Entomology, Ornithology & Herpetology

Raghav Entomol Ornithol Herpetol 2012, 1:1
DOI: 10.4172/2161-0983.1000102

Research Article Open Access

Anopheles gambiae (Diptera: Culicidae) Cytochrome P450 (P450) Supergene Family: Phylogenetic Analyses and Exon-Intron Organization

Raghavendra K[1]*, Niranjan Reddy BP[1,2] and Prasad GBKS[2]

[1]*National Institute of Malaria Research (ICMR), Sector 8, Dwarka, New Delhi, India*
[2]*School of Studies in Biotechnology, Jiwaji University, Gwalior, MP, India*

Abstract

The cytochrome P450 superfamily is involved mainly in developmental processes and xenobiotic metabolism in insects. Analysis of the *Anopheles gambiae* genome has shown 105 putatively active P450 genes that are distributed in four major clans, namely mitochondrial, CYP2, CYP3, and CYP4. In the present study, phylogenetic analysis using multiple methodologies, exon-intron organization, correlation between genes in gene clusters and their gene organizations were analyzed. Further to this, usability of intronic positions in deciphering the evolutionary relatedness among the members of *Ag*P450 supergene family was studied. The results show that the *Ag*P450 supergene family is evolved through the complex process of duplications followed by structural-functional evolution. This supergene family might have undergone numerous intron-losses and gains during the process of evolution. However, this process is closely related with the evolutionary relationship among the members of the *Ag*P450 supergene family. Furthermore, this study identifies the need of in-depth study to elucidate the functional importance of the conserved intron in CYP6 family.

Keywords: Cytochrome P450 (P450); Phylogenetic analysis; Dollo parsimony; Exon-intron (gene) organization; Intron loss-gain

Introduction

The *Anopheles gambiae* (Diptera: Culicidae) is the principal vector of malaria in Africa accountable for nearly 86% of total cases reported [1]. The genome sequence analysis has revealed a strong evidence for the presence of ~14,000 protein encoding transcripts [1]. This particular information facilitates the understanding of various aspects of the malaria mosquito biology, including the dynamics of insecticide resistance, and also aid-in design and development of new insecticide molecules for the effective vector control. It is known that there are three major supergene families, namely cytochrome P450s (CYPs), esterases, and glutathione *S*-transferases (GSTs) involved in the metabolization of xenobiotics, and thus are primarily responsible for developing the insecticide resistance [2].

With more evolutionarily diverse genome sequences being completed, the field of evolutionary genomics in the analysis of individual gene families has become a main-stream of research. It is now possible to study the nature of relationships between gene (exon-intron) structure and gene expression by focusing on specific differences and similarities in the gene organization across genes belonging to different gene families of a supergene family [3]. Cytochrome P450 mono-oxygenases (P450s) also known as mixed function oxidases are a super-family of haem-thiolate proteins that are involved in the metabolism of a wide variety of chemical reactions important for both developmental processes and in the detoxification of foreign compounds, including insecticides [4]. The P450 enzymes are found in all domains of life [5] that include prokaryotes (archaea and bacteria), lower eukaryotes (fungi) and higher eukaryotes (plants and animals) [4]. Albeit, the P450s are not absolutely necessary for some prokaryotes and protists, this superfamily has a very old origin, certainly earlier than the emergence of eukaryotes [6], and even before the accumulation of molecular oxygen in the atmosphere [4,7]. Genes encoding the P450 enzymes are designated with the abbreviation "CYP", followed by an Arabic numeral indicating the gene family, a capital letter indicating the subfamily, and another numeral for the individual gene (For example: CYP6Z1, where CYP=CYtochrome oxidase P450; 6=gene family; Z=subfamily; 1=number of an individual gene). The present norms for classification and nomenclature are based on the sequence similarities; gene family members should share >40% amino acids identity, while subfamily members should share >55% identity [8,9]. Based on the sequence similarities to the known mammalian genes, the *Anopheles gambiae* P450s (*Ag*P450s) have been classified into four clans, namely mitochondrial, CYP2, CYP3, and CYP4 [10]. The CYP2 clan consists of CYP15, CYP303, CYP304, CYP305, CYP306, CYP307 gene families. The CYP3 and CYP4 clan consists of CYP6, CYP9, CYP329, and CYP4, CYP325 gene families, respectively. Whereas the mitochondrial clan consists of CYP12, CYP49, CYP301, CYP302, CYP314, CYP315 gene families [10]. Cytochrome P450 supergene family has been studied extensively for insecticide resistance mechanisms in malaria and other disease vectors [9,11,12].

Considering the ubiquitous and diverse nature of the superfamily and its importance in developing insecticide resistance in insects, here, we made a comprehensive study to understand the phylogenetic relationship using different approaches, exon-intron (gene) organization and its correlation with the evolutionary relationship among the members of *Ag*P450s. By searching the public domain sequence databases, we found 105 putatively active P450 gene sequences. The study revealed extensive diversity, yet, greater degree of gene organization conservation within the P450 gene families. The CYP3 clan members have exceptionally conserved gene organization that is essentially biased towards single phase '1' class of introns.

