Supporting Online Material

Table of Contents

Taxonomy and mosquito strain selection

Mosquitoes within the present *C. quinquefasciatus* species have previously been classified as *C. fatigans* and *C. pipiens quinquefasciatus*, the southern house mosquito. For a taxonomic review see Mattingly *et al.* (*1*).

We sought to verify that no contamination of the strain had occurred since it had been established (March 2001) by sequencing a 500 bp. fragment of the *white* gene from JHB and 6 other *Culex* strains it has been housed near. This showed that all the sequenced strains had distinct mutations that distinguished them from the JHB colony. These mutations were not matched in the *C. quinquefasciatus* contig or trace files, suggesting that no laboratory contamination had occurred since the original field collection. Therefore, while these data are not conclusive, because of the small data set and because lost colonies could not be examined, there appeared to be no indication that the difficulty in genome assembly (see "Assembly fragmentation" below) was due to contamination since the colony's inception.

The JHB colony used in this study is maintained at the University of California, Riverside, USA; and at the University of California, Davis, USA. Contact P. Atkinson or A. Cornel for colony maintenance records and tissue availability.

Genome sequencing and assembly

Sequencing and assembly of the 579 million base pair (Mbps) *C. quinquefasciatus* genome was performed through a collaboration of the Broad Institute (Broad) and the J. Craig Venter Institute (JCVI) with 6.14X average sequencing coverage. Assembly of shotgun sequencing trace files was performed using the ARACHNE 2 program (*2*). This resulted in an assembly containing 48,671 contig fragments with N50 contig size of 28.55 Kbps. These were assembled into 3,171 scaffold sequences with N50 scaffold size of 486.76 Kbps. Contig sizes ranged from 201 bp. to 11,094 bp. and scaffold sizes ranged from 1,197 bp. to 3,873,010 bp. This genome had a GC percentage of 37.42.

Assembly fragmentation

Sequencing and assembly of the *C. quinquefasciatus* genome was performed by the same two sequencing centers as the *Ae. aegypti* genome project, with no large quality differences in sequencing output detected between the two centers. Therefore, there appeared to be no *a-priori* reason to suspect that technical difficulties with the sequencing or assembly could explain the unexpectedly high level of fragmentation of the *C. quinquefasciatus* genome. High levels of repeated sequences in a genome could make it difficult for the assembly software to create large scaffolds. However, a detailed analysis of transposable elements and other repeated sequences did not indicate an unexpectedly high diversity of such sequences compared to the *An. gambiae* and *Ae. aegypti* genomes (see "Transposable Elements" section below). Total assembly size was not very different from an estimate based on reassociation kinetics (540 Mbps, (*3*)), suggesting that the possibility of a significant portion of the genome not having been incorporated into the assembly was unlikely.

We examined if the presence of two or more haplotypes, resulting from interbreeding of genetically distinct individuals, could have contributed to the fragmentation problem by assessing the duplication status of 26 markers uniquely present in this species and 7 expected single copy genes. All the markers, as well as three expected single copy genes, match a single location in the *C. quinquefasciatus* genome. The remaining single copy genes only had weak matches in the genome thus could not lead to any conclusion (Table S2). An abundance of paralogs with similar intronic regions would also be an indication of a potential haplotype problem. Thus, we estimated the degree of similarity in intronic regions of *C. quinquefasciatus* paralogs. We found that 1% of the paralogs intronic regions were similar at more than 50% to another paralog intronic regions, suggesting that the majority of the paralogs had not been artificially created by an haplotype issue. To further quantify the haplotype problem, we identified *C. quinquefasciatus* genes having twice as many paralogs as their *Ae. aegypti* and *An. gambiae* counterparts (2:1:1, 4:2:2 and 6:3:3 categories), and compared it to the same calculation in *Ae. aegypti. C. quinquefasciatus* had less paralogs than *Ae. aegypti* in category "2:1:1" (583 vs.683), the same number in category "4:2:2" (134) and 36 in category "6:3:3" (none for *Ae. aegypti*). The small differences observed in the figures between these species, suggests that, while some duplications are observed, the problem is not more significant here than it was for the *Ae. aegypti* genome assembly (*4*).

To examine whether any assembly fragmentation due to allelic variation had resulted in the assembly of haplotype scaffolds that could inflate the number of predicted genes, we examined the percent protein sequence identities among identified paralogous genes in *C. quinquefasciatus* compared to *Ae. aegypti* and *An. gambiae*. Employing GeneTrees defined using the Ensembl Compara pipeline (*5*) at VectorBase, the percent protein identities for all pairs of within-species paralogs were compared among the three mosquito species (Fig. S4). 8,009 *C. quinquefasciatus* within-species paralogs from 2,225 GeneTrees result in 43,940 pairwise identities, 6,987 *Ae. aegypti* within-species paralogs from 2,158 GeneTrees result in 37,246 pairwise identities, and 3,854 *An. gambiae* within-species paralogs from 1,124 GeneTrees result in 19,298 pairwise identities. The mean percent identities of paralog pairs were very similar for *C. quinquefasciatus* (36.2%) and *Ae. aegypti* (37.4%), and lower for *An. gambiae* (33.1%). The proportion of paralog pairs with very high percent identities (i.e. those that could possibly be haplotypes) is slightly higher in *C. quinquefasciatus* (Fig. S4A). However, partitioning the pairwise identities into those between paralogs on the same versus those on different supercontigs (chromosomes for *An. gambiae*) revealed that the majority of the very closely-related pairs of *C. quinquefasciatus* paralogs are found on the same supercontigs (Fig. S4) rather than on different supercontigs (Fig. S4). Such closely-related paralogs located on the same sequence region are more likely to be real genes resulting from recent tandem duplications while those found on different sequence regions (especially very short ones) could be haplotypes. Incompatible overlaps between different sequencing reads in the same region, caused by different haplotypes, tend to create short assembled regions that end up artificially separated in the assembly. Highly similar paralogues on different sequences could thus be haplotypes, but being on the same sequence provides more assurance that they are more likely to be recent tandem duplicates rather than haplotypes. Examining the numbers of pairs of paralogs on different sequence regions compared to the number of pairs of paralogs on the same sequence regions (different/same ratio) revealed very similar ratios for *C. quinquefasciatus* (5.73) and *Ae. aegypti* (5.72) which have similar numbers of supercontigs with paralogous genes, 1,594 and 1,429 respectively, compared to the lower *An. gambiae* ratio (1.71) which has only 6 different sequence regions. Thus, the possible fragmented assembly of *C. quinquefasciatus* haplotype regions does not appear to be a major issue, and is comparable to that of the *Ae. aegypti* genome.

DNA/DNA comparative analysis

Among sequenced mosquito genomes *C. quinquefasciatus* and *Ae. aegypti* are most closely related phylogenetically (subfamily Culicinae, Fig. 1A in main text). Therefore, it would be expected that the sequences of these two genomes should be more similar to each other overall than to *An. gambiae*. We tested this expectation by running pairwise translated DNA comparisons using BLAT (*6*) between these three mosquito genomes, as well as with the distantly related *D. melanogaster* genome. *C. quinquefasciatus* had 2.6 times more DNA alignments with *Ae. aegypti* than with *An. gambiae*. *C. quinquefasciatus* also had 5 times more DNA alignments with *Ae. aegypti* than with *D. melanogaster*, confirming our expectation. Average identity percentages and block lengths are shown in Table S3.

Automated gene annotation

Three automated gene prediction pipelines were run independently by the two sequencing centers and Vectorbase. These were later merged by Broad into a single initial consensus gene set CpipJ1.1, later updated to gene set CpipJ1.2 (see “Merging of gene sets from the three institutions” below). The three centers used different approaches to generate each gene set in order to improve the gene discovery rate. The methodologies used by each center are described below. Updates to the gene set are curated by VectorBase (*7*). Gene, intron, and exon statistics of the most recent gene set (CpipJ1.2) are shown in Table S4, along with similar statistics for the *An. gambiae*, *Ae. aegypti* and *D. melanogaster* genomes.

J. Craig Venter Institute gene prediction pipeline methodology

A repeat library was generated using the program RepeatScout (*8*). Repeat family members that occurred more than 50 times in the genome or that had detectable homology to known transposable elements were compiled into a repeat library. This library was then used with the RepeatMasker program (*9*) to identify and mask repeat instances in the genome. An initial set of gene predictions was generated based on protein homologies, by aligning GenBank dipteran proteins onto the genome using the AAT (*10*) and GeneWise (*11*) programs. Concurrent with this analysis, *C. quinquefasciatus* ESTs were aligned to the genome and high quality alignments were used for automated gene structure annotations using the software packages PASA (stringent condition) (*12*) and AAT (paralog predictions). Finally, five *ab initio* gene prediction programs were run on the genome: SNAP (*13*), Phat (*14*), Augustus (*15*), GlimmerHMM (*16*), and Twinscan (*17*). *Ab initio* gene prediction models generated by Broad (see below) were also added to this gene set. These *ab initio* gene sets were combined into one set using the EVidenceModeler utility (*18*). A total of 23,165 gene models were generated.

The Broad Institute gene prediction pipeline methodology

Supercontig sequences were masked using the repeat libraries generated by JCVI (described above) and VectorBase (described below). Additional transposon and other repeat sequences were identified and masked using the BLAST algorithm with a data set of approximately 37,650 transposons and repeat sequences from GenBank. Five *ab-initio* prediction programs, Augustus (*15*), SNAP (*13*), GeneID (*19*), FgeneSH (*20*) and GeneWise (*11*), were trained with existing gene sets from the *Ae. aegypti* and *C. quinquefasciatus* genomes. Gene models predicted by JCVI were also integrated. Non-coding RNAs were identified by running the RFAM (*21*) and tRNAScan (*22*) programs on the entire genome. Overlapping gene predictions were clustered into loci and a custom gene caller program was executed to evaluate each prediction and select the most likely gene model based on splice sites and similarity to known proteins. A total of 18,673 genes were identified.

VectorBase gene prediction pipeline methodology

VectorBase's approach to gene prediction differed from the other two pipelines by focusing on similarity to known genes rather than *ab initio* gene model prediction. The Ensembl pipeline (*23*) was used to predict protein coding and non-coding genes using mRNA, EST/cDNA and protein evidence. Supercontig sequences were initially masked using the RepeatMasker program (*9*) with a library of *C. quinquefasciatus*, *Ae. aegypti* and *An. gambiae* repeat sequences from public databases, as well as repeats identified using the RECON (*24*) and RepeatScout (*8*) programs. UniProt protein sequences (*25*) were mapped to the supercontig sequences using the Genewise program (*11*). Two gene sets were then built based on the taxonomic origin of the proteins: 1) a "targeted" gene set from *C. quinquefasciatus* proteins only, with strict criteria, and 2) a "similarity" gene set from the remaining proteins. In the "similarity" gene set, gene predictions were prioritized according to protein origin: genes based on *D. melanogaster* proteins were placed first on the genome, then additional non-overlapping models were added based on mosquito, diptera, eukaryota and finally metazoa proteins. Independently, the *C. quinquefasciatus* EST and mRNA sequences were mapped to the supercontig sequences using the Exonerate program (*26*), generating a third gene set. Finally, an *ab initio* gene set was built by running the SNAP program (*13*) on the supercontig sequences and retaining only predictions containing a Pfam domain. The four gene sets were then merged into a single gene set containing 14,207 genes.

Merging of gene sets from the three institutions

The gene sets generated by JCVI, Broad, and VectorBase were merged into a single consensus gene set by Broad, using the same procedure as for the *Ae. aegypti* annotation (*4*). Statistics of the merging are given in Table S5. In average, between 12% and 31% of the genes were in common between the sets, with JCVI and Broad being the most similar, as would be expected since both sets were largely based on the same approach. These two pipelines, as any *ab initio*-based prediction methods, had a tendency to over-predict genes, while the very conservative, similarity-based, VectorBase pipeline is likely to have missed some. The combination of the three methods ensured a higher rate of gene discovery. Once consolidated, the merged set was considered as the reference set and the intermediate gene sets were discarded.

This final gene set, CpipJ1.1, contained 20,394 protein-coding gene models and 4,030 non-protein coding genes. Following a manual review of some of these gene models (see "Gene number overestimate" below), 1,511 gene models were found to be invalid and were removed from the gene set, with no new models added. This new set of 18,883 genes was called CpipJ1.2 and is the basis of all subsequent analysis.

Expressed sequence tags (ESTs)

A total of 75,848 EST sequences from whole tissue adults samples were used to inform automated gene predictions. Among protein coding genes in the CpipJ1.2 data set 4,257 genes (22.5%) matched at least one of these EST sequences, the majority of these matches (4,114) were to gene coding regions. Sequences were deposited in the dbEST database (*27*).

Quality of protein-coding gene predictions

In an effort to estimate the quality of the protein-coding gene predictions for *C. quinquefasciatus*, *Ae. aegypti* and *An. gambiae*, we examined the lengths of single-copy orthologs between these genomes and the well annotated *D. melanogaster* gene set. Single-copy orthologs are likely to experience strong evolutionary constraints on gene structure and function. While natural variations in the encoded lengths of such single-copy orthologs are to be expected (mainly due to genomic insertions and deletions) they should nevertheless exhibit strong positive correlations among different species. We found generally good concordance between the mosquito and fruitfly gene lengths, suggesting that the *C. quinquefasciatus* gene set was of good quality.

The results from the orthology delineation procedures among the three mosquito species and twelve Drosophilids from the OrthoDB resource (*28*) were interrogated to identify all single-copy orthologs among the mosquitoes and two fruitflies, *D. melanogaster* and *D. mojavensis*. The amino acid lengths of 4,269 strict single-copy orthologs (one member in each of the five species) were compared using the *D. melanogaster* proteins as the baseline. The scatter plots in Fig. S5 show the *D. melanogaster* protein length (x) against the orthologous protein length (y) for each species: the dashed lines show a linear regression, and the solid lines show a robust linear regression. The concordance of x and y are given with 95% confidence limits (CL), perfect concordance (1.0) would require all points to fall on the 45 degree line (x=y) of perfect agreement falling on the border between the shaded and un-shaded regions. To examine the distributions of evident deviations from perfect agreement, the density of data points falling at each degree below and above 45 degrees were plotted (solid colored curves). These density distributions were compared to normal fittings of the data (dotted colored curves) with means fixed at 45 degrees. The areas representing the positive differences between the observed data and the normal fitted data below and above one standard deviation from the mean of the normal fitted data (σ, dashed gray vertical lines) are filled with the respective colors for each species. The values of these proportions of significantly shorter proteins (<σ) and significantly longer proteins (>σ) are enumerated for quantitative comparisons. By way of comparison to the mosquitoes, the results from the same analyses with *D. mojavensis* (much more closely-related yet still one of the most distantly related of the sequenced Drosophilids) are also shown. The *D. mojavensis* proteins achieve a concordance value of 0.96, while the mosquitoes exhibit lower concordance values in agreement with their larger evolutionary distance from *D. melanogaster. An. gambiae* achieves concordance of 0.92 and *Ae. aegypti* of 0.93, while *C. quinquefasciatus* was only slightly less consistent at 0.90. *D. mojavensis* protein lengths were only slightly skewed towards shorter predictions compared to *D. melanogaster.* This trend was more evident in all three mosquito species.

Employing homology-based approaches conserved single-copy orthologs are often the simplest genes to predict, and as such this analysis likely examined a subset of some of the most accurately predicted proteins in each species. Nevertheless, the results provide a clear indication of the good quality of the *C. quinquefasciatus* protein-coding gene predictions relative to *An. gambiae* and *Ae. aegypti*.

Analysis of gene numbers

Because of the unexpectedly large number of gene models predicted for *C. quinquefasciatus* by the automated gene prediction pipelines compared to other mosquitoes, a number of additional analyses were undertaken to understand the nature of this increase and are described in detail below. Taken together these analyses showed that while the automated consensus gene set likely overestimated the number of genes by as much as 14% (overestimates of gene numbers are likely to have occurred in the other two mosquito genome annotations as well) this increase was unlikely to be caused in large part by the mis-annotation as genes of random genome sequences. Furthermore, they indicate that the *C. quinquefasciatus* genome contains significantly more expanded gene families than the other two mosquitoes, supporting the conclusion that the observed gene increase in *C. quinquefasciatus* is rooted in biological reality.

Partial manual gene re-annotation

In an effort to quantify the accuracy of the initial automated consensus gene set (CpipJ1.1), a detailed manual examination of 841 automated gene predictions from random supercontigs files was undertaken. After review 419 of these genes (50%) required no modification, 171 genes (20%) required structural modifications to their annotation without affecting the total gene number, 123 genes (15%) were merged together into 54 new genes, 12 genes (1.4%) were split into 30 genes, 91 genes (11%) were deleted, and 25 new genes were added (3%). This resulted in a net reduction of 117 genes (14%) from the consensus gene set. Applied to the CpipJ1.2 *C. quinquefasciatus* gene dataset this would yield an estimated gene number of 16,239 genes, still larger than either of the other two mosquito genomes. This could be considered a conservative estimate because this manual review was performed prior to the update of the gene set to version CpipJ1.2, where 1,511 gene models were removed. Overestimates in the number of *An. gambiae* genes have also been observed (*29*).

Examination of singletons in gene clusters

Spurious gene predictions (random open reading frames) could contribute to the larger set of predicted protein-coding genes in *C. quinquefasciatus*. However, such miss-predictions would not be expected to exhibit homology with other proteins and therefore would augment the proportion of singletons in sequence clustering analyses. Using the procedure described below we examined the proportion of singletons across a range of sequence clustering stringencies and found that the augmented total number of genes is unlikely to be due to the inclusion of many spurious gene predictions.

The National Center for Biotechnology Information's (NCBI) Blastclust utility allows clustering of protein sequences based on all-against-all BLAST comparisons. Selection of variable cut-offs of sequence lengths and identities of the pairwise BLAST matches can build clusters of varying stringency. Groups of proteins with shared domains exhibiting sufficient sequence identity will cluster together into protein families according to the criteria applied. We carried out this clustering analysis on the three mosquito proteomes using length restrictions of 50% and 70% along one of the two proteins being compared and sequence identity cut-offs of 20% to 90% (in 10% increments). The results of this analysis on individual proteomes are shown in the Fig. S6 bar charts, and the results on combined proteomes are shown in the Fig. S6 pie charts (only 50% identity cut-off is shown).

An. gambiae exhibited the lowest proportions of proteins that formed part of multi-gene families (i.e. higher proportions of singletons) while *C. quinquefasciatus* and *Ae. aegypti* both showed similar and larger proportions of clustered sequences (i.e. lower proportions of singletons). This indicated that the larger proteomes of *C. quinquefasciatus* and *Ae. aegypti* contained more members of multi-gene families rather than spurious gene predictions based on random open reading frames. When all three proteomes were analyzed together (Fig. S6 pie charts) a dramatic reduction in the proportions of singletons was observed as well as a substantial increase in the proportion of clusters with three members. These shifts represent the proteins that have no homologs within each individual proteome but have likely one-to-one-to-one orthologs in the other mosquito proteomes. This analysis supports an overall trend of *An. gambiae* having the smallest cluster sizes and *C. quinquefasciatus* the largest. This conclusion was also supported by evidence from the gene family expansion analyses (see below). The proportions of singletons at the 50% length cut-off were comparable among the three mosquito species (left pie chart, gray slices). However, at the more stringent length cut-off of 70% the proportion of *C. quinquefasciatus* singletons increased when compared to the two other mosquitoes (right pie chart, gray slices). This could also be seen in the individual proteome analyses, as the sequence identity stringencies increased the difference in the number of singletons between *C. quinquefasciatus* and *Ae. aegypti* increased as well, with an excess of *C. quinquefasciatus* singletons. This could arise from divergent *C. quinquefasciatus* duplicates appearing as singletons (falling out of clusters) at higher sequence identity stringencies.

At the most stringent sequence identity cut-off of 90% *C. quinquefasciatus* exhibited a small yet distinct proportion of clusters with more than 20 members, this was not the case for either *An. gambiae* or *Ae. aegypti*. Investigating the nature of these groups revealed several large clusters of histone proteins which are known to occur at high copy-numbers and exhibit high sequence similarities. Therefore, these large clusters are likely to be the result of genuine gene expansions. However, a small number of likely contaminants were also identified: viral attachment proteins (IPR009013) which are not found in Metazoa (Length70-Identity90: 60 proteins).

