *GR* families, similar to most extant anophelines (Fig. 4C, fig. S20). Estimated gain/loss rates of *OR* and *GR* genes per million years (error-corrected $\lambda = 1.3\times10^{-3}$ for *OR*s and 2.0×10^{-4} for *GR*s) were much lower than the overall level of anopheline gene families. Similarly, we found almost the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

same number of antennae-expressed *IR*s (~20) in all anopheline genomes. Despite overall conservation in chemosensory gene numbers, we observed several examples of gene gain and loss in specific lineages. Notably, there was a net gain of at least 12 *OR*s in the common ancestor of the *An. gambiae* complex (Fig. 4C).

OR and *GR* gene repertoire stability may derive from their roles in several critical behaviors. Host preference differences are likely to be governed by a combination of functional divergence and transcriptional modulation of orthologs. This model is supported by studies of antennal transcriptomes in the major malaria vector *An. gambiae* (35), and comparisons between this vector and its morphologically identical sibling *An. quadriannulatus* (36), a very closely related species that plays no role in malaria transmission (despite vectorial competence) because it does not specialize on human hosts. Furthermore, we found that many subfamilies of ORs and GRs showed evidence of positive selection (19 of 53 *ORs;* 17 of 59 *GRs*) across the genus, suggesting potential functional divergence.

Several blood feeding-related behaviors in mosquitoes are also regulated by peptide hormones (37). These peptides are synthesized, processed and released from nervous and endocrine systems and elicit their effects through binding appropriate receptors in target tissues (38). In total, 39 peptide hormones were identified from each of the sequenced anophelines (fig. S21). Interestingly, no ortholog of the well-characterized head peptide (HP) hormone of the culicine mosquito *Aedes aegypti* was identified in any of the assemblies. In *Ae. aegypti*, HP is responsible for inhibiting host seeking behavior following a blood meal (39). As anophelines broadly exhibit similar behavior (40), the absence of HP from the entire clade suggests they may have evolved a novel mechanism to inhibit excess blood feeding. Similarly, no ortholog of insulin growth factor 1 (*IGF1*) was identified in any anophelines even though *IGF1* orthologs have been identified in other dipterans, including *D. melanogaster* (41) and *Ae. aegypti* (42). *IGF1* is a key component of the insulin/insulin growth factor 1 signaling (IIS) cascade, which regulates processes including innate immunity, reproduction, metabolism and lifespan (43). Nevertheless, other members of the IIS cascade are present, and four insulin-like peptides are found in a compact cluster with gene arrangements conserved across anophelines (fig. S22). This raises questions regarding the modification of IIS signaling in the absence of *IGF1* and the functional importance of this conserved genomic arrangement.

Epigenetic mechanisms impact many biological processes via modulation of chromatin structure, telomere remodeling and transcriptional control. Of the 215 epigenetic regulatory genes in *D. melanogaster* (44), we identified 169 putative *An. gambiae* orthologs (table S32), suggesting the presence of mechanisms of epigenetic control in *Anopheles* and *Drosophila.* We find, however, that retrotransposition may have contributed to the functional divergence of at least one gene associated with epigenetic regulation. The ubiquitin-conjugating enzyme *E2D* (orthologous to *effete* (45) in *D. melanogaster*) duplicated via retrotransposition in an early anopheline ancestor, and the retrotransposed copy is maintained in a subset of anophelines. Although the entire amino acid sequence of *E2D* is perfectly conserved between *An. gambiae* and *D. melanogaster*, the retrogenes are highly divergent (Fig. 5A), and may contribute to functional diversification within the genus.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Saliva is integral to blood feeding – it impairs host hemostasis and also affects inflammation and immunity. In *An. gambiae* the salivary proteome is estimated to contain the products of at least 75 genes, most being expressed solely in the adult female salivary glands. Comparative analyses indicate that anopheline salivary proteins are subject to strong evolutionary pressures, and these genes exhibit an accelerated pace of evolution, as well as a very high rate of gain/loss (Fig. 3, fig. S23). Polymorphisms within *An. gambiae* populations from limited sets of salivary genes were previously found to carry signatures of positive selection (46). Sequence analysis across the anophelines shows that salivary genes have the highest incidence of positively selected codons among the seven gene classes (fig. S24), indicating that co-evolution with vertebrate hosts is a powerful driver of natural selection in salivary proteomes. Moreover, salivary proteins also exhibit functional diversification through new gene creation. Sequence similarity, intron/exon boundaries, and secondary structure prediction point to the birth of the *SG7/SG7-2* inflammation-inhibiting (47) gene family from the genomic region encoding the C-terminus of the 30 kDa protein (Fig. 5B), a collagen-binding platelet inhibitor already present in the blood-feeding ancestor of mosquitoes and black flies (48). Based on phylogenetic representation, these events must have occurred before the radiation of anophelines but after separation from the culicines.

Resistance to insecticides and other xenobiotics has arisen independently in many anopheline species, fostered directly and indirectly by anthropogenic environmental modification. Metabolic resistance to insecticides is mediated by multiple gene families, including cytochrome P450s and glutathione-S-transferases (GSTs), which serve to generally protect against all environment stresses, both natural and anthropogenic. We manually characterized these gene families in seven anophelines spanning the genus. Despite their large size, gene numbers (87–104 P450 genes, 27–30 GST genes) within both gene families are highly conserved across all species, though lineage-specific gene duplications and losses are often seen (tables S33-S34). As with the *OR* and *GR* olfaction-related gene families, P450 and GST repertoires may be relatively constant due to the large number of roles they play in anopheline biology. Orthologs of genes associated with insecticide resistance either via up-regulation or coding variation (e.g., *Cyp6m2, Cyp6p3* [*Cyp6p9* in *An. funestus*], *Gste2, Gste4*) were found in all species, suggesting that virtually all anophelines likely have genes capable of conferring insecticide resistance through similar mechanisms. Unexpectedly, one member of the P450 family (*Cyp18a1*) with a conserved role in ecdysteroid catabolism (and consequently development and metamorphosis (49) appears to have been lost from the ancestor of the *An. gambiae* species complex, but is found in the genome and transcriptome assemblies of other species, indicating that the *An. gambiae* complex may have recently evolved an alternate mechanism for catabolizing ecdysone.