***Corresponding author:** Raghavendra K, (Scientist 'E'), National Institute of Malaria Research (ICMR), Sector 8, Dwarka, New Delhi, PIN- 110077, India, Tel: +91-11-25307 205; E-mail: kamarajur2000@yahoo.com

Received April 17, 2012; **Accepted** August 27, 2012; **Published** August 30, 2012

Citation: Raghavendra K, Niranjan Reddy BP, Prasad GBKS (2012) *Anopheles gambiae* (Diptera: Culicidae) Cytochrome P450 (P450) Supergene Family: Phylogenetic Analyses and Exon-Intron Organization. Entomol Ornithol Herpetol 1:102. doi:10.4172/2161-0983.1000102

The mitochondrial and CYP2 gene families have diversified gene organizations, except for the CYP12 gene family. The study revealed that the CYP4, CYP6, CYP9, and CYP325 gene families that are majorly implicated in causing the insecticide resistance are expanded largely through the extensive local duplications. Furthermore, the study failed to identify any signals that qualify the presence of concerted evolution among the locally duplicated gene clusters. The outcome from these analyses could be used to study intron-loss and gain studies and to find out ancient introns through comparative genomic analyses given the fact that this supergene family is one of the oldest and is present in almost all domains of life.

Materials and Methods

Sequence data retrieval, multiple sequence alignment and NJ based phylogenetic analysis

The cytochrome P450 gene and protein sequences of *Anopheles gambiae* were downloaded from different databases, namely, The Insect P450 Site (http://p450.sophia.inra.fr/), Ensembl (AgamP3, February 2006) (http://metazoa.ensembl.org/Anopheles_gambiae/Info/Index), and from NCBI (http://www.ncbi.nlm.nih.gov/). The P450 genes are highly diverse in their sequences and functions, but are characterized by a conserved region towards the C-terminus, the heme-binding domain containing the axial Cys ligand to the heme [13]. To verify the P450 gene annotation, the downloaded putatively active monoxygenases genes were searched for the P450 supergene family specific amino acid sequence pattern from PROSITE database (http://www.expasy.ch/prosite/). The Scan Prosite tool available online at (http://prosite.expasy.org/scanprosite/) was used to scan the sequences using the [FW] - [SGNH] - x - [GD] - {F - [RKHPT] - {P} - C - [LIVMFAP] - [GAD] [Prosite accession number PS00086] cytochrome P450 cysteine heme-iron ligand signature. All the 105 downloaded sequences are having the conserved FXXGXXXCXG (where X represents the non conserved amino acid) sequence at the C terminal end verifying the cytochrome P450 candidature.

The multiple sequence alignments (MSA) of the protein sequences were performed using the Muscle tool from EBI (http://www.ebi.ac.uk/Tools/msa/muscle/). MUSCLE stands for MUltiple Sequence Comparison by Log-Expectation. Muscle is claimed to achieve better average accuracy and better speed than ClustalW2 or T-Coffee alignment programs [14]. The alignment was saved in FASTA format and used to construct the neighbor-joining (NJ) [15] phylogenetic tree in MEGA 4.0 software [16] with 100 bootstrappings. The genetic distances were estimated using the Poisson correction and complete deletion option was selected for excising the large gaps in the sequence alignments from the analysis. The multiple sequence alignment of 105 genes was summarized by Plotcon (http://bioweb2.pasteur.fr/docs/EMBOSS/plotcon.html) based sequence similarity graph which represents the similarity along a set of aligned sequences. In addition to the NJ method, character based phylogenetic analyses (maximum parsimony and maximum likelihood) were also performed to ascertain the phylogenetic status of the genes and gene families. The Max-mini Branch & Bound method was used to search the best fit MP tree while Jones-Taylor-Thornton probability model assuming the constant rate of change rate variation among the sites were used to infer the maximum parsimony (MP) and maximum likelihood (ML) based phylogenetic analysis. Both analyses were performed using the PHYLIP 3.5c software (the *PHYL*ogeny *I*nference *P*ackage) which is freely available at

http://evolution.genetics.washington.edu/phylip.html

Principal coordinates analysis (PCO)

Principal Coordinates analysis facilitates for building sequence ordinations using sets of aligned sequences [17]. These ordinations can be used to complement the phylogenetic analyses, especially, when a large number of sequences are considered (http://pbil.univ-lyon1.fr/Rwe/). As our analysis includes 105 P450 gene sequences, we performed PCO analysis using the R-software freely available at http://pbil.univ-lyon1.fr/Rwe/ to complement the phylogenetic analyses. The analysis was performed by providing the Clustal W-MSA P450 superfamily of protein sequences. The analysis assumes that all of the sequences in the superfamily are homologous and that different subsets have developed distinctive features such as different substrate specificities or specific interactions with different macromolecules. The analysis is intended to identify the set of sequences that are most likely to be responsible for the distinctive features and group them accordingly on the F1 × F2 coordinate map.

Construction of Blocks, Blocks based phylogeny and sequence logos

As the functional annotation of *Ag*P450s is not completed, gene families are grouped according to sequence similarities. The sequence similarity of the proteins in a gene family or even in a subfamily is far from uniform; some regions are clearly conserved while others display a little sequence similarity [18]. Therefore, to quantify the conserved sequence motifs the Blocks were constructed [19] for the *Ag*P450 superfamily using Blocks construction tool available at http://bioinformatics.weizmann.ac.il/blocks/about_logos.html. Blocks based phylogeny was constructed using the tool available at http://blocks.fhcrc.org/ to see whether inclusion of only highly conserved sequence motifs will affect the phylogenetic relationship determined using complete amino acid sequences.