Gene family expansion

The large *C. quinquefasciatus* gene set could be due to elevated gene duplication events that have created multiple copies of many genes. We examined this possibility by looking at the proportionate sizes of multi-gene families among the mosquito proteomes. These analyses supported the existence of significant gene expansions in *C. quinquefasciatus* when compared to *Ae. aegypti* and *An. gambiae.*

Employing the Blastclust utility from the NCBI (see "Examination of singletons in gene clusters" above) all proteins from the three mosquito species were clustered with pairwise BLAST matches of all-against-all sequence comparisons. This clustering analysis was repeated eight times using two sequence length and four sequence identity cut-offs. The sequence length cut-offs required that at least 50% or 70% of the length of one of the pair of sequences formed part of the match. The identity cut-offs required the pairwise match to have a sequence identity of at least 30%, 40%, 50%, or 60%. In order to focus on multi-gene families only clusters with more than 10 members were retained. In addition, to strictly examine gene expansions clusters were required to have at least one member protein from each of the three mosquito species. The results of these clustering analyses are shown as boxplots in Fig. S7 and Table S6 where values for paired Wilcoxon signed-rank tests showing significant differences at each cut-off level are shown.

The advantage of this approach was that it did not require knowledge of protein domains to cluster protein families since it only used sequence similarities. Because different gene families likely evolved at different rates, the analysis had to be performed over a range of different cut-offs to make sure that the trend was the same all the way through different levels of sequence conservation. In all the clustering analyses *C. quinquefasciatus* clusters were larger than those of *Ae. aegypti*. In turn, all *Ae. aegypti* clusters were larger than those of *An. gambiae*. This result strongly supported the conclusion that significant gene family expansions in *C. quinquefasciatus* led to the increased total predicted gene count. These data also suggest that gene family expansions were partially responsible for the larger predicted gene set in *Ae. aegypti* over *An. gambiae*.

Synteny analysis

Homology pipeline

Orthologs and paralogs were determined by running the Ensembl GeneTree pipeline (*5*) between *C. quinquefasciatus* (genebuild CpipJ1.2), *Ae. aegypti* (genebuild AaegL1.1) and *An. gambiae* (genebuild AgamP3.4), using *D. melanogaster* (genebuild FlyBase 4.3), *Homo sapiens* (genebuild NCBI36) and *Caenorhabditis elegans* (genebuild WB170) as outgroups. The homology relationships derived from by this pipeline were the basis for all the subsequent analyses.

Microsynteny

Microsynteny blocks were defined so that they contained at least two single-copy orthologous genes that have maintained their local gene neighborhood in each pair of genomes, allowing only a limited number of intervening genes. Close to a quarter of the *C. quinquefasciatus* genome fell into microsynteny blocks with the other two mosquito genomes and encompassed 79% of the orthologs between *C. quinquefasciatus* and *Ae. aegypti* and 70% of those between *C. quinquefasciatus* and *An. gambiae* (Table S7).

A map of conserved local rearrangements was generated by identifying genomic blocks that satisfied the following conditions: 1) each block contained at least two neighboring one-to-one orthologs in each pair of genomes, 2) in each block 33% or fewer of the genes did not have one-to-one orthologs in each pair of genomes, 3) if orthologous genes from a pair genomes were on different chromosomes then only two such genes were allowed between orthologous pairs on the same chromosome, and 4) no more than 5 genes with no one-to-one orthologs were allowed between any pair of orthologous genes. Approximately 3% of these microsynteny blocks contained intervening genes with orthologs to another chromosome and approximately 13% of blocks contained genes with no orthologs. The results are shown in Table S7. As expected, the *C. quinquefasciatus* and *Ae. aegypti* genomes showed greater synteny than either the *C. quinquefasciatus* and *An. gambiae* pair or the *C. quinquefasciatus* and *D. melanogaster* pair. *C. quinquefasciatus* and *Ae. aegypti* had longer syntenies, included more genes, and a larger proportion of the *C. quinquefasciatus* genome (in base pairs and in genes) was found to be in synteny.

Macrosynteny

Macrosynteny was estimated by counting the number of orthologous genes shared between *C. quinquefasciatus* or *Ae. aegypti* scaffolds and *An. gambiae* and *D. melanogaster* chromosome arms. A higher level of conservation was observed between *C. quinquefasciatus* scaffolds and *An. gambiae* and *D. melanogaster* chromosome arms than between *Ae. aegypti* scaffolds and their *An. gambiae* and *D. melanogaster* counterparts (Table S8). Moreover, the small size of the synteny blocks also suggested that significant gene shuffling had taken place with increased levels of rearrangements observed in *Ae. aegypti* compared to *C. quinquefasciatus*. Finally, a detailed analysis of the distribution of embedded and overlapping gene pairs among all three mosquito genomes showed that many of these pairs were not conserved between *C. quinquefasciatus* and the other two mosquitoes, indicating that gene relocation events had occurred. Taken together these data strongly suggest that gene shuffling within the same chromosome arms has occurred during the evolution of these mosquitoes, and that among the Culicinae the genome of *C. quinquefasciatus* has remained more stable than that of *Ae. aegypti*.

The extent of *C. quinquefasciatus* macrosynteny with *An. gambiae* and *D. melanogaster* was assessed in two ways. For each *C. quinquefasciatus* scaffold the number of genes, or synteny blocks, with an ortholog to a given *An. gambiae* or *D. melanogaster* chromosome arm was counted. Scaffolds with less than 2 blocks were excluded from the synteny blocks analysis. The same analysis was performed between *Ae. aegypti, An. gambiae* and *D. melanogaster* (Table S8). Comparison between the two culicinae was not possible due to the current fragmented states of their genomes. The percentage of scaffolds with genes (or syntenies) with orthologs on a single *An. gambiae* or *D. melanogaster* chromosome arm was found to be higher in *C. quinquefasciatus* than in *Ae. aegypti*, while the percentage was lower for genes (or syntenies) with orthologs on two or more *An. gambiae* or *D. melanogaster* chromosome arms. This supported the conclusion that the *Ae. aegypti* chromosomes had undergone more extensive rearrangements than the *C. quinquefasciatus* chromosomes, which was also supported by the chromosomal location analysis.

Orthology relationships

The Ensembl GeneTree pipeline (*5*) was employed to delineate orthology relations among protein-coding genes of the three mosquitoes and three outgroup species: *D. melanogaster*, *C. elegans* and

Homo sapiens. Nearly two thirds of *C. quinquefasciatus* genes exhibited orthologous relations to genes in both of the other two mosquitoes, with a conserved core of 4,744 genes maintained as strict single-copy orthologs (Fig. 1C main text). Those with *D. melanogaster* single-copy orthologs facilitated codon-based estimation of DNA substitutions in the three mosquitoes and were used to construct a phylogenetic tree of mosquito relationships (Fig. 1A main text). A further 10% of *C. quinquefasciatus* genes shared orthology exclusively with *Ae. aegypti*, likely representing genes specific to the Culicinae subfamily, and only 2% with *An. gambiae*, highlighting missing annotations or possible losses in the *Aedes* lineage. The larger total number of *C. quinquefasciatus* genes was mirrored in all categories of orthologous groups with multi-copy orthologs: *C. quinquefasciatus* had more genes than *Ae. aegypti,* and genes common to the Culicinae were more numerous than those specific to *An. gambiae* (Table S9).

Mosquito InterPro domains

Comparison of InterPro domains identified in the three mosquito genomes with those of *D. melanogaster* revealed that the largest expansions within the mosquitoes were among genes linked to olfaction (IPR006625) and blood clotting and platelet aggregation (IPR002181) (Fig. S8A and Table S10). A similar analysis performed using Gene Ontology (GO) terms showed that terms involved with iron transport (GO:0020037, GO:0009239), olfaction (GO:0004984, GO:0005549), and exoskeleton (GO:0042302, GO:0006032) were expanded (Fig. S8B and Table S11).

Chromosomal assignment

Using 34 mapped *C. quinquefasciatus* and *Ae. aegypti* markers (*30*, *31*) as well as unpublished *C. quinquefasciatus* marker data (*32*), a chromosomal location was assigned to 38 *C. quinquefasciatus* genes by aligning these markers to *C. quinquefasciatus* supercontig sequences and looking for markers overlapping with genes (Table S12). These results were then extrapolated to all the genes on the supercontig sequences, assuming that if one gene on a supercontig had been located to a chromosome then all the genes from this same supercontig sequence should map to that same chromosome. Using this assumption one marker (CX61) was mapped to a supercontig sequence even though it did not map to a gene. A total of 1,768 genes were placed on the three *C. quinquefasciatus* chromosomes. Orthology analysis with *An. gambiae* and *D. melanogaster* were based on the orthologs/paralogs predicted by the Ensembl GeneTree pipeline previously described (*5*). We looked for potential correlations between *C. quinquefasciatus* chromosomes with *An. gambiae* and *D. melanogaster* chromosomes (Table S13). These results were compared with a similar analysis in *Ae. aegypti* (*4*). This indicated that there is likely whole chromosome conservation between *C. quinquefasciatus*, *An. gambiae*, and *D. melanogaster*, whereas *Ae. aegypti* would have undergone a chromosome arm exchange (Fig. S1).

Phylogenetic analysis

The phylogenetic tree represented in Figure 1A was derived from alignments of single-copy orthologs between *C. quinquefaciatus*, *Ae. aegypti*, *An. gambiae*, and *D. melanogaster* that were analysed using PAML's baseml implementation with the "G3" model allowing separate rates for each codon position (*33*). Dates of divergence were sourced from previous studies (*34*, *35*, *36*).

Repeated sequences and transposable elements

Transposable element annotation

Transposable element (TE) discovery methods were purposefully similar to methods used for the *An. gambiae* and *Ae. aegypti* genomes in order to facilitate comparisons between these three genomes (*4*, *37*). The following automated repeated sequence and TE discovery methods were used: RECON (*24*), RepeatScout (*8*), a general TE discovery algorithm developed by J. M. C. R., as well as individual search algorithms designed for specific TE families. These last TE family specific algorithms are available upon request from individual researchers listed in the TEFam database (described below). The output of the automated TE and repeated sequence discovery methods were used to generate a preliminary list of TEs. Using this preliminary list expert research groups conducted a thorough search for specific TE families. The decision of which criteria to use to define each TE family was left to the expert groups but these criteria generally required the presence of terminal inverted repeats, presence of open reading frames (except in the case of MITE sequences), as well as similarity at the nucleotide level of over 75% for all members of a single TE family. Representative sequences of all identified TE families were deposited into the TEFam database (*38*) along with contact names for expert research groups.

Genomic coverage by transposable element derived sequences in three mosquito genomes

Frequency and genomic coverage of *C. quinquefasciatus* TE sequences was estimated using RepeatMasker (*9*). TE copy number and percentage genome coverage was estimated using the same parameters as those used for the *Ae. aegypti* genome (*4*). TE sequences were manually screened for simple repeat sequences (in addition to the built-in screen of the RepeatMasker program) to avoid over-representation of their genomic coverage and copy number. Percentage of the genome occupied by single/low copy DNA, simple and tandem repeats, and unclassified repeats were calculated by parsing the outputs of the RECON, RepeatScout, RepeatMasker programs. In addition to *C. quinquefasciatus* the genomic coverage by TEs in the *Ae. aegypti* and *An. gambiae* genomes was also estimated using the same methods as for *C. quinquefasciatus*. TE libraries for these genomes were generated using the TEFam database (*38*). However, because TEFam does not contain all known TEs for *An. gambiae* the library for this species was supplemented with *An. gambiae* specific sequences from the RepBase database (*39*), unpublished MITE sequences used by Holt *et al.* (*37*), and with novel MITE sequences discovered in the course of this analysis. All MITE sequences were deposited in the TEFam database.

Overall the TE percentage estimated for *Ae. aegypti* and *An. gambiae* did not diverge substantially from the estimates reported by Holt *et al.* (*37*) and Nene *et al.* (*4*), but a few differences merit examination. Holt *et al.* (*37*) reported that approximately 16% of the euchromatic portion of the *An. gambiae* genome was composed of TEs; we estimated that 12% of the genome sequences were derived from TEs. While our estimate included both euchromatic and heterochromatic sequences, it is likely that TEs in the poorly assembled heterochromatin were substantially underrepresented in our estimate since we relied on sequence similarity to infer the presence of TEs. Nene *et al.* (*4*) estimated that MITE sequences occupied 16% of the *Ae. aegypti* genome while the present estimate is only 10%. The difference between these estimates is largely explained by the presence in Nene *et al.* (*4*) of a category of MITEs labeled “otherMITEs”. To our knowledge these “otherMITEs” were not deposited in a public database and could not be included in the present study.

Retrotransposons

At least 72% of the 171 LTR retrotransposon elements had full-length insertions with intact open reading frames (ORFs) into the genome of *C. quinquefasciatus*. Interestingly, there was evidence of trans-mobilization of LTR-retrotransposons. This was suggested by 1) the presence of elements containing only a gag ORF and long terminal repeats (TEfam accession numbers TF001486, TF001487, TF001562, and TF001564), and 2) the presence of an element with long terminal repeat sequences and a long internal non-coding sequence resembling the Large Retrotransposon Derivative elements (LARDs) described previously in several plant genomes (*40, 41*) (TEfam accession TF001656).

Eleven of the 17 known non-LTR retrotransposon clades were identified in the *C. quinquefasciatus* genome based on their reverse transcriptase (RT) domain. These included two unique gag-only nonautonomous CM-gag retroposons that lack an RT domain (TEFam accessions TF001657 and TF001658). These were placed in the Jockey clade based on gag-domains similarity. Full-length copies of the Jockey, CR1, L1, L2, R1, LOA, Loner and I clades were found the *C. quinquefasciatus* genome. In addition EST sequences were mapped to some of the full-length Jockey, CR1, L1, CM-gag, R1 and I clades. A few clades showed substantial divergence within *C. quinquefasciatus*, for example the L1 clade includes 57 families and the CR1 clade 39 families. Overall, non-LTR retrotransposon are highly diversified in mosquitoes with the Loner and Outcast clades unique to mosquito genomes.

Miniature Inverted Terminal Repeat elements (MITES)

Miniature inverted repeat elements (MITEs), sequences that lack coding potential and are believed to be mobilized by the transposase encoded by other DNA transposons, made up a large percentage of the assembled genome of *C. quinquefasciatus* (17%). This was a larger percentage than for *Ae. aegypti* (10%) and a substantially larger percentage than for *An. gambiae* (1%). These data suggest that large numbers of MITE-like sequences could be characteristic of the culicinae. Nene *et al.* (*4*) suggested that the high number of MITEs in *Ae. aegypti* could be evidence that they contributed to the large size of that genome compared to *An. gambiae*. The observation of a similarly large number of MITEs in *C. quinquefasciatus*, along with a larger genome size than *An. gambiae*, supports this view. To better understand the dynamics of MITE-like sequences in the *C. quinquefasciatus* genome, many MITEs were linked to the presumed transposase responsible for their movement (Table S14). This indicated that the largest number of *C. quinquefasciatus* MITE sequences resemble full length *hAT* TEs.

Transposable element age distribution in mosquitoes

Age distribution of TE classes was estimated using the methodology described in Waterson *et al.* (*42*). MITE sequences were not included in this analysis because sequences internal to the terminal inverted repeats of many mosquito MITEs appear to be evolving mostly in a non-neutral manner, mostly by internal rearrangements and segmental duplications. Percent divergences from consensus sequences reported by RepeatMasker were converted to nucleotide distance measures using the Jukes-Cantor formula to correct for multiple hits. Results were pooled into bins of single unit distances (Fig. S3). Absolute ages could not be assigned to the TE distance measures because we lack an appropriate understanding of the rate of evolution of these sequences. However, distance measures could be used for comparison of relative ages between the three genomes. Sequences in Fig. S3 were ordered from left to right from most similar to consensus sequences (youngest) to most distant (oldest).

Whilst numbers of base pairs occupied by TE derived sequences (Fig. S9) and percent of the genome occupied by TEs (Table S14) varied substantially between mosquito genomes, there was surprising uniformity in the relative age distribution of the various TE classes. LTR and non-LTR retroelements dominated the most recent relative age classes, with a gradual reduction of abundance over time. This pattern was consistent with the presence of recently active retrotransposons and with gradual degradation of these sequences. Conversely, DNA elements showed a more uniform age distribution pattern with smaller percentages of these elements in the most recent relative age classes in *Ae. aegypti* and *C. quinquefasciatus*. This pattern could be explained by the ability of some DNA elements to move horizontally into new host genomes followed by rapid increase in copy numbers (e.g. *43*). The similarity of the relative age distributions between the three genomes suggests that the large percentage of *Ae. aegypti* genome composed of TE sequences was unlikely to be due to higher fixation probabilities of TE sequences in this genome, as would be expected from historical fragmentations of the mosquito populations.

Gene family annotation by expert groups

Expert groups were provided the results of the automated gene annotation to assist in the annotation of specific gene families. Manually curated gene models were deposited in VectorBase (*7*).

Olfactory receptors

Insect olfactory receptors (ORs) are a highly divergent group of sensory receptors. With 180 identified OR-related sequences (162 with complete open reading frames) *C. quinquefasciatus* has the largest number of such sequences of all dipteran species examined to date; 62 sequences in *D. melanogaster* (*44*, *45*), 79 in *An. gambiae* (*46*) and 131 in *Ae. aegypti* (*47*). The apparent expansion of the OR gene repertoire appears to be characteristic of culicine mosquitoes and may reflect culicine olfactory behavioural diversity, particularly surrounding host-choice; *C. quinquefasciatus* feeds on both birds and humans, and some *Culex* populations appear to switch host preference in a seasonally directed manner (*48*). Previously, Bohbot *et al.* (*47*) identified 12 potentially orthologous OR gene families, consisting of 18 OR genes, between *Ae. aegypti* and *An. gambiae*, including the ubiquitous insect OR gene homologous to the Dm*Or83b* gene. Within these 12 putative 'mosquito' OR families, phylogenetic analyses (Fig. S10) revealed that three of the OR families maintain strict microsynteny conservation among all three sequenced mosquito genomes (OR7/40, OR6, and OR66); three families maintained strict microsynteny between two of the three species and displayed an expansion of the family in the other (OR25, 69 and 58); and there were five instances of gene family expansion and/or duplication (Fig. S10). In only one instance (OR43/44), microsynteny appears not to have been conserved in *C. quinquefasciatus*. There were ten apparent culicine OR families consisting of 59 ORs in total (23 *Ae. aegypti* and 36 *C. quinquefasciatus*). In all cases but one (Cq32-34; AaOR71) the *Ae. aegyypti* ORs were basal to *C. quinquefasciatus* (Fig. S11).

Gustatory Receptors

In *D. melanogaster* gustatory receptors (GRs) mediate perception of both odorants and tastants, for example a highly conserved lineage is known to mediate perception of carbon dioxide, while others are implicated in perception of sugars, bitter compounds, and cuticular hydrocarbons (*49*, *50*). Only the carbon dioxide receptors have been functionally characterized in mosquitoes (*52*), and as expected *C. quinquefasciatus* has all three of these conserved GRs (*52*, *53*). The sugar receptors are another highly divergent and reasonably well conserved lineage of GRs, and the *C. quinquefasciatus* genome encodes 14 of them. Only three other "simple" orthologous relationships of mosquito GRs and *Drosophila* GRs exist; the DmGr66a relatives, which presumably act as heteromeric partners for other bitter taste receptors (*54*), DmGr43a orthologs of unknown function, and orthologs of the DmGr28a/b genes with unusual expression patterns (*55*). The remaining relationships were of four kinds, as shown in Fig. S12A-D. First, there were a series of eight orthologous relationships of, usually, single mosquito GRs with no orthologs evident in *Drosophila* or beyond. Second, there were five apparent multiple independent duplications in each mosquito lineage where orthologous relationships remain unclear. Third, there were several instances of gene losses from one or more lineages. Fourth, there were four large alternatively-spliced loci, one of which is expanded within *C. quinquefasciatus* (CpGr76a-ii) with the potential to encode 35 GRs that differ in their N-terminal sequences, as is typical for these alternatively-spliced GR loci, for example in *Ae. aegypti* (*56*).

