Letter

A Comparative Genomic Analysis of Two Distant Diptera, the Fruit Fly, *Drosophila melanogaster*, and the Malaria Mosquito, *Anopheles gambiae*

Viacheslav N. Bolshakov,[1] Pantelis Topalis,[1] Claudia Blass,[2] Elena Kokoza,[2,3] Alessandra della Torre,[4] Fotis C. Kafatos,[2,5] and Christos Louis[1,5,6]

[1]*Genome Research Laboratory, Institute of Molecular Biology and Biotechnology, FORTH, 71110 Heraklion, Crete, Greece;* [2]*European Molecular Biology Laboratory, 69117 Heidelberg, Germany;* [3]*Institute of Cytology and Genetics, 630090 Novosibirsk, Russia;* [4] *Dipartimento di Scienze di Sanità Pubblica, Sez. di Parassitologia, Università "La Sapienza", 00185 Roma, Italy; and* [5]*Department of Biology, University of Crete, 71110 Heraklion, Crete, Greece*

Genome evolution entails changes in the DNA sequence of genes and intergenic regions, changes in gene numbers, and also changes in gene order along the chromosomes. Genes are reshuffled by chromosomal rearrangements such as deletions/insertions, inversions, translocations, and transpositions. Here we report a comparative study of genome organization in the main African malaria vector, *Anopheles gambiae*, relative to the recently determined sequence of the *Drosophila melanogaster* genome. The ancestral lines of these two dipteran insects are thought to have separated ~250 Myr, a long period that makes this genome comparison especially interesting. Sequence comparisons have identified 113 pairs of putative orthologs of the two species. Chromosomal mapping of orthologous genes reveals that each polytene chromosome arm has a homolog in the other species. Between 41% and 73% of the known orthologous genes remain linked in the respective homologous chromosomal arms, with the remainder translocated to various nonhomologous arms. Within homologous arms, gene order is extensively reshuffled, but a limited degree of conserved local synteny (microsynteny) can be recognized.

Modern genomics have revolutionized genetics and, consequently, biology. The enormous acceleration of data acquisition, in fields such as whole genome sequence determination and genome-wide gene expression profiling, has opened novel possibilities for the study of model organisms and organisms for which, until recently, only rudimentary biological knowledge was available (orphan organisms). For example, until a decade ago only a few tens of genes had been identified in important insect disease vectors such as *Anopheles gambiae* or *Aedes aegypti*, which now number ~24,000 and 1700 entries, respectively, in the nucleic acid databases. Many of these represent partial genomic sequences, sequence tagged sites (STSs), and anonymous cDNAs or expressed sequence tags (ESTs; for review, see Louis 1999). Such genetic and molecular genetic information may prove helpful in designing new schemes to fight the diseases transmitted by these mosquitoes, such as malaria and dengue (James et al. 1999). Progress in elucidating the genomic information of formerly orphan insect organisms can be considerably accelerated by using the closest available model organism, in this case *Drosophila melanogaster*, as a guide.

A. gambiae s.s. (sensu stricto) is a member of the African *A. gambiae* species complex that consists of six distinct sibling species and itself can be distinguished into a series of taxa or incipient species (Coluzzi et al. 1985), all differing in vectorial capacity (see Touré et al. 1998). The pioneering studies of Coluzzi and his collaborators on the construction of polytene maps for this species complex and the documentation of both fixed and polymorphic inversions can be viewed as the start of genomic research on the malaria mosquito.

Over the past decade, knowledge of the molecular biology and genetics of *A. gambiae* s.s. has vastly improved. For example, numerous molecular studies of the interactions between *Anopheles* and *Plasmodium* have radically improved our understanding of this vector-parasite system (for review, see Sinden 1999). The molecular study of the genome was initiated with the construction of a first low-resolution physical map, linked to the polytene chromosomes (Zheng et al. 1991), followed by the construction of a detailed, microsatellite-based recombination map (Zheng et al. 1993, 1996). Integration of the genetic (recombinational), cytogenetic (polytene), and molecular (clone and sequence) maps has progressed rapidly; it entails the genetic and cytogenetic mapping of random amplified polymorphic DNA (RAPD) markers (Dimopoulos et al. 1996a), the recombinational mapping of microsatellites, and the assignment of both microsatellites and anonymous DNA markers to specific chromosomal locations, using in situ hybridization to polytene chromosomes (della Torre et al. 1996; Dimopoulos et al. 1996a; Zheng et al. 1996; Wang et al. 1999). Microsatellites have been used successfully both for gene mapping (Collins et al. 1997; Zheng et al. 1997; Ranson et al. 2000) and for studies of population biology (e.g., see Lanzaro et al. 1998; Kamau et al. 1999; Wang et al. 1999, 2001). Finally, routine germline transformation and thus reverse genetic studies of *A. gambiae* can be expected soon, judging by the recent success in transforming both anopheline (*A. stephensi*; Catteruccia et al. 2000) and aedine mosquitoes (*Ae. aegypti*; Coates et al. 1998; Jasinskiene et al. 1998).

[6]Corresponding author.
E-MAIL louis@imbb.forth.gr; FAX 30-81-391104.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.196101.

Important additional tools for comparative genomic studies of *A. gambiae* have become available recently. They include a collection of ESTs that may represent ~10% of the mosquito genes (Dimopoulos et al. 2000), and ~17,500 sequence-tagged ends of a bacterial artificial chromosomes (BAC) chromosomal library representing 14.5 Mb or 7% of the expected euchromatic DNA sequence (http://bioweb.pasteur.fr/BBMI; C. Roth and F.H. Collins, pers. comm.). An experimental strategy that combines the identification of orthologs by sequence similarity searches and their mapping to the chromosomes or linkage groups of different species has proven to be very informative in comparative genomic studies of both animals (O'Brien et al. 1999) and plants (Terryn et al. 1999). An important type of information derived from such studies is the degree of conserved synteny: to what extent the chromosomal dynamics in evolution permit linkage group conservation, that is, persistent linkage of most genes in a given chromosome between compared species (long-range synteny in homologous chromosomes). A second important issue is to what extent originally neighboring genes remain clustered (local conserved synteny, or microsynteny) rather than becoming randomized in terms of their order within the homologous chromosome.

Here we used the essentially complete sequence information on the *D. melanogaster* genome (Adams et al. 2000), together with the available *A. gambiae* genomic resources, to address the questions of sequence conservation, long-range synteny, and local microsynteny between the genomes of the mosquito and the fruit fly, two distantly related diptera.

RESULTS

Chromosomal Distribution of *A. gambiae* Orthologs of Genes From Two *D. melanogaster* Chromosomal Regions

In a first set of experiments aiming at exploring long-range synteny and microsynteny, we identified, among the currently available *A. gambiae* sequences, putative orthologs of genes in which in *D. melanogaster* are clustered within two well-studied chromosomal regions, each nearly 3 Mb long. We then determined the genomic locations of these putative orthologs by hybridization to the mosquito polytene chromosomes.

The fruit fly genomic regions that were chosen for these experiments have been completely sequenced and annotated, both in clone-by-clone sequencing projects and as part of whole-genome shotgun sequencing. One of these *Drosophila* regions is the autosomal *Adh* region, covering 2.9 Mb on both sides of the *Adh* gene, in divisions 34B–35F of chromosomal arm *2L* (Ashburner et al. 1999). The other is the tip of chromosome *X*, encompassing 2.6 Mb in polytene divisions 1–3 (Benos et al. 2000, 2001). Both of these regions were also covered by whole-genome shotgun sequencing (Adams et al. 2000).

The 256 genes from the tip of the *X* and the 219 genes from the *Adh* region of *Drosophila* were used to query, by TBLASTN, collections of both STSs and ESTs of *Anopheles*: the 17,506 STSs representing end sequences of BAC clones, and the 6012 ESTs that correspond to 2380 potential genes (cDNA clone clusters from a subtracted normalized library; Dimopoulos et al. 2000). To define genes as putative orthologs, hits that satisfied criteria of high score of >40, probability *P*(N) of $<1\ e^{-10}$, and percentage of identical amino acid residues >30 over a long range were selected in a first round. From them, all spurious hits that were caused by the presence of low complexity segments were eliminated, and the remaining hits were confirmed by BLASTX analysis as best bidirectional hits against a database of 14,080 amino acid sequences of known and predicted *Drosophila* genes (release 1.0; Adams et al. 2000). Those that passed this test were further verified by direct comparison to the corresponding *Drosophila* entry, taking into account potential intron-exon boundaries. Henceforth, these validated genes will be referred to as orthologs for convenience (see also Discussion). These procedures (see also Methods) identified 19 mosquito orthologs of unique genes found in the tip region of the *Drosophila X* chromosome and 31 orthologs of unique genes found in the *Drosophila Adh* region. For greater accuracy, we eliminated from consideration additional probable orthologs (18 showing hits to *X*-tip and nine showing hits to *Adh* region genes), because they belong to chromosomally dispersed multigene families. This was necessary because the true ortholog can not be chosen among the different members of a given gene family until both genomes are fully sequenced.