Susceptibility to malaria parasites is a key determinant of vectorial capacity. Dissecting the immune repertoire (50, 51) (table S35) into its constituent phases reveals that classical recognition genes and genes encoding effector enzymes exhibit relatively low levels of sequence divergence. Signal transducers are more divergent in sequence but are conserved in representation across species and rarely duplicated. Cascade modulators, while also divergent, are more lineage-specific and generally have more gene duplications (Fig. 3, fig.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

S25). A rare duplication of an immune signal transduction gene occurred through the retrotransposition of the signal transducer and activator of transcription, *STAT2*, to form the intronless *STAT1* after the divergence of *An. dirus* and *An. farauti* from the rest of the subgenus *Cellia* (Fig. 5C, table S36). Interestingly, an independent retroposition event appears to have independently created another intronless STAT gene in the *An. atroparvus* lineage. In *An. gambiae*, *STAT1* controls the expression of *STAT2* and is activated in response to bacterial challenge (52, 53), and the STAT pathway has been demonstrated to mediate immunity to *Plasmodium* (53, 54), so the presence of these relatively new immune signal transducers may have allowed for rewiring of regulatory networks governing immune responses in this subset of anophelines.

Conclusion

Since the discovery over a century ago by Ronald Ross and Giovanni Battista Grassi that human malaria is transmitted by a narrow range of blood-feeding female mosquitoes, the biological basis of malarial vectorial capacity has been a matter of intense interest. Inasmuch as previous successes in the local elimination of malaria have always been accomplished wholly or in part through effective vector control, an increased understanding of vector biology is crucial for continued progress against malarial disease. These 16 new reference genome assemblies provide a foundation for additional hypothesis generation and testing to further our understanding of the diverse biological traits that determine vectorial capacity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Daniel E. Neafsey[1,*,†], Robert M. Waterhouse[2,3,4,5], Mohammad R. Abai[6], Sergey S. Aganezov[7], Max A. Alekseyev[7], James E. Allen[8], James Amon[9], Bruno Arcà[10], Peter Arensburger[11], Gleb Artemov[12], Lauren A. Assour[13], Hamidreza Basseri[6], Aaron Berlin[1], Bruce W. Birren[1], Stephanie A. Blandin[14,15], Andrew I. Brockman[16], Thomas R. Burkot[17], Austin Burt[18], Clara S. Chan[2,3], Cedric Chauve[19], Joanna C. Chiu[20], Mikkel Christensen[8], Carlo Costantini[21], Victoria L.M. Davidson[22], Elena Deligianni[23], Tania Dottorini[16], Vicky Dritsou[24], Stacey B. Gabriel[25], Wamdaogo M. Guelbeogo[26], Andrew B. Hall[27], Mira V. Han[28], Thaung Hlaing[29], Daniel S.T. Hughes[8,30], Adam M. Jenkins[31], Xiaofang Jiang[32,27], Irwin Jungreis[2,3], Evdoxia G. Kakani[33,34], Maryam Kamali[35], Petri Kemppainen[36], Ryan C. Kennedy[37], Ioannis K. Kirmitzoglou[16,38], Lizette L. Koekemoer[39], Njoroge Laban[40], Nicholas Langridge[8], Mara K.N. Lawniczak[16], Manolis Lirakis[41], Neil F. Lobo[42], Ernesto Lowy[8], Robert M. MacCallum[16], Chunhong Mao[43], Gareth Maslen[8], Charles Mbogo[44], Jenny McCarthy[11], Kristin Michel[22], Sara N. Mitchell[33], Wendy Moore[45], Katherine A. Murphy[20], Anastasia N. Naumenko[35], Tony Nolan[16], Eva M. Novoa[2,3], Samantha O'Loughlin[18], Chioma Oringanje[45], Mohammad A. Oshaghi[6], Nazzy Pakpour[46], Philippos A. Papathanos[16,24], Ashley N. Peery[35], Michael Povelones[47], Anil Prakash[48], David P. Price[49,50], Ashok Rajaraman[19], Lisa J. Reimer[51], David C.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Rinker[52], Antonis Rokas[52,53], Tanya L. Russell[17], N'Fale Sagnon[26], Maria V. Sharakhova[35], Terrance Shea[1], Felipe A. Simão[4,5], Frederic Simard[21], Michel A. Slotman[54], Pradya Somboon[55], Vladimir Stegniy[12], Claudio J. Struchiner[56,57], Gregg W.C. Thomas[58], Marta Tojo[59], Pantelis Topalis[23], José M.C. Tubio[60], Maria F. Unger[42], John Vontas[41], Catherine Walton[36], Craig S. Wilding[61], Judith H. Willis[62], Yi-Chieh Wu[2,3,63], Guiyun Yan[64], Evgeny M. Zdobnov[4,5], Xiaofan Zhou[53], Flaminia Catteruccia[33,34], George K. Christophides[16], Frank H. Collins[42], Robert S. Cornman[62], Andrea Crisanti[16,24], Martin J. Donnelly[51,65], Scott J. Emrich[13], Michael C. Fontaine[42,66], William Gelbart[67], Matthew W. Hahn[68,58], Immo A. Hansen[49,50], Paul I. Howell[69], Fotis C. Kafatos[16], Manolis Kellis[2,3], Daniel Lawson[8], Christos Louis[41,23,24], Shirley Luckhart[46], Marc A.T. Muskavitch[31,70], José M. Ribeiro[71], Michael A. Riehle[45], Igor V. Sharakhov[35,27], Zhijian Tu[32], Laurence J. Zwiebel[72], and Nora J. Besansky[42,†]

Affiliations

[1]Genome Sequencing and Analysis Program, Broad Institute, 415 Main Street, Cambridge, MA 02142, USA. [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA. [3]The Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA. [4]Department of Genetic Medicine and Development, University of Geneva Medical School, rue Michel-Servet 1, 1211, Geneva, Switzerland. [5]Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland. [6]Department of Medical Entomology and Vector Control, School of Public Health and Institute of Health Researches, Tehran University of Medical Sciences, Tehran, Iran. [7]George Washington University, Department of Mathematics and Computational Biology Institute, 45085 University Drive, Ashburn, VA 20147, USA. [8]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [9]National Vector Borne Disease Control Programme, Ministry of Health, Tafea Province, Vanuatu. [10]Department of Public Health and Infectious Diseases, Division of Parasitology, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy. [11]Department of Biological Sciences, Cal Poly Pomona, 3801 West Temple Avenue, Pomona, CA 91768, USA. [12]Tomsk State University, 36 Lenina avenue, Tomsk, Russia. [13]Department of Computer Science and Engineering, Eck Institute for Global Health, 211B Cushing Hall, University of Notre Dame, Notre Dame, IN 46556, USA. [14]Inserm, U963, Immune responses in the malaria vector A. gambiae, F-67084, Strasbourg, France. [15]CNRS, UPR9022, Immune responses and development in insects, IBMC, F-67084, Strasbourg, France. [16]Department of Life Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [17]Faculty of Medicine, Health and Molecular Science, Australian Institute of Tropical Health Medicine, James Cook University, Cairns 4870, Australia. [18]Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK. [19]Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby BC, V5A1S6, Canada. [20]Department of Entomology and Nematology, One Shields Avenue, UC Davis, Davis, CA 95616, USA. [21]Institut de