Salivary gland genes

Saliva of blood sucking arthropods contain a complex cocktail of pharmacologically active components that disarms their host's hemostasis, the physiological process responsible for stopping blood loss following vessel damage and comprised by redundant processes leading to platelet aggregation, blood clotting, and vessel constriction (*57*). The salivary glands of mosquitoes additionally serve a role in nectar feeding, and have sugar hydrolytic enzymes. Antimicrobial agents are ubiquitously found in saliva of these animals as well. Previous silotranscriptome analysis of anophelines (*An. gambiae, An. stephensi, An. funestus* and *An. darlingi*) (*58*, *59*, *60*, *61*) and culicine (*Ae. aegypti, Ae. albopictus* and *C. quinquefasciatus*) (*62*, *63*, *64*) mosquitoes revealed that there are 75-150 different secreted proteins associated with the salivary function, in many cases consisting of expanded gene families. Perhaps because the vertebrate host exercises immune pressure neutralizing the salivary activities of hematophagous arthropods, these genes are at fast pace of evolution, leading to the expression of novel protein families uniquely found in this organ. There are common gene families to all mosquitoes, such as the D7 protein family, distantly related to the odorant binding family (8 genes in *An. gambiae*) or the Aegyptin/30 kDa antigen family, unique to blood sucking Nematocera (single gene in anophelines, but two genes in culicines), as well as genus, or even subgenus specific families, such as the sG1 family unique to *Anopheles*, or the gSG6 protein found in anopheline subgenera *Celia* or *Anopheles*, but not on *Nyssorhynchus*. A large protein family named the 16.7 kDa family, unique to *Culex*, was previously discovered following salivary transcriptome analysis. The genome of *C. quinquefasciatus* reveals additional members of this family, totaling 28 genes, 13 of which have EST representation. Interestingly most of these genes are uniexonic, suggesting an expansion by retrotransposition. The function of these proteins is still unknown. The proteome annotation also allowed retrieval of protein families that were found by similarity to proteins identified in a more detailed transcriptome analysis of *Ae. aegypti* and *An. gambiae* such as members of the 62 kDa family, additional members of the D7 family as well as of the Aegyptin/30 kDa family, as well as other proteins. The annotated hyperlinked table of all these putative salivary proteins can be retrieved from the author's web site (*65*).

Selenoproteins

Selenoproteins are a diverse family of proteins containing Selenium (Se) in the form of the non-canonical amino acid selenocysteine (Sec). Selenocysteine, the 21st amino acid is similar to cysteine (Cys) but with Se replacing Sulphur. Selenocysteine is coded by UGA, normally a stop codon, and a number of factors combine to achieve the co-translational recoding of UGA to Sec (*66*). Selenoproteins exist in all domains of life, Eukarya, Eubacteria, and Archaea. However, no selenoproteins have been found in higher plants or fungi. Only three selenoproteins have been so far reported in insects, SPS2, SelH and SelK. Interestingly, some dipteran species (i.e. *Drosophila willistonii*) seem to have lost selenoproteins and the capacity to synthesize them (*67*).

Because of the non-standard usage of the UGA codon, selenoproteins are usually misannotated in eukaryote genomes. Here we used Selenoprofiles, a selenoprotein-oriented gene prediction pipeline to search the *C. quinquefasciatus* genome for selenoproteins and for proteins involved in selenocysteine synthesis and metabolism. We identified the three known insect selenoprotein, all of which are also present in the other sequenced mosquito genomes. We also identified all genes known to be involved in selenoprotein metabolism (SPS1, SPS2, secp43, eEFsec, pstk, SecS). Selenoprofile predictions were used to refine the initial gene structure of the selenoprotein genes predicted by our computational pipeline.

Figure Legends

Fig. S1. Cladogram of *C. quinquefasciatus*, *Ae. aegypti*, *An. gambiae* and *D. melanogaster* showing chromosome arm similarities (colors indicate syntenic chromsome arms). The double lines indicate a potential chromosomal arm exchange.

Fig. S2. Percentage occupancy of major genomic elements in *C. quinquefasciatus*. Retrotransposons (class I TEs) and DNA transposoable elements (class II TEs) are grouped together.

Fig. S3. Relative age distribution of transposable element classes in the three mosquito genomes using the methodology described in Waterson *et al*. (*42*). Jukes-Cantor corrected divergence measure from consensus TE sequences are shown along the horizontal axis, with sequences grouped into bins of 1 unit distance. Percent of the genome occupied by each TE class is shown along the vertical axis with classes stacked to improve readability.

Fig. S4. Analysis of percent protein sequence identities among within-species paralogs. Within-species paralogs for *C. quinquefasciatus* (*Cq*), *Ae. aegypti* (*Aa*), and *An. gambiae* (*Ag*) were identified using the Ensembl GeneTrees pipeline (*5*). The frequencies normalized by the number of paralog pairs (Density) of percent protein sequence identities between pairs of paralogs are plotted for all paralog pairs (**A**), paralog pairs on the same supercontig or chromosome (**B**), or paralog pairs on different supercontigs or chromosomes (**C**). Numbers next to species abbreviations indicate the number of paralog pairs in each species.

Fig. S5. Analysis of single copy ortholog protein lengths between *D. melanogaster, D. mojavensis* and the three mosquito species. Scatterplots show concordence between lengths of *D. melanogaster* and orthologs in the compared species. Distributions of deviations from perfect length agreement are shown as density distributions. See text (Quality of protein-coding gene predictions) for full details.

Fig. S6. Analysis of proteome clustering for the three mosquito species. Proportions of each mosquito proteome clustering into groups of varying sizes (from 2 to >20 members) as well as the remaining singletons are shown as bars for a range of clustering stringencies (20% identity to 90%). Bars are grouped in triads with the first bar showing *C. quinquefasciatus* (*Cq*) data, the second *Ae. aegypti* (*Aa*), and the third *An. gambiae* (*Ag*). Analyses using length restrictions of 50% are shown on the left, 70% on the right. Pie charts indicate similar analyses to the bar graphs with length restrictions of 50% and 70% and percent identity of 50%, but with all three proteomes combined.

Fig. S7. Analysis of mosquito gene family expansions. The number of members in each cluster, converted to percentages, is shown along the horizontal axis. For example, a cluster of 40 genes composed 8 genes from *An. gambiae*, 12 genes from *Ae. aegypti*, and 20 genes from *C. quinquefasciatus*, would be converted into cluster sizes of 20%, 30% and 50% respectively. Clusters were required to contain a total of more than 10 member proteins and to have at least one member

protein from each of the three mosquito species. Gene sequences with at least 30%, 40%, 50%, or 60% sequence identity along at least 50% (bottom half of the figure) or 70% (top half) along the length of one of the pairs of sequences forming the match are shown along the vertical axis. Boxplots, colored by species (*C. quinquefasciatus Cq* blue, *Ae. aegypti Aa* green, *An. gambiae Ag* red), show the median values (central lines), quartiles (left and right ends of the boxes), whiskers (dashed lines), and outliers (circles) for each dataset. The vertical size of each box is proportional to the number of clusters in each stringency dataset (#, given on right of figure). If the notched areas of two boxplots do not overlap they were considered to be significantly different and Table S6 gives values for paired Wilcoxon signed-rank tests showing significant differences at each stringency level.

Fig. S8. Top-50 Interpro domains (**A**) and Gene Ontology (GO) terms (**B**) present at least 4 times in each of the species: *C. quinquefasciatus*, *Ae. aegypti*, *An. gambiae* and *D. melanogaster*, ordered by ratio domain (or term) occurrence in mosquito vs. *D. melanogaster.* The ratio variation is indicated by the line. Left scale represents the domain (or the term) occurrence and right scale shows the ratio value.

Fig. S9. Number of base pairs occupied by transposable element derived sequences in the three mosquito genomes.

Fig. S10A-B. Phylogenetic relationships of the *C. quinquefasciatus*, *Ae. aegypti*, *An. gambiae* and *D. melanogaster* odorant receptor (OR) families. The tree was generated using distances calculated with the Jones-Taylor-Thornton amino acid exchange matrix (*68*) using Protdist v3.6 (*69*). The tree was rooted through the highly conserved orthologous *D. melanogaster* OR83b family (DmOr83b, Agor7, AaOR7 and CqOR7). This tree was supported by bootstrapping with 100 replicates via neighbor-joining. These uncorrected distances are shown in the appropriate branch points. Species/genera-specific gene expansions are indicated to the right of the tree by vertical lines. Protein names are abbreviated to CqOR, AaOR, GPRor and DmOr for *C. quinquefasciatus* (green), *Ae. aegypti* (red), *An. gambiae* (blue) and *D. melanogaster* (orange) respectively.

Fig. S11. Analysis of microsynteny between orthologous and paralogous *C. quinquefasciatus* (Cx), *An. gambiae* (Ag) and *Ae. aegypti* (Aa) odorant receptor (OR) encoding regions (red arrows). Neighboring genes sharing >50% amino acid identity are shaded according to interspecific homology (grey, light grey, blue and white arrow). Sequence orientation is indicated by the arrow direction. Non-OR genes are labeled according their VectorBase identifiers, namely the prefixes for *C. quinquefasciatus*, *Ae. aegypti* and *An. gambiae* are CPJI00(0), AAEL0(0) and AGAP0(0) respectively. Overall genome location is identified by chromosome location for *An. gambiae* and by supercontig number for *C. quinquefasciatus* and *Ae. aegypti* to the right. Centromeres are indicated by a dot. Figure not drawn to scale.

Fig. S12A-D. Phylogenetic relationships of the mosquito gustatory receptors (Grs) analysed using corrected distances. *An. gambiae* Grs (AgGr) are shown in pink, *Ae. aegypti* (AaGr) in blue, and *C. quinquefasciatus* (CpGr) in green. Branches considered to be orthologous relationships with substantial bootstrap support from 100 replications of uncorrected distance analysis are highlighted as thicker lines. Bootstrap support in percentages is only shown for selected major branches. Subfamilies,

conserved lineages, gene losses, and other groupings discussed in the text are indicated on the right. The figure is broken up into **A**, **B**, **C**, and **D** panels, with the carbon dioxide receptors in panel **A** designated the outgroup based on their basal groupings in larger analyses including more basal insects.

Fig. S1.

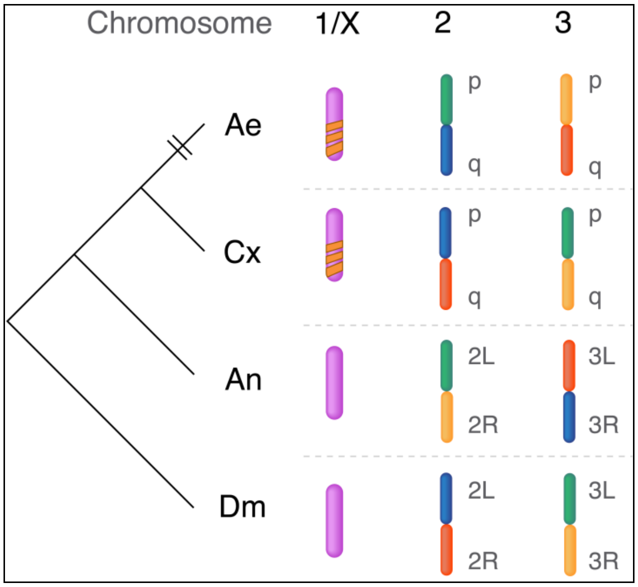


Fig. S2.

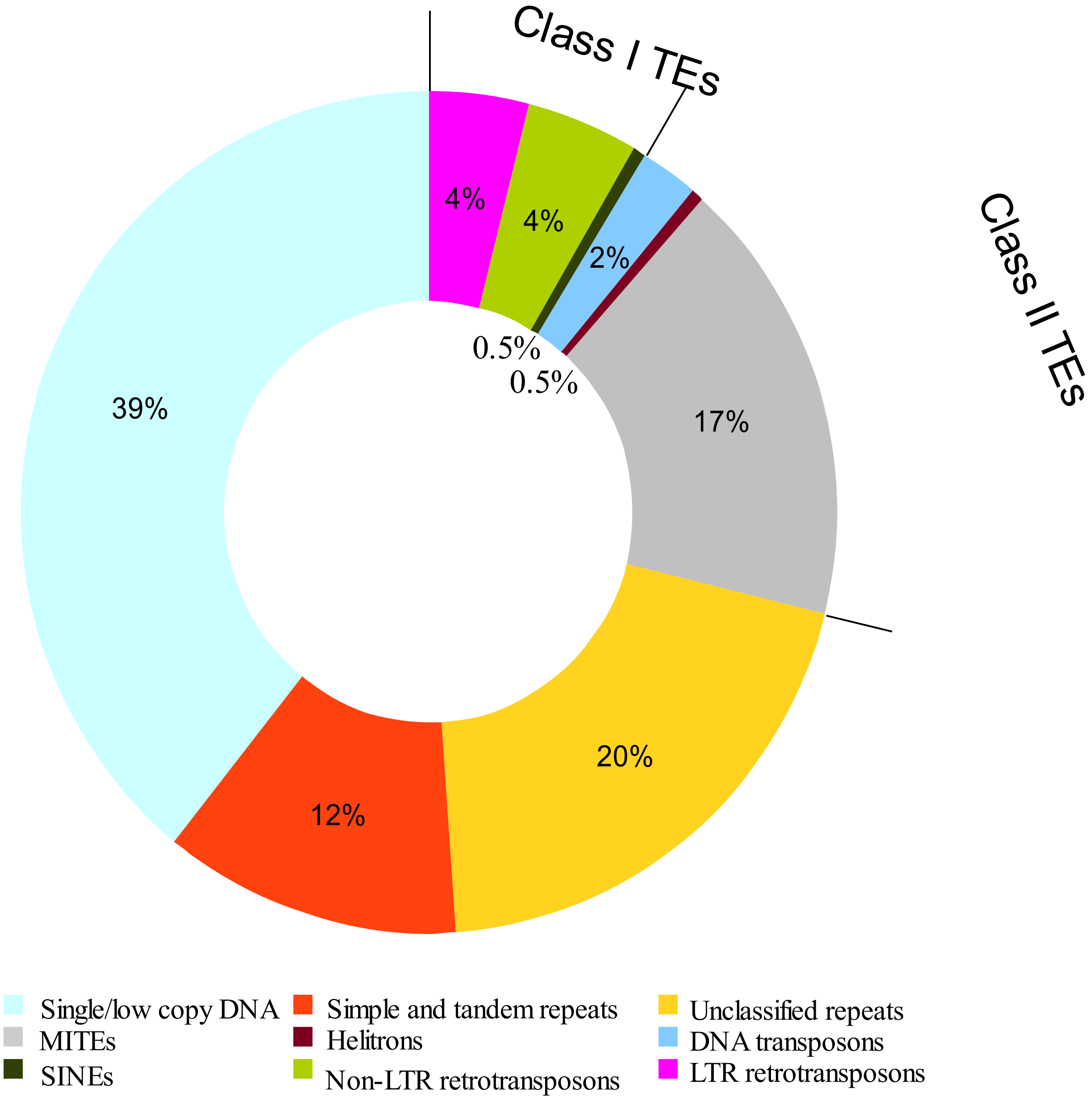

Class I TEs
Class II TEs
4%
4%
2%
0.5%
0.5%
17%
39%
20%
12%
Single/low copy DNA
Simple and tandem repeats
Unclassified repeats
MITEs
Helitrons
DNA transposons
SINEs
Non-LTR retrotransposons
LTR retrotransposons

Fig. S3.

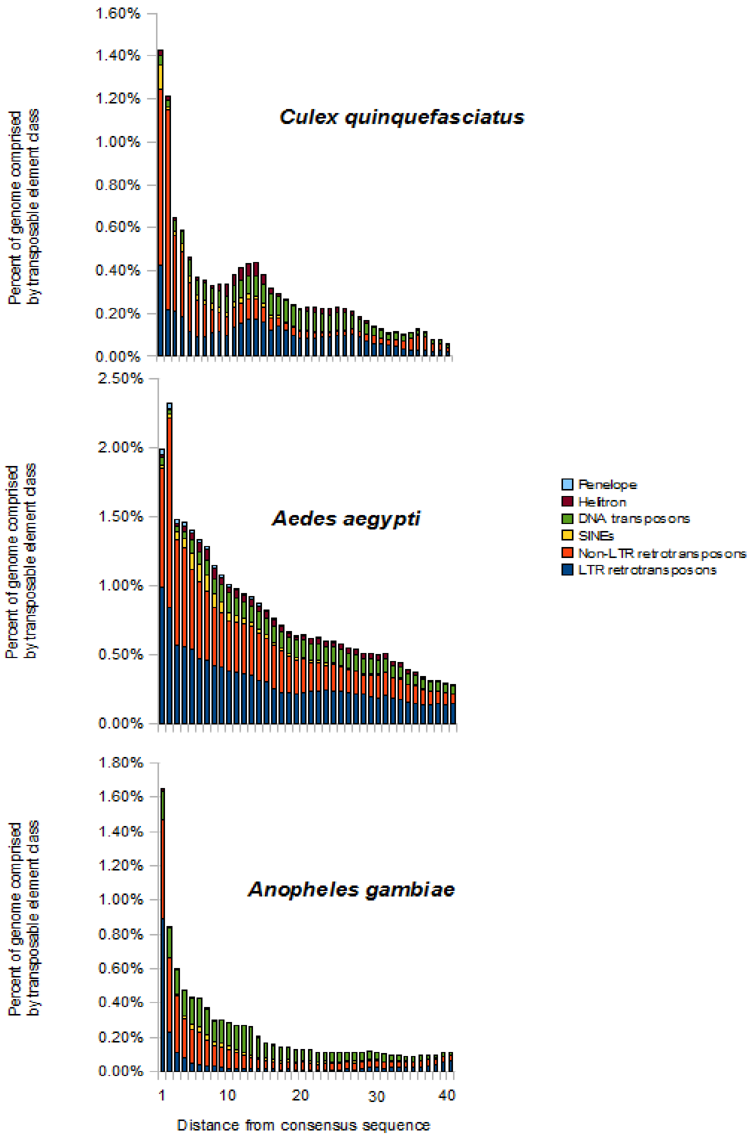

Culex quinquefasciatus
Aedes aegypti
Anopheles gambiae
Percent of genome comprised by transposable element class
1.60%
1.40%
1.20%
1.00%
0.80%
0.60%
0.40%
0.20%
0.00%
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
1.80%
1.60%
1.40%
1.20%
1.00%
0.80%
0.60%
0.40%
0.20%
0.00%
1
10
20
30
40
Distance from consensus sequence
Penelope
Helitron
DNA transposons
SINEs
Non-LTR retrotransposons
LTR retrotransposons

Fig. S4.