Introns as phylogenetic markers

Dollop version 3.5c estimates phylogenies by the Dollo [20] or polymorphism parsimony criteria for discrete character data with two states (0 and 1). Dollo parsimony analysis is mainly used to construct the phylogenies based on restriction fragment length polymorphism data. However, the analysis also used in reconstructing intron based phylogenetic relationship among the taxa [21]. The analysis assumes that each derived character state is originated only once on the tree and the intron loss is more frequent than intron gain. To see whether introns are useful in phylogenetic reconstruction of the P450 supergene family, the binary matrix based on intron presence and absence was created assuming '0' for the absence of intron and '1' for the presence of intron at the homologous positions on the amino acid multiple sequence alignment. The data was analyzed in Phylip (the *PHYL*ogeny *I*nference *P*ackage) using DOLLOP and CONSENSE modules (Felsenstein 2002 http://evolution.genetics.washington.edu/phylip.html). CONSENSE program was used to construct the most parsimony tree using 100 tied trees as an input. The program computes consensus trees by the majority-rule consensus tree method, which also helps in finding the strict consensus tree. The consensus tree contains a number at each node that represents the number of times that particular node repeated in 100 tied trees.

Determining and mapping of intron position, length and intron phase

Intron position, length and intron phases were determined by

aligning the gene sequences to the protein sequences using the Wise 2 software with default parameters (http://www.ebi.ac.uk/). Multiple sequence alignment of protein sequences was performed using the ClustalX2 and Muscle software to map the intron positions and to determine the total consensus length of the aligned P450 genes. CorelDRAW.X3 program was used to develop the photomap of the exon-intron organization and position of the intron phases of P450 superfamily. Intron density per gene was calculated by dividing the total number of introns to the total number of genes present in the gene family. The correlation coefficient between the total gene length and total intron length, the coding DNA sequence (CDS) and frequency of exons, were analyzed using the SPSS software along with regression equations for each of the relationships.

Intron losses and gains in *An. gambiae* P450 supergene family

Intron loss and gain studies were performed using MALIN: *M*aximum *L*ikelihood *A*nalysis of *I*ntron *E*volution software (http://www.iro.umontreal.ca/~csuros/introns/malin/). The introns gains and losses were analyzed using Dollo parsimony (first proposed by Farris 1977) and posteriors module that follow probabilistic rate models that can be directly used for estimation of ancestral intron presence, as well as intron gains and losses. Contrary to the Dollo parsimony, probabilistic models consider ambiguous intron sites for the calculation of intron-gain and loss. Dollo parsimony [21,22] and probabilistic models [23] have been used in determining the intron losses and gains.

Results

Protein multiple sequence alignment and phylogenetic tree analysis

The average length of the retrieved 105 *Ag*P450 amino acid sequences is around 512 ± 1.6 (S.E.) amino acids (a.a), whereas the consensus length of the multiple sequence alignment is 868 a.a residues (Supplementary figure 2a). The excessive amount of alignment gaps are essentially due to poor conservation of the sequence at N-terminal region of *CYP307A1*, *CYP307B1*, *CYP12F3*, and *CYP314A1* genes (Supplementary figure 2a). Summary of *Ag*P450s' multiple sequences alignment is shown in (Supplementary figure 2a) (PLOTCON sequence similarity graph). The same alignment was used to construct the Neighbor Joining (NJ) based phylogenetic tree (Supplementary figure 2b). One thousand bootstrapping iterations were performed and the corresponding bootstrapped values of each node are shown in the (Supplementary figure 2b). The analysis revealed four unambiguously distinguishable clans, namely mitochondrial, CYP2, CYP3, and CYP4 with higher bootstrapped values (>90). The phylogenetic tree based on the sets of conserved sequences called Blocks also showed similar topology to the NJ based phylogenetic tree (data not shown). In addition to these, the maximum parsimony and maximum likelihood (Figure 1) based phylogenetic inferences further corroborated the phylogenetic inferences drawn from the NJ and Blocks based phylogenetic relationships (data not shown). In order to obtain the most consensus phylogenetic relationship among the members of P450 supergene family, a consensus tree was generated by combining the inferred trees from NJ, MP, ML, Bayesian, and Blocks using CONSENSE program