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Recherche pour le Développement, UMR MIVEGEC, 911, Avenue Agropolis, BP 64501 Montpellier, France. [22]Division of Biology, Kansas State University, 271 Chalmers Hall, Manhattan, KS 66506, USA. [23]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Hellas, Nikolaou Plastira 100 GR-70013, Heraklion, Crete, Greece. [24]Centre of Functional Genomics, University of Perugia, Perugia, Italy. [25]Genomics Platform, Broad Institute, 415 Main Street, Cambridge, MA 02142, USA. [26]Centre National de Recherche et de Formation sur le Paludisme, Ouagadougou 01 BP 2208, Burkina Faso. [27]Program of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA 24061, USA. [28]School of Life Sciences, University of Nevada, Las Vegas, NV 89154, USA. [29]Department of Medical Research, No. 5, Ziwaka Road, Dagon Township, Yangon 11191, Myanmar. [30]Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA. [31]Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. [32]Department of Biochemistry, Virginia Tech, Blacksburg, VA 24061, USA. [33]Harvard School of Public Health, Department of Immunology and Infectious Diseases, Boston, MA 02115, USA. [34]Dipartimento di Medicina Sperimentale e Scienze Biochimiche, Università degli Studi di Perugia, Perugia, Italy. [35]Department of Entomology, Virginia Tech, Blacksburg, VA 24061, USA. [36]Computational Evolutionary Biology Group, Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, M13 9PT, UK. [37]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94143, USA. [38]Bioinformatics Research Laboratory, Department of Biological Sciences, New Campus, University of Cyprus, CY 1678, Nicosia, Cyprus. [39]Wits Research Institute for Malaria, Faculty of Health Sciences, and Vector Control Reference Unit, National Institute for Communicable Diseases of the National Health Laboratory Service, Sandringham 2131, Johannesburg, South Africa. [40]National Museums of Kenya, P.O. Box 40658-00100, Nairobi, Kenya. [41]Department of Biology, University of Crete, 700 13 Heraklion, Greece. [42]Eck Institute for Global Health and Department of Biological Sciences, University of Notre Dame, 317 Galvin Life Sciences Building, Notre Dame, IN 46556, USA. [43]Virginia Bioinformatics Institute, 1015 Life Science Circle, Virginia Tech, Blacksburg, VA 24061, USA. [44]KEMRI-Wellcome Trust Research Programme, CGMRC, PO Box 230-80108, Kilifi, Kenya. [45]Department of Entomology, 1140 E. South Campus Drive, Forbes 410, University of Arizona, Tucson, AZ 85721, USA. [46]Department of Medical Microbiology and Immunology, School of Medicine, University of California Davis, One Shields Avenue, Davis, CA 95616, USA. [47]Department of Pathobiology, University of Pennsylvania School of Veterinary Medicine, 3800 Spruce Street, Philadelphia, PA 19104, USA. [48]Regional Medical Research Centre NE, Indian Council of Medical Research, PO Box 105, Dibrugarh-786 001, Assam, India. [49]Department of Biology, New Mexico State University, NM 88003, USA. [50]Molecular Biology Program, New Mexico State University, Las Cruces, NM 88003, USA. [51]Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L35QA, UK. [52]Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

37235, USA. [53]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA. [54]Department of Entomology, Texas A&M University, College Station, TX 77807, USA. [55]Department of Parasitology, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand. [56]Fundação Oswaldo Cruz, Av Brasil 4365, RJ Brazil. [57]Instituto de Medicina Social, UERJ, Rio de Janeiro, Brazil. [58]School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA. [59]Department of Physiology, School of Medicine, CIMUS, Instituto de Investigaciones, Sanitarias, University of Santiago de Compostela, Spain. [60]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK. [61]School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, L3 3AF, UK. [62]Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA. [63]Department of Computer Science, Harvey Mudd College, Claremont, CA 91711, USA. [64]Program in Public Health, College of Health Sciences, University of California, Irvine, Hewitt Hall, Irvine, CA 92697, USA. [65]Malaria Programme, Wellcome Trust Sanger Institute, Cambridge, CB10 1SJ, UK. [66]Centre of Evolutionary and Ecological Studies (CEES-MarECon group), University of Groningen, Nijenborgh 7, NL-9747 AG Groningen, The Netherlands. [67]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Ave, Cambridge, MA 02138, USA. [68]Department of Biology, Indiana University, Bloomington, IN 47405, USA. [69]Centers for Disease Control and Prevention, 1600 Clifton Road NE MSG49, Atlanta, GA 30329, USA. [70]Biogen Idec, 14 Cambridge Center, Cambridge, MA 02142, USA. [71]Laboratory of Malaria and Vector Research, NIAID, 12735 Twinbrook Parkway, Rockville MD 20852, USA. [72]Departments of Biological Sciences and Pharmacology, Institutes for Chemical Biology, Genetics and Global Health, Vanderbilt University and Medical Center, Nashville, TN 37235, USA.

Acknowledgements

All sequencing reads and genome assemblies have been submitted to NCBI (umbrella BioProject ID = PRJNA67511). Genome and transcriptome assemblies are also available from VectorBase (http://vectorbase.org) and the Broad Institute (http://olive.broadinstitute.org/collections/anopheles.4).

The authors wish to acknowledge the NIH Eukaryotic Pathogen and Disease Vector Sequencing Project Working Group for guidance and development of this project. Sequence data generation was supported at the Broad Institute by the National Human Genome Research Institute (U54 HG003067). We would like to thank the many members of the Broad Institute Genomics Platform and Genome Sequencing and Analysis Program who contributed to sequencing data generation and analysis.