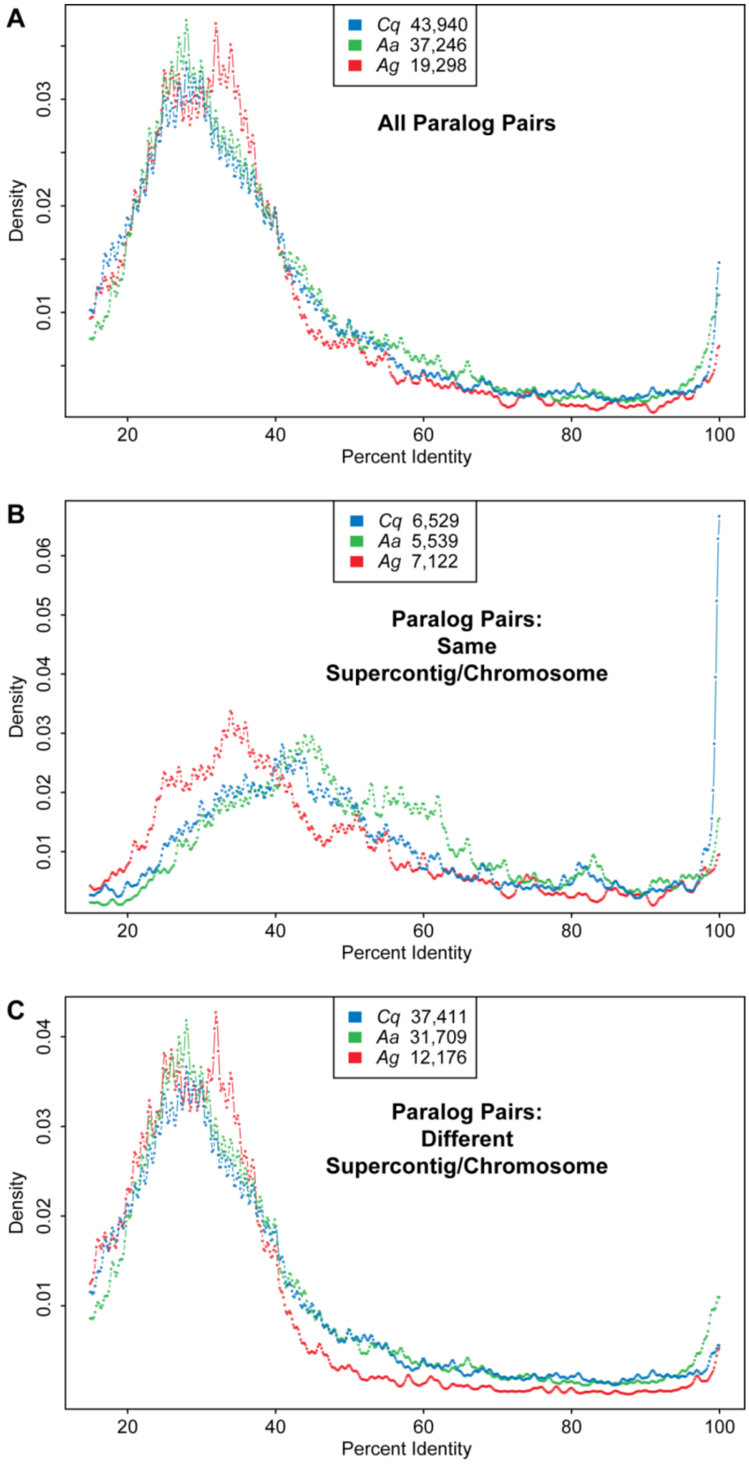

A
Cq 43,940
Aa 37,246
Ag 19,298
All Paralog Pairs
Density
0.01
0.02
0.03
20
40
60
80
100
Percent Identity
B
Cq 6,529
Aa 5,539
Ag 7,122
Paralog Pairs:
Same
Supercontig/Chromosome
Density
0.01
0.02
0.03
0.04
0.05
0.06
20
40
60
80
100
Percent Identity
C
Cq 37,411
Aa 31,709
Ag 12,176
Paralog Pairs:
Different
Supercontig/Chromosome
Density
0.01
0.02
0.03
0.04
20
40
60
80
100
Percent Identity

Fig. S5.

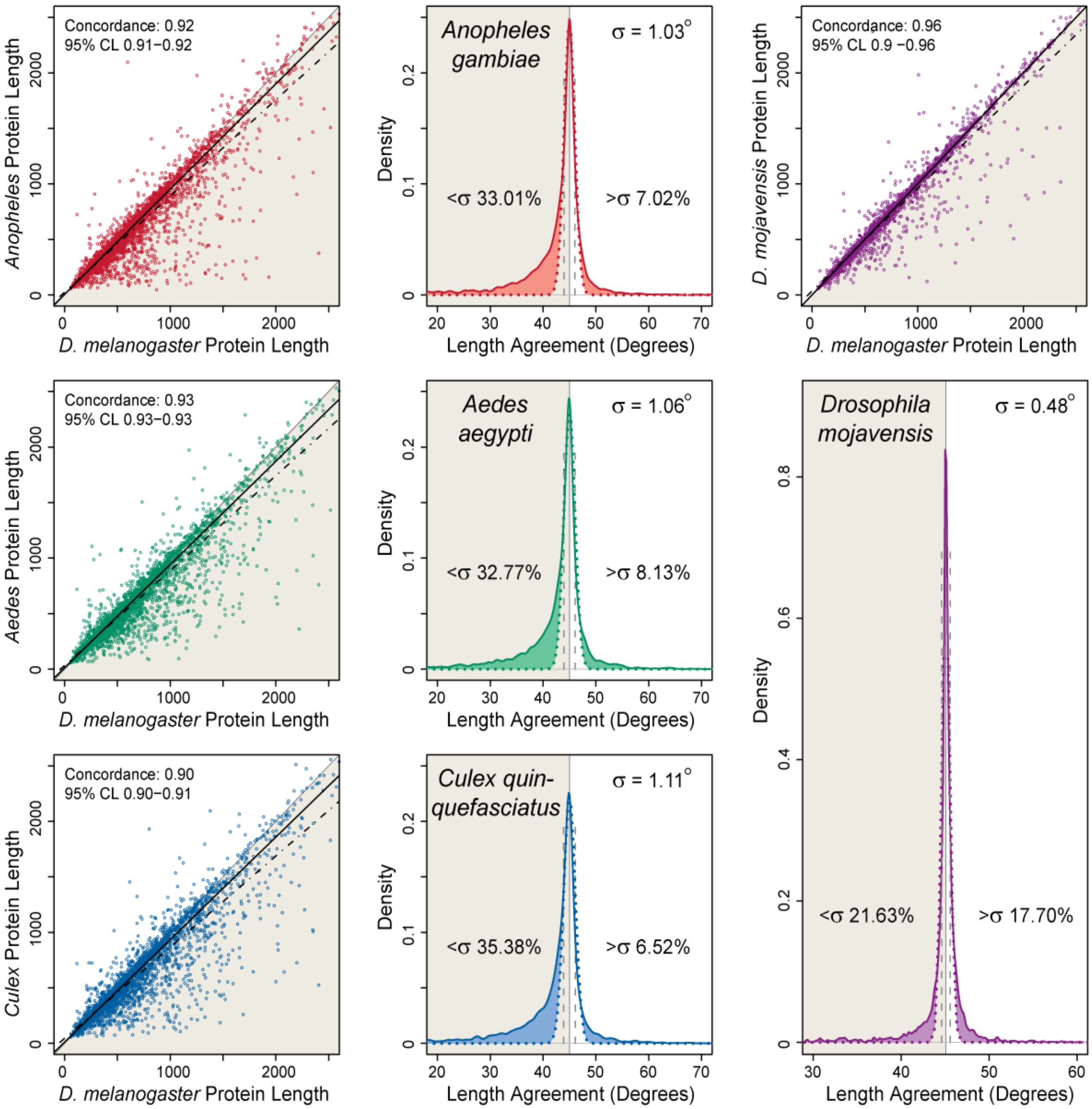

Concordance: 0.92
95% CL 0.91−0.92
Anopheles Protein Length
D. melanogaster Protein Length
0
1000
2000
Anopheles gambiae
σ = 1.03°
<σ 33.01%
>σ 7.02%
Density
0
0.1
0.2
Length Agreement (Degrees)
20
30
40
50
60
70
Concordance: 0.96
95% CL 0.9 −0.96
D. mojavensis Protein Length
Concordance: 0.93
95% CL 0.93−0.93
Aedes Protein Length
Aedes aegypti
σ = 1.06°
<σ 32.77%
>σ 8.13%
Drosophila mojavensis
σ = 0.48°
<σ 21.63%
>σ 17.70%
0.2
0.4
0.6
0.8
Concordance: 0.90
95% CL 0.90−0.91
Culex Protein Length
Culex quin-quefasciatus
σ = 1.11°
<σ 35.38%
>σ 6.52%

Fig. S6.

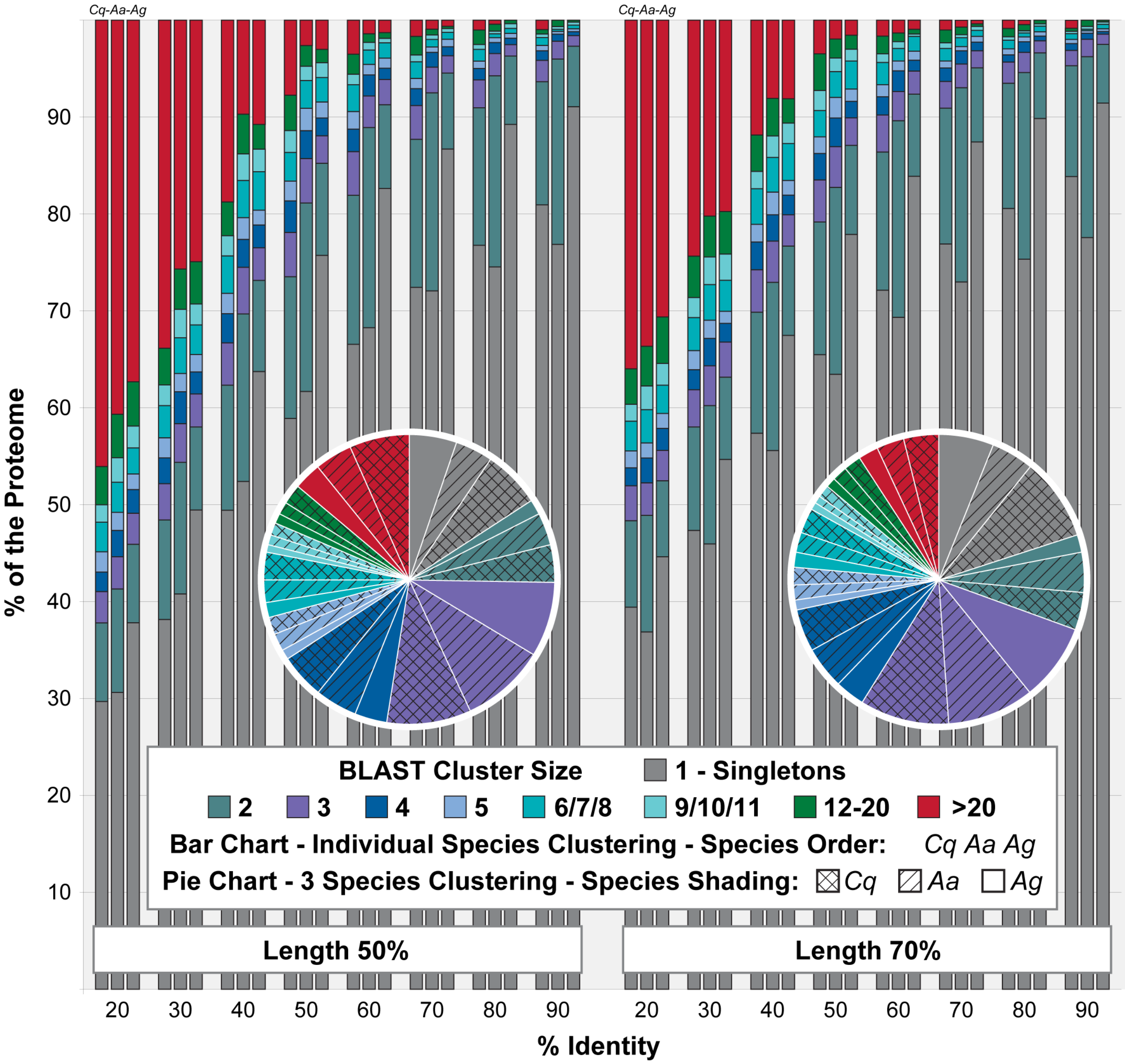

Cq-Aa-Ag
Cq-Aa-Ag
% of the Proteome
90
80
70
60
50
40
30
20
10
BLAST Cluster Size
1 - Singletons
2
3
4
5
6/7/8
9/10/11
12-20
>20
Bar Chart - Individual Species Clustering - Species Order: Cq Aa Ag
Pie Chart - 3 Species Clustering - Species Shading: Cq Aa Ag
Length 50%
Length 70%
20
30
40
50
60
70
80
90
20
30
40
50
60
70
80
90
% Identity

Fig. S7.

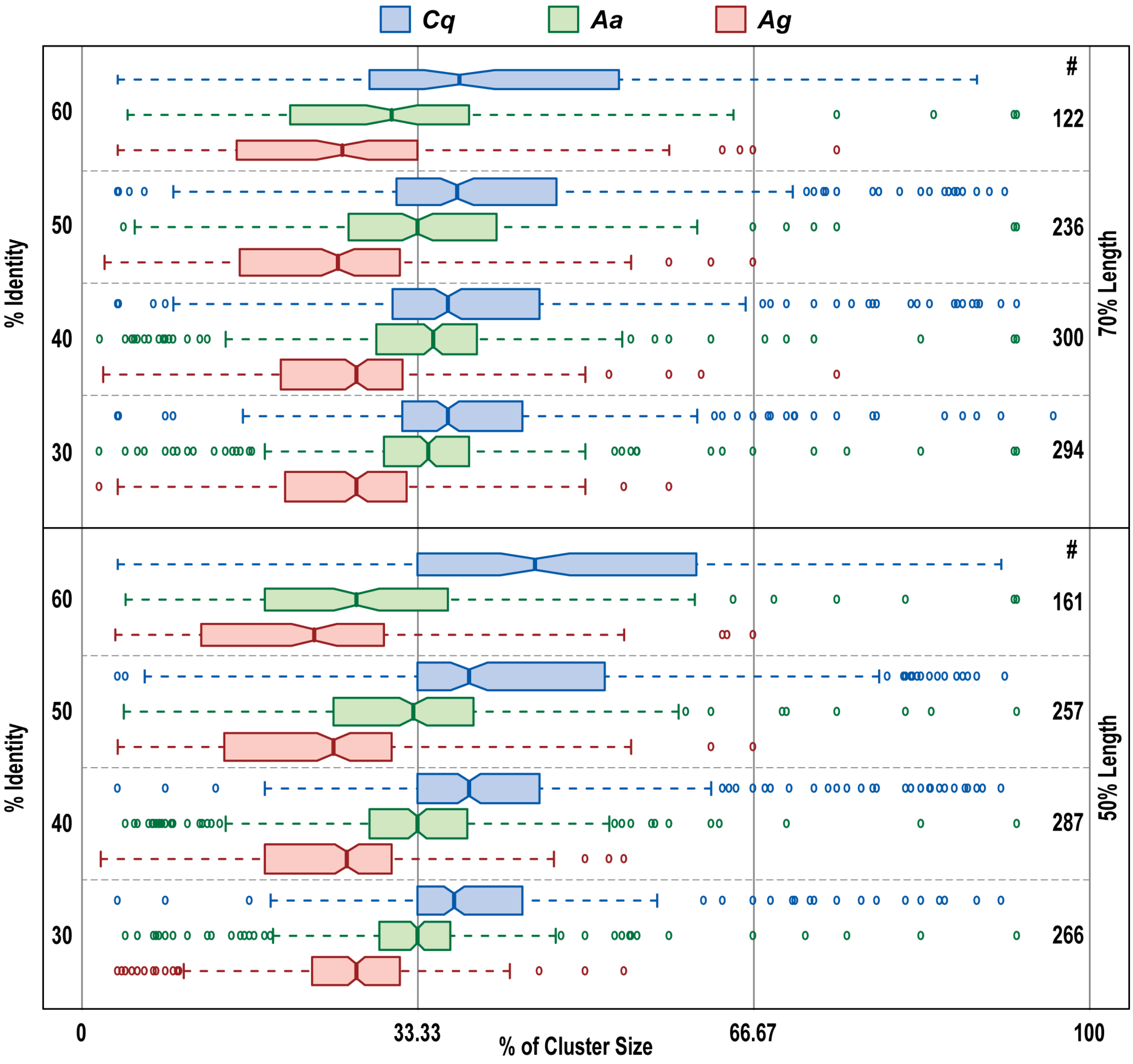
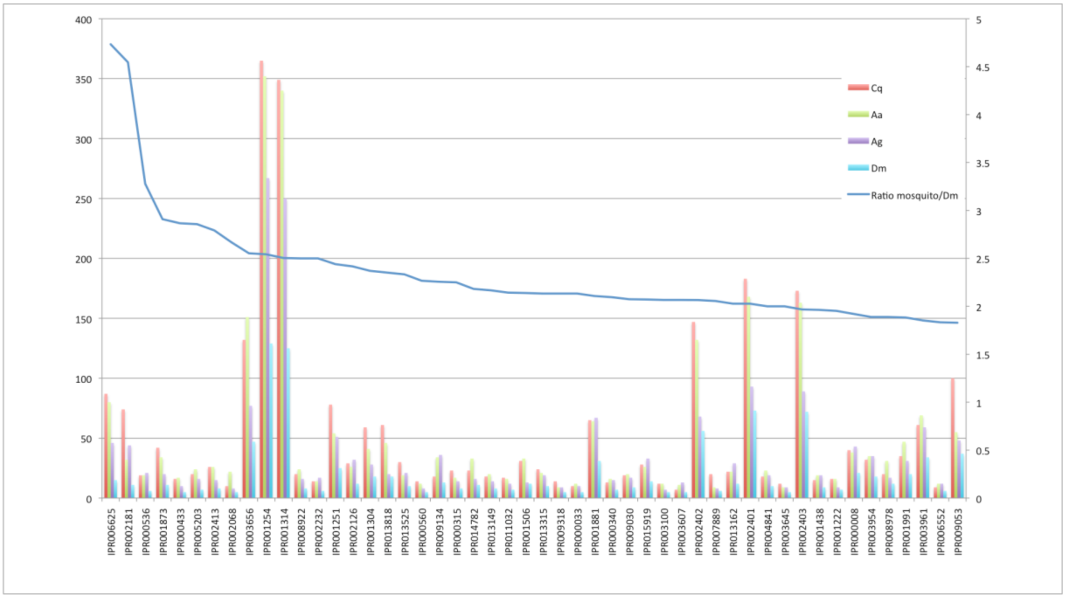

Cq
Aa
Ag
#
60
50
40
30
122
236
300
294
% Identity
70% Length
#
60
50
40
30
161
257
287
266
% Identity
50% Length
0
33.33
66.67
100
% of Cluster Size

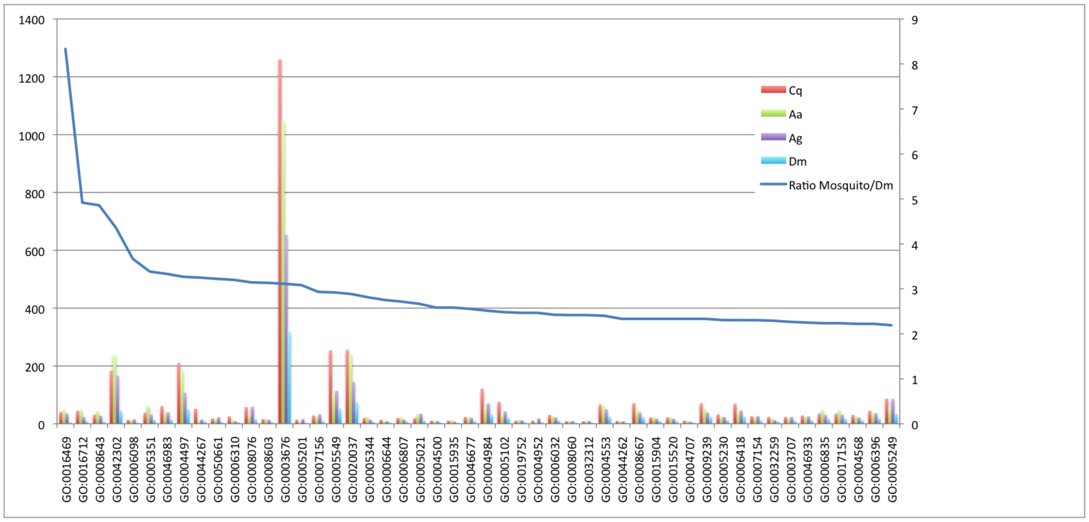
Fig. S8.