The 50 orthologs that were retained for further analysis were present in 33 BAC and 37 cDNA clones (a number of them were detected by both STSs and ESTs). Representative clones were used as probes for in situ hybridization analysis to *A. gambiae* polytene chromosomes. Tables 1 and 2 include the results of this analysis for the *X*-tip orthologs and *Adh* region orthologs, respectively. Notably, the Tables show cytogenetic and molecular locations of the 50 *Drosophila* genes and the sequence identifiers and cytogenetic locations of the corresponding *A. gambiae* orthologs. The distribution of the *X*-Tip and *Adh* region orthologs among the five polytene chromosome arms of *A. gambiae* are tabulated in Table 3, together with the results of statistical analysis of these distributions using the binomial test of significance, confirmed by the χ^2 test.

For the statistical analysis, we compared the number of orthologs corresponding to each *Drosophila* region that were observed in each chromosomal arm of the mosquito to the number expected if the association were random according to chromosomal arm length. To calculate the expected numbers, the lengths of the five mosquito chromosomal arms were estimated according to the number of their lettered subdivisions, as recognized in the map of Coluzzi and associates (22 subdivisions for *X*, 54 for *2R*, 40 for *2L*, 37 for *3R*, and 31 for *3L*; or 12.0%, 29.35%, 21.7%, 20.1%, and 16.85% of the total, respectively; the map is accessible at http://www.anodb.gr/AnoDB/Cytomap/). The binomial test is an exact probability test that is used to examine the distribution of a single dichotomy in conditions when only a relatively small sample is available, as is the case here. It provides a one-sample test of the difference between the sampled distribution and a given distribution. In this case, the given distribution is based on the null hypothesis that the genes of each *Drosophila* chromosomal region are randomly redistributed across all five chromosome arms of *A. gambiae* according to their lengths. As shown in Table 3 for the gene probes derived from the tip of chromosome *X*, all *P* values are >0.05, and thus, the null hypothesis cannot be rejected. Similarly, the χ^2 statistic (not shown) is equal to 2.415, lower than the critical value $\chi^2_4[0.05] = 9.49$; therefore, the null hypothesis can not be rejected. By these criteria, none of the five mosquito chromosomal arms is significantly enriched for orthologs of the *X*-tip genes of *Drosophila*.

In contrast, the results of the same analysis strongly indicate that the *Drosophila* arm *2L* (at least its *Adh* region) corresponds to the chromosome arm *3R* of *A. gambiae* (*P* = 9.939

Table 1. Cytological Mapping of Putative *Anopheles gambiae* Orthologues of Genes in the *Drosophila melanogaster* Divisions 1–3 of Chromosome X

D. melanogaster			% Identical amino acids	*A. gambiae*		
Cytology	Mol. map	Gene		Sequence	Accession no.	Cytology
1B4-5	263	*svr*	61	4A3B-AAA-E-12-R	AJ284634	*24:* 12B
1B5-7	294	*EG:65F1.1*	34–46	02A19-t7	AL140406	*2L:* 21C
1B10	394	*RpL36*	83	4A3A-P7A4-F	AJ281575	*3R:* 32A
1B10-11	412	*EG:115C2.11*	97	4A3B-AAX-A-12-R	AJ285778	*2L:* 21F
1B10-11	420	*spkA*	92	4A3B-AAI-C-06-F	AJ284038	*3R:* 35C
1B12-C1	451	*sdk*	58–78	06M18-sp6	AL143155	*X:* 1D
1C5-D1	704	*EG:BACR7A4.5*	48–62	21I19-sp6	AL154783	*3L:* 38B
1D4-E1	797	*EG:BACR7A4.8*	51	4A3A-AAT-H-11-R	AJ282437	*3L:* 44D
2A2-3	1137	*EG:132E8.3*	78	4A3A-P3A1-F	AJ281385	*2R:* 15A
2A4-B1	1188	*EG:49E4.1*	90	25H04-sp6	AL153026	*2R:* 7A
2B1-2	1239	*sta*	75	4A3A-AAT-D-12-R	AJ282394	*3L:* 44B
2B6-7	1417	*EG:171E4.4*	48	21P07-t7	AL151067	*X:* 2C
2B9-10	1521	*Adar*	49–87	27D16-sp6	AL154010	*X:* 4A
2B13-14	1621	*EG:63B12.4*	61	4A3A-AAT-B-06-R	AJ282366	*2R:* 12C
2B13-15	1637	*trr*	45–95	32G18-sp6	AL157001	*3L:* 41D
2B14-16	1648	*arm*	76	4A3A-AAQ-E-05-F	AJ280532	multiple
2D4-E1	1893	*Pgd*	29–68	30A12-sp6	AL155664	*3R:* 31D
2F1-2	2022	*EG:25E8.1*	67	4A3A-ABB-H-01-F	AJ281139	*2R:* 9B
2F3	2077	*EG:BACH48C10.2*	40	4A3B-AAW-F-07-R	AJ285740	*3R:* 33B

For *D. melanogaster,* gene names and cytological localization are derived from FlyBase; the genes are ordered according to the molecular map, which represents the approximate positions (in kilobases) of the 5′ ends of the genes (Benos et al. 2001). For *A. gambiae,* the cytological localizations on polytene chromosomes are according tot he maps of M. Coluzzi, A. Sabbatini, M.A. Di Deco, and V. Petrarca (pers. comm.); the names of the sequences are as they appear in corresponding database entries with the indicated accession numbers. The *A. gambiae* 4A3A and 4A3B entries are expressed sequence tags, and entries ending with sp6 or t7 are sequence tagged sites (STSs) from chromosomal bacterial artificial chromosomes (BACs), % identities refer to one or more exons.

e^{-12}). That mosquito arm includes nearly fourfold as many genes as expected: It contains 24 (77%) of the currently available orthologs of the *Drosophila Adh* region genes, whereas only 7 (23%) orthologs are scattered over three other mosquito autosomal arms. Furthermore, in three out of four remaining mosquito chromosomal arms (*2R*, *2L*, and *3L*), the prevalence of orthologs of *Drosophila 2L* genes is statistically significantly lower than expected. Thus, the binomial test clearly rejects the null hypothesis of random redistribution of *Adh* region genes, in terms of both positive and negative correlations. Rejection is also supported by the χ^2 analysis, in which the statistic (not shown) is equal to 64.12 with the same critical value as before ($\chi^2_4[0.05] = 9.49$).

It should be noted from Table 3 that the mosquito orthologs of the *Adh* region genes are not evenly distributed within the mosquito *3R* arm: Half of them are located within four chromosomal subdivisions (29C, 31C, 32B, and 33A), whereas the other half are scattered among the other 33 subdivisions of *3R*. This apparent clustering may correspond to microsynteny, as will be discussed below.

Distribution of Randomly Selected *A. gambiae* Sequences and Their *D. melanogaster* Orthologs

A similar but reverse method was used in a second experiment addressing the question of long-range synteny. In this case, we started by mapping random *A. gambiae* STSs mapped on the polytene chromosomes and determined their orthologs and the respective cytogenetic locations in *D. melanogaster*.

Randomly selected BAC clones of *A. gambiae* from the library that had been used to determine STS end sequences (http://bioweb.pasteur.fr/BBMI; C. Roth and F.H. Collins, pers. comm.) were mapped by in situ hybridization to mosquito polytene chromosomes. A total of 1217 STS were available from 720 cytogenetically mapped clones, and they were used for a BLASTX search of the protein sequences corresponding to the 14,080 known and predicted *D. melanogaster* genes (release 1.0; Adams et al. 2000). This search led to the identification of 49 mapped STS that were putative orthologs of unique *D. melanogaster* genes. In addition, *A. gambiae* genes of known cytogenetic location were used to search the same *Drosophila* database, yielding 21 additional hits. This number also included *cecropin* and *ADP/ATP*, two *A. gambiae* genes, each of which is homologous to a corresponding small multigene family in *Drosophila*, clustered at a single cytogenetic location. Table 4 lists these 70 mosquito gene sequences by cytogenetic location, together with their *Drosophila* orthologs and their locations. Table 5 summarizes and correlates the chromosomal locations of corresponding sequences in the two species. As in the previous experiment, the binomial test and the confirming χ^2 analysis (not shown) used the numbers of orthologs expected on each *Drosophila* chromosomal arm, in this case according to a random distribution calculated on the basis of the respective known DNA content of the *Drosophila* arms (Adams et al. 2000).