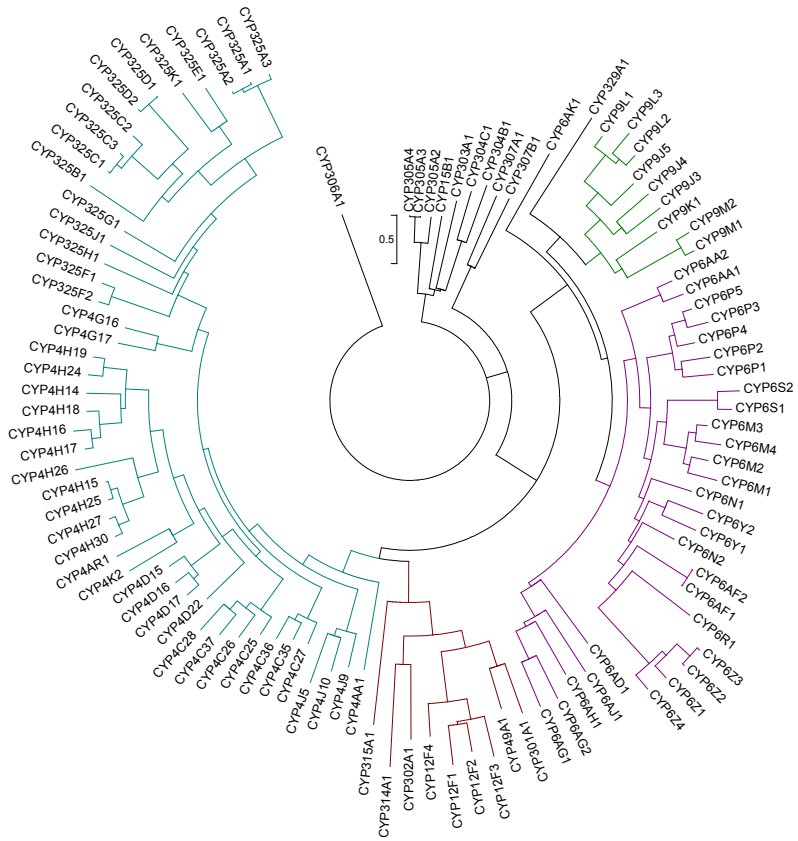

Figure 1: Diagrammatic representation of ML based phylogenetic relationship among the members of An. gambiae P450 sequences.

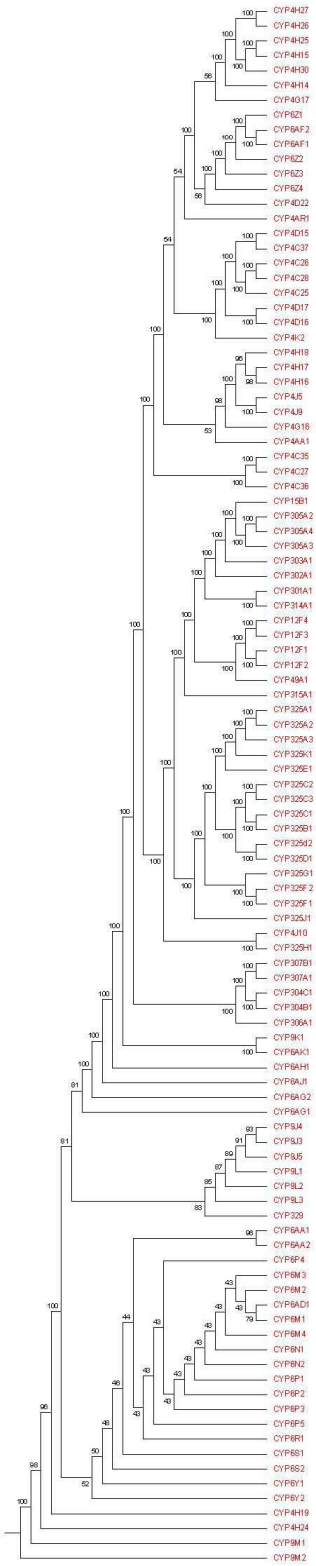

Figure 2: Dollo parsimony based phylogenetic analysis using introns as phylogenetic informative markers. The observed phylogenetic incongruencies are represented with red markings. Considering the fact that intronic positions were used in the Dollo parsimony analysis, it is not surprising to observe monophyletic groupings of the intron less genes (boxed in blue rectangles). Surprisingly, the analysis could successfully resolve the phylogenetic relationship between mitochondrial and CYP2 clans (marked with dotted line green color boxes).

with 1000 bootstrapping replicates (Figure 1) (data not shown here). In addition, the Principle Coordinate (PCO) analysis was performed to supplement the understanding of the phylogenetic analysis (Online Resource 1). The analysis revealed a very close relationship of mitochondrial clan members with the CYP2 clan members.

Furthermore, an attempt was made to assess the usability of intronic positions as phylogenetic informative markers using the DOLLO parsimony. The motivation for this analysis was drawn from the observation that the intronic positions and phases are highly conserved among the members within a gene family, while sufficiently diverged between the gene families. DOLLO parsimony based analysis resulted in a phylogeny which is not congruent with the multi-analysis based consensus phylogeny (Figure 2, red markings are made to represent the phylogenetic incongruencies). Interestingly, the analysis could successfully resolve the phylogenetic relationships of the CYP2 and mitochondrial clan members. From these observations it is evident that intronic positions could be used for resolving phylogenetic relationship among the conserved/ortholog genes and may not be for the gene families that are evolving through extensive local duplication. In a more simplified way, intronic positions could be potential phylogenetic markers for resolving evolutionary relationships among orthologs and not between the paralogs. Furthermore, these observations suggest that the intron-losses and gains are not correlated with the evolutionary distance among the members of the supergene family. The reason we hypothesize that the protein sequences of some genes, especially which are under tremendous xenobiotic stress, might have evolved faster than the intron loss and gain events.

In order to understand the relationship between gene clusters and phylogenetic relationship, the comparative analysis is performed. The results show a perfect positive association between the occurrences of genes in the gene clusters and their monophyletic phylogenetic relationships (Supplementary figure 2b). In other words, the genes that are found to occur in gene clusters in the genome show a monophyletic phylogenetic relationship. Nevertheless, the clustered genes have also shared intronic positions and intron phases (Supplementary figures 2b and 2c); where clustered genes are represented with the same color; implying that the genes within the gene clusters are nothing but a result of local gene duplications.