A



400
350
300
250
200
150
100
50
0
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Cq
Aa
Ag
Dm
Ratio mosquito/Dm
IPR006625
IPR002181
IPR000536
IPR001873
IPR000433
IPR005203
IPR002413
IPR002068
IPR003656
IPR001254
IPR001314
IPR008922
IPR002232
IPR001251
IPR002126
IPR001304
IPR013818
IPR013525
IPR000560
IPR009134
IPR000315
IPR014782
IPR013149
IPR011032
IPR001506
IPR013315
IPR009318
IPR000033
IPR001881
IPR000340
IPR009030
IPR015919
IPR003100
IPR003607
IPR002402
IPR007889
IPR013162
IPR002401
IPR004841
IPR003645
IPR002403
IPR001438
IPR001222
IPR000008
IPR003954
IPR008978
IPR001991
IPR003961
IPR006552
IPR009053

B



1400
1200
1000
800
600
400
200
0
9
8
7
6
5
4
3
2
1
0
Cq
Aa
Ag
Dm
Ratio Mosquito/Dm
GO:0016469
GO:0016712
GO:0008643
GO:0042302
GO:0006098
GO:0005351
GO:0046983
GO:0004497
GO:0044267
GO:0050661
GO:0006310
GO:0008076
GO:0008603
GO:0003676
GO:0005201
GO:0007156
GO:0005549
GO:0020037
GO:0005344
GO:0006644
GO:0006807
GO:0005021
GO:0004500
GO:0015935
GO:0046677
GO:0004984
GO:0005102
GO:0019752
GO:0004952
GO:0006032
GO:0008060
GO:0032312
GO:0004553
GO:0044262
GO:0008667
GO:0015904
GO:0015520
GO:0004707
GO:0009239
GO:0005230
GO:0006418
GO:0007154
GO:0032259
GO:0003707
GO:0046933
GO:0006835
GO:0017153
GO:0004568
GO:0006396
GO:0005249

Fig. S9

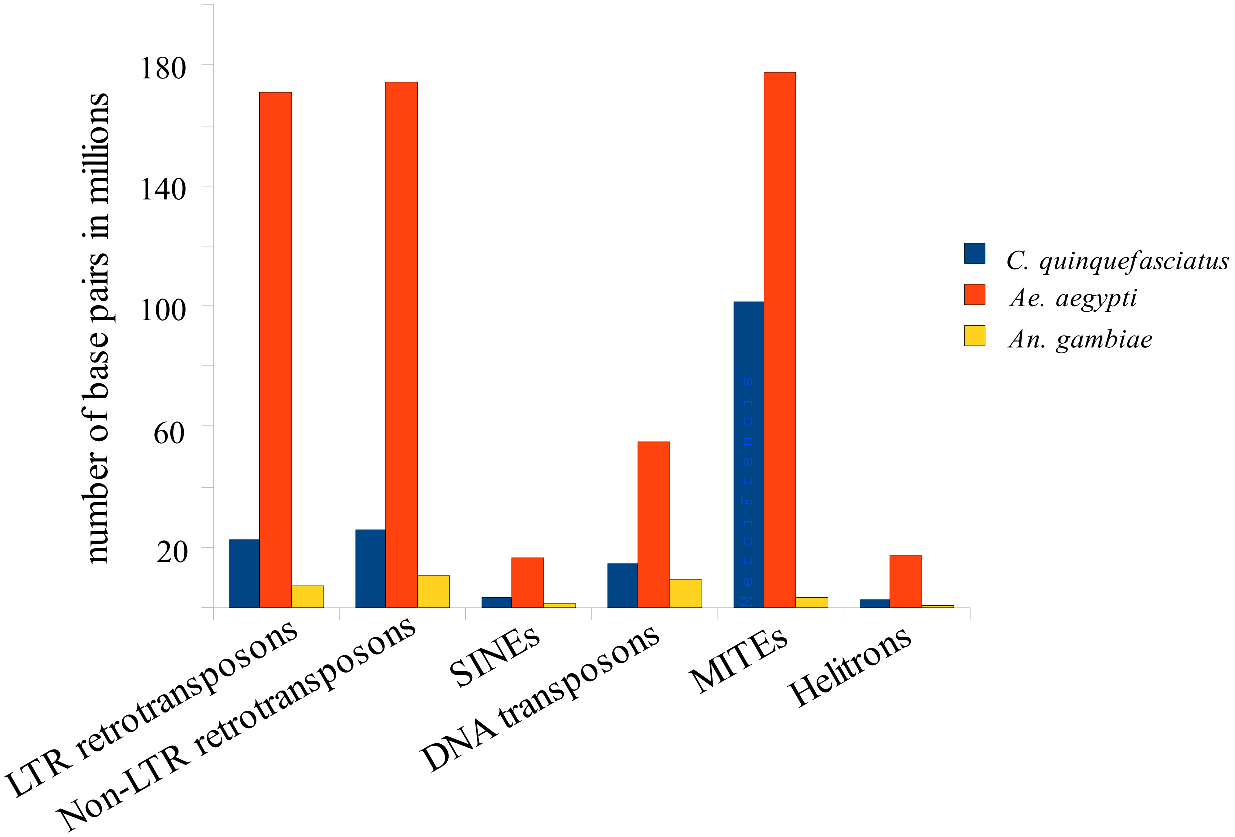

number of base pairs in millions
180
140
100
60
20
C. quinquefasciatus
Ae. aegypti
An. gambiae
MerciFrancis
LTR retrotransposons
Non-LTR retrotransposons
SINEs
DNA transposons
MITEs
Helitrons

Fig. S10A

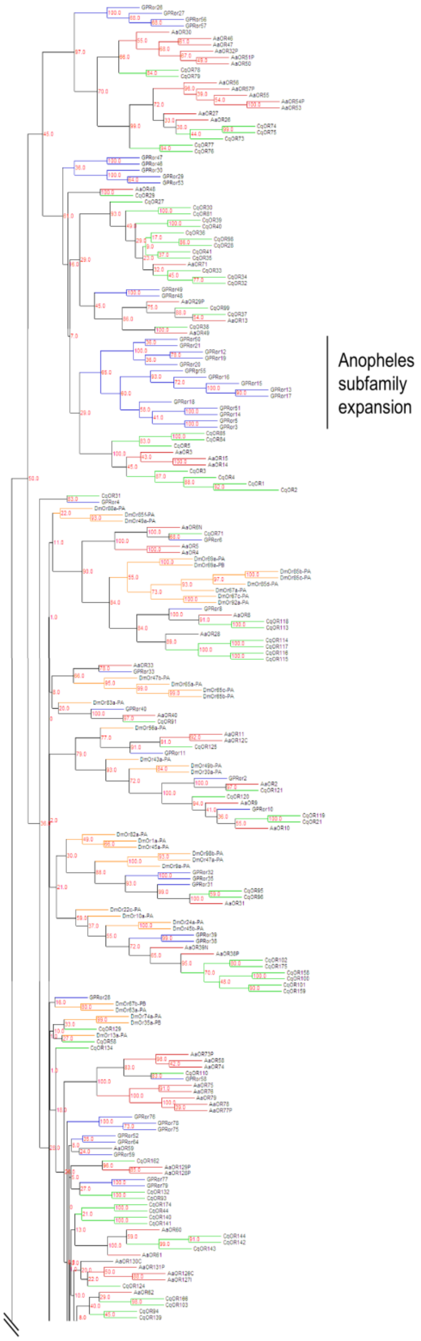


Fig. S10B

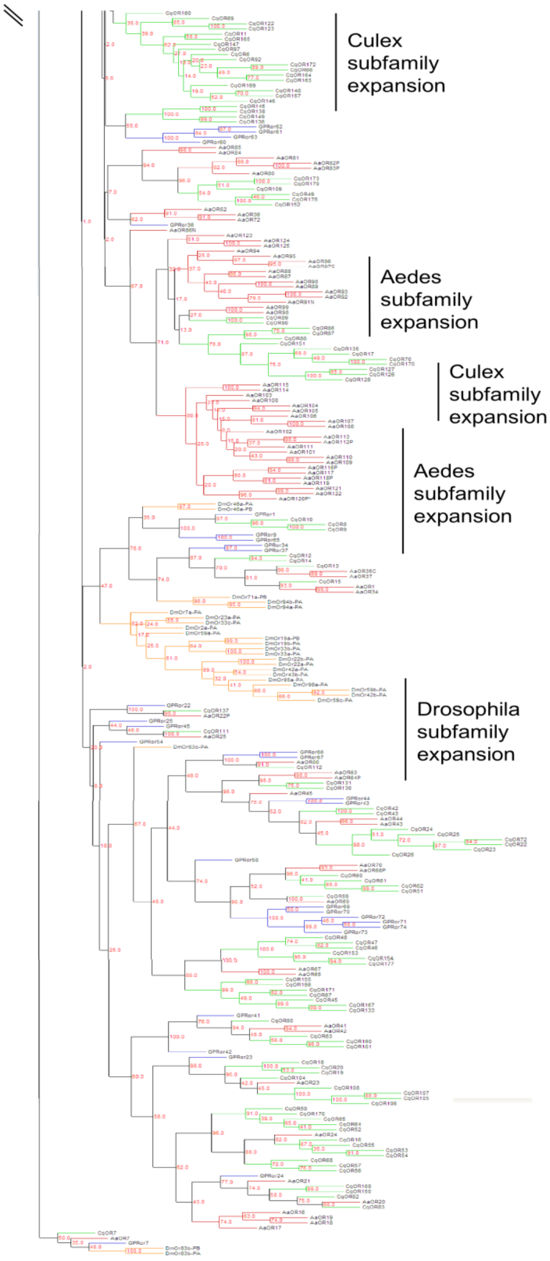

Culex subfamily expansion
Aedes subfamily expansion
Culex subfamily expansion
Aedes subfamily expansion
Drosophila subfamily expansion

Fig. S11

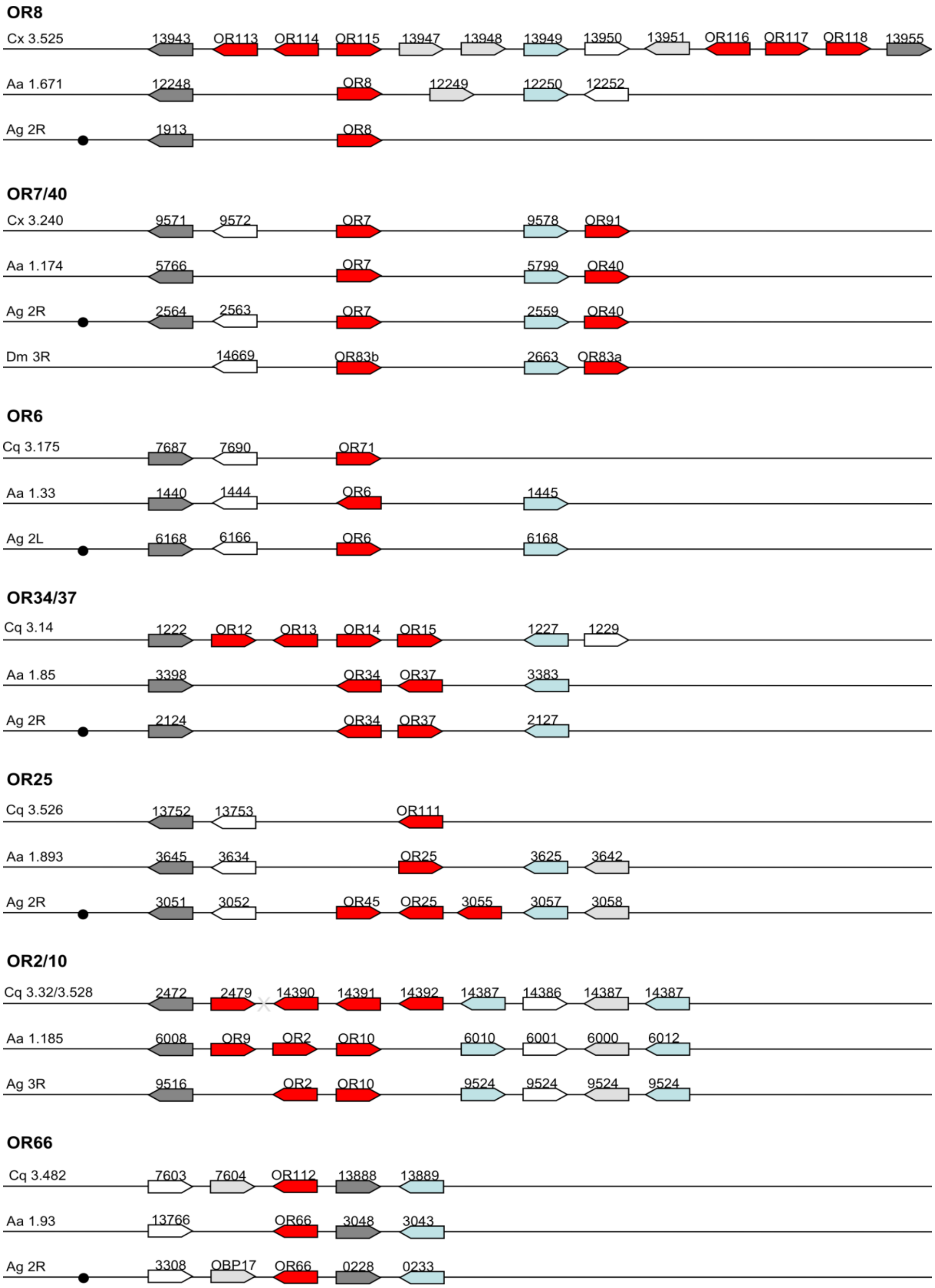

OR8
Cx 3.525
13943
OR113
OR114
OR115
13947
13948
13949
13950
13951
OR116
OR117
OR118
13955
Aa 1.671
12248
OR8
12249
12250
12252
Ag 2R
1913
OR8
OR7/40
Cx 3.240
9571
9572
OR7
9578
OR91
Aa 1.174
5766
OR7
5799
OR40
Ag 2R
2564
2563
OR7
2559
OR40
Dm 3R
14669
OR83b
2663
OR83a
OR6
Cq 3.175
7687
7690
OR71
Aa 1.33
1440
1444
OR6
1445
Ag 2L
6168
6166
OR6
6168
OR34/37
Cq 3.14
1222
OR12
OR13
OR14
OR15
1227
1229
Aa 1.85
3398
OR34
OR37
3383
Ag 2R
2124
OR34
OR37
2127
OR25
Cq 3.526
13752
13753
OR111
Aa 1.893
3645
3634
OR25
3625
3642
Ag 2R
3051
3052
OR45
OR25
3055
3057
3058
OR2/10
Cq 3.32/3.528
2472
2479
14390
14391
14392
14387
14386
14387
14387
Aa 1.185
6008
OR9
OR2
OR10
6010
6001
6000
6012
Ag 3R
9516
OR2
OR10
9524
9524
9524
9524
OR66
Cq 3.482
7603
7604
OR112
13888
13889
Aa 1.93
13766
OR66
3048
3043
Ag 2R
3308
OBP17
OR66
0228
0233

Fig. S12A

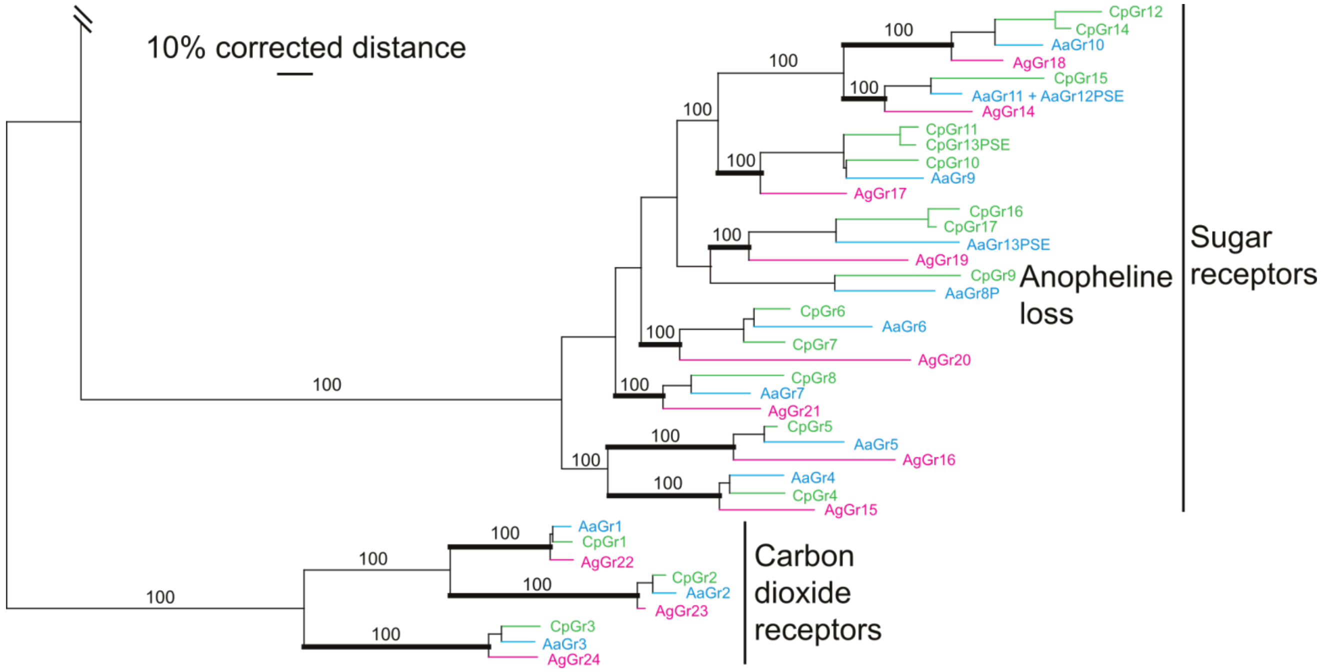

10% corrected distance
100
CpGr12
CpGr14
AaGr10
AgGr18
100
100
CpGr15
AaGr11 + AaGr12PSE
AgGr14
100
CpGr11
CpGr13PSE
100
CpGr10
AaGr9
AgGr17
CpGr16
100
CpGr17
AaGr13PSE
AgGr19
CpGr9
AaGr8P
Anopheline loss
Sugar receptors
CpGr6
AaGr6
100
CpGr7
AgGr20
100
CpGr8
AaGr7
AgGr21
CpGr5
100
AaGr5
100
AgGr16
AaGr4
100
CpGr4
AgGr15
AaGr1
100
CpGr1
AgGr22
100
CpGr2
100
AaGr2
AgGr23
100
Carbon dioxide receptors
CpGr3
100
AaGr3
AgGr24