References and Notes

1. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 2005; 77:171–192. [PubMed: 16001361]
2. Cohuet A, Harris C, Robert V, Fontenille D. Evolutionary forces on Anopheles: what makes a malaria vector? Trends Parasitol. 2010; 26:130–136. [PubMed: 20056485]
3. Manguin, S.; C, P.; Mouchet, J.; E, JL. Biodiversity of Malaria in the World. John Libbey Eurotext; 2008.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4. Holt RA. The genome sequence of the malaria mosquito Anopheles gambiae. Science. 2002; 298:129–149. [PubMed: 12364791]
5. Marinotti O. The genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids Res. 2013; 41:7387–7400. [PubMed: 23761445]
6. Zhou D. Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. BMC Genomics. 2014; 15:42. [PubMed: 24438588]
7. Jiang X. Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. Genome Biol. 2014; 15:459. [PubMed: 25244985]
8. Neafsey DE. The evolution of the Anopheles 16 genomes project. G3 Bethesda Md. 2013; 3:1191–1194.
9. Fontaine MC. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. in press.
10. Moreno M. Complete mtDNA genomes of Anopheles darlingi and an approach to anopheline divergence time. Malar. J. 2010; 9:127. [PubMed: 20470395]
11. Gnerre S. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 2011; 108:1513–1518. [PubMed: 21187386]
12. Materials and methods are available as supplementary material on Science Online.
13. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12:491. [PubMed: 22192575]
14. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. A polytene chromosome analysis of the Anopheles gambiae species complex. Science. 2002; 298:1415–1418. [PubMed: 12364623]
15. Kamali M. Multigene phylogenetics reveals temporal diversification of major african malaria vectors. PloS One. 2014; 9:e93580. [PubMed: 24705448]
16. Toups MA, Hahn MW. Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. Genetics. 2010; 186:763–766. [PubMed: 20660646]
17. Baker DA, Russell S. Role of Testis-Specific Gene Expression in Sex-Chromosome Evolution of Anopheles gambiae. Genetics. 2011; 189:1117–1120. [PubMed: 21890740]
18. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. 2013; 30:1987–1997. [PubMed: 23709260]
19. Hahn MW, Han MV. S.-G. Han, Gene Family Evolution across 12 Drosophila Genomes. PLoS Genet. 2007; 3:e197. [PubMed: 17997610]
20. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 2007; 24:1586–1591. [PubMed: 17483113]
21. Dottorini T. 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. 2007; 104:16215–16220. [PubMed: 17901209]
22. Baldini F, Gabrieli P, Rogers DW, Catteruccia F. Function and composition of male accessory gland secretions in Anopheles gambiae: a comparison with other insect vectors of infectious diseases. Pathog. Glob. Health. 2012; 106:82–93. [PubMed: 22943543]
23. Assis R, Zhou Q, Bachtrog D. Sex-Biased Transcriptome Evolution in Drosophila. Genome Biol. Evol. 2012; 4:1189–1200. [PubMed: 23097318]
24. Grath S, Parsch J. 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. 2012; 4:346–359. [PubMed: 22321769]
25. Perry JC, Harrison PW, Mank JE. The Ontogeny and Evolution of Sex-Biased Gene Expression in Drosophila melanogaster. Mol. Biol. Evol. 2014; 31:1206–1219. [PubMed: 24526011]
26. Rogers DW. Transglutaminase-mediated semen coagulation controls sperm storage in the malaria mosquito. PLoS Biol. 2009; 7:e1000272. [PubMed: 20027206]
27. Baldini F. The interaction between a sexually transferred steroid hormone and a female protein regulates oogenesis in the malaria mosquito Anopheles gambiae. PLoS Biol. 2013; 11:e1001695. [PubMed: 24204210]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

28. Shaw WR. Mating activates the heme peroxidase HPX15 in the sperm storage organ to ensure fertility in Anopheles gambiae. Proc. Natl. Acad. Sci. U. S. A. 2014; 111:5854–5859. [PubMed: 24711401]
29. Willis JH. Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. Insect Biochem. Mol. Biol. 2010; 40:189–204. [PubMed: 20171281]
30. Cornman RS. Annotation and analysis of a large cuticular protein family with the R&R Consensus in Anopheles gambiae. BMC Genomics. 2008; 9:22. [PubMed: 18205929]
31. Cornman RS, Willis JH. Extensive gene amplification and concerted evolution within the CPR family of cuticular proteins in mosquitoes. Insect Biochem. Mol. Biol. 2008; 38:661–676. [PubMed: 18510978]
32. Cornman RS, Willis JH. Annotation and analysis of low-complexity protein families of Anopheles gambiae that are associated with cuticle. Insect Mol. Biol. 2009; 18:607–622. [PubMed: 19754739]
33. Cornman RS. Molecular evolution of Drosophila cuticular protein genes. PloS One. 2009; 4:e8345. [PubMed: 20019874]
34. Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH. Developmental expression patterns of cuticular protein genes with the R&R Consensus from Anopheles gambiae. Insect Biochem. Mol. Biol. 2008; 38:508–519. [PubMed: 18405829]
35. Rinker DC. Blood meal-induced changes to antennal transcriptome profiles reveal shifts in odor sensitivities in Anopheles gambiae. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:8260–8265. [PubMed: 23630291]
36. Rinker DC. Antennal transcriptome profiles of anopheline mosquitoes reveal human host olfactory specialization in Anopheles gambiae. BMC Genomics. 2013; 14:749. [PubMed: 24182346]
37. Altstein M, Nässel DR. Neuropeptide signaling in insects. Adv. Exp. Med. Biol. 2010; 692:155–165. [PubMed: 21189678]
38. Goetze JP, Hunter I, Lippert SK, Bardram L, Rehfeld JF. Processing-independent analysis of peptide hormones and prohormones in plasma. Front. Biosci. Landmark Ed. 2012; 17:1804–1815. [PubMed: 22201837]
39. Stracker TH, Thompson S, Grossman GL, Riehle MA, Brown MR. Characterization of the AeaHP gene and its expression in the mosquito Aedes aegypti (Diptera: Culicidae). J. Med. Entomol. 2002; 39:331–342. [PubMed: 11931033]
40. de Oliveira CD, Tadei WP, Abdalla FC, Paolucci Pimenta PF, Marinotti O. Multiple blood meals in Anopheles darlingi (Diptera: Culicidae). J. Vector Ecol. 2012; 37:351–358. [PubMed: 23181859]
41. Okamoto N. A fat body-derived IGF-like peptide regulates postfeeding growth in Drosophila. Dev. Cell. 2009; 17:885–891. [PubMed: 20059957]
42. Riehle MA, Fan Y, Cao C, Brown MR. Molecular characterization of insulin-like peptides in the yellow fever mosquito, Aedes aegypti: expression, cellular localization, and phylogeny. Peptides. 2006; 27:2547–2560. [PubMed: 16934367]
43. Antonova, Y.; Arik, AJ.; Moore, W.; Riehle, MM.; Brown, MR. Insect Endocrinology. Gilbert, LI., editor. Academic Press; 2012. p. 63-92.
44. Swaminathan A, Gajan A, Pile LA. Epigenetic regulation of transcription in Drosophila. Front. Biosci. Landmark Ed. 2012; 17:909–937. [PubMed: 22201781]
45. Cipressa F. Effete, a Drosophila chromatin-associated ubiquitin-conjugating enzyme that affects telomeric and heterochromatic position effect variegation. Genetics. 2013; 195:147–158. [PubMed: 23821599]
46. Arcà B. Positive selection drives accelerated evolution of mosquito salivary genes associated with blood-feeding. Insect Mol. Biol. 2014; 23:122–131. [PubMed: 24237399]
47. Isawa H, Yuda M, Orito Y, Chinzei Y. A mosquito salivary protein inhibits activation of the plasma contact system by binding to factor XII and high molecular weight kininogen. J. Biol. Chem. 2002; 277:27651–27658. [PubMed: 12011093]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

