Evolution of Supergene Families Associated with Insecticide Resistance

Hilary Ranson,[1] Charles Claudianos,[2,3] Federica Ortelli,[1] Christelle Abgrall,[4] Janet Hemingway,[1] Maria V. Sharakhova,[5] Maria F. Unger,[5] Frank H. Collins,[5] René Feyereisen[4]*

The emergence of insecticide resistance in the mosquito poses a serious threat to the efficacy of many malaria control programs. We have searched the *Anopheles gambiae* genome for members of the three major enzyme families—the carboxylesterases, glutathione transferases, and cytochrome P450s—that are primarily responsible for metabolic resistance to insecticides. A comparative genomic analysis with *Drosophila melanogaster* reveals that a considerable expansion of these supergene families has occurred in the mosquito. Low gene orthology and little chromosomal synteny paradoxically contrast the easily identified orthologous groups of genes presumably seeded by common ancestors. In *A. gambiae*, the independent expansion of paralogous genes is mainly a consequence of the formation of clusters among locally duplicated genes. These expansions may reflect the functional diversification of supergene families consistent with major differences in the life history and ecology of these organisms. These data provide a basis for identifying the resistance-associated enzymes within these families. This will enable the resistance status of mosquitoes, flies, and possibly other holometabolous insects to be monitored. The analyses also provide the means for identifying previously unknown molecules involved in fundamental biological processes such as development.

Insecticides form a central component of most malaria control programs [see accompanying paper by Hemingway *et al.* in this issue (*1*)]. Mechanisms of resistance, such as reduced penetration, increased sequestration, and increased detoxification, all contribute to decrease the effective dose of insecticide, whereas a decreased target site sensitivity or modification of target site may render a dose of insecticide ineffective. Three protein families are largely responsible for insecticide metabolism: the cytochrome P450s, carboxylesterases (COEs), and glutathione transferases (GSTs). The proteins of these families are also involved in the synthesis and breakdown of a multitude of endogenous metabolic compounds, the protection against oxidative stress, the transmission of nerve signals, and the transportation of compounds through cells (*2–4*). Determining the identity of the enzymes involved in insecticide metabolism has been complicated by our lack of knowledge of the complexity of these families and the difficulties of identifying truly orthologous genes between different insect species.

To identify putative COE, GST, and P450 genes in *A. gambiae*, we conducted a BLAST search of the genome with consensus regions or sequences of previously identified insect and mammalian members of these gene families. When significant matches were observed in the genomic sequence, the region was manually annotated to identify the putative transcripts and translation products. Conserved regions, such as the heme binding region of P450s (*4*), a catalytic triad of residues that included the nucleophilic ("catalytic") serine of COE enzymes (*5*), the SNAIL/TRAIL motif of GSTs (*3*), and protein length (about 550, 200, and 500 amino acids for COEs, GSTs, and P450s, respectively) were used to confirm membership of the gene family, and, where available, the predicted protein sequences were checked against expressed sequence tag (EST) translations to confirm the manual annotations. This process resulted in the identification of 51 COE, 31 GST, and 111 P450 gene sequences.

A difficulty arose in distinguishing recently duplicated genes from allelic variants. For example, two P450s (*CYPq3l1* and *CYPq3l2*) differ by a single amino acid and are 99.8% identical at the nucleotide level. These genes were assembled within a single scaffold, and we concluded that they were duplicated genes. This phenomenon is not unique to *A. gambiae*. In *Papilio polyxenes*, *CYP6B4* and *CYP6B5* are 99.3% identical at the nucleotide level (*6*), and several sequence polymorphisms within the *CYP4* "family," which were originally attributed to allelic variation (*7*, *8*), may represent recently diverged P-450 genes. In other instances, we found evidence of alternative haplotypes within the genome sequence, including a large cluster of *CYP6* genes on chromosome 3R (30A) that is represented on two different scaffolds (*9*).

A recognized P450 nomenclature system has been in place since 1987, and the *A. gambiae* P450 genes were classified and will be named according to these rules. As in *Drosophila*, the largest groups of genes fall into the *CYP4* and *CYP6* families. The large number of *CYP4* genes was expected because 17 *CYP4* genes had been identified by polymerase chain reaction methods in *A. albimanus* (*7*) and 18 in *A. gambiae* (*8*).