The data from this second experiment (Table 5) completely confirm and extend the conclusions from the first experiment. They identify statistically significant and unique chromosomal arm homologies with the *P* values ranging from 0.0193 to 0.0009, as follows: X^{Ag}/X^{Dm}, $2R^{Ag}/3R^{Dm}$, $2L^{Ag}/3L^{Dm}$, $3R^{Ag}/2L^{Dm}$, and $3L^{Ag}/2R^{Dm}$. Except for these, no other pairs even approach statistical significance as homologs. However, the dot chromosome *4* of *Drosophila* does not exist in the mosquito, and the single known *Anopheles* homolog of a chromosome *4* gene is found on the mosquito *X* chromosome. In

Table 2. Cytological Mapping of Putative *Anopheles gambiae* orthologs of Genes in the *Drosophila melanogaster Adh* Region of Chromosome *2L*

D. melanogaster			% Identical amino acids	*A. gambiae*		
Cytology	Mol. map	Gene		Sequence	Accession no.	Cytology
34B11-C1	44	*B4*	46–76	31F12-sp6	AL156399	*3R:* 29A
34C3-4	130	*BG:BACR48E02.4*	79	4A3B-AAC-G-09-R	AJ284828	*3R:* 31C(1)
34D1	294	*BG:DS00797.7*	80	4A3A-AAK-H-12-F	AJ280101	*3R:* 32B(2)
34D1-3	306	*adat*	48–53	05M01-sp6	AL142526	*3R:* 33C(3)
34D1-3	316	*b*	47–63	12K15-sp6	AL146601	*3R:* 32B(2)
34D2-4	324	*Spo2*	88	4A3A-ABA-B-10-R	AJ282629	*3R:* 32B(2)
34D3-4	328	*Rpll33*	84	4A3B-AAC-H-08-R	AJ284838	*3L:* 38A
34E2-4	508	*bgm*	55–76	31H08-sp6	AL156466	*3R:* 30E
34F1-2	728	*smi35A*	82	26G10-t7	AL153535	*3R:* 29B
34F5-A1	848	*Rab14*	66–95	32D17-sp6	AL156885	*3R:* 30D
35A1	853	*l(2)35Aa*	57	4A3A-AAN-C-07-R	AJ281979	*2R:* 19B
35A1	857	*spel1*	60–75	25H11-sp6	AL153035	*3R:* 37D
35B8	1433	*BG:SD01219.3*	65	23G11-sp6	AL151883	*3R:* 35B
35B9	1498	*BG:DS00929.3*	88	4A3B-AAE-D-12-R	AJ284976	*3L:* 38A
35C1	1548	*ck*	46–60	27M17-sp6	AL154364	*3R:* 30C
35CD1-2	1599	*BG:DS04929.3*	51–85	02O08-sp6	AL140935	*3R:* 36C
35D3	1975	*lace*	58–79	11J20-t7	AL146045	*3R:* 29C(4)
35D4	2146	*BG:DS07108.2*	58–72	16K02-sp6	AL148101	*3R:* 34D
35D4	2185	*CycE*	54–70	24K22-sp6	AL152578	*3R:* 30A
35E1-2	2411	*beat-B*	49–77	03I12-t7	AL141289	*3R:* 33C(3)
35F1	2500	*beat*	62–69	03I12-sp6	AL141288	*3R:* 33C(3)
35F6-8	2739	*BG:DS02740.4*	55	4A3A-ABC-D-11-F	AJ281189	*3R:* 29C(4)
35F7-8	2746	*Cyp303a1*	51	09O07-t7	AL145116	*3R:* 36D-37A
35F7-8	2750	*heix*	70	4A3A-ABA-E-06-F	AJ281029	*2R:* 17A
35F7-8	2757	*CG5861*	78	4A3B-AAX-D-01-F	AJ284498	*3R:* 31C(1)
35F7-8	2759	*Sed5*	83	4A3B-AAF-G-04-R	AJ285095	*3R:* 31C(1)
35F7-8	2761	*cni*	82	4A3A-AAY-B-08-R	AJ282455	*2R:* 9A
35F8-9	2776	*cact*	51	4A3A-AAT-A-11-R	AJ282359	*3R:* 29C(4)
35F8-9	2780	*l(2)35Fe*	68	4A3A-AAS-G-11-R	AJ282339	*3R:* 29C(4)
35F9-11	2807	*BG:DS09218.4*	63–78	27H11-t7	AL154163	*2R:* 9C
35F10-11	2816	*BG:DS09218.5*	51–58	32O04-t7	AL157296	*2L:* 21F

Legend as in Table 1. Underlined *Drosophila* genes show transposition to chromosomes of *A. gambiae* other than *3R*. The superscripts 1–4 mark clones that are co-localized within subdivision 31C, 32B, 33C, and 29C, respectively (see Results).

this second experiment, as much as in the first, the relative order of orthologous genes within the corresponding chromosomal arms of the two species appeared to be scrambled. Again, however, some residual microsynteny was detected (see below). For an additional statistical analysis of the same data, we took as a starting point the chromosomal distribution of the *Drosophila* orthologs and compared the observed and expected distributions of *Anopheles* genes; this inverse comparison corresponds to that of the first experiment. As shown in Table 6, the inverse *P* values are all significant, con-

Table 3. Chromosomal Distribution of *A. gambiae* Orthologs of *D. melanogaster* Genes

Drosophila regions, no. of observed and expected genes, and P values of binomial test	*A. gambiae* polytene chromosome arms				
	X	*2R*	*2L*	*3R*	*3L*
X-tip, divisions 1–3 (19 genes)					
Observed genes	3	5	2	4	5
Expected genes	2.27	5.58	4.13	3.82	3.20
P value of binomial test	0.719	0.81	0.283	1	0.35
Adh region of *2L* (31 genes)					
Observed genes	0	4	1	24	2
Expected genes	3.71	9.10	6.74	6.23	5.22
P value of binomial test	**(0.028)**	**(0.048)**	**(0.0079)**	**9.939 e^{-12}**	0.15

The binomial test was applied as follows. For each of the two *D.melanogaster* regions analyzed, the number of putative orthologs observed in each *A. gambiae* polytene chromosome arm was listed. The corresponding expected number of orthologs was calculated according to the number of lettered subdivisions per *Anopheles* chromosomal arm (see Results), and the P value was determined by the binomial test. Statistically significant results are shown in bold, within or without parantheses depending on whether genes are underrepresented or overrepresented in that arm, respectively. Note that *Adh* region orthologs are significantly overrepresented in the mosquito *3R* arm and underrepresented in *X, 2R,* and *2L*.

vincingly confirming the chromosomal arm homologies established from Table 5.

Local Synteny of *Adh* Region Orthologs

As noted above, many genes are scrambled within the respective homologous chromosomal arms. However, a careful analysis of gene order between genes of the *Adh* region in the *D. melanogaster 2L* and their orthologs in the *A. gambiae 3R* gave a clear indication that a significant proportion, ~30%, remain locally clustered with the same neighboring gene. This local synteny may also be called microsynteny, in that it apparently only entails two or three genes at a time. The patterns of both gene scrambling and microsynteny are best displayed graphically, as in Figure 1. It should be noted that because of the availability of the genome sequence, the *Drosophila Adh* region genes are placed on both cytogenetic and DNA sequence scales; their orthologs in *Anopheles* can only be placed on the cytogenetic scale for now.

Of the 31 recognized mosquito orthologs of *Adh* region genes, 24 map to the *Anopheles 3R* chromosome, and 13 of these are found clustered in just four subdivisions, forming four cytogenetic clusters that are at least partially microsyntenic. In contrast, the remaining 11 mosquito orthologs are scattered individually amongst the remaining 33 chromosomal subdivisions of the *Anopheles 3R* chromosome.