Correlation between the phylogenetic relationship and gene (exon-intron) organization

The correlation between the phylogenetic relationships of the genes and their exon-intron organizations were performed to know the underlying evolutionary processes that might be influencing the present day organization of this supergene family. It is known that the genes with shared evolutionary relationships have conserved/comparable gene organizations. Expectedly, further corroborating the important role of gene duplications in the evolution of the *Ag*P450 supergene family, genes within the monophyletic relationship showed the conserved gene organizations (Supplementary figure 2c).

For instance, the intron less genes, namely *CYP4H19, CYP4H24* and *CYP9M1, CYP9M2* have shown monophyletic phylogenetic relationship. Similarly, the CYP6Z subfamily has four genes with a single phase '1' introns. All the introns are at a similar position with the same intronic phase but are placed at a different position in relation to the conserved phase '1' intron of the CYP6 family (Supplementary figure 2c). Supporting this observed difference in the gene organization, interestingly, CYP6Z subfamily has formed a separate monophyletic clade (100% bootstrapped values) (Supplementary figure 2b). This

suggests that in CYP6Z sub family, four genes might have evolved from a single common ancestor through duplications. Moreover, five out of 30 total genes (CYP6AG1, AG2, AH1, AJ1, and AK1) in the CYP6 family have ≥ 3 intron numbers. Similar with the above observation, they have also formed a different but monophyletic clade from the rest of the CYP6 genes, suggesting the strong correlation between the gene organization and phylogenetic relationship. These observations imply that the *Ag*P450 supergene family has expanded through the extensive duplications that followed by functional divergence through independent divergent evolution of the genes .

Gene clusters and evolution of the exon-intron organization

In order to assess the correlation between the genes that exists in a gene cluster and their gene organization patterns, the gene clusters and the exon-intron organizations of the members of the *Ag*P450 supergene family were inferred. Two or more P450 genes that uninterruptedly exist in particular loci are reckoned as a gene cluster. It is expected that, if a gene family is evolving through local gene duplications by forming gene clusters, the members within the gene clusters must have conserved gene organization patterns. Following this definition, a total of 17 gene clusters (each consists of ≥ 2 genes) were identified (Online Resource 2). The CYP325 gene family on the chromosome 2R (13 genes) and CYP6 gene family on the chromosome 3R (14 genes) were identified as two largest gene clusters. Out of 30 genes of the CYP6 family, most of them (27/30 genes) were found to occur in gene clusters that are majorly distributed on the right arm of the 2nd and 3rd chromosomes. Out of 16 total genes of the CYP325 family, 12 genes are localized as a single gene cluster on the chromosome 2R. Of 30 members of CYP4 family, 24 genes are localized into the various gene clusters that are distributed across the genome except on left arm of the chromosome 3. Eight out of 9 genes of the CYP9 family were localized into a gene cluster. In total, 21 singleton genes were identified, and are found to be distributed throughout the genome (Online Resource 2). The comparison of gene clusters formation and gene organization revealed a perfect correlation between them (Supplementary figures 2b and 2c) i.e. the gene family members within a gene cluster have conserved intron-exon organizations. The analysis further revealed that the mitochondrial and CYP2 clan gene families are not expanded much. Together, mitochondrial and CYP2 clans, have shared 8/21 singletons. The CYP12 family of the mitochondrial clan is having four gene copies that are organized into single gene cluster on the chromosome 3R with conserved gene organizations (Supplementary figure 2c and Online Resource 2).

Intron number and intron phase distribution in AgP450 supergene family

Out of 105 *Ag450s*, four genes are intron less-*CYP4H19, CYP4H24, CYP9M1,* and *CYP9M2.* Of a total of 296 introns, 39 genes have single introns, while 13, 12, 18, 5, 4, 9 and 1 gene has 2, 3, 4, 5, 6, 7, and 10 numbers of introns, respectively (Figure 3a). A maximum number of 10 introns were found in *CYP49A1* gene (Supplementary figure 2c and Online Resource 3). The CYP4 clan consists of 157 (53%) of the total number of introns, while 63 (21%), 53 (18%), and 23 (8%) of total introns were contributed by mitochondrial, CYP3, and CYP2 clans, respectively. Considering the observed fact that the P450 gene families are probably involved in insecticide resistance are expanded through the gene duplications, only CYP4, 6, 9, 12, and 325 gene families were considered for intron evolution analysis.