For the GST and COE families, however, rules for classification have not been clearly established. Sequence identity and immunological relations are the major criteria for the assignment of GSTs to a particular subfamily ("class"), because substrate specificities are broad and often overlapping. In many cases, the true biological functions of these proteins are unknown (*3*). Seven of the GSTs belong to the zeta, omega, theta, and microsomal classes, which are represented in a diverse range of species, including mammals (*3*). The majority of insect GSTs, however, cannot be assigned to recognized mammalian classes. Two classes of insect-specific GSTs have been described (*10*), and we have classified 12 and 7 of the *A. gambiae* GSTs to the delta and epsilon classes, respectively. This still left four GST-like proteins that may represent previously unknown insect-specific GST classes (fig. S1).

The majority of the COE gene sequences were subdivided into eight subfamilies: α-esterases, juvenile hormone esterases, β-esterases, gliotactins, acetylcholinesterases, neurotactins, neuroligins, and glutactin type. The juvenile hormone esterase, α-esterase, and β-esterase families form one large ancestral clade (fig. S2). These enzymes account for the majority of the catalytically active COEs. Neurotactin, neuroligin, gliotactin, and glutactin are cell surface proteins whose extracellular regions show considerable sequence homology to acetylcholinesterase. They are generally considered to be noncatalytic with a variety of functions essential to development and neurogenesis (*11*). We have identified predicted proteins (COEglt1I-7I and COEglt1J-2J) related to *D. melanogaster* glutactin and (COEnrl2H) neuroligin that atypically contain a catalytic serine. The closest *D. melanogaster* relatives of these *A.*

[1]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK. [2]Centre for the Molecular Genetics of Development, Research School of Biological Sciences, Australian National University, GPO Box 475, ACT 2601 Australia. [3]Commonwealth Scientific and Industrial Research Organization (CSIRO) Entomology, GPO Box 1700, Canberra ACT 2601, Australia. [4]Institut National de la Recherche Agronomique Centre de Recherche d'Antibes, 1382 Route de Biot, 06560 Valbonne, France. [5]Center for Tropical Disease Research and Training, University of Notre Dame, Notre Dame, IN 46556–0369, USA.

*To whom correspondence should be addressed. E-mail: rfeyer@salis.antibes.inra.fr

gambiae proteins, CG7529 and CG31146 (fig. S2), also contain a catalytic serine.

A striking feature of the COE, GST, and P450 families in insects is the number of duplicated genes. These paralogs tend to be clustered rather than dispersed throughout the genome. Only 9 of 31 GST genes, 16 of 51 COE genes, and 22 of the 111 P450 genes are present as singletons. There are 16 clusters of 4 or more genes. Most of these are composed of paralogous genes, but clusters composed of genes from more than one family are also present. The clusters are dispersed on all chromosome arms (Table 1), although a "super hyphenated cluster" is found on chromosome 2L bands 23 to 25 that contains 23 COE, a single GST, and 11 P450 gene sequences interrupted by putative coding sequences from other gene families. The degree of clustering appears more prominent in *A. gambiae* than in *D. melanogaster* (fig. S3), where over 50% of these three gene families are found singly or in pairs. This may be partly explained by the enhanced capacity of *D. melanogaster* to lose DNA and expel pseudogenes (*12*, *13*). *D. melanogaster* has undergone few gene duplications in the recent past and has fewer gene families than the nematode *Caenorhabditis elegans* (*13*), an observation consistent with our results from the mosquito.

Our comparison indicates a considerable expansion of the COE (51 versus 36) and P450 (111 versus 90) gene families in *A. gambiae* as compared to those families in *D. melanogaster* (*2*, *14*). For P450 genes, this mainly reflects an expansion of the *CYP4* family. For the COE family, this is largely due to an increase in the number of genes associated with development and neuronal function (fig. S2). This expansion is most notable in the glutactin subfamily, which contains nine members in *A. gambiae* versus four in *D. melanogaster* (*2*), but increased numbers of genes in the juvenile hormone esterase and β-esterase subfamilies were also found. The expansion of these gene families may confer an increased ability in neurosensory perception, possibly aiding in chemical detection of nutrient sources, natural enemy avoidance, or oviposition site recognition by the mosquito.

We do not know whether all the members of the COE, GST, and P450 gene families that we have identified encode functionally active proteins. Full-length cDNA sequences have only been obtained for about 25% of these genes, and several members of these gene families may turn out to be transcriptionally silent or encode aberrant proteins. Five members of the *A. gambiae* P450 family are pseudogenes and one of the GST genes (*GSTd6*) may also be nonfunctional, but no obvious pseudogenes were found among the esterases.