48. Calvo E. 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. 2007; 282:26928–26938. [PubMed: 17650501]

49. Guittard E. CYP18A1, a key enzyme of Drosophila steroid hormone inactivation, is essential for metamorphosis. Dev. Biol. 2011; 349:35–45. [PubMed: 20932968]

50. Waterhouse RM. Evolutionary Dynamics of Immune-Related Genes and Pathways in Disease-Vector Mosquitoes. Science. 2007; 316:1738–1743. [PubMed: 17588928]

51. Bartholomay LC. Pathogenomics of Culex quinquefasciatus and meta-analysis of infection responses to diverse pathogens. Science. 2010; 330:88–90. [PubMed: 20929811]

52. Barillas-Mury C, Han YS, Seeley D, Kafatos FC. Anopheles gambiae Ag-STAT, a new insect member of the STAT family, is activated in response to bacterial infection. EMBO J. 1999; 18:959–967. [PubMed: 10022838]

53. Gupta L. The STAT pathway mediates late-phase immunity against Plasmodium in the mosquito Anopheles gambiae. Cell Host Microbe. 2009; 5:498–507. [PubMed: 19454353]

54. Bahia AC. The JAK-STAT pathway controls Plasmodium vivax load in early stages of Anopheles aquasalis infection. PLoS Negl. Trop. Dis. 2011; 5:e1317. [PubMed: 22069502]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

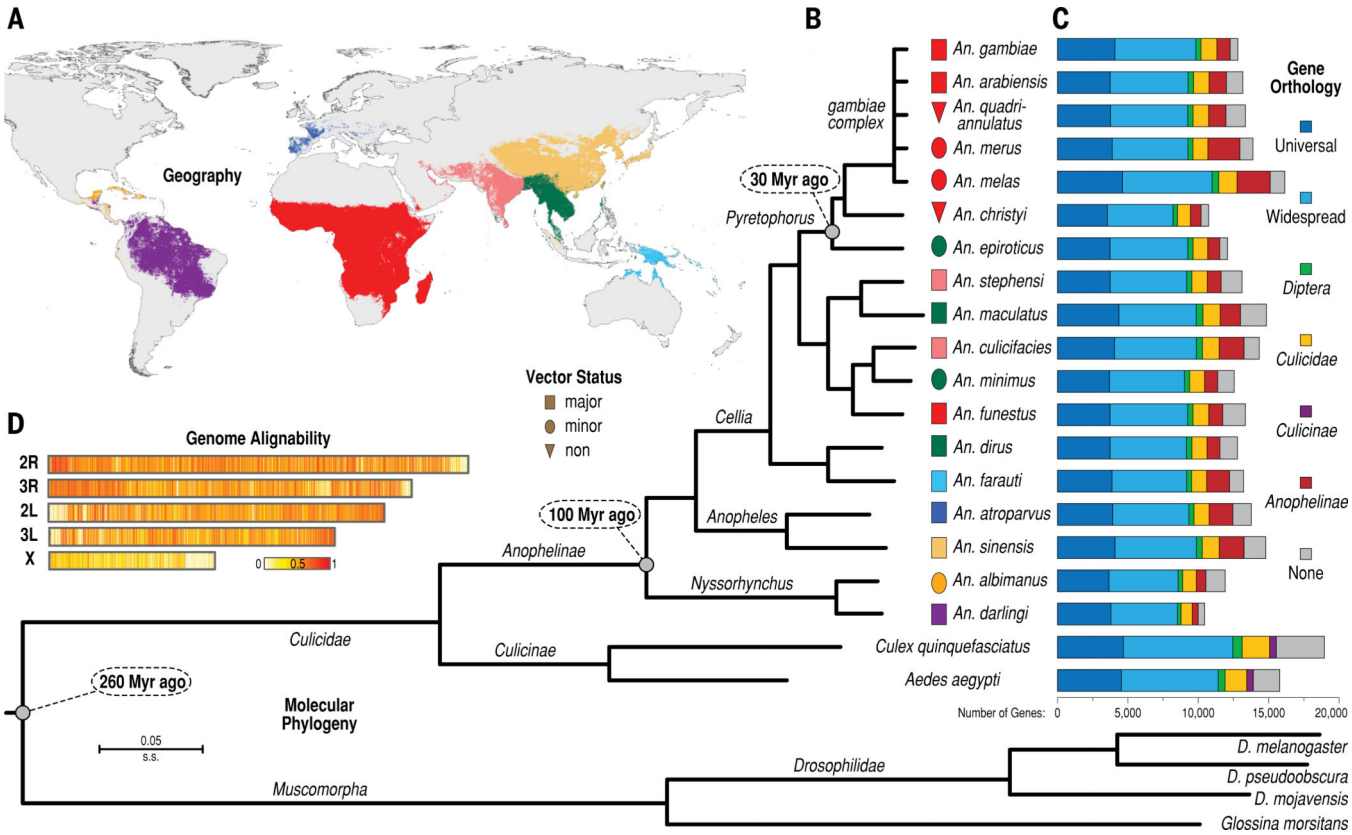

Figure 1. Geography, vector status, molecular phylogeny, gene orthology, and genome alignability of the 16 newly sequenced anopheline mosquitoes and selected other dipterans