Fig. S12B

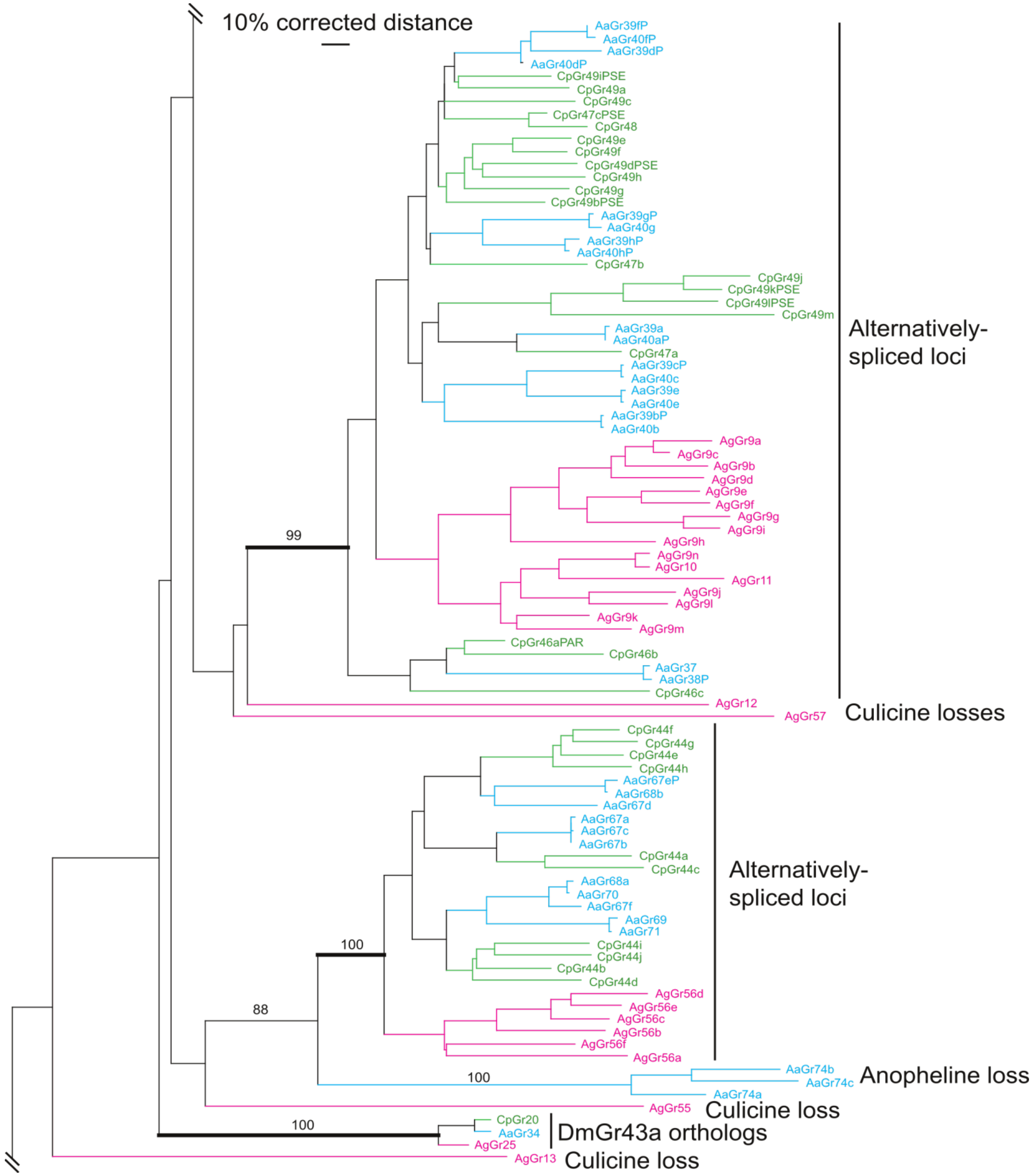

10% corrected distance
AaGr39fP
AaGr40fP
AaGr39dP
AaGr40dP
CpGr49iPSE
CpGr49a
CpGr49c
CpGr47cPSE
CpGr48
CpGr49e
CpGr49f
CpGr49dPSE
CpGr49h
CpGr49g
CpGr49bPSE
AaGr39gP
AaGr40g
AaGr39hP
AaGr40hP
CpGr47b
CpGr49j
CpGr49kPSE
CpGr49lPSE
CpGr49m
AaGr39a
AaGr40aP
CpGr47a
AaGr39cP
AaGr40c
AaGr39e
AaGr40e
AaGr39bP
AaGr40b
AgGr9a
AgGr9c
AgGr9b
AgGr9d
AgGr9e
AgGr9f
AgGr9g
AgGr9i
AgGr9h
AgGr9n
AgGr10
AgGr11
AgGr9j
AgGr9l
AgGr9k
AgGr9m
CpGr46aPAR
CpGr46b
AaGr37
AaGr38P
CpGr46c
AgGr12
AgGr57
99
Alternatively-
spliced loci
Culicine losses
CpGr44f
CpGr44g
CpGr44e
CpGr44h
AaGr67eP
AaGr68b
AaGr67d
AaGr67a
AaGr67c
AaGr67b
CpGr44a
CpGr44c
AaGr68a
AaGr70
AaGr67f
AaGr69
AaGr71
CpGr44i
CpGr44j
CpGr44b
CpGr44d
AgGr56d
AgGr56e
AgGr56c
AgGr56b
AgGr56f
AgGr56a
100
88
Alternatively-
spliced loci
AaGr74b
AaGr74c
AaGr74a
100
Anopheline loss
AgGr55
Culicine loss
CpGr20
AaGr34
AgGr25
100
DmGr43a orthologs
AgGr13
Culicine loss

Fig. S12C

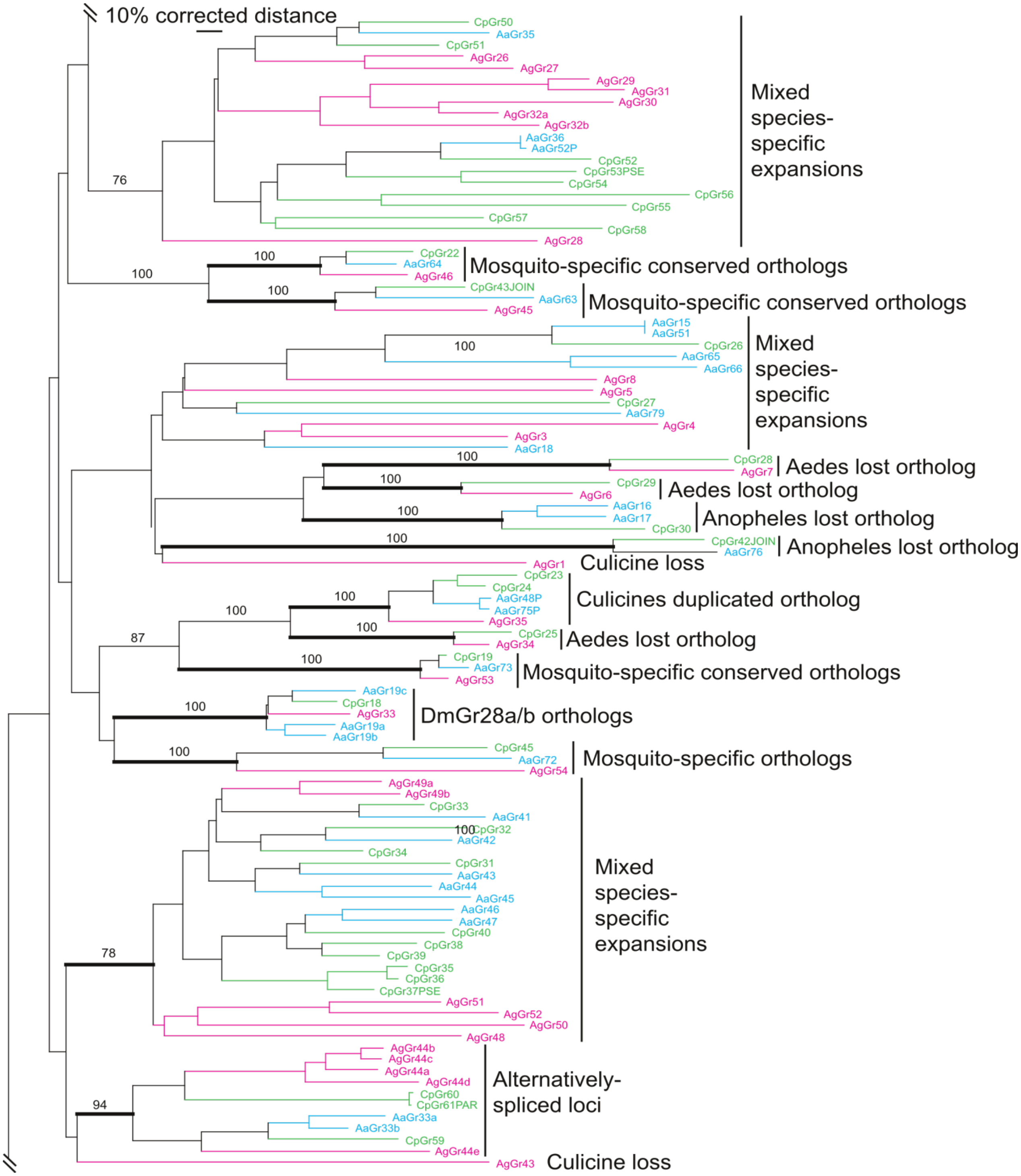

10% corrected distance
CpGr50
AaGr35
CpGr51
AgGr26
AgGr27
AgGr29
AgGr31
AgGr30
AgGr32a
AgGr32b
AaGr36
AaGr52P
CpGr52
CpGr53PSE
CpGr54
CpGr56
CpGr55
CpGr57
CpGr58
AgGr28
76
Mixed species-specific expansions
100
CpGr22
AaGr64
AgGr46
Mosquito-specific conserved orthologs
100
100
CpGr43JOIN
AaGr63
AgGr45
Mosquito-specific conserved orthologs
AaGr15
AaGr51
CpGr26
AaGr65
AaGr66
100
AgGr8
AgGr5
CpGr27
AaGr79
AgGr4
AgGr3
AaGr18
Mixed species-specific expansions
100
CpGr28
AgGr7
Aedes lost ortholog
100
CpGr29
AgGr6
Aedes lost ortholog
100
AaGr16
AaGr17
CpGr30
Anopheles lost ortholog
100
CpGr42JOIN
AaGr76
Anopheles lost ortholog
AgGr1
Culicine loss
CpGr23
CpGr24
AaGr48P
AaGr75P
AgGr35
100
100
Culicines duplicated ortholog
100
CpGr25
AgGr34
Aedes lost ortholog
87
100
CpGr19
AaGr73
AgGr53
Mosquito-specific conserved orthologs
AaGr19c
CpGr18
AgGr33
AaGr19a
AaGr19b
100
DmGr28a/b orthologs
100
CpGr45
AaGr72
AgGr54
Mosquito-specific orthologs
AgGr49a
AgGr49b
CpGr33
AaGr41
100CpGr32
AaGr42
CpGr34
CpGr31
AaGr43
AaGr44
AaGr45
AaGr46
AaGr47
CpGr40
CpGr38
CpGr39
CpGr35
CpGr36
CpGr37PSE
AgGr51
AgGr52
AgGr50
AgGr48
78
Mixed species-specific expansions
AgGr44b
AgGr44c
AgGr44a
AgGr44d
CpGr60
CpGr61PAR
AaGr33a
AaGr33b
CpGr59
AgGr44e
94
Alternatively-spliced loci
AgGr43
Culicine loss

Fig. S12D

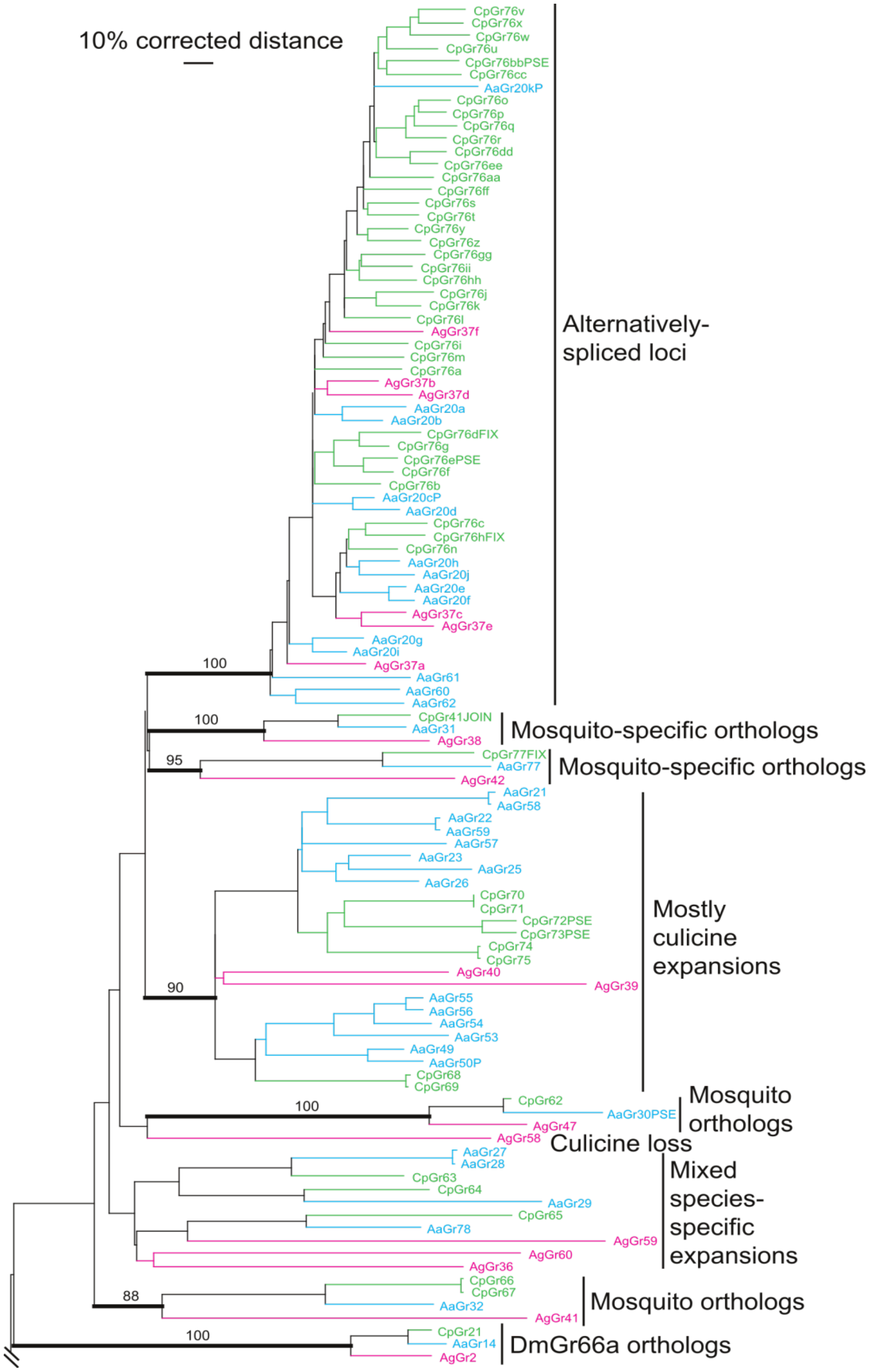

10% corrected distance
CpGr76v
CpGr76x
CpGr76w
CpGr76u
CpGr76bbPSE
CpGr76cc
AaGr20kP
CpGr76o
CpGr76p
CpGr76q
CpGr76r
CpGr76dd
CpGr76ee
CpGr76aa
CpGr76ff
CpGr76s
CpGr76t
CpGr76y
CpGr76z
CpGr76gg
CpGr76ii
CpGr76hh
CpGr76j
CpGr76k
CpGr76l
AgGr37f
CpGr76i
CpGr76m
CpGr76a
AgGr37b
AgGr37d
AaGr20a
AaGr20b
CpGr76dFIX
CpGr76g
CpGr76ePSE
CpGr76f
CpGr76b
AaGr20cP
AaGr20d
CpGr76c
CpGr76hFIX
CpGr76n
AaGr20h
AaGr20j
AaGr20e
AaGr20f
AgGr37c
AgGr37e
AaGr20g
AaGr20i
AgGr37a
AaGr61
AaGr60
AaGr62
100
Alternatively-
spliced loci
100
CpGr41JOIN
AaGr31
AgGr38
Mosquito-specific orthologs
95
CpGr77FIX
AaGr77
AgGr42
Mosquito-specific orthologs
AaGr21
AaGr58
AaGr22
AaGr59
AaGr57
AaGr23
AaGr25
AaGr26
CpGr70
CpGr71
CpGr72PSE
CpGr73PSE
CpGr74
CpGr75
AgGr40
AgGr39
90
AaGr55
AaGr56
AaGr54
AaGr53
AaGr49
AaGr50P
CpGr68
CpGr69
Mostly
culicine
expansions
100
CpGr62
AaGr30PSE
AgGr47
AgGr58
Mosquito
orthologs
Culicine loss
AaGr27
AaGr28
CpGr63
CpGr64
AaGr29
CpGr65
AaGr78
AgGr59
AgGr60
AgGr36
Mixed
species-
specific
expansions
88
CpGr66
CpGr67
AaGr32
AgGr41
Mosquito orthologs
100
CpGr21
AaGr14
AgGr2
DmGr66a orthologs

Table S1. Abundance of selected gene families in *C. quinquefasciatus*, *Ae. aegypti*, and *An. gambiae*. Abundance numbers for *Ae. aegypti* and *An. gambiae* were taken from published reports (*45*, *46*, *56*, *70*, *71*), and numbers of immune-related genes are presented in full in (*72*).

	C. quinquefasciatus	*Ae. aegypti*	*An. gambiae*
Carboxy / cholinesterases			
Alpha esterases	30	22	16
Hormone processing			
Beta esterases	3	2	5
Juvenile hormone	14	6	4
Esterases			
Glutactin	6	10	9
Acetylcholinesterasese	2	2	2
Others	9	7	4
Transferases			
Cytosolic glutathione	32	27	28
Microsomal glutathione	5	5	3
Cytochrome P450s	170	158	102
Olfactory receptors	180	113	79
Gustatory receptors	123	95	89
Immune-related			
Antimicrobial peptides (attacins, cecropins,defensins, diptericins)	6	16	10
Caspases	16	11	14
CLIP-domain serine proteases (Classes A,B,C,D & E)	80	67	56
C-type Lectins	55	39	24
Fibrinogen-related proteins	87	35	55
Galectins	15	12	10
Peroxidases (glutathione, heme, thioredoxin)	19	20	26
Gram-negative binding proteins	11	7	7

(continued)

Table S1. Continued

Inhibitors of apoptosis	6	5	8
Lysozymes	4	7	8
MD2-like proteins	19	24	15
Peptidoglycan recognition proteins	8	8	7
Prophenoloxidases	9	10	9
Scavenger receptors (Classes A, B & C)	20	20	19
Serine protease inhibitors	32	23	17
Spaetzle-like proteins	7	9	6
Thio-ester containing proteins	10	8	13
Toll receptors	9	12	10

Table S2. 26 single markers and single copy genes in *Ae. aegypti* and their occurrence in *C. quinquefasciatus*.

	Name	GenBank ID	Occurrence in Culex
Markers	Tsf	AF019117	Single
	Hexam2	U86080	Single
	BMIOP	U84248	Single
	LF204	BM378050	Single
	LF397	BM378051	Single
	LF150	BM005476	Single
	nAcBP	AY040341	Single
	RT6	BH214544	Single
	VCP	L46594	Single
	LF342	BM005512	Single
	Hsp83	X03910	Single
	MUCI	AF308862	Single
	AEG12	AY038041	Single
	Chym	AY038039	Single
	RpL17a	AF315597	Single
	DDC	U27581	Single
	BA67	AI561370	Single
	LF253	T58331	Single
	MalI	M30442	Single
	TrypLate	X64363	Single
	LF106	BM005490	Single
	LF417	BM005499	Single
	LF296	BM005501	Single
	LF386	BM005497	Single
	para	AF468968	Single
	AspSyn	U84118	Single
Genes	RpS3	XM_321155	Single
	AsNOS	AY583529	No reliable match
	As60A	AF284816	No reliable match
	ferritin	XM_001652682	Single
	White	U88851	Single
	Sia I	AF108099	No reliable match
	HEX1A	XM_315780	No reliable match

Table S3. Average identity percentages and block lengths of DNA/DNA comparisons between *C. quinquefasciatus* (*Cq*), and *Ae. aegypti* (*Aa*), *An. gambiae* (*Ag*) and *D. melanogaster* (*Dm*) genomes.

Similarity blocks	**Average %ID**	**Average block length in kilobases (maximum size)**
Cq.-Aa.	82 %	119 (11.4 Kb)
Cq.-Ag.	80 %	126 (5.6 Kb)
Cq.-Dm.	78 %	131 (4.5 Kb)

Table S4. Genome, gene, exon, and intron annotation statistics for *C. quinquefasciatus, Ae. aegypti, An. gambiae* and *D. melanogaster*. Abbreviation: gigabases (Gb), megabases (Mb), base pairs (bps).