Eventual deflations in the final estimate of the number of functionally active proteins caused by the identification of pseudogenes may be offset by the unearthing of multiple transcripts from a single gene generated by alternative splicing. The open reading frames of two predicted *A. gambiae* juvenile hormone esterase genes overlap, and further analysis may reveal that *COEjhe2F* and *COEjhe3F* are alternatively spliced transcripts of a single gene. In the GST family, the gene *GSTd1* produces four alternative, mature GST transcripts, each of which contains a common 5′ exon spliced to one of four distinct 3′ exons (*15*). Our analysis of the genome, combined with EST data (*16*), showed that the *GSTs1* gene (formerly known as *aggst2-1*) (*17*) is also the product of an alternatively spliced gene: A common 5′ exon is joined to either two or three distinct downstream exons to produce alternative mature transcripts. The NH_2-terminus of GSTs is the most highly conserved region because it contains residues important in the binding and activation of glutathione, whereas the COOH-terminus contains the majority of the residues conferring substrate specificity. Thus, splicing an exon encoding a common NH_2-terminal domain to alternative exons encoding variable COOH-termini is an efficient means of expanding the diversity of substrates recognized by GSTs with a minimal increase in gene duplication.

With whole-genome sequencing projects now largely complete for two insect species, we can use comparative genomics to ask questions related to the evolution of these supergene families. Secure orthologs between *D. melanogaster* and *A. gambiae*, identified by careful analysis of phylogenetic trees, comprise less than 15% of the full complement of the supergene families. A deficit of true orthologs was also found in an analysis of the innate immune system in the two Diptera genomes (*18*). The products of the secure orthologs that do remain presumably perform essential physiological functions and therefore divergent evolution is restrained, allowing their recognition as orthologs.

In the majority of cases, rather than true orthologs, we can identify orthologous groups of paralogous genes, showing that these gene families have radiated independently, from common antecedent genes, in lower and higher Diptera (Fig. 1 and figs. S1 and S2). In many instances, the paralogous genes are clustered on the chromosome, indicating that expansion has occurred by physically local gene duplications or amplifications. In most cases of substantial gene expansion between the two species, it is not possible to identify the ancestral genes. This may be a consequence of a number of processes, including functional diversification or displacement, concerted evolution, problems with homoplasy, or lack of structural constraint.

The continual process of gene duplication and diversification in function or regulation has allowed these enzymes to expand into new biochemical niches within the same time frames in which their host organisms respond to environmental and ecological changes. A key factor facilitating the functional diversification of these enzyme families is their permissiveness to alterations in their primary structure: Dramatic changes in substrate specificity can be achieved by single or small numbers of amino acid substitutions (*5*, *19*, *20*). Gene duplications or point mutations are sources of variation subject to selection. Although physiological constraints would tend to select against such variation, environmental stresses may be overcome by positive selection of new variants. *D. melanogaster* and *A. gambiae* occupy very different ecological niches and are thus exposed to different ranges of exogenous compounds, favoring the independent radiation of the principal enzyme families involved in xenobiotic detoxification.

The mitochondrial P450s, a clearly distinguishable clade within the P450 family, illustrates a typical case of all three gene families. The phylogeny (Fig. 1) distinguishes two types of sequences: the five pairs of true orthologs (*CYP49A1*, *301A1*, *302A1*, *314A1*, and *315A1*) and two groups of paralogous *CYP12* genes in each species. The *CYP12* family has expanded independently in *D. melanogaster* and *A. gambiae* to give rise to six and four paralogous genes, respectively. The *CYP12* family is similar to the *CYP6* and *CYP9* families of microsomal P450 in that they are inducible by xenobiotics, contain many members arranged in clusters, and are involved in xenobiotic metabolism (*21*). This

Table 1. Cytological location of *A. gambiae* and *D. melanogaster* GST, COE, and P450 genes. References for *D. melanogaster* genes (*2*, *10*, *14*, *26*).

Species	Gene class	Chromosome arm					
		X	2L	2R	3L	3R	Total
A. gambiae	GSTs	5	1	14	1	10	31
	COEs	1	23	11	15	1	51
	P450s	10	11	47	11	32	111
D. melanogaster	GSTs	3	1	17	5	13	39
	COEs	1	8	6	4	16	35*
	P450s	15	14	34	9	17	89*

*One P450 gene and one COE gene are unmapped.

mitochondrial family is distinct from the five pairs of mitochondrial orthologs that probably represent enzymes involved in key metabolic pathways. *D. melanogaster Cyp302a1* and *Cyp315a1* are synonymous with *disembodied* and *shadow*, the genes encoding the C22 and C2 hydroxylases of the ecdysteroid biosynthetic pathway from cholesterol, respectively (*22*).