(**A**) Global geographic distributions of the 16 sampled anophelines and the previously sequenced *An. gambiae* and *An. darlingi*. Ranges are colored for each species or group of species as shown in panel B, e.g. light blue for *An. farauti*. (**B**) The maximum likelihood molecular phylogeny of all sequenced anophelines and selected dipteran outgroups. Shapes between branch termini and species names indicate vector status (rectangles, major vectors; ellipses, minor vectors, triangles, non-vectors) and are colored according to geographic ranges shown in panel A. (**C**) Barplots show total gene counts for each species partitioned according to their orthology profiles; from ancient genes found across insects to lineage-restricted and species-specific genes. (**D**) Heat map illustrating the density (in 2 kb sliding windows) of whole genome alignments along the lengths of *An. gambiae* chromosomal arms: from white where *An. gambiae* aligns to no other species, to red where *An. gambiae* aligns to all the other anophelines.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

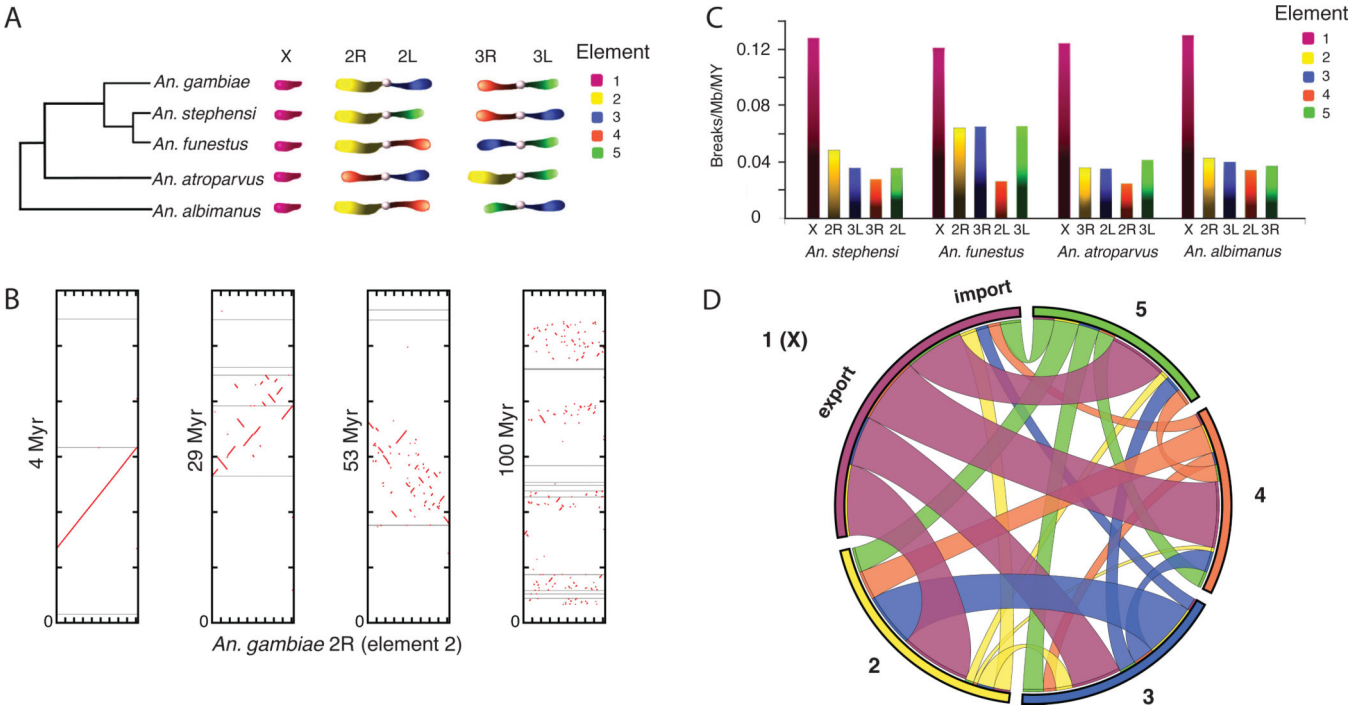

Figure 2. Patterns of anopheline chromosomal evolution

(**A**) Anopheline genomes have conserved gene membership on chromosome arms ('elements'; colored and labeled 1–5). Unlike *Drosophila*, chromosome elements reshuffle between chromosomes via translocations as intact elements, and do not show fissions or fusions. The tree depicts the supported molecular topology for the species studied. (**B**) Conserved synteny blocks decay rapidly within chromosomal arms as the phylogenetic distance increases between species. Moving left to right, the dotplot panels show gene-level synteny between chromosome 2R of *An. gambiae* (x axis) and inferred ancestral sequences (y axes; inferred using PATHGROUPS) at increasing evolutionary timescales (MYA = million years ago) estimated via an ultrametric phylogeny. Gray horizontal lines represent scaffold breaks. Discontinuity of the red lines/dots indicates rearrangement. (**C**) Anopheline X chromosomes exhibit higher rates of rearrangement ($P < 1\times10^{-5}$), measured as breaks per megabase (Mb) per million years (MY), compared with autosomes, despite a paucity of polymorphic inversions on the X. (**D**) The anopheline X chromosome also displays a higher rate of gene movement to other chromosomal arms than any of the autosomes. Chromosomal elements are labeled around the perimeter; internal bands are colored according to the chromosomal element source and match element colors in panels A and C. Bands are sized to indicate the relative ratio of genes imported versus exported for each chromosomal element, and the relative allocation of exported genes to other elements.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