The CYP325 family consists of a total of 16 genes (Table 1) with

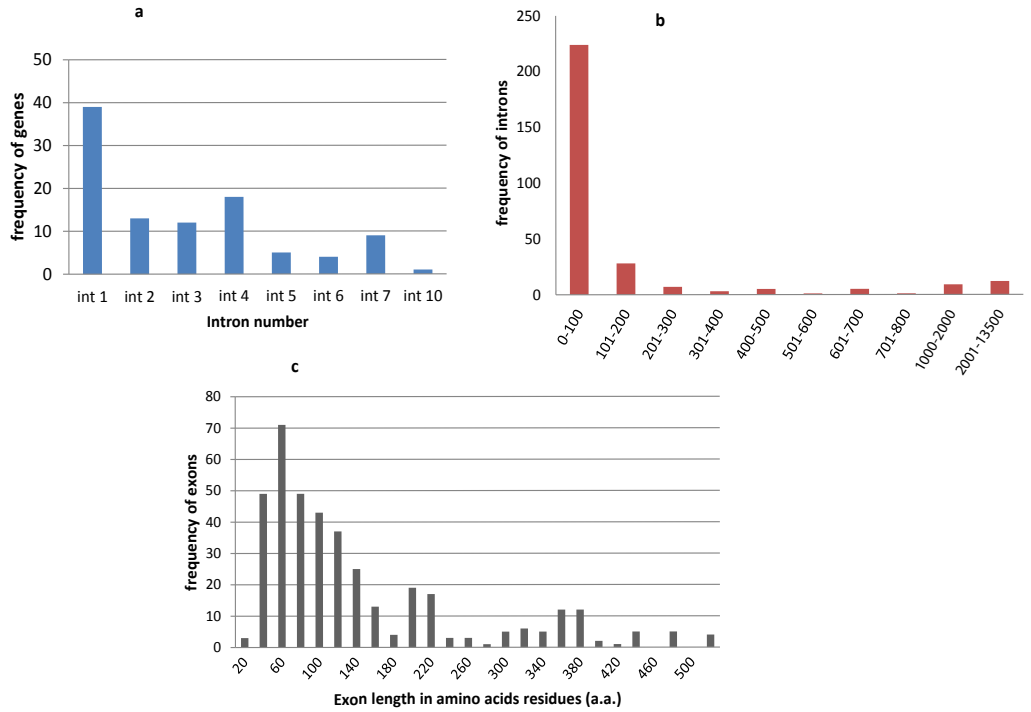

Figure 3a: Presence of intron number dominance in the *Ag*P450 supergene family.
Figure 3b: Frequency histogram showing the different length of introns in the *Ag*P450 supergene family.
Figure 3c: Frequency histogram showing the different length of exons in the cytochrome P450 supergene family of *An. Gambiae*.

Name of the Clan	Name of the Gene family	Phase 0	Phase 1	Phase 2	Total number of introns	Total number of genes	Gene family wise intron density*	Unique introns	Conserved introns**
CYP4 Clan	**CYP325**	23	22	11	56	16	3.5	--	7
	CYP4	21	56	24	101	28	3.6	5	10
CYP3 Clan	**CYP6**	10	29	6	45	30	1.5	5	8
	CYP9	0	1	6	7	7	1	1	1
	CYP329	0	0	1	1	1		--	--
CYP2 Clan	**CYP15**	1	1	0	2	1		--	--
	CYP303	1	0	1	2	1		--	--
	CYP304	0	3	2	5	2	2.5	1	2
	CYP305	0	3	3	6	3	2	2	0
	CYP306	2	1	1	4	1		--	--
	CYP307	0	3	1	4	2	2	2	2
Mitochondrial Clan	**CYP12**	12	4	12	28	4	7	4	--
	CYP49	4	2	4	10	1		--	--
	CYP301	3	0	3	6	1		--	--
	CYP302	1	3	2	6	1		--	--
	CYP314	3	1	2	6	1		--	--
	CYP315	3	2	2	7	1		--	--
		84	**131**	**81**	**296**	101			

Note: *intron density per gene family was calculated by dividing the total number of introns to the total number of genes. The gene families with only one gene are not considered for the intron density analysis

**conserved intron designation was given when an intron found to be conserved in more than one gene of a respective gene family. And thus, the conserved and unique introns were mentioned for only the gene families which have more than one gene. Some of the introns are found to be conserved in more than one gene family. Such intronic positions are highlighted using vertical lines in the supplementary figure 2c

Table 1: Table showing intron phase, intron number and family-wise intron density distribution across different gene families of P450 superfamily of *Anopheles gambiae*.

7 conserved intronic positions (Supplementary figure 2b), of which 2nd, 4th, 5th, and 6th intron positions (Supplementary figure 2c) for line markings that identify the conserved intron positions) are highly conserved spanning 13 (81%), 11 (68%), 12 (75%) and 11 genes (68%) of total CYP325 family genes, respectively. Nine genes (56% of total genes of CYP325 family) have all the conserved four introns in their gene organizations (Supplementary figure 2c).

The CYP4 gene family consists of a total of 30 genes. As this family consists of 2 intron less genes, 28 is considered as the total number of genes in the data analysis. Out of 15 (10 conserved and 5 variable) intronic positions of the CYP4 family, five positions are highly conserved, namely 2nd position {20 genes (71% of total 28 genes)}-, 3rd (32%), 5th (68%), 6th (68%), and 12th (32%) positions (for line makings Supplementary figure 2c).

The CYP6 gene family consists of 30 genes. Of a total 13 (8 conserved and 5 variable) intronic positions, 9th and 12th position's phase '1' introns are highly conserved; contribute 20 (67%) and 7 (23%) genes of a total of 30 genes, respectively. Twenty five genes out of 30 genes (83%) are having single, phase '1' intron. The CYP12 family has a total of four genes with seven conserved intron positions.

The CYP9 gene family consists of a total 9 genes (*CYP9M1* and *CYP9M2* are intron less) with two (1 conserved and 1 variable) intronic positions. Of which 1st positions phase '2' intron is highly conserved spanning 6 genes (86%) of a total 7 genes (two genes are intron less).

Out of 296 total introns, 84, 131, and 81 are belonging to the phase 0, 1, and 2 introns, respectively (Table 1). The P450 superfamily has more phase 1 introns than phase 0 and 2 introns. Phase 0 and 2 introns contribute almost equally to the gene organization of the *Ag*P450s. The gene family-wise distribution of the intron phases is given in table 1.