The two distal-most mosquito cytogenetic clusters, on divisions 29C and 31C, are both derived from a tight cluster of 27 *Drosophila* genes that are located within ~150 kb at cytogenetic location 35F6-11 (Ashburner et al. 1999). Of these 27 genes, 10 have known mosquito orthologs, and seven of these map to the mosquito chromosome arm *3R*; five are microsyntenic. The latter include two adjacent genes (the *CG5861* and *Sed5* orthologs) that map to the 31C cytogenetic cluster. The 29C cytogenetic cluster includes two adjacent genes (the *cact* and *l(2)35Fe* orthologs) plus one outlier (the *BG:DS02740.4* ortholog). Each of these clusters additionally encompasses one ortholog of a distant *Adh* region gene (*BG:BACR48E04.2* and *lace*, respectively).

Similarly, the mosquito 32B cytogenetic cluster includes three *Anopheles* orthologs of genes *BG:DS00797.7*, *b*, and *Sop2* that in *Drosophila* are part of an 16-gene cluster located within ~65 kb at 34D1-4 (Ashburner et al. 1999). Two orthologs of other genes from the same cluster, *adat* and *RpII33*, are known in the mosquito but do not map at 32B; the orthologs of the 11 remaining genes in the 34D1-4 *Drosophila* cluster are as yet unknown.

Finally, the fourth mosquito cytogenetic cluster at 33C includes the orthologs of *adat* from the *Drosophila* 34D1-4 region (see above) plus two genes, *beat-B* and *beat* from the *Drosophila* 35E1-F1 region. In *Drosophila*, the latter two genes are paralogs with the same exon-intron structure and show 53% identity at the amino acid level. They are separated by ~100 kb, a region that encompasses three other genes, *BG:DS07486.2*, *beat-C* (also a paralog of *beat-B* and *beat*), and *Bic-C* (Ashburner et al. 1999); the orthologs of these three genes are not yet known in the mosquito. Interestingly, the orthologs of *beat-B* and *beat* are from the STSs at the two ends of the same mosquito BAC clone (03I12), and thus are also separated by ~120 kb. It would be interesting to sequence this clone and thus discover whether the mosquito orthologs of the *BG:DS07486.2*, *beat-C*, and *Bic-C* genes are also located in this interval.

DISCUSSION

The analysis presented here was made possible by the availability of the essentially complete sequence of the *D. melanogaster* genome (Adams et al. 2000) and is a clear example of comparative genomic research. It illustrates how full genomic information from a model species can help provide considerable insight into the genomic structure of even a rather distantly related and little-studied orphan organism, when combined with bioinformatics analysis of partial sequence information and physical mapping of clones representing ESTs and STSs. It should be recalled that the fruit fly and the mosquito are estimated to have diverged ~250 Myr (Yeates and Wiegmann, 1999). The study addresses three main questions.

The question of sequence divergence between orthologous genes of *Drosophila* and *Anopheles* relates to our ability to detect such genes. We have used rather stringent similarity criteria to accept genes as orthologs, and thus we expect that our reported collection of orthologs includes few if any false positives and excludes some widely divergent orthologs. Consistent with these expectations, the STS resource of BAC ends represents ~7% of the estimated euchromatic DNA of *A. gambiae* and yielded 26 (5.5%) orthologs of the 475 *Drosophila* genes present at the tip of the *X* and the *Adh* region of *Drosophila*. The EST resource includes 2380 cDNA clone clusters, but it is difficult to say how many actual genes are represented, because of the possibility of undetected overlaps. The EST resource yielded 24 of the orthologs or 5.1% of the genes in the *Adh* region and the tip of the *X* in *Drosophila*. Accepting the orthology of all genes shown in Tables 1, 2, and 4, we note that the detected orthologous exons show a range of 26% to 97% sequence identity at the amino acid level, with an average of 61.6% identity. If we consider only the most similar available exons, the orthologous genes have 31% to 97% local sequence identity, or 65.4% on average. This indicates that in most future cases, it should be possible to recognize orthologous genes in the two species using our criteria or to clone them by sequence homology.

The second question concerns the gross homology of chromosomes between the fruit fly and the mosquito. It is striking that both species have two major metacentric autosomes as well as an apparently telocentric *X* chromosome in the euchromatic polytene genome (five chromosomal arms in total). Only the very minor chromosome *4* (~1% of the genome in *Drosophila*) is absent from *Anopheles*. Taken together, our data show unequivocally that the five *A. gambiae* chromosome arms can be assigned a distinct homolog in the chromosomal complement of the fruit fly, and vice versa.

From Table 6, it can be seen that in different chromosomal arms, between 27 and 59% of the genes have undergone interchromosomal translocation to nonhomologous arms since the last common ancestor of *D. melanogaster* and *A. gambiae*. The extent to which translocations occur varies for different arms (Table 6) and also apparently for different chromosomal regions. Comparison between Tables 3 and 5 indicates that translocations have occurred more frequently for genes that are now at the *X*-tip of *Drosophila* than for the *X* as a whole; whereas translocations have occurred less frequently for the *Adh* region than for that arm as a whole. Overall, using Muller's definition of the chromosomal elements of *Drosophila* (Muller 1940), the *A. gambiae* chromosome arms *X*, *2R*, *2L*, *3R*, and *3L* are homologous to the *Drosophila* elements *A*, *E*, *D*, *B*, and *C*, respectively. Interestingly, in both species the arrangement of paired elements is the same (A, B + C, D + E). The *A. gambiae* chromosomes *2* and *3* are homologous to the *D. melanogaster* chromosomes *3* and *2* respectively.

A dense collection of DNA markers from *Aedes aegypti*

Table 4. Cytological Position of Randomly Selected *A. gambiae* Sequences and Their *D. melanogaster* Orthologs With Unique Localization in the Genome