The specific functions of insect GSTs are largely unknown, but patterns of orthology may provide clues to in vivo functions. All the insect GSTs that have been implicated thus far in xenobiotic metabolism belong to either the delta or epsilon classes (*10*). A similar association occurs with the α-esterase subfamily of COEs (*2*, *5*). Insect-specific subfamilies of GSTs, COEs, and P450s have expanded and radiated independently in *A. gambiae* and *D. melanogaster*, presumably in response to environmental change. GSTs and COEs from clades represented in nonarthropod taxa typically contain orthologous genes and are predicted to be involved in essential physiological pathways. Two GST genes (*GSTu1* and *GSTu4*) and one COE gene (*COE11o*) that are currently unassigned to any subfamily have clear orthologs in *D. melanogaster*, suggesting that these genes perform a vital metabolic function within insects (figs. S1 and S2).

To identify genes encoding enzymes involved in insecticide resistance, one approach is to extrapolate from examples in other insects where individual genes have been convincingly implicated in resistance and search for their orthologs in *A. gambiae*. This was successfully used to identify mutations in esterase genes in the housefly that are associated with resistance to organophosphates (*5*). Orthologous α-esterase clusters composed of paralogous duplications or gene amplifications often share spatial and temporal expression patterns that predispose these genes to selection by insecticides. Amplified Esterase A and Esterase B of mosquitoes (COEae1G and COEae2G) are notable examples (*2*) (fig. S2).

Unfortunately, this approach appears unlikely to succeed in identifying the members of the GST and P450 families involved in resistance, either because the resistance-associated genes in other species have not yet been clearly defined or because the orthology is either too tenuous or nonexistent. For example, in *D. melanogaster* two P450 genes, *CYP6A2* and *CYP6G1*, are involved in DDT resistance (*23*, *24*), but these are each contained within paralogous sets of genes in *Drosophila* and neither set has secure *Anopheles* orthologs. Nonetheless, knowledge of the full complement of these gene families will rapidly allow the design of specific DNA microarrays to identify those genes overexpressed in insecticide-resistant *A. gambiae* strains.

The physical mapping of the majority of the *A. gambiae* genes enables us to identify genes found within the boundaries of known resistance-associated loci as determined by genetic mapping studies (*1*, *25*). So far, two loci associated with resistance to DDT and two loci associated with resistance to pyrethroids have been genetically mapped in *A. gambiae*. One of the major DDT-resistance loci colocalizes with two members of the epsilon GST class, one of which is able to detoxify DDT (*10*). We have now identified six more GST genes sequentially arranged within this region of the genome on 3R division 33B, all of which are now under scrutiny for their role in DDT resistance. As for the two major pyrethroid resistance loci, one is probably an altered allele of the sodium channel gene encoding the pyrethroid-target site, but biochemical data suggest that the second locus is involved in oxidative metabolism of the insecticide. We can now confirm that several large clusters of cytochrome P450 genes coincide with this resistance locus. The colocalization of the insecticide resistance–associated loci with large clusters of structural genes may still be coincidental, especially as the mapping resolution is very low at this point, but the data from the genome sequencing project has given a credible starting point for the elucidation of the roles that these gene families play in insecticide resistance (*1*).

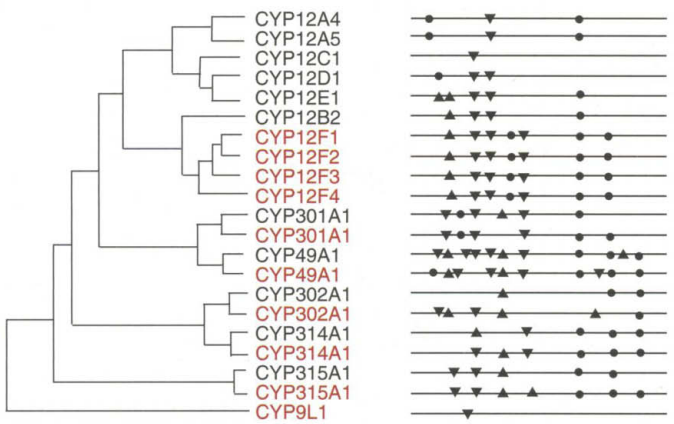

Fig. 1. Phylogenetic tree of the mitochondrial P450 proteins of *D. melanogaster* and *A. gambiae*. Alignment was by the program CLUSTALX, followed by Phylip's Protpars. *CYP9L1* was used as the outgroup. The mitochondrial P450s are recognized as a group by their amphipathic NH_2-terminus (mitochondrial-targeting sequence) and by two conserved basic residues involved in the interactions with the redox partner (adrenodoxin). *A. gambiae* proteins are shown in red, *D. melanogaster* in black. The schematic alignment next to the tree shows the position of introns in the genes: phase 0 (●), phase 1 (▲), and phase 2 (▼). On the 20 sequences representing the mitochondrial P450s of both species, there are 22 different intron positions. Multiple instances of intron loss and gain in the 250 million years of divergence are evidenced on the pairs of orthologs.