GENOME	Genome size	Gene number	Transcript number	Exon number (gene wise)	Average number of exons / gene	Number of introns (gene wise)
C. quinquefasciatus	579 Mb	18,883	18,883	71,094	3.7	52,211
Ae. aegypti	1.38 Gb	15,419	16,789	63,650	4.1	51,076
An. gambiae	278 Mb	12,457	13,133	52,595	4.2	38,051
D. melanogaster	168 Mb	14,039	19,789	65,706	4.6	43,588

GENES	Gene number	Average gene size (bps.)	Maximum gene size (bps.)
C. quinquefasciatus	18,883	5,673	154,128
Ae. aegypti	15,419	15,488	428,674
An. gambiae	12,457	5,145	365,622
D. melanogaster	14,039	5,253	279,927

EXONS	Exon number	Average exon size (bps.)	Maximum exon size (bps.)
C. quinquefasciatus	71,094	356	12,993
Ae. aegypti	63,650	405	13,140
An. gambiae	52,595	378	14,041
D. melanogaster	65,706	464	27,725

INTRONS	Intron number	Average intron size (bps.)	Maximum intron size (bps.)	Median intron size (bps.)	Mode intron size (bps.)
C. quinquefasciatus	52,211	1,043	88,658	4,720	59
Ae. aegypti	51,076	3,793	329,294	9,920	62
An. gambiae	38,051	875	174,432	3,331	72
D. melanogaster	43,588	788	185,510	3,192	58

Table S5. Statistics of the merging process for the three independent gene sets. Sets were compared two-by-two, on a locus basis. Using the tool developed for the *Ae.aegypti* annotation (*4*), a single gene model was selected at each locus.

	Same[1]	**Different**[2]		**No map**[3]		**Merge/split**[4]		**Compatible**[5]			**Complex**[6]
		diff	**extreme diff**	**no_map**	**isoform nomap**	**merge**	**split**	**compat-endOK**	**compat-staggered**	**compat-encaps**	
JCVI vs. BROAD	5915 25% JCVI 31%broad	4904	217	JCVI: 7,496 BROAD: 2,313	-	648	1158	1597	72	139	140
JCVI vs. VB	3316 14% JCVI 23% VB	4572	109	JCVI: 9,521 VB: 1,167	14	843	372	2690	210	408	55
BROAD vs. VB	2345 12% broad 16% VB	4955	82	BROAD: 5,460 VB: 2,282	14	1232	264	2120	181	378	49

[1] Same*:* Same locus, same gene structure

[2] Different:

- *Different*: Same locus, different gene structure
- *Extreme diff*: Same locus, different reading frames

[3] No map:

- *No map*: Gene from Set-1 lacks counterpart in Set-2
- *Isoform nomap*: Gene from Set-1 lacks counterpart in Set-2 – but other isoforms might be correct

[4] Merge/Split:

- *Merge*: multiple genes from Set-1 merged into a single gene in Set-2
- *Split*: Single gene from Set-1 split into multiple genes in Set-2

[5] Compatible:

- *Compat-endOK*: Same structure in region of overlap and share the same end or start
- *Compat-stagerred*: Same structure in region of overlap but staggered boundaries
- *Compat-encaps*: Gene from Set-1 entirely consumes gene from Set-2 and is identical in region of overlap

[6] Complex: many-to-many gene mapping

Table S6. Statistics of gene family clustering analyzes. Individual clusters are identified in the first column with (in order) species abbreviation, sequence length cut-off (50% or 70%), and percent identity (30%, 40%, 50%, or 60%). Various statistics are provided, including number of sequences, and value of paired Wilcoxon signed-rank tests for pairs of species (last three columns).

Cut-Offs	Num.	Min.	1stQu.	Median	Mean	3rdQu.	Max.	p Cq-Aa	p Cq-Ag	p Aa-Ag
Ag.70.60	122	3.57	15.38	25.87	25.51	32.89	75.00	2.332e-04	1.871e-08	6.232e-03
Aa.70.60		4.55	20.88	30.77	31.00	38.33	92.86			
Cq.70.60		3.57	29.24	37.50	43.48	53.33	88.89			
Ag.70.50	236	2.27	15.85	25.43	24.41	31.58	66.67	9.418e-04	< 2.2e-16	8.244e-14
Aa.70.50		4.17	26.58	33.33	33.82	41.18	92.86			
Cq.70.50		3.57	31.25	37.27	41.76	47.02	91.67			
Ag.70.40	300	2.13	19.88	27.27	25.78	31.78	75.00	7.501e-04	< 2.2e-16	< 2.2e-16
Aa.70.40		1.79	29.27	34.89	34.27	39.18	92.86			
Cq.70.40		3.57	30.91	36.36	39.95	45.45	92.86			
Ag.70.30	294	1.75	20.35	27.27	26.08	32.20	58.33	2.711e-04	< 2.2e-16	< 2.2e-16
Aa.70.30		1.75	30.00	34.41	34.53	38.46	92.86			
Cq.70.30		3.57	31.86	36.36	39.39	43.75	96.49			
Ag.50.60	161	3.33	11.86	23.08	23.12	30.00	66.67	3.149e-10	< 2.2e-16	3.228e-04
Aa.50.60		4.35	18.18	27.27	28.82	36.36	92.86			
Cq.50.60		3.57	33.33	45.00	48.06	61.02	91.30			
Ag.50.50	257	3.57	14.15	25.00	23.30	30.77	66.67	1.848e-08	< 2.2e-16	1.576e-15
Aa.50.50		4.17	25.00	32.92	32.12	38.89	92.86			
Cq.50.50		3.57	33.33	38.46	44.58	51.92	91.67			
Ag.50.40	287	1.89	18.18	26.32	24.69	30.77	53.85	1.984e-08	< 2.2e-16	< 2.2e-16
Aa.50.40		4.35	28.57	33.33	33.06	38.28	92.86			
Cq.50.40		3.57	33.33	38.46	42.25	45.45	91.30			
Ag.50.30	266	3.57	22.92	27.27	26.36	31.58	53.85	3.533e-07	< 2.2e-16	< 2.2e-16
Aa.50.30		4.35	29.57	33.33	33.58	36.53	92.86			
Cq.50.30		3.57	33.33	36.98	40.06	43.75	91.30			

Table S7. Characteristics of synteny blocks (microsynteny) between *C. quinquefasciatus, Ae. aegypti*, *An. gambiae* and *D. melanogaster*. Abbreviations: kilobase pairs (Kbps), megabase pairs (Mbps).

	Cq/Aa synteny blocks	**Cq/Ag synteny blocks**	**Cq/Dm synteny blocks**
number of one-to-one orthologs	7,973	7,550	6,278
number of syntenies	1,546	1,514	692
number of one-to-one orthologs in synteny	6,318	5,287	1,653
Average number of genes in synteny	4	3.5	2.4
Average synteny length (Kbps)	Cq: 91 Aa: 235	Cq: 90 Ag: 54	Cq: 23 Dm: 65
Total length of synteny (Mbps)	Cq: 141 Aa: 363	Cq: 137 Ag: 81	Cq: 45 Dm: 16
% genome in synteny	Cq: 25% Aa: 26%	Cq: 23% An: 35%	Cq: 7.8% Dm: 13%
% genes in synteny	Cq: 33% Aa: 41%	Cq: 28% Ag: 42%	Cq: 9% Dm: 12%
Expansion factor	0.4	1.7	2.8

Table S8. Genome macrosynteny. Percent of *C. quinquefasciatus* and *Ae. aegypti* scaffolds with orthologs to *An. gambiae* and *D. melanogaster* chromosome arms. Calculation performed considering the individual genes ("By gene") or the synteny blocks ("By synteny").

		C. quinquefasciatus	*Ae. aegypti*
By gene	Percent of scaffolds with orthologs to 1, 2, 3 or more than 3 *An. gambiae* chromosome arms, respectively.	75%, 19%, 3%, 2%	68%, 22%, 8%, 2%
	Percent of scaffolds with orthologs to 1, 2, 3 or more than 3 *D. melanogaster* chromosome arms, respectively.	48%, 21%, 14%, 16%	41%, 23%, 17%, 18%
By synteny block	Percent of scaffolds with orthologs to 1, 2 or 3 *An. gambiae* chromosome arms, respectively.	90%, 9.5%, 0.5%	71%, 26%, 2.5%
	Percent of scaffolds with orthologs to 1, 2 or 3 *D. melanogaster* chromosome arms, respectively.	56%, 40%, 4%	54%, 42%, 3%

Table S9. Number of orthologous and paralogous relationships between *C. quinquefasciatus* (Cq), *Ae. aegypti* (Aa) and *An. gambiae* (Ag) gene sets based on the Ensembl GeneTree pipeline. Order Cq:Aa:Ag.

	N:1:1	**1:N:1**	**1:1:N**	**1:N:N**	**N:1:N**	**N:N:1**	**N:N:N**
Number of relationships (trees)	656	717	173	65	66	334	650
Number of paralogs involved	Cq: 1,555	Aa: 1,597	Ag: 359	Cq: 65 Aa: 158 Ag: 128	Cq: 279 Aa: 66 Ag: 148	Cq: 980 Aa: 906 Ag: 334	Cq: 3,674 Aa: 3,561 Ag: 2,389

Table S10. Occurrence of the 50 most over-represented InterPro domains in the three mosquitoes, *C. quinquefasciatus*, *Ae. aegypti, An. gambiae*, and *D. melanogaster* genomes.

InterPro	Cq	Aa	Ag	Dm	Description
IPR006625	87	80	46	15	Insect pheromone/odorant binding protein PhBP
IPR002181	74	32	44	11	Fibrinogen, alpha/beta/gamma chain, C-terminal globular
IPR000536	19	19	21	6	Nuclear hormone receptor, ligand-binding, core
IPR001873	42	34	20	11	Na+ channel, amiloride-sensitive
IPR000433	16	17	10	5	Zinc finger, ZZ-type
IPR005203	20	24	16	7	Hemocyanin, C-terminal
IPR002413	26	26	15	8	Ves allergen
IPR002068	10	22	8	5	Heat shock protein Hsp20
IPR003656	132	151	77	47	Zinc finger, BED-type predicted
IPR001254	365	352	267	129	Peptidase S1/S6, chymotrypsin/Hap
IPR001314	349	340	250	125	Peptidase S1A, chymotrypsin
IPR008922	20	24	16	8	Di-copper centre-containing
IPR002232	14	14	17	6	5-Hydroxytryptamine 6 receptor
IPR001251	78	54	51	25	Cellular retinaldehyde-binding/triple function, C-terminal
IPR002126	29	26	32	12	Cadherin
IPR001304	59	41	28	18	C-type lectin
IPR013818	61	46	20	18	Lipase, N-terminal
IPR013525	30	19	21	10	ABC-2 type transporter
IPR000560	14	12	8	5	Histidine acid phosphatase
IPR009134	18	34	36	13	Tyrosine-protein kinase, vascular endothelial growth factor receptor, N-terminal
IPR000315	23	17	14	8	Zinc finger, B-box
IPR014782	23	33	16	11	Peptidase M1, membrane alanine aminopeptidase, N-terminal
IPR013149	18	20	14	8	Alcohol dehydrogenase, zinc-binding
IPR011032	17	16	12	7	GroES-like
IPR001506	31	33	13	12	Peptidase M12A, astacin
IPR013315	24	21	19	10	Spectrin alpha chain, SH3 domain
IPR009318	14	9	9	5	Trehalose receptor
IPR000033	10	12	10	5	Low-density lipoprotein receptor, class B (YWTD) repeat

(continued)

Table S10. Continued.

IPR000340	13	16	15	7	Dual specificity phosphatase, catalytic domain
IPR009030	19	20	17	9	Growth factor, receptor
IPR015919	28	26	33	14	Cadherin-like
IPR003100	12	12	7	5	Argonaute/Dicer protein, PAZ
IPR003607	7	11	13	5	Metal-dependent phosphohydrolase, HD domain
IPR002402	147	132	68	56	Cytochrome P450, E-class, group II
IPR007889	20	9	8	6	Helix-turn-helix, Psq
IPR013162	22	22	29	12	CD80-like, immunoglobulin C2-set
IPR002401	183	168	93	73	Cytochrome P450, E-class, group I
IPR004841	18	23	19	10	Amino acid permease domain
IPR003645	12	9	9	5	Follistatin-like, N-terminal
IPR002403	173	163	89	72	Cytochrome P450, E-class, group IV
IPR001438	15	19	19	9	EGF-like, type 2
IPR001222	16	16	9	7	Zinc finger, TFIIS-type
IPR000008	40	38	43	21	C2 calcium-dependent membrane targeting
IPR003954	32	35	35	18	RNA recognition, domain 1
IPR008978	20	31	17	12	HSP20-like chaperone
IPR001991	35	47	31	20	Sodium:dicarboxylate symporter
IPR003961	61	69	59	34	Fibronectin, type III
IPR006552	9	12	12	6	VWC out
IPR009053	100	55	48	37	Prefoldin

Table S11. Occurrence of the 50 most over-represented GO terms in the three mosquitoes, *C. quinquefasciatus*, *Ae. aegypti, An. gambiae*, and *D. melanogaster* genomes.

InterPro	Cq	Aa	Ag	Dm	Description
GO:0019752	11	14	12	5	carboxylic acid metabolic process
GO:0004952	11	7	19	5	dopamine receptor activity
GO:0006032	31	27	22	11	chitin catabolic process
GO:0008060	9	10	10	4	ARF GTPase activator activity
GO:0032312	9	10	10	4	regulation of ARF GTPase activity
GO:0004553	67	63	50	25	hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0044262	10	9	9	4	cellular carbohydrate metabolic process
GO:0008667	71	51	39	23	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase activity
GO:0015904	22	23	18	9	tetracycline transport
GO:0015520	22	23	18	9	tetracycline:hydrogen antiporter activity
GO:0004707	11	10	7	4	MAP kinase activity
GO:0009239	71	51	39	23	enterobactin biosynthetic process
GO:0005230	33	26	24	12	extracellular ligand-gated ion channel activity
GO:0006418	70	43	46	23	tRNA aminoacylation for protein translation
GO:0007154	26	23	27	11	cell communication
GO:0032259	24	18	13	8	methylation
GO:0003707	24	20	24	10	steroid hormone receptor activity
GO:0046933	29	26	26	12	hydrogen ion transporting ATP synthase activity, rotational mechanism
GO:0006835	35	48	31	17	dicarboxylic acid transport
GO:0017153	35	48	31	17	sodium:dicarboxylate symporter activity
GO:0004568	31	27	22	12	chitinase activity
GO:0006396	45	38	37	18	RNA processing
GO:0005249	87	57	86	35	voltage-gated potassium channel activity

(continued)

Table S11. Continued.

GO:0019752	11	14	12	5	carboxylic acid metabolic process
GO:0004952	11	7	19	5	dopamine receptor activity
GO:0006032	31	27	22	11	chitin catabolic process
GO:0008060	9	10	10	4	ARF GTPase activator activity
GO:0032312	9	10	10	4	regulation of ARF GTPase activity
GO:0004553	67	63	50	25	hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0044262	10	9	9	4	cellular carbohydrate metabolic process
GO:0008667	71	51	39	23	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase activity
GO:0015904	22	23	18	9	tetracycline transport
GO:0015520	22	23	18	9	tetracycline:hydrogen antiporter activity
GO:0004707	11	10	7	4	MAP kinase activity
GO:0009239	71	51	39	23	enterobactin biosynthetic process
GO:0005230	33	26	24	12	extracellular ligand-gated ion channel activity
GO:0006418	70	43	46	23	tRNA aminoacylation for protein translation
GO:0007154	26	23	27	11	cell communication
GO:0032259	24	18	13	8	methylation
GO:0003707	24	20	24	10	steroid hormone receptor activity
GO:0046933	29	26	26	12	hydrogen ion transporting ATP synthase activity, rotational mechanism
GO:0006835	35	48	31	17	dicarboxylic acid transport
GO:0017153	35	48	31	17	sodium:dicarboxylate symporter activity
GO:0004568	31	27	22	12	chitinase activity
GO:0006396	45	38	37	18	RNA processing
GO:0005249	87	57	86	35	voltage-gated potassium channel activity

Table S12. Chromosomal assignment of 38 *C. quinquefasciatus* genes.

C. quinquefasciatus chromosome number	*C. quinquefasciatus* chromosome position	Marker Name	GenBank accession(s)	*C. quinquefasciatus* supercontig	*C. quinquefasciatus* gene	*An. gambiae* ortholog gene(s)	*An. gambiae* ortholog chromosome	*D. melanogaster* ortholog gene(s)	*D. melanogaster* ortholog chromosome
1	0	CX60	FD664718	supercont3.127	CPIJ006671	AGAP000541 AGAP009572	X (9 Mb) 3R (36 Mb)	CG12324 CG2033	2R (6.7 Mb) X (13 Mb)
1	5.9	LF188 (*)	BM005472	supercont3.660	CPIJ014663	AGAP000883	X (16 Mb)	CG11901	3R (25 Mb)
1	11.7	LF284	BM005502	supercont3.56	CPIJ003890	AGAP002306	2R (19 Mb)	CG5502	3R (23.7 Mb)
1	12.5	TY7	R19560	supercont3.213	CPIJ009089	AGAP001617	2R (6.8 Mb)	CG5915	3R (19.8Mb)
2	0	LF334	BM005506	supercont3.32	CPIJ002431	AGAP009491	3R (34.7 Mb)	CG6105	2L (10.9 Mb)
2	9.6	LF335	BM005505	supercont3.68	CPIJ004343	AGAP009604	3R (36.9 Mb)	CG10527	2R (16.9 Mb)
2	15.9	CX90	FD664719 FD664720	supercont3.65	CPIJ004272	AGAP003857	2R (44.6 Mb)	CG2013	3R (0.7 Mb)
2	27.3	DDC	U27581	supercont3.474	CPIJ013307	AGAP009091	3R (25.5 Mb)	CG10697	2L (19 Mb)
2	29.2	CX40	FD664709	supercont3.48	CPIJ003470	none	none	CG7203 CG7214 CG7216	2L (7.7 Mb)
2	36.1	CX44	FD664710 FD664711	supercont3.5 supercont3.1074	CPIJ000470 CPIJ017879	AGAP007769	3R (0.5 Mb)	CG3326	2L (3.3 Mb)
2	41.2	LF129	BM005504	supercont3.175	CPIJ007698	AGAP009920	3R (45 Mb)	CG5827	2L (5 Mb)
2	42.3	CX61	FD664712 FD664713	supercont3.134	none				
2	42.9	LF203	BM005503	supercont3.95	CPIJ005613	AGAP010163	3R (49.7 Mb)	CG18001	2R (0.4 Mb)
2	54.3	CX107	FD664723 FD664724	supercont3.129	CPIJ006471	AGAP011828	3L (33.8 Mb)	CG6692	2R (9.8 Mb)
2	55.9	LF108	T58322 T58321	supercont3.67	CPIJ004482	AGAP010933	3L (13.3 Mb)	CG4464	X (16.5 Mb)
2	59.1	LF168	R47184 R47183	supercont3.129 supercont3.951	CPIJ006480 CPIJ016775	AGAP012100 AGAP012100	3L (37.6 Mb) 3L (37.6 Mb)	CG10305 CG10305	2L (18.4 Mb) 2L (18.4 Mb)
2	65.8	CX22	FD664703 FD664704	supercont3.626	CPIJ015038	AGAP010792	3L (10.4 Mb)	CG6020	3L (20.4 Mb)
2	65.9	CX35	FD664707 FD664708	supercont3.66	CPIJ004396	AGAP004789 AGAP004791	2L (3.5 Mb)	CG9418	2R (17 Mb)

(continued)