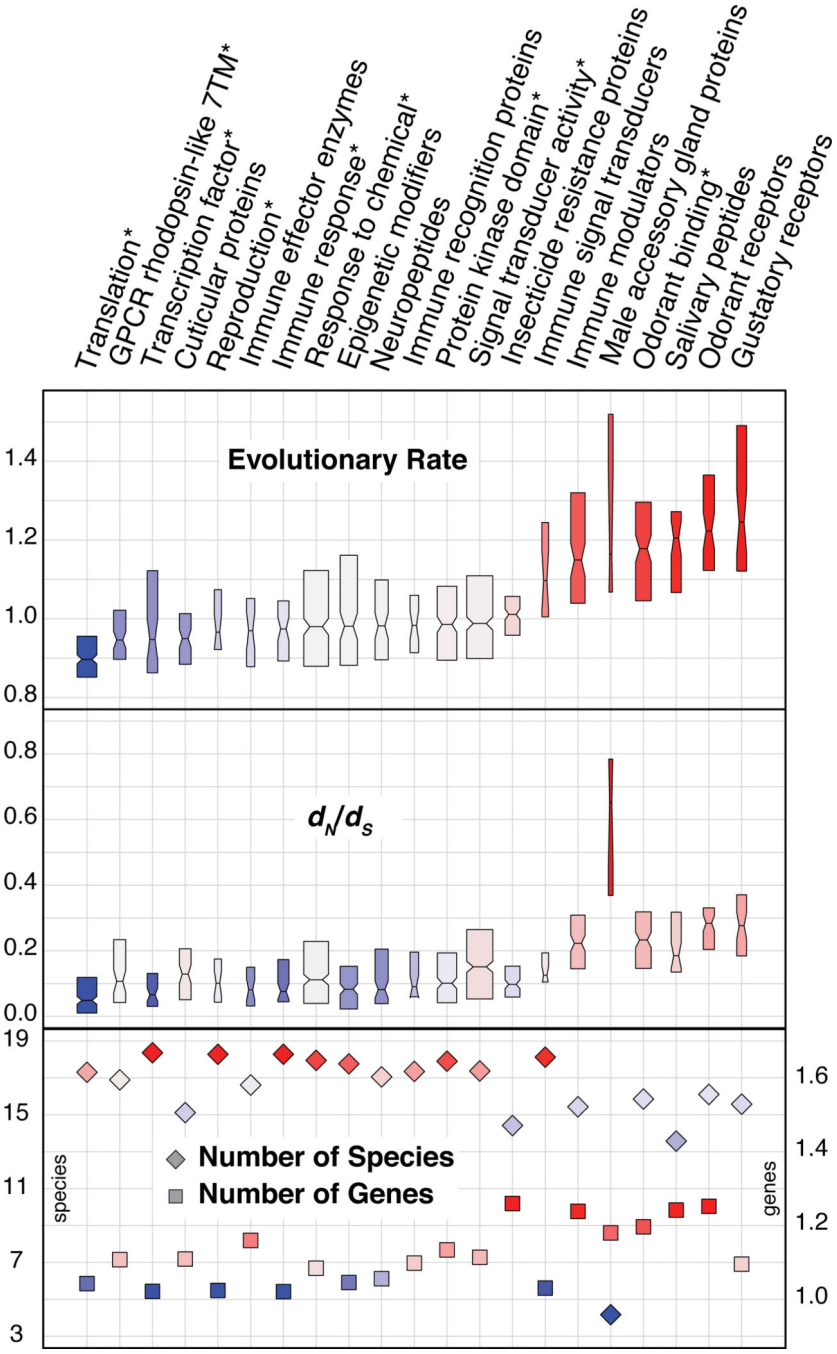

Figure 3. Contrasting evolutionary properties of selected gene functional categories

Examined evolutionary properties of orthologous groups of genes include: a measure of amino acid conservation/divergence (evolutionary rate), a measure of selective pressure (d_N/d_S), a measure of gene duplication in terms of mean gene copy-number per species (number of genes), and a measure of ortholog universality in terms of number of species with orthologs (number of species). Notched boxplots show medians, extend to the first and third quartiles, and their widths are proportional to the number of orthologous groups in each functional category. Functional categories derive from curated lists associated with various

functions/processes as well as annotated Gene Ontology or InterPro categories (denoted by asterisks).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