Intron and exon length distribution

The conserved single introns in the CYP6 gene family have 58-103 bp nucleotide length (Online Resource 3). Nonetheless, in addition to the CYP6 gene family, the CYP4, CYP9 gene families have also single intron genes with more or less similar nucleotide lengths. Out of 296 introns, ~76% (224) have intron sequence length ≤ 100 bp (Figure 3b). The minimum and maximum lengths of the introns observed were 32 and 13216 bp of *CYP6AG2* and *CYP306A1* genes, respectively (Online Resource 3). The average length of an intron in the *Ag*P450 supergene family is 368 ± 75 bp.

Exon length peaks at ~200 and ~400 a.a and are mostly contributed by single intron genes belonging to the CYP4, 6, 9, and 325 gene families. Whereas, <200 amino acid exons were dominated by multiple intronic genes. The exons with >500 amino acids are intron less genes (Figure 3c). The *CYP325D1* and *CYP9M2* genes are having the smallest (33 a.a) and largest (529 a.a) exons, respectively. The average length of an exon is 135 ± 5.7 a.a.

Intron, exon density

The intron and exon density per gene of *Ag*P450 superfamily is determined as 2.93 and 3.81. Interestingly, the *Drosophila melanogaster* P450s have also similar intron density as 2.97. Intron density for each of the gene family was calculated; provided at least the gene family should consist of a minimum of two genes. The minimum intron density '1' was observed in the CYP9 gene family, while highest intron density '7' was observed with the CYP12 gene family. The CYP325, CYP4, CYP6, CYP9, CYP12, CYP304, CYP305, and CYP307 gene families have intron density 3.5, 3.57, 1.5, 1.0, 7.0, 2.5, 2.0, and 2.0, respectively (Table 1). Single intron genes dominated the *Ag*P450s (Figure 3a). Average number of exons and introns per gene is 3.8 ± 0.207 and 2.8 ± 0.207, respectively.

Intron loss and gain events

Intron loss and gain studies were performed using the 49 informative intronic positions. The maximum number of ambiguous characters per site was 2. A total of 14 intron gains and 67 intron loss events were calculated using the DOLLO parsimony. While using

the probabilistic model, 22 and 177 intron gain and loss events were calculated, respectively, at the terminal taxa. For a branch point, a total of 26,112 for DOLLO parsimony and 178, 287 for probabilistic model were predicted as intron gains and losses.

Correlation between intron length and total gene size, and CDS and frequency of exons

The correlation coefficient between total intron length and total gene length was calculated as r=0.9. The *CYP4C28*, *CYP301A1, CYP305A2*, *CYP306A1*, *CYP307A1*, *CYP325C2*, *CYP325E1*, *CYP325F2, CYP49A1*, *CYP4C25*, *CYP4C26*, *CYP4C27*, *CYP4C37*, and *CYP4G16* genes are having a total gene length of more than 3000 bp, while rest (91 genes) have ~2500 bp total gene size. Among all, the genes from CYP6 and CYP9 gene families have almost similar length, i.e. ~1500 bp. The perfect correlation was observed between the total intron length and total gene length (Online Resource 4a). This observation suggests that although total gene length is varied dramatically among the members of P450 supergene family, it does not have any influence on total length of the coding DNA sequence (CDS). In other words, the total intron length of a gene is the major cause for increase in the total gene length and not the CDS. A weak correlation coefficient (r^2=0.008) was observed between CDS length and frequency of exons, implies that increase in the length of CDS does not affect the frequency of exons occurrence in this supergene family (Online Resource 4b).

Discussion

Insect P450s are a superfamily with many genes that are evolving rapidly [24]. Large scale genome sequencing of the insect species has provided the data to conduct studies on gene organizations that may provide useful information to infer the evolutionary relatedness of the P450 gene(s). Subsequently, the information may help in the functional annotation and basic understanding of evolution of the P450 genes and its implications in insecticide resistance. Compared to plants and mammals, much less is known about the functional aspects of the insect P450s [24].

In spite of expansion of the *Ag*P450 supergene family through extensive duplications followed by a divergent evolution, the total protein length is highly conserved, i.e.500 a.a. The same observation reported from different domains of life [25], suggests that strict constraints exists on the supergene family to maintain the conserved sequence motifs thereby the common 3D structure [25]. The *Ag*P450s contributes ~0.75% to the total predicted genes while, *Drosophila* P450s contribute 0.61%. Analysis of the *Ag*P450s for their evolutionary relatedness using the multiple methods (NJ, MP, ML, and Bayesian) revealed four unambiguously distinguishable clades, namely mitochondrial, CYP2, CYP3, and CYP4 with higher bootstrapped values (Supplementary figure 2b). Phylogenetic analyses using MP, ML, and Bayesian inferences revealed similar topologies.