Table S12. Continued

Cx quinqefasciatus chromosome number	*Cx. quinqefasciatus* chromosome position	Marker Name	GenBank accession(s)	*Cx. quinqefasciatus* supercontig	*Cx. quinquefasciatus* gene	*An. gambiae* ortholog gene(s)	*An. gambiae* ortholog chromosome	*D. melonogaster* ortholog gene(s)	*D. melanogaster* ortholog chromosome
2	67.4	LF386	BM005497	supercont3.111	CPIJ005878	AGAP011687	3L (31.3 Mb)	CG4759	3R (21.5 Mb)
2	70.8	LF377	BM005496						
2	73.7	LF323	BM005507	supercont3.177	CPIJ008264	AGAP011423	3L (24.4 Mb)	CG12161	3R (1 Mb)
								CG18341	X (5.7 Mb)
								CG3329	3L (14.9 Mb)
2	76.4	CX114	FD664728	supercont3.177	CPIJ008265	AGAP011424	3L (24.4 Mb)	CG4046	2R (18.4 Mb)
			FD664729						
2	85.9	CX111	FD664725	supercont3.1217	CPIJ018569	AGAP012397	3L (41.7 Mb)	CG11246	3L (21.7 Mb)
			FD664726	supercont3.108	CPIJ005206	AGAP012397	3L (41.7 Mb)	CG11246	3L (21.7 Mb)
3	0	LF115	R67978	supercont3.3	CPIJ000205	AGAP007580	2L (47.8 Mb)	CG8615	3L (7.2 Mb)
			R67977						
3	17.9	CX112	FD664727	supercont3.99	CPIJ005652	AGAP007157	2L (43.6 Mb)	CG10423	3R (21 Mb)
3	18.5	CX17	FD664699	supercont3.205	CPIJ008759	AGAP005659	2L (18.3 Mb)	none	none
			FD664700						
3	21	LF250	T58311	supercont3.309	CPIJ010827	AGAP005991	2L (24.5 Mb)	CG6253	3L (8.6 Mb)
			T58310						
3	21.7	LF124	BM005518	supercont3.99	CPIJ005652	AGAP007157	2L (43.6 Mb)	CG10423	3R (21 Mb)
			T58324						
3	25.4	LF264	BM005463						
3	26	CX53	FD664714	supercont3.139	CPIJ007044	AGAP004904	2L (5.7 Mb)	CG6871	3L (18.8 Mb)
			FD664715					CG9314	2L (8.7 Mb)
3	26.3	LF99c (!)	BM005477	supercont3.208	CPIJ008915	none	none	CG4264	3R (11 Mb)
				supercont3.126	CPIJ006534	AGAP004944	2L (6.5 Mb)	CG8937	3L (13.9 Mb)
3	28	LF272	BM005484	supercont3.73	CPIJ004532	AGAP004887	2L (4.9 Mb)	CG3922	3L (9.4 Mb)
3	45.8	CX51							
3	51.5	LF106	BM005490	supercont3.550	CPIJ013966	AGAP004462	2R (56.7 Mb)	CG6684	3R (7 Mb)
						AGAP005092	2L (9.9 Mb)		
3	52.1	CX59							
3	62.2	LF99a (!)	BM005477	supercont3.208	CPIJ008915	none	none	CG4264	3R (11 Mb)
				supercont3.126	CPIJ006534	AGAP004944	2L (6.5 Mb)	CG8937	3L (13.9 Mb)
3	73	CX11	FD664697	supercont3.446	CPIJ013141	none	none	CG4994	3L (14.5 Mb)

(continued)

Table S12. Continued

C. quinqefasciatus chromosome number	*C. quinqefasciatus* chromosome position	Marker Name	GenBank accession(s)	*C. quinqefasciatus* supercontig	*C. quinquefasciatus* gene	*An. gambiae* ortholog gene(s)	*An. gambiae* ortholog chromosome	*D. melonogaster* ortholog gene(s)	*D. melanogaster* ortholog chromosome
3	79.2	LF128	BM005494	supercont3.185	CPIJ008366	AGAP001903	2R (11.9 Mb)	CG10748 CG10749 CG7998	3L (12.9 Mb) 3R (14 Mb)

(*) LF188 maps as an RFLP locus to two loci in *Aedes aegypti*: chr.1 and chr.2

(!) LF99 maps as an RFLP locus to two loci in *Aedes aegypti*: chr.1 and chr.3

Table S13. Correlations between *C. quinquefasciatus*, *Ae. aegypti*, *An. gambiae* and *D. melanogaster* chromosomes. Correlations between *C. quinquefasciatus*, *An. gambiae* and *D. melanogaster* were obtained using markers and orthologs between the three genomes. Correlations between *Ae. aegypti*, *An. gambiae* and *D. melanogaster* were taken from the literature (*4*). Correlation between *C. quinquefasciatus* and *Ae. aegypti* were extrapolated from the first two analyses.

C. quinquefasciatus	***Ae. aegypti***	***An. gambiae***	***D. melanogaster***
Chr.1 p	1	X	X;3R
Chr.1 q	1	2R	3R
Chr.2p	2q	3R	2L
Chr.2q	3q	3L	2R;3L
Chr.3p	2p	2L	3L;2R
Chr.3q	3p	2R	3R (3L)

Table S14. Number of elements in families and copy number of transposable elements (TE) for *C. quinquefasciatus*, as well as percentage of the genome occupied by TE sequences for three mosquito genomes. MITE sequences were only classified for the *C. quinquefasciatus* genome. Names and statistics of transposable elements not present in *C. quinquefasciatus* are italicized. Totals are shown in bold.

TE family	*C. quinquefasciatus*			*Ae. aegypti*	*An. gambiae*
	Num. Ele.	**copy num.**	**% genome**	**%genome**	**% genome**
Class I					
LTR retrotransposons	**171**	**1,886**	**3.89%**	**12.41%**	**2.64%**
Ty1_copia	32	328	0.18%	5.51%	0.15%
Ty3_gypsy	57	749	0.87%	2.51%	1.51%
Pao_Bel	81	631	1.82%	4.39%	0.98%
LARD	1	178	1.02%	0.00%	0.00%
Non-LTR retrotransposons	**209**	**16,869**	**4.45%**	**12.67%**	**3.75%**
CR1	31	821	0.28%	0.94%	1.61%
I	11	62	0.02%	0.67%	0.13%
Jockey	16	4,714	1.77%	3.32%	0.58%
L1	57	589	0.15%	0.34%	0.07%
L2	9	1,020	0.61%	0.17%	0.04%
LOA	9	176	0.09%	1.02%	0.00%
Loner	2	128	0.12%	0.88%	0.10%
Outcast	4	8	0.00%	0.06%	0.09%
R1	32	243	0.14%	1.76%	0.18%
RTE	8	907	0.38%	3.46%	0.79%
R4	*0*	*0*	*0.00%*	*0.05%*	*0.05%*
unclassified LINE	30	8,201	0.88%	0.00%	0.11%
SINEs					
tRNA-related Sines	**11,758**	**11,460**	**0.52%**	**1.22%**	**0.46%**

(continued)

Table S14. Continued.

	C. quinquefasciatus			*Ae. aegypti*	*An. gambiae*
Class II					
DNA transposons	**129**	**151,300**	**19.40%**	**13.97%**	**4.54%**
hAT	8	53	0.43%	0.51%	0.43%
P	1	1	0.19%	0.15%	0.23%
PIF	2	7	0.11%	1.88%	0.20%
piggyBac	1	23	0.35%	0.02%	0.07%
pogo	7	38	0.30%	0.71%	0.00%
Tc1	1	23	0.02%	0.30%	1.83%
gambol	3	21	0.01%	0.00%	0.27%
Transib	5	1,327	0.57%	0.00%	0.14%
Mutator	4	397	0.29%	0.00%	0.00%
mariner	*0*	*0*	*0.00%*	*0.00%*	*0.11%*
DD41D	*0*	*0*	*0.00%*	*0.00%*	*0.00%*
ItmD37D	*0*	*0*	*0.00%*	*0.10%*	*0.00%*
ItmD37E	*0*	*0*	*0.00%*	*0.32%*	*0.01%*
MITEs	**97**	**149410**	**17.12%**	**9.97%**	**1.25%**
hAT-linked	35	31,200	4.00%	N/A	N/A
P-linked	1	1,527	0.12%	N/A	N/A
PIF-linked	1	1,228	0.41%	N/A	N/A
Tc1-linked	11	17,663	1.71%	N/A	N/A
Transib-linked	3	738	0.12%	N/A	N/A
mutator-linked	1	300	0.50%	N/A	N/A
Chapaev-linked	3	9,811	1.18%	N/A	N/A
Sola-linked	5	6,297	1.00%	N/A	N/A
Joey-linked	0	0	0.00%	N/A	0.00%
m2bp-unclassifed	9	6,124	0.72%	0.00%	0.00%
m3bp-unclassifed	1	209	0.11%	0.48%	0.18%
m4bp-unclassifed	9	58,428	6.27%	2.18%	0.00%

(continued)

Table S14. Continued.

	C. quinquefasciatus			*Ae. aegypti*	*An. gambiae*
m5bp-unclassifed	1	2,079	0.13%	0.00%	0.03%
m7bp-unclassifed	6	1,182	0.11%	0.00%	0.00%
m8bp-unclassifed	4	1,728	0.09%	1.44%	0.82%
m9bp-unclassifed	0	0	0.00%	0.09%	0.00%
mTA-unclassifed	7	10,896	0.64%	5.79%	0.22%
Helitron	6	822	**0.49%**	**1.26%**	**0.11%**
Penelope					
Penelope-like	*0*	***0***	***0***	***0.41%***	***0.00%***
Totals	**12,611**	**350,506**	**28.75%**	**41.94%**	**11.49%**

References

1. P. F. Mattingly, L. E. Rozeboom, K. L. Knight, H. Laven, F. H. Drummond, S. R. Christophers, P. G. Shute, The *Culex pipiens* complex. *Trans. Roy. Ent. Soc.* London **102**, 331-382 (1951).

2. S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, E. S. Lander, ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177-189 (2002).

3. P. N. Rao, K. S. Rai, Genome evolution in the mosquitoes and other closely related members of the superfamily Culicoidea. *Hereditas* **113**, 139-144 (1990).

4. V. Nene, *et al.*, Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**, 1718-1723 (2007).

5. A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, E. Birney, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **2**, 327-335 (2009).

6. W. J. Kent, BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

7. D. Lawson, P. Arensburger, P. Atkinson, N. J. Besansky, R. V. Bruggner, R. Butler, K. S. Campbell, G. K. Christophides, S. Christley, E. Dialynas, M. Hammond, C. A. Hill, N. Konopinski, N. F. Lobo, R. M. MacCallum, G. Madey, K. Megy, J. Meyer, S. Redmond, D. W. Severson, E. O. Stinson, P. Topalis, E. Birney, W. M. Gelbart, F. C. Kafatos, C. Louis, F. H. Collins. VectorBase: a data resource for invertebrate host genomics. Nucleic Acids Res. **37**, D583-587 (2009).

8. A. L. Price, N. C. Jones, P. A. Pevzner, De novo identification of repeat families in large genomes. *Bionformatics* **Suppl 1**, 351-358 (2005).

9. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0 http://www.repeatmasker.org (1996-2004).

10. X. Huang, M.D. Adams, H. Zhou, A.R. Kerlavage, A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37-45 (1997).

11. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res.* **14**, 988-995 (2004).

12. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith, L. I. Hannick, R. Maiti, C. M. Ronning, K. B. Rush, C. D. Town, S. L. Salzberg, O. White, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666. (2003)

13. I. Korf, Gene finding in novel Genomes. *BMC Bioinformatics* **5**, 59 (2004).

14. S. E. Cawley, A. I. Wirth, T. P. Speed, Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Bio. Parasit.* **118**, 167-174 (2001).

15. M. Stanke, R. Steinkamp, S. Waack, B. Morgenstern, AUGUSTUS: a web server for gene finding in eukaryotes. Nuc. Acids Res. **32**, W309-W312. (2004)

16. W. H. Majoros, M. Pertea, S. L. Salzberg, TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879 (2004).

17. A. E. Tenney , H. B. Randall, C. Vaske, J. K. Lodge, T. L. Doering, M. R. Brent, Gene prediction and verification in a compact genome with numerous small introns. *Genome Res.* **14**, 2330-2335 (2004).

18. B. J. Haas, S. L. Salzberg, W. Zhu1, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell1, J. R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

19. E. Blanco, G. Parra, R. Guigó, in: *Curr. Protocols in Bioinfor.*, A. Baxevanis Ed. (Wiley, New York, 2002) Unit 4.3.

20. A. A. Salamov, V. V. Solovyev, *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **11**, 817-832 (2000).

21. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S. R. Eddy, Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439-441 (2003).

22. 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).

23. P. Flicek *et al.,* Ensembl's 10th year. *Nucleic Acids Res.* **38**, D557-D562 (2010).

24. Z. Bao, S. R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269-1276 (2002).

25. http://www.uniprot.org

26. G. S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

27. http://www.ncbi.nlm.nih.gov/dbEST/

28. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* **36**, D271-D275 (2008).

29. Personal communication, K. S. Campbell.

30. A. Mori, D. W. Severson, B. M. Christensen, Comparative linkage maps for the mosquitoes (*Culex pipiens* and *Aedes aegypti*) based on common RFLP loci. *J. Hered.* **90**, 160-164 (1999).

31. A. Mori, J. Romero-Severson, D. W. Severson, Genetic basis for reproductive diapause is correlated with life history traits within the *Culex pipiens* complex. *Insect Mol. Biol.* **16**, 515-524 (2007).

32. Personal communication, A. Mori and D. Severson.

33. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 586-1591 (2007).

34. D. H. Foley, J. H. Bryan, D. Y. Yeates, A. Saul, Evolution and systematics of Anopheles: insights from phylogeny of australasian mosquitoes. *Mol. Phyl. Evol.* **9**, 262-275 (1998).

35. M. W. Gaunt, M. A. Miles, An insect molecular clock dates the origin of the insects and accords with the palaeontological and biogeographic landmarks. *Mol. Biol. Evol.* **19**, 748-761 (2002).

36. J. Krzywinski, O. G. Grushko, N. J. Besansky, Analysis of the complete mitochondrial DNA from *Anopheles funestus*: An improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. *Mol. Phyl. Evol.* **39**, 417-423 (2006).

37. R. A. Holt *et al.*, The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129-49 (2002).

38. http://tefam.biochem.vt.edu/tefam

39. http://www.girinst.org/repbase/

40. R. Kalendar, C. M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy, A. H. Schulman, Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* **166**, 1437-1450 (2004).

41. F. Sabot, A. H. Shulman, Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* **97**, 381-388 (2006).

42. R. H. Waterson *et al.,* Initial sequencing and comparative analysis of the mouse genome. *Nature* **450**, 520-562 (2002)

43. J. K. I. Pace, C. Glibert, M. S. Clark and C. Feshotte, Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc. Natl. Acad. Sci. USA* **105**, 17023-17028 (2008).

44. L. B. Vosshall, A. M. Wong, R. Axel, An olfactory sensory map in the fly brain. *Cell* **102**, 147-159 (2000).

45. K. F. Störtkuhl, R. Kettler, Functional analysis of an olfactory receptor in *Drosophila melanogaster.* P*roc. Natl. Acad. Sci. USA.* **98**, 9381 (2001).

46. C. A. Hill, A. N. Fox, R. J. Pitts, L. B. Kent, P. L. Tan, M. A. Chrystal, A. L. Cravchik, F. H. Collins, H. M. Robertson, L. J. Zwiebel, G Protein–Coupled Receptors in *Anopheles gambiae*. *Science* **298**, 176-178 (2002).

47. J. Bohbot, R. J. Pitts, H.-W. Kwon, M. Rutzler, H. M. Robertson, L. J. Zwiebel, Molecular characterization of the *Aedes aegypti* odorant receptor gene family. *Insect Mol. Biol.* **16**, 525-537

(2007).

48. C. R. Williams, M. J. Kokkinn, B. P. Smith, Intraspecific variation in odor-mediated host preference of the mosquito *Culex annulirostris*. *J. Chemical Ecol.* **29**, 1889-1903 (2003).

49. E. A. Hallem, A. Dahanukar, J. R. Carlson, Insect odor and taste receptors. *Annu. Rev. Entomol.* **51**, 113-35 (2006).

50. L. B. Vosshall , R. F. Stocker, Molecular architecture of smell and taste in *Drosophila*. *Annu. Rev. Neurosci.* **30**, 505-33 (2007).

51. T. Lu, Y. T. Qiu, G. Wang, J. Y. Kwon, M. Rutzler, H. W. Kwon, R. J. Pitts, J. J. van Loon, W. Takken, J. R. Carlson, L. J. Zwiebel, Odor coding in the maxillary palp of the malaria vector mosquito *Anopheles gambiae*. *Curr. Biol.* **17**, 1533-44 (2007).

52. H. M. Robertson, L. B. Kent, Evolution of the gene lineage encoding the carbon dioxide receptor in insects. *J. Insect Sci.* **9**, 19 (2009).

53. L. B. Kent , H. M. Robertson, Evolution of the sugar receptors in insects. *BMC Evol. Biol.* **9**, 41 (2009).

54. S. J. Moon, M. Köttgen, Y. Jiao, H. Xu, C. Montell, A taste receptor required for the caffeine response *in vivo*. *Curr. Biol.* **16**, 1812-1817 (2006).

55. N. Thorne, H. Amrein, Atypical expression of *Drosophila* gustatory receptor genes in sensory and central neurons. *J. Comp Neurol.* **506**, 548-568 (2008).

56. L. B. Kent, K. K. Walden, H. M. Robertson, The Gr family of candidate gustatory and olfactory receptors in the yellow-fever mosquito *Aedes aegypti*. *Chem. Senses.* **33**, 79-93 (2008).

57. J. M. C. Ribeiro, I. M. B. Francishetti, Role of arthropod saliva in blood feeding: Sialome and post-sialome perspectives. *Ann. Rev. Entomol.* **48**, 73-88 (2003).

58. B. Arca, F. Lombardo, I. M. B. Francishetti, O. Marinotti, M. Coluzzi, J. M. C. Ribeiro, An updated catalogue of salivary gland transcirpts in the adult female mosquito, *Anopheles gambiae*. *J. Experimental Biol.* **208**, 3971-3986 (2005).

59. E. Calvo, A. Dao, V. M. Phan, J. M. C. Ribeiro, An insight into the sialome of *Anopheles funestus* reveals an emerging pattern in anopheline savivary protein families. *Insect Biochem. Mol. Biol.* **37**, 164-175 (2007).

60. E. Calvo, V. M. Phan, O. Marinotti, J. F. Andersen, J. M. C. Ribeiro, The salivary gland transcriptome of the neotropical malaria vector *Anopheles darlingi* reveals accelerated evolution of genes relevant to hematophagy. *BMC Genomics* **10**, 57 (2009).

61. J. G. Valenzuela, I. M. B. Francischetti, V. M. Pham, M. K. Garfield, J. M. C. Ribeiro, Exploring the salivary gland transcriptome and proteome of the *Anopheles stephensi* mosquito. *Insect Biochem.*

Mol. Biol. **33**, 717-732 (2003).

62. B. Arca, F. Lombardo, I. M. B. Francishetti, V. M. Pham, M. Mestres-Simon, J. F. Andersen, J. M. C. Ribeiro, An insight into the sialome of the adult female mosquito *Aedes albopictus*. *Insect Biochem. Mol. Biol.* **37**, 107-127 (2007).

63. J. M. C . Ribeiro, R. Charlab, V. M. Pham, M. Garfield, J. G. Valenzuela, An insight into the salivary transcriptome and proteome of the adult female mosquito *Culex pipiens quinquefasciatus*. *Insect Biotech. Mol. Biol.* **34**, 543-563 (2004).

64. J. M. C. Ribeiro , B. Arca, F. Lombardo, E. Calvo, V. M. Phan, P. K. Chandra, S. K. Wikel, An annotated catalogue of salivary gland transcripts in the adult female mosquito, *Aedes aegypti*. *BMC genomics* **8**, 102 (2007).

65. http://exon.niaid.nih.gov/transcriptome/Cp_genome/cp-spitome-web.xls

66. C. Allmang, L. Wurth, A. Krol, The selenium to selenoprotein pathway in eukaryotes: more molecular partners than anticipated. *Biochim Biophys Acta.* **1790**, 1415-23 (2009).

67. C. E. Chapple, R. Guigó, Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PLoS One* **13**;3(8):e2968 (2008).

68. D. T. Jones, W. R. Taylor and J. M. Thornton, The rapid generation of mutation data matrices from protein sequences. *Computer applications in the Biosciences* **8**, 275-282 (1992).

69. J. Felsenstein, PHYLIP (phylogeny inference package) Seattle, DC: Department of Genome Sciences, University of Washington (2005).

70. C. Strode et al. Genomic analysis of detoxification genes in the mosquito *Aedes aegypti*. *Insect Biochem. Mol. Biol.*, **38**,113-23 (2008)

71. H. Ranson, C. Claudianos, F. Ortelli, C. Abgrall, J. Hemingway, M. V. Sharakhova, M. F. Unger, F. H. Collins, R. Feyereisen, Evolution of supergene families associated with insecticide resistance. *Science* **298**,179-181 (2002).

72. Bartholomay *et al.*, Pathogenomics of *Culex quinquefasciatus* and meta-analysis of infection responses to diverse pathogens. *Science* (this issue).