A. gambiae			% Identical amino acids	*D. melanogaster*		
Cytology	Sequence/gene	Accession no.		Gene	Cytology	Reference
X: 1C	*cec*	AF200686	45–52	*CecA1*	*3R:* 99E3-4	Vizioll et al. 2000
				CecA2	*3R:* 99E3-4	
				CecB	*3R:* 99E3-4	
				CecC	*3R:* 99E3-4	
X: 1D	*06M18-sp6*	AL143155	58	*sdk*	*X:* 1B12	This study
X: 2A	*w*	U29484	54–72	*w*	*X:* 3C2	Besansky et al. 1995
X: 2C	*21P07-t7*	AL151067	45	*EG: 171E4.4*	*X:* 2B6-7	This study
X: 3D	*20I10-t7*	AL150270	47–77	*CG1472*	*2R:* 46C1-4	This study
X: 3D-4A	*pKM42*	U50467	92	*Cycbeta100B*	*3R:* 100B5-8	della Torre et al. 1996
X: 4A	*27D16-sp6*	AL154010	87	*Adar*	*X:* 2B9	This study
X: 4B	*10A05-sp6*	AL145180	43–75	*CG11428*	*X:* 12A7-8	This study
X: 4C	*10F12-sp6*	AL145357	36–80	*ey*	*4:* 102D5	This study
X: 5B	*10I08-t7*	AL145443	87–97	*sqh*	*X:* 5D6	This study
X: 5D	*10A22-t7*	AL145201	36	*CG15054*	*X:* 17B1	This study
X: 5D	*10L08-sp6*	AL145544	58	*CG6803*	*3R:* 88E5	This study
X: 6	*17N20-sp6*	AL148805	32–46	*CG8772*	*2R:* 49B8-9	This study
2R: 8C	*17I13-sp6*	AL148612	44–61	*CG7593*	*3R:* 99B11-C1	This study
2R: 11B	*30E10-t7*	AL155810	74	*ESTS:149B10S*	*3L:* 66A8-10	This study
2R: 12A-B	*pKM50*	U50471	66	*CG7808*	*3R:* 99C6-7	della Torre et al. 1996
2R: 12C	*17M11-t7*	AL148757	39	*mod(r)*	*X:* 1B9-10	This study
2R: 14D	*10G21-t7*	AL145401	88	*CG9492*	*3L:* 70E3	This study
2R: 14E	*pKM52*	U50472	57	*CG6783*	*3R:* 86E13-15	della Torre et al. 1996
2R: 16A	*10D21-t7*	AL145303	70–73	*slo*	*3R:* 96A18-20	This study
2R: 19B-E	*Scr*	AF080564	35–54	*Scr*	*3R:* 84A5-B1	Devenport et al. 2000
2R: 19D	*17P11-t7*	AL148868	47	*ninaC*	*2L:* 27F5-6	This study
2R: 19D	*Antp*	AF080565	53	*Antp*	*3R:* 84B2-4	Devenport et al. 2000
2R: 19D	*Ubx*	AF080562	62	*Ubx*	*3R:* 89D6-E2	Devenport et al. 2000
2R: 19E	*abd-A*	AF080566	61–67	*abd-A*	*3R:* 89E3-4	Devenport et al. 2000
2L: 20C	*05B09-t7*	AL142182	47	*EG:BACH7M4.1*	*X:* 2F5	This study
2L; 20D	*10F01-sp6*	AL145340	26–53	*CG10173*	*3L:* 65A7-8	This study
2L: 21C	*02A19-t7*	AL140406	34	*EG:65F1.1*	*X:* 1B4	This study
2L;21F	*32O04-t7*	AL157296	52	*BG:DS09218.5*	*2L:* 35F10	This study
2L: 22C	*21G04-t7*	AL150716	42	*CG18289*	*2R:* 41F3	This study
2L: 22D	*19D01-sp6*	AL149568	55	*CG3790*	*2R:* 49D2-4	This study
2L: 22D	*18E22-t7*	AL149075	66	*CG10483*	*3L:* 64F4-5	This study
2L: 22E	*16A02-t7*	AL147784	31	*AAF49108.1*	*3L:* 76D2-3	This study
2L: 23B	*21I15-t7*	AL150818	37	*RecQ4 gene*	*3L:* 66B11	This study
2L; 23C	*pKM2*	U50477	70	*CG7991*	*3L:* 62B1	della Torre et al. 1996
2L: 23D	*07F24-t7*	AL143530	39	*CG13076*	*3L:* 72D1	This study
2L: 24C	*16L07-sp6*	AL148140	83	*CG12215*	*2R:* 46EF	This study
2L: 26A	*ADP/ATP*	L11617	76–80	*sesB*	*X:* 9E4-7	Beard et al. 1994
				Ant2	*X:* 9E4-7	
2L; 27A	*32J02-sp6*	AL157092	31–85	*Ptpmeg*	*3L:* 61C1	This study
2L: 27A	*27H17-sp6*	AL154170	60	*EG:52C10.1*	*2R:* 54E9	This study
2L: 27D	*pKM134*	U50479	90	*RpS13*	*2L:* 29B2-3	della Torre et al. 1996
2L: 28A	*10F06-t7*	AL145346	48–66	*CG15373*	*X:* 16F3	This study
2L: 28C	*lanB2*	AJ271193	60	*LanB2*	*3L:* 68B10	Viachou et al. 2001
2L: 28D	*22L01-t7*	AL151486	41–55	*CG4324*	*2R:* 60A16-B1	This study
3R: 29B	*pKM3*	U50480	43–72	*Sin3A*	*2R:* 49B2-3	della Torre et al. 1996
3R: 30C	*27M17-sp6*	AL154364	46–60	*ck*	*2L:* 35C1	This study
3R: 31B	*10G09-sp6*	AL145389	84	*CG9313*	*2R:* 57B11-12	This study
3R: 33A	*cc41*	AF002238	50–73	*CG17489*	*2L:* 40B-D	A. Cornel, pers. comm.
3R: 33B	*bd*	AF042732	51–58	*CG10655*	*2L:* 37B11-13	Romans et al. 1999
3R: 33B	*TU37B2*	AF042732	76–96	*CG10470*	*2L:* 37B11-13	Romans et al. 1999
3R: 33B	*Dox*	AF042732	74	*Dox-A2*	*2L:* 37B11-13	Romans et al. 1999
3R: 33B	*Ddc*	AF063021	74–80	*Ddc*	*2L:* 37C1	Romans et al. 1999
3R: 34A	*25P12-sp6*	AL153311	68–90	*Cbp53E*	*2R:* 53E6-11	This study
3R: 34C	*10F08-t7*	AL145350	56	*CG18252*	*2L:* 25C8-9	This study
3R: 34C	*Gambif1*	X95911	44–73	*dl*	*2L:* 36C2-3	Barillas-Mury et al. 1996
3R: 35C	*10B06-t7*	AL145213	36–48	*CG11003*	*3L:* 69E1	This study
3R: 38B	*2BI19-sp6*	AL154783	56	*EG:BACR7A4.5*	*X:* 1C5-D1	This study
3L: 40B	*24C03-t7*	AL152258	62	*CG1698*	*2R:* 46B7	This study
3L: 40B	*17M09-sp6*	AL148752	41	*CG4464*	*3L:* 67A1-3	This study
3L: 40B	*32L10-t7*	AL157191	54	*CG3252*	*X:* 4F5	This study
3L: 41D	*20D20-sp6*	AL150137	41	*CG18289*	*2R:* 41F3	This study

(Table continued on facing page.)

Table 4. (*Continued*)

A. gambiae			% Identical amino acids	*D. melanogaster*		
Cytology	Sequence/gene	Accession no.		Gene	Cytology	Reference
3L: 42A	*defensin*	X93562	56	*Def*	*2R:* 46D7-9	Richman et al. 1996
3L: 42A	*26B23-t7*	AL153383	43	*CG8642*	*2R:* 44D2	This study
3L: 42A	*30G20-sp6*	AL155891	42–88	*CG3186*	*2R:* 60B2-3	This study
3L: 42B	*10D11-t7*	AL145288	28–71	*CG6501*	*2R:* 54D1	This study
3L: 43A	*20B11-t7*	AL150058	41	*snk*	*3R:* 87D11	This study
3L: 45C	*03G02-sp6*	AL141206	67	*CG2103*	*3R:* 63A1-2	This study
3L: 46A	*19L20-t7*	AL149862	44	*CG7169*	*3L:* 78E4-5	This study
3L: 46A	*28J05-sp6*	AL154795	53	*CG8706*	*2R:* 44B3	This study
3L: 46CD	*19N23-sp6*	AL149940	40	*CG14196*	*X:* **18B6-8**	This study

Legend as in Table 1. *A. gambiae* sequence names are as they appear in the corresponding references or database entries.

Table 5. Chromosomal Distribution of Putative *D. melanogaster* Orthologs of *A. gambiae* Genes

A. gambiae polytene chromosome arms, observed and expected genes, and P values of the binomial test	D. melanogaster polytene chromosome arms					
	X	*2L*	*2R*	*3L*	*3R*	*4*
X (13 genes)						
Observed genes	7	0	2	0	3	1
Expected genes	2.39	2.52	2.35	2.67	3.07	0.13
P value of the binomial test	0.0036	0.0508	0.2844	0.0503	0.2547	0.1152
2R (12 genes)						
Observed genes	1	1	0	2	8	0
Expected genes	2.21	2.33	2.17	2.47	2.83	0.12
P value of the binomial test	0.2351	0.2169	0.0913	0.2791	0.0016	0.8864
2L (19 genes)						
Observed genes	4	2	5	8	0	0
Expected genes	3.49	3.68	3.43	3.91	4.49	0.19
P value of the binomial test	0.2101	0.1651	0.1373	0.0193	(0.0060)	0.8262
3R (13 genes)						
Observed genes	1	8	3	1	0	0
Expected genes	2.39	2.52	2.35	2.67	3.07	0.13
P value of the binomial test	0.2089	0.0009	0.2300	0.1692	(0.0298)	0.8775
3L (13 genes)						
Observed genes	2	0	7	2	2	0
Expected genes	2.39	2.52	2.35	2.67	3.07	0.13
P value of the binomial test	0.2822	0.0608	0.0033	0.2624	0.2238	0.8775

Legend as in Table 3, except that expected numbers and *P* values were calculated for *D. melanogaster* chromosome arms. The actual DNA content of the *D. melanogaster* chromosome arms (Adams et al. 2000) was used for the calculations.

Table 6. Chromosomal Distribution of Putative *A. gambiae* Orthologs of *D. melanogaster* Genes

Drosophila chromosome arm	Total observed genes	*A. gambiae* homolog	Orthologs on homolog	Expected orthologs	% of genes on homolog	Expected percentage	Inverse *P* value	*P* value
X	15	*X*	7	1.79	46.7	12.0	0.0008	0.0036
2L	11	*3R*	8	3.52	72.7	29.4	0.0063	0.0016
2R	17	*3L*	7	3.04	41.2	21.7	0.0037	0.0193
3L	13	*2L*	8	2.41	61.5	20.1	0.0005	0.0009
3R	13	*2R*	8	3.03	61.5	16.9	0.0168	0.0033

The binomial test was applied as follows. For each *D. melanogaster* chromosome arm, the number of observed orthologs of *A. gambiae* genes was tallied (numbers are from the vertical columns in Table 5). The *Anopheles* arm (column 3) is from Table 5. Expected orthologs and percentages for each *A. gambiae* chromosome arm were calculated on the basis of the length these arms (see Results). The *P* value determined is indicated as inverse *P* value whereas the corresponding *P* value in column 9 is taken from Table 5. Column 6 is calculated from columns 2 and 4.