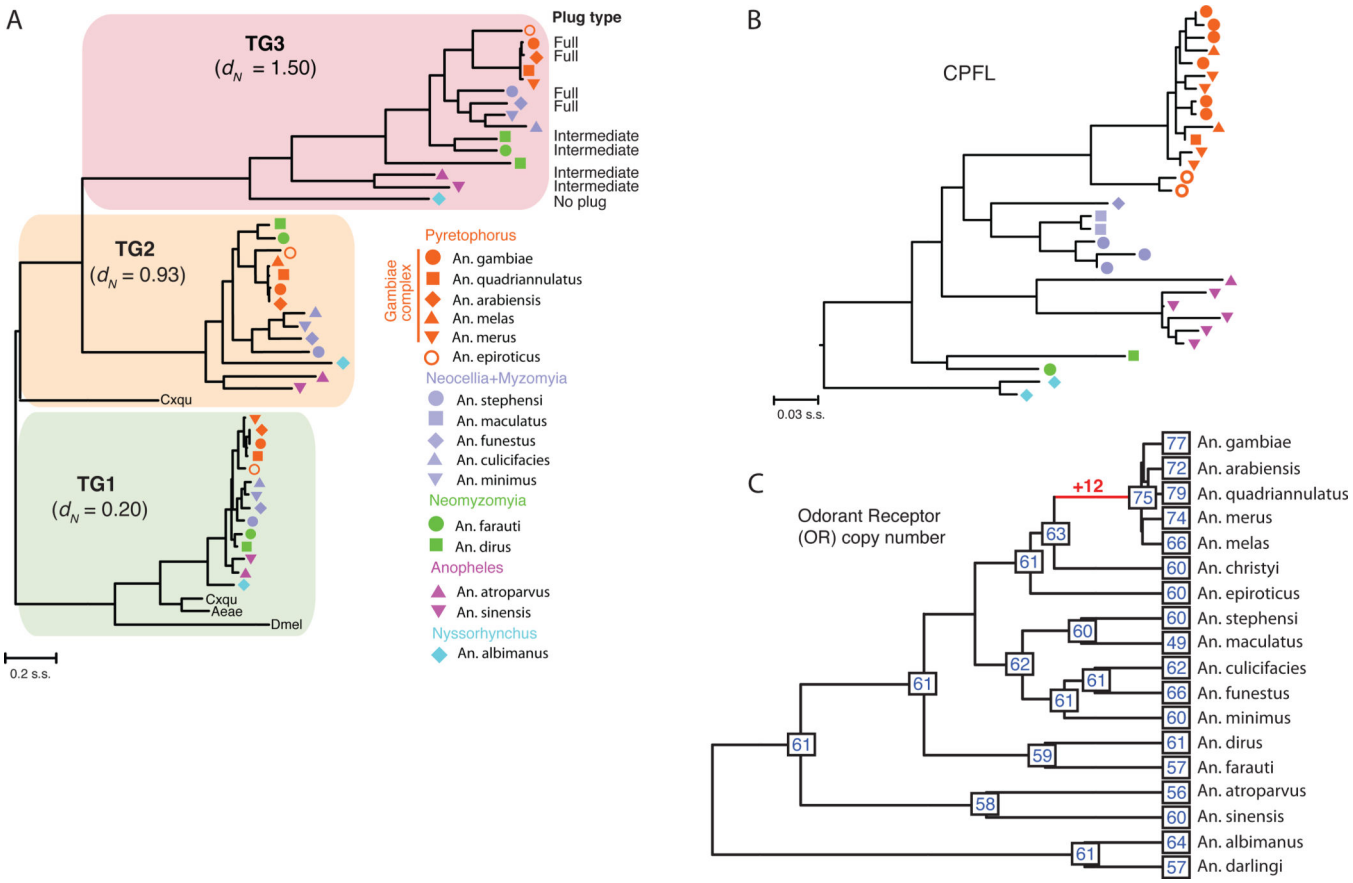

Figure 4. Phylogeny-based insights into anopheline biology

(**A**) Maximum-likelihood amino acid based phylogenetic tree of three transglutaminase enzymes (*TG1* (green), *TG2* (yellow) and *TG3* (red)) in 14 anopheline species with *Culex quinquefasciatus* (Cxqu), *Ae. aegypti* (Aeae) and *D. melanogaster* (Dmel) serving as outgroups. *TG3* is the enzyme responsible for the formation of the male mating plug in *An. gambiae*, acting upon the substrate Plugin, the most abundant mating plug protein. Higher rates of evolution for plug-forming *TG3* are supported by elevated levels of d_N. Mating plug phenotypes are noted where known within the *TG3* clade. (**B**) Concerted evolution in CPFL cuticular proteins. Species symbols used are the same as in panel a. In contrast to the *TG1/TG2/TG3* phylogeny, CPFL paralogs cluster by sub-generic clades rather than individually recapitulating the species phylogeny. Gene family size variation among species may reflect both gene gain/loss and variation in gene set completeness. (**C**) Odorant receptor (*OR*) observed gene counts and inferred ancestral gene counts on an ultrametric phylogeny. At least 10 *OR* genes were gained on the branch leading to the common ancestor of the *An. gambiae* species complex, though the overall number of *OR* genes does not vary dramatically across the genus.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

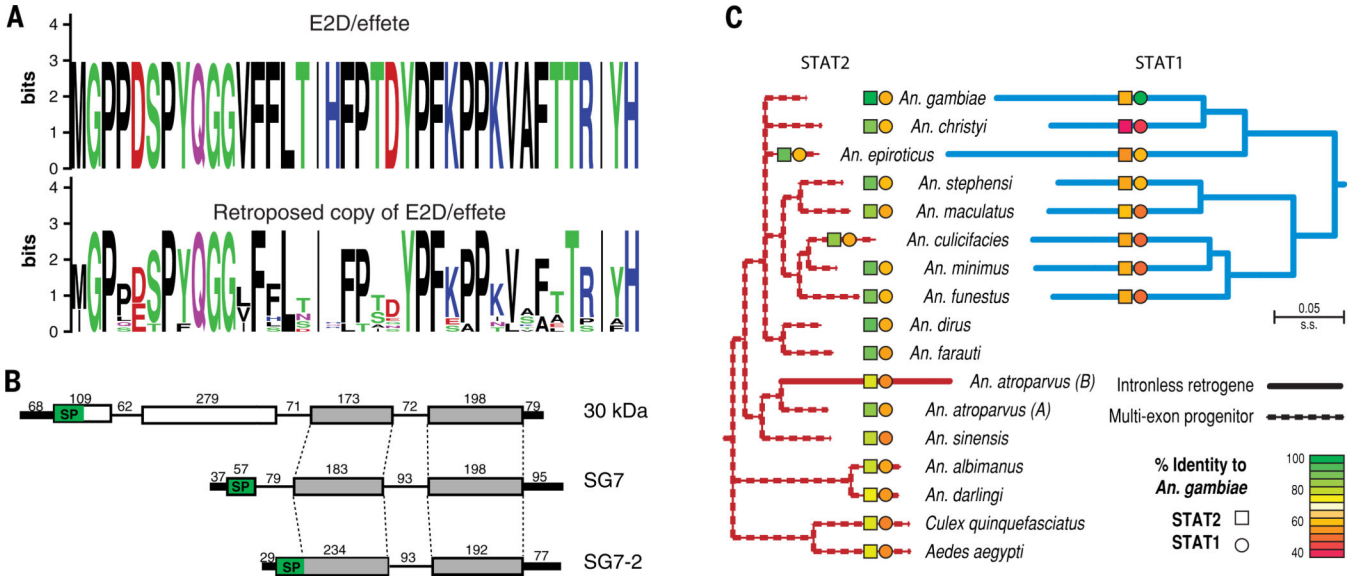

Figure 5. Genesis of novel anopheline genes

(**A**) Retrotransposition of the *E2D/effete* gene generated a ubiquitin-conjugating enzyme at the base of the genus, which exhibits much higher sequence divergence than the original multi-exon gene. WebLogo plots contrast the amino acid conservation of the original *effete* gene with the diversification of the retrotransposed copy (residues 38–75; species represented are *An. minimus*, *An. dirus*, *An. funestus*, *An. farauti*, *An. atroparvus*, *An. sinensis*, *An. darlingi*, and *An. albimanus*). (**B**) The *SG7* salivary protein-encoding gene was generated from the C-terminal half of the 30 *KDa* gene. *SG7* then underwent tandem duplication and intron loss to generate another salivary protein, *SG7-2*. Numerals indicate length of segments in base pairs. (**C**) The origin of *STAT1*, a signal transducer and activator of transcription gene involved in immunity, occurred through a retrotransposition event in the *Cellia* ancestor after divergence from *An. dirus* and *An. farauti*. The intronless *STAT1* is much more divergent than its multi-exon progenitor, *STAT2*, and has been maintained in all descendent species. An independent retrotransposition event created a retrogene copy in *An. atroparvus*, which is also more divergent than its progenitor.