References and Notes

1. J. Hemingway, L. Field, J. Vontas, *Science* **298**, 96 (2002).
2. C. Claudianos, E. Crone, C. Coppin, R. Russell, J. Oakeshott, in *Agrochemical Resistance Extent, Mechanism and Detection*, J. M. Clark, I. Yamaguchi, Eds. (ACS Symposium Series no. 808, American Chemical Society, Washington, DC, 2001), pp. 90–101.
3. D. Sheehan, G. Meade, V. M. Foley, C. A. Dowd, *Biochem. J.* **360**, 1 (2001).
4. D. Werck-Reichhart, R. Feyereisen, *Genome Biol.* **1**, 3003 (2000).
5. J. G. Oakeshott, C. Claudianos, R. J. Russell, G. C. Robin, *Bioessays* **21**, 1031 (1999).
6. C.-F. Hung, R. Holzmacher, E. Connolly, M. R. Berenbaum, M. A. Schuler, *Proc. Nat. Acad. Sci. U.S.A.* **93**, 12200 (1996).
7. J. A. Scott, F. H. Collins, R. Feyereisen, *Biochem. Biophys. Res. Commun.* **205**, 1452 (1994).
8. H. Ranson *et al.*, *Insect Mol. Biol.* **11**, 409 (2002).
9. Both scaffolds, CRA_9xP1GAV57C1 (GenBank accession no. AAAB01008911.1) and CRA_x9P1GAV5CRW (AAAB01008964.1), map to 3R division 30A and contain allelic variants of the same 12 P-450 genes. In this and other cases where multiple haplotypes were present in the assembled genome, the scaffolds constituting the minimal tiling scaffold of the entire genome were used as templates for gene mining. In three cases, the presumed alternate haplotype contained two paralogs for the single version of a gene in the reference haplotype.
10. H. Ranson *et al.*, *Biochem. J.* **359**, 295 (2001).
11. D. Grisaru, M. Sternfeld, A. Eldor, D. Glick, H. Soreq, *Eur. J. Biochem.* **264**, 672 (1999).
12. D. A. Petrov, *Genetica* **115**, 81 (2002).
13. Z. Gu, A. Cavalcanti, F.-C. Chen, P. Bouman, W.-H. Li, *Mol. Biol. Evol.* **19**, 256 (2002).
14. N. Tijet, C. Helvig, R. Feyereisen, *Gene* **262**, 189 (2001); available at http://P450.antibes.inra.fr
15. H. Ranson, F. H. Collins, J. Hemingway, *Proc. Nat. Acad. Sci. U.S.A.* **95**, 14284 (1998).
16. A. Dana, F. H. Collins, unpublished data.
17. R. A. Reiss, A. A. James, *Insect Mol. Biol.* **2**, 25 (1993).
18. G. K. Christophides *et al.*, *Science*, in press.
19. T. L. Domanski, J. R. Halpert, *Curr. Drug Metabol.* **2**, 117 (2001).
20. A. M. Caccuri *et al.*, *J. Biol. Chem.* **276**, 5427 (2001).
21. V. M. Guzov, G. C. Unnithan, A. A. Chernogolov, R. Feyereisen, *Arch. Biochem. Biophys.* **359**, 231 (1998).
22. J. T. Warren *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11043 (2002).
23. J. B. Bergé, R. Feyereisen, M. Amichot, *Phil. Trans. R. Soc. Biol. Sci* **353**, 1701 (1998).
24. P. J. Daborn *et al.*, *Science* **297**, 2253 (2002).
25. H. Ranson *et al.*, *Insect Mol. Biol.* **9**, 499 (2000).
26. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).

27. Financial support was provided to the *Anopheles gambiae* Genome Consortium by NIH grant U01 AI50687 to Celera Genomics, NIH grant U01 AI48846 to the University of Notre Dame, and funds from the French Government to Génoscope. C.C. was supported by a grant from the National Health and Medical Research Council of Australia (grant 997038), H.R. was supported by a fellowship from the Royal Society.

Supporting Online Material

www.sciencemag.org/cgi/content/full/298/5591/179/DC1

Figs. S1 to S3

30 July 2002; accepted 6 September 2002