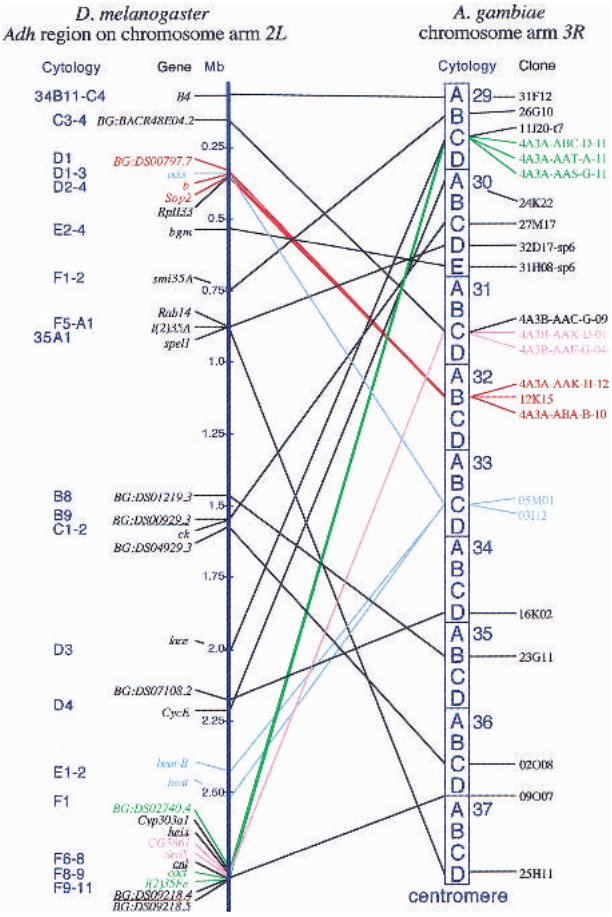

Figure 1 Distribution of *Drosophila melanogaster* genes of the *Adh* region and their *A. gambiae* orthologs on chromosome *3R*. The left vertical dark blue bar represents the *Drosophila Adh* region, listing the names of the genes, their cytological location, and their positions on the molecular map (in megabases) according to Ashburner et al. (1999). Underlined are the genes for which corresponding *Anopheles* orthologs are not localized on chromosome *3R* (see Table 2). The dark blue vertical boxes on the right represent the chromosome arm *3R* of *Anopheles gambiae*, with numbered divisions and lettered subdivisions. Clones that start with 4A3A or 4A3B are cDNAs, and the rest are STSs from chromosomal BACs. The solid lines point to the cytogenetic locations of these clones. The positions of orthologs are indicated by connecting lines between the *Drosophila Adh* region and *Anopheles* chromosome *3R*. Colored lines connect *A. gambiae* cytogenetic gene clusters (green, 29C; pink, 31C; red, 32B; and light blue, 33C) and their respective fruit fly orthologs, also shown in the same color. The two chromosomes are not drawn to scale.

(restriction fragment length polymorphisms) was used by Severson et al. (1994) to evaluate genetic diversity and synteny among aedine mosquitoes and *A. gambiae*; however, synteny with *Drosophila* was not examined. In a valuable earlier study, Matthews and Munsterman (1994) used 29 enzyme loci to study linkage conservation amongst lower diptera (13 species of mosquitoes, not including *A. gambiae*) and higher diptera (*D. melanogaster*). In different mosquito species five to 19 loci were mapped. The investigators concluded that mosquito chromosomes are modified by paracentric inversions and interchromosomal translocations. They also noted that several amall groups of two to four enzyme loci have been conserved in linkage in both mosquitoes and the fruit fly, "although most traces of homology between the two dipteran linkages have disappeared." In the present study, a much larger number of orthologous gene sequences, mapped by in situ hybridization to polytene chromosomes, permitted firmer conclusions: pairwise identification of homologous polytene chromosomes in *A. gambiae* and *D. melanogaster* and quantification of the extent of nonhomologous arm translocations between the fruit fly and the mosquito.

The third and final issue is the distribution of genes within broadly homologous chromosomal arms, and the length of locally syntenic regions conserved between these two dipteran species. Previous studies have compared different distant *Drosophila* species to one another by in situ hybridization of gene-specific probes or larger genomic fragments usually derived from *D. melanogaster*. These studies included a cross-comparison of *D. melanogaster* (as a reference species) and several other species, including *D. obscura*, *D. madeirensis*, *D. virilis*, *D. repleta*, *D. buzzattii*, and *D. hydei* (Loukas and Kafatos 1988; Whiting et al. 1989; Segarra and Aguade 1992; Lozovskaya et al. 1993; Segarra et al. 1995; Nurminsky et al. 1996; Vieira et al. 1997; Ranz et al. 1999, 2000; Gonzales et al. 2000). These *Drosophila* species were separated from *D. melanogaster* 25 to 60 Myr (Beverley and Wilson 1984; Russo et al. 1995). The homologous chromosome arms are usually easily identified by their gene content, but the relative order and distances of the genes are considerably reshuffled in the different species. Observed sizes of chromosomal fragments conserved between species range from 20 to 600 kb (Ranz et al. 1999, 2000; Gonzales et al. 2000), although one cannot exclude undetected small rearrangements within the larger fragments. Calculations that take into consideration the number of inversion breakpoints in several selected genomic regions and the divergence time between species indicate that the frequency of breakpoints occurring in the genus *Drosophila* may be as high as 0.05 to 0.08 per megabase of sequence per million years (Ranz et al. 2000). The lower estimate of this frequency would imply that in the genome of *A. gambiae*, calculated to have a size of ~260 Mb, we may expect microsyntenic regions conserved relative to *Drosophila* to have an average DNA length of 50 to 80 kb of DNA. This is in striking contrast to the frequency of breakpoints computed for a mouse-human comparison (divergence time ~112 Myr; Kumar and Hedges 1998), which is about two orders of magnitude lower (Ranz et al. 2000). The sizes of conserved segments in these two species are estimated to be 24 kb to 90.5 Mb in length, averaging 15.6 Mb (Lander et al. 2001). We have detected microsyntenic blocks of two to three genes each by cytological co-localization of these genes in the same *Anopheles* polytene chromosome lettered subdivision. It must be stressed that this evidence neither establishes nor excludes that the genes are located next to each other in the genome. As yet, we have a DNA distance estimate for only one microsyntenic pair, *beat* and *beat-B*: 100 kb in *Drosophila* and a BAC length (~120 kb average) in *Anopheles*. However, our evidence strongly argues that locally syntenic regions between the mosquito and the fruit fly are not long.

Microsynteny between *Anopheles* and *Drosophila* was also detected by Romans et al. (1999), who isolated and characterized a 4.2-kb genomic fragment containing the *Anopheles Bb*, *TU37B2*, and *Dox-A2* genes. These are orthologs of the *Drosophila* genes *CG10655*, *CG10470*, and *Dox-A2*, respectively, all located within a 4.5-kb genome region in the fruit fly (Adams et al. 2000). Analysis of the molecular organization of two mosquito chromosomal regions indicated the occurrence of

several rearrangements that changed both the position and orientation of *Bb* and *TU37B2* in comparison to their *Drosophila* orthologs. We have confirmed these results and found that the syntenic area does not extend much beyond the genes mentioned (data not shown).

Taking these results together, the degree of observed microsynteny between *Drosophila* and *Anopheles* is not high and may be even lower than predicted. The degree of microsynteny is an important parameter for future efforts to use the *D. melanogaster* gene order to identify mosquito orthologs definitively, leading to functional hypotheses and to assays of these proposed functions in the genetically tractable fruit fly. Firm elucidation of the degree of microsynteny will be one of the major benefits expected from full sequencing of the *A. gambiae* genome, which is expected to begin shortly.

METHODS

Source of Sequence Data

Amino acid sequences of the genes in divisions 1–3 of chromosome *X* of *D. melanogaster* can be obtained by anonymous FTP from ftp://ftp.ebi.ac.uk/pub/databases/edgp/misc/ashburner/EG_genes.991229.pep.fa.gz (Benos et al. 2000, 2001), whereas amino acid sequences of the genes identified in the *Adh* region are found in http://www.fruit fly.org/sequences/aa_Adh.dros (Ashburner et al. 1999). Amino acid sequences of all genes identified through the whole genome sequence (release 1.0) are available at http://www.fruitfly.org/sequence/dlMfasta.html (Adams et al. 2000). For *A. gambiae*, nucleotide sequences of ESTs from immune-competent cell line cDNA libraries (Dimopoulos et al. 2000) and STSs from the BAC genomic library (C. Roth and F.C. Collins, pers. comm.), as well as other mosquito sequences with known cytological location, can be BLAST-searched at AnoDB, the *Anopheles* database (http://konops.anodb.gr/cgi-bin/blast2.pl).

Computational Methods and Analysis of Results

For similarity searches, a locally installed WU-BLAST, version 2.0a, suite of programs (Altschul et al. 1990; W. Gish, unpubl.) was used. *D. melanogaster* amino acid sequences of genes from selected regions were compared to *A. gambiae* STS and EST databases using TBLASTN with standard default parameters. STS and EST sequences showing similarity with a high score of >40, a probability *P*(N) of $< e^{-10}$, and a percentage of identical amino acids >30, were selected and checked as best bidirectional hits after confirming the hit using BLASTX with standard default parameters against a database of 14,080 amino acid sequences of known and predicted *Drosophila* genes (release 1.0, http://www.fruitfly.org/sequence/dlMfasta.html#rel1; Adams et al. 2000). Only STSs and ESTs that passed these criteria were selected, and their alignments were further verified using the available exon-intron structure of the corresponding *D. melanogaster* genes, as shown in the National Center for Biotechnology Information (NCBI) version of the *D. melanogaster* database (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/7227.html). The same BLASTX search criteria were also used in the reciprocal experiment, comparing *A. gambiae* nucleotide sequences of known cytological location to protein encoding genes in *D. melanogaster* genes. The names and cytological locations of *D. melanogaster* genes were taken from FlyBase (http://flybase.bio.indiana.edu; The FlyBase Consortium 1999); additional information and literature references on genes can also be found there.

In Situ Hybridization to *A. gambiae* Polytene Chromosomes

BAC and cDNA clones were hybridized to preparations of *A. gambiae* polytene chromosomes essentially as described in Kumar and Collins (1994). The hybridization signals were localized according to the cytological map of M. Coluzzi, A. Sabbatini, M.A. Di Deco, and V. Petrarca (unpubl., accessible at http://www.anodb.gr/AnoDB/Cytomap/).

ACKNOWLEDGMENTS

We are indebted to Drs. Frank Collins and Charles Roth for submitting their data to public databases before publication and to Drs. Mario Coluzzi and Igor Zhimulev for their support of the participation of their laboratories in the in situ hybridization analysis of *A. gambiae* sequences. We would also like to acknowledge the invaluable assistance of Drs. Poulikos Prastakos and Yannis Kamarianakis in the statistical analysis. This research was supported by grants from the UNDP/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases (TDR), the INCO programme of the European Union, the National Institutes of Health, the Hellenic Secretariat General for Research and Technology, and the John D. and Catherine T. McArthur Foundation.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Ashburner, M., Misra, S., Roote, J., Lewis, S.E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., et al. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: The *Adh* region. *Genetics* **153:** 179–219.

Barillas-Mury, C., Charlesworth, A., Gross, I., Richman, A., Hoffmann, J.A., and Kafatos, F.C. 1996. Immune factor Gambif1, a new rel family member from the human malaria vector, *Anopheles gambiae*. *EMBO J.* **15:** 4691–4701.

Beard, C.B., Crews-Oyen A.E., Kumar V.K., and Collins F.H. 1994. A cDNA encoding an ADP/ATP carrier from the mosquito *Anopheles gambiae*. *Insect Mol. Biol.* **3:** 35–40.

Benos, P.V., Gatt, M.K., Ashburner, M., Murphy, L., Harris, D., Barrell, B., Ferraz, C., Vidal, S., Brun, C., Demailles, J., et al. 2000. From sequence to chromosome: The tip of the *X* chromosome of *D. melanogaster*. *Science* **287:** 2220–2222.

Benos, P.V., Gatt, M.K., Murphy, L., Harris, D., Barrell, Ferraz, C., Vidal, S., Brun, C., Demailles, J., Cadieu, E. et al. 2001. From first base: The sequence of the tip of the *X* chromosome of *D. melanogaster*, a comparison of two sequencing strategies. *Genome Res.* **11:** 710–730.

Besansky, N.J., Bedell, J.A., Benedict, M.Q., Mukabayire, O., Hilfiker, D., and Collins, F.H. 1995 Cloning and characterization of the white gene from *Anopheles gambiae*. *Insect Mol. Biol.* **4:** 217–231.

Beverley, S.M. and Wilson, A.C. 1984. Molecular evolution in *Drosophila* and higher diptera, II: A time scale for fly evolution. *J. Mol. Evol.* **21:** 1–13.

Catteruccia, F., Nolan, T., Loukeris, T.G., Blass, C., Savakis, C., Kafatos, F.C., and Crisanti, A. 2000. Stable germline transformation of the malaria mosquito *Anopheles stephensi*. *Nature* **405:** 959–962.

Coates, C.J., Jasinskiene, N., Miyashiro, L., and James, A.A. 1998. Mariner transposition and transformation of the yellow fever mosquito, *Aedes aegypti*. *Proc. Natl. Acad. Sci.* **95:** 3748–3751.

Collins, F.C., Zheng, L., Paskewitz, S.M., and Kafatos, F.C. 1997. Progress in map-based cloning of the *Anopheles gambiae* genes responsible for the encapsulation of malarial parasites. *Ann. Trop. Med. Parasitol.* **91:** 517–521.

Coluzzi, M., Di Deco, M.A., and Petrarca, V. 1985. Chromosomal inversion intergradation and incipient speciation in *Anopheles gambiae*. *Biol. Zool.* **52:** 45–63.

della Torre A., Favia G., Mariotti, G., Coluzzi, M., and Mathiopoulos, K.D. 1996 Physical map of the malaria vector *Anopheles gambiae*. *Genetics* **143:** 1307–1311.

Devenport, M., Blass, C., and Eggleston, P. 2000. Characterization of

the *Hox* gene cluster in the malaria vector mosquito, *Anopheles gambiae*. *Evol. Devel.* **2:** 326–339.
Dimopoulos, G., Richman, A., della Torre, A., Kafatos, F.C., and Louis, C. 1996a. Identification and characterization of differentially expressed cDNAs of the vector mosquito, *Anopheles gambiae*. *Proc. Natl. Acad. Sci.* **93:** 13066–13071.
Dimopoulos, G., Zheng, L.B., Kumar, V., della Torre, A., Kafatos, F.C., and Louis, C. 1996b. Integrated genetic map of *Anopheles gambiae*: Use of RAPD polymorphisms for genetic, cytogenetic and STS landmarks. *Genetics* **143:** 953–960.
Dimopoulos, G., Casavant, T.L., Chang, S., Scheetz, T., Roberts, C., Donohue, M., Schultz, J., Benes, V., Bork, P., Ansorge, W., et al. 2000. *Anopheles gambiae* pilot gene discovery project: Identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. *Proc. Natl. Acad. Sci.* **97:** 6619–6624.
The FlyBase Consortium 1999. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucl. Acid Res.* **27:** 85–88.
Gonzales, J., Bertran, E., Ashburner, M., and Ruiz, A. 2000. Molecular organisation of the *Drosophila melanogaster Adh* chromosomal region in *D. repleta* and *D. buzzatii*, two distantly related species of the *Drosophila* subgenus. *Chrom. Res.* **8:** 375–385.
James, A.A., Beerntsen, B.T., del Capurro, M., Coates, C.J., Coleman, J., Jasinskiene, N., and Krettli, A.U. 1999. Controlling malaria transmission with genetically-engineered, *Plasmodium*-resistant mosquitoes: Milestones in a model system. *Parassitologia* **41:** 461–471.
Jasinskiene, N., Coates, C.J., Benedict, M.Q., Cornel, A.J., Rafferty, C.S., James, A.A., and Collins, F.H. 1998. Stable transformation of the yellow fever mosquito, *Aedes aegypti*, with the Hermes element from the housefly. *Proc. Natl. Acad. Sci.* **95:** 3743–3747.
Kamau, L., Mukabana, W.R., Hawley, W.A., Lehmann, T., Irungu, L.W., Orago, A.A., and Collins, F.H. 1999. Analysis of genetic variability in *Anopheles arabiensis* and *Anopheles gambiae* using microsatellite loci. *Insect Mol. Biol.* **8:** 287–297.
Kumar, V. and Collins F.H. 1994. A technique for nucleic acid in situ hybridization to polytene chromosomes of mosquitoes in the *Anopheles gambiae* complex. *Insect Mol. Biol.* **3:** 41–47.
Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh,W., et al. 2001. Initial sequencing and analysis of human genome. *Nature* **409:** 860–921.
Lanzaro, G.C., Toure, Y.T., Carnahan, J., Zheng, L., Dolo, G., Traore, S., Petrarca, V., Vernick, K.D., and Taylor, C.E. 1998. Complexities in the genetic structure of *Anopheles gambiae* populations in West Africa as revealed by microsatellite DNA analysis. *Proc. Natl. Acad. Sci.* **95:** 14260–14265.
Louis, C. 1999. Malaria entomology: Can genomics help? *Parassitologia*. **41:** 489–492.
Loukas, M. and Kafatos, F.C. 1988. Chromosomal locations of actin genes are conserved between the melanogaster and obscura groups of *Drosophila*. *Genetica* **76:** 33–41.
Lozovskaya, E.I., Petrov, D.A., and Hartl, D.L. 1993. A combined molecular and cytogenetic approach to *Drosophila* evolution using large-fragment DNA cloning. *Chromosoma* **102:** 253–266.
Matthews, T.C. and Munsterman, L.E. 1994. Chromosomal repatterning and linkage group conservation in mosquito karyotypic evolution. *Evolution* **48:** 146–154.
Muller, H.J. 1940. Bearings of the *Drosophila* work on systematics. In *The new systematics* (ed. J. Huxley), pp. 185–268. Clarendon Press, Oxford.
Nurminsky, D.I., Moriyama, E.N., Lozovskaya, E.R., and Hartl, D.L. 1996. Molecular phylogeny and genome evolution in the *Drosophila virilis* species group: Duplications of the *Alcohol Dehydrogenase* gene. *Mol. Biol. Evol.* **13:** 132–149.
O'Brien S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N., Womack, J.E., and Marshal Graves, J.A. 1999. The promise of comparative genomic in mammals. *Science* **286:** 458–481.
Ranson, H., Jensen, B., Wang, X., Prapanthadara, L., Hemingway, J., and Collins, F.H. 2000. Genetic mapping of two loci affecting DDT resistance in the malaria vector *Anopheles gambiae*. *Insect Mol. Biol.* **9:** 499–507.
Ranz, J.M., Caceres, M., and Ruiz, A. 1999. Comparative mapping of cosmid and gene clones from a 1.6 Mb chromosomal region of *Drosophila melanogaster* in three species of the distantly related subgenus *Drosophila*. *Chromosoma* **108:** 32–43.
Ranz, J.M., Casalis, F., and Ruiz, A. 2000. How malleable is the eucaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11:** 230–239.
Richman, A.M., Bulet, P., Hetru, C., Barillas-Mury, C., Hoffmann, J.A., and Kafatos, F.C. 1996. Inducible immune factors of the vector mosquito *Anopheles gambiae*: Biochemical purification of a defensin antibacterial peptide and molecular cloning of preprodefensin cDNA. *Insect Mol. Biol.* **5:** 203–210.
Romans, P., Black, W.C., Sakai, R.K., and Gwadz, R.W. 1999. Linkage of gene causing malaria refractoriness to the Diphenol oxidase-2 on chromosome 3 of *Anopheles gambiae*. *Am. J. Trop. Med. Hyg.* **60:** 22–29.
Russo, C.A.M., Takezaki, N., and Nei, M. 1995. Molecular phylogeny and divergence time of Drosophilid species. *Mol. Biol. Evol.* **12:** 391–404.
Segarra, C. and Aguade, M. 1992. Molecular orgenization of the *X* chromosome in different species of *obscura* group of *Drosophila*. *Genetics* **130:** 513–521.
Segarra, C., Lozovskaya, E.R., Ribo, G., Aguade, M., and Hartl., D.L. 1995. P1 clones from *Drosophila melanogaster* as markers to study the chromosomal evolution of Muller's *A* element in two species of *obscura* group of *Drosophila*. *Chromosoma* **104:** 129–136.
Severson, D.W., Mori, A., Zhang, Y., and Christensen, B.M. 1994. The suitability of restriction fragment length polymorphism markers for evaluating genetic diversity among and synteny between mosquito species. *Am. J. Trop. Med. Hyg.* **50:** 425–432.
Sinden, R.E. 1999. *Plasmodium* differentiation in the mosquito. *Parassitologia* **41:** 139–148.
Terryn, N., Rouze, P., and Van Montagu, M. 1999. Plant genomics. *FEBS Lett.* **452:** 3–6.
Touré, Y.T., Petrarca, V., Traoré, S.F., Coulibaly, A., Maiga, H.M., Sankaré, O., Sow, M., Di Deco, M.A., and Coluzzi, M. 1998. The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia* **40:** 477–511.
Vieira, J., Vieira, C.P., Hartl, D.L., and Lozovskaya, E.R. 1997. A framework physical map of *Drosophila virilis* based on P1 clones: Applications in genome evolution. *Chromosoma* **106:** 99–107.
Vizioli, J., Bulet, P., Charlet, M., Lowenberger, C., Blass, C., Mueller, H.-M., Dimopoulos, G., Hoffmann, J., Kafatos, F.C., and Richman,A. 2000. Cloning and analysis of a cecropin gene from the malaria vector mosquito, *Anopheles gambiae*. *Insect Mol. Biol.* **9:** 75–84.
Vlachou, D., Lycett, G., Sidén-Kiamos, I., Blass, C., Sinden, R.E. and Louis, C. 2001. *Anopheles gambiae* laminin interacts with the P25 surface protein of *Plasmodium berghei* ookinetes. *Mol. Biochem. Parasit.* **112:** 229–237.
Wang, R., Kafatos, F.C., and Zheng, L.B. 1999. Microsatellite markers and genotyping procedures for *Anopheles gambiae*. *Parasitol. Today* **15:** 33–37.
Wang, R., Zheng, L., Touré, Y.T., Kafatos, F.C., and Dandekar, T. 2001. When genetic distance matters: Measuring genetic differentiation at microsatellite loci in whole genome scans of recent and incipient mosquito species. *Proc. Natl. Acad. Sci.* **98:** 10769–10774.
Whiting, J.H., Pliley, M.D., Farmer, J.L., and Jeffery, D.E. 1989. In situ hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis*. *Genetics* **122:** 99–109.
Yeates, D.K. and Wiegmann, B.M. 1999. Congruence and controversy: Towards the higher-level phylogeny of Diptera. *Ann. Rev. Entomol.* **44:** 397–428.
Zheng, L.B., Saunders, R.D., Fortini, D., della Torre, A., Coluzzi, M., Glover, D.M., and Kafatos, F.C. 1991. Low-resolution genome map of the malaria mosquito *Anopheles gambiae*. *Proc. Natl. Acad. Sci.* **88:** 11187–11191.
Zheng, L.B., Collins, F.H., Kumar, V., and Kafatos, F.C. 1993. A detailed genetic map for the *X* chromosome of the malaria vector, *Anopheles gambiae*. *Science* **261:** 605–608.
Zheng, L.B., Benedict, M.Q., Cornel, AJ., Collins, F.H., and Kafatos, F.C. 1996. An integrated genetic map of the African human malaria vector mosquito, *Anopheles gambiae*. *Genetics* **143:** 941–952.
Zheng, L., Cornell, A.J., Wang, R., Erfle, H., Voss, H., Ansorge, W., Kafatos, F.C., and Collins, F.C. 1997. Quantitative trait loci for refractoriness of *Anopheles gambiae* to *Plasmodium cynomolgi* B. *Science* **276:** 425–428.

Received May 11, 2001; accepted in revised form October 26, 2001.

A Comparative Genomic Analysis of Two Distant Diptera, the Fruit Fly, *Drosophila melanogaster*, and the Malaria Mosquito, *Anopheles gambiae*

Viacheslav N. Bolshakov, Pantelis Topalis, Claudia Blass, et al.

Genome Res. 2002 12: 57-66
Access the most recent version at doi:10.1101/gr.196101