The multiple sequence alignment (MSA) used for the construction of the phylogenetic tree that includes less conserved regions along with the highly conserved regions, may mask the actual phylogenetic relationship among the members of the supergene family. Moreover, conserved and functionally important sequence motifs have a direct correlation with their evolutionary relationship and the restriction to undergo variations. This is the reason why we have also constructed another NJ based phylogenetic tree using sequence Blocks that are specific to the *Ag*P450 supergene family. Interestingly, the resulted phylogenetic tree revealed a similar topology that is consistent with the most consensus tree (data not shown here) drawn from multiple analysis. The analysis of other insect species (*Drosophila melanogaster, Apis mellifera, Aedes aegypti*) cytochrome P450 superfamily have also revealed identical results by categorizing the P450 superfamily into four distinguishable clades [26,27]. This suggests that the ancestors of these clades might have existed during the early radiation of insects. Probably this may be the first study to test the feasibility of use of only conserved sequence blocks, instead of whole protein sequences, in inferring the evolutionary relationship among the members of P450 supergene family. The PCO analysis revealed very close evolutionary relationship between mitochondrial and CYP2 clans. The higher branch lengths and absence of paralogs for mitochondrial and CYP2 members suggest the absence of recent radiations (Supplementary figure 2b). In contrast, the CYP3 and CYP4 clan gene families revealed small branch lengths with numerous paralogs implies occurrence of numerous recent radiations. The diversity of the P450 supergene family arose by an extensive process of gene duplications [28] and by probable, but less well documented cases of gene amplifications, gene conversions, genome duplications, gene loss and lateral gene transfers [4,7].

In addition to the above phylogenetic analysis, intronic positions were also tested for their suitability in resolving the evolutionary relatedness among the members of the gene family using the DOLLO parsimony. However, the analysis resulted into a phylogeny which is not congruent to the consensus phylogeny that is inferred from using the different methodology (amino acids sequences based NJ, MP and ML, sequence blocks). The DOLLO parsimony method has been successfully used to resolve the deep evolutionary scenario in three major lineages, namely arthropods, nematodes, and deuterostomes [21]. Further, the introns as a phylogenetic marker has been successfully used to infer the phylogenetic relationships having the shorter evolutionary distances [22]. Ambiguity in our analysis may be due to a complex evolutionary scenario associated with this supergene family. In general it has also been noticed that the protein sequences evolve faster than the intronic positions. Thus, we hypothesize that the members of *Ag*P450 sequences might have evolved faster than the intronic sequence, and thus this resulted in failure of intron positions in revealing the observed most consensus phylogeny.

A group of two or more related genes that encode a similar protein in the genome are called as gene clusters. The major evolutionary force behind the formation of gene clusters is gene duplications. Analysis of correlation between the genes in a gene cluster and their phylogenetic relationships along with exon-intron organizations revealed that the genes belonging to gene clusters have monophyletic phylogenetic relationship with conserved gene organizations (Supplementary figures 2b and 2c). Occurrence of genes in gene clusters and monophyletic phylogenetic relationship among the members of gene clusters highlights the importance of gene duplications in the P450 gene families' expansion. Two largest gene clusters that are formed by the CYP325 and CYP6 gene families are involved in the xenobiotic metabolism [12]. The genes within the gene clusters may undergo various evolutionary selections like gene conversions, concerted evolution, and divergent evolution etc. It is known that some sequence tracks of P450 sequences undergo the gene conversions. In order to confirm any influence of the concerted evolution on the evolution of the *Ag*P450s, we have calculated the net sequence diversity within each gene family using MEGA v4.0 software. The assumption is that if there is any active role of a concerted evolution, the gene family members should be homologous and thus the sequence diversity should be converging (i.e. less). It is calculated only if the gene family has more than two members; irrespective of their location in the genome i.e. the diversity between two genes belonging to the same gene family but are apart in the genome is also calculated.

Name of the gene family	Amino acids substitution rates per site	S. E.
cyp304	0.8541	0.09351
cyp305	0.3113	0.122111
cyp325a	0.0987	0.017162
cyp325d	0.0308	0.054962
cyp325c	0.0984	0.081309
cyp325f	0.1386	0.025258
cyp4g	0.4310	0.058096
cyp4j	0.6516	0.073114
cyp4h	0.6134	0.057189
cyp4d	0.4985	0.101897
cyp4c	0.5898	0.058382
cyp9m	0.1718	0.031889
cyp9j	0.6878	0.068146
cyp9l	0.3725	0.065971
cyp6ag	0.9946	0.107993
cyp6z	0.3147	0.05455
cyp6aa	0.3170	0.067171
cyp6s	0.3821	0.076487
cyp6p	0.4251	0.070205
cyp6y	0.6158	0.061595
cyp6n	0.5877	0.113919
cyp6m	0.3804	0.036492
cyp12	0.6097	0.135246

Table 2: Estimates of average evolutionary divergence over sequence pairs within groups.

The number of amino acid differences per sequence from an estimation of net average between the groups of sequences is shown in table 2. The analysis involved 94 amino acids sequences and a total of 230 positions in the final dataset. All positions containing gaps and missing data were eliminated. The analysis failed to see any convergence (sequence homogenization) and thus we hypothesize that this supergene family which is expanded majorly through CYP3 and CYP4 clans may be due to adaptive evolutionary mechanism to cope with the numerous xenobiotics to which the insects are continuously get exposed. And thus the duplicated genes might never have undergone homogenization after their duplication to preserve their diverse substrate specificities to effectively metabolize numerous classes of xenobiotics. Analysis of P450 supergene family from vertebrates' revealed birth and death, and divergent evolution are the two main forces that are driving the evolution of the P450 gene families.

The mitochondrial, CYP2, CYP3, and CYP4 clans are having 6, 6, 3 and 2 families respectively. Although CYP2 and mitochondrial clans are contributing only ~18% of total P450 genes, they are more diverse (6 gene families each) than CYP3 (3 gene families) and CYP4 (2 gene families) clans. The close sequence similarity of the mitochondrial P450s to CYP2 clan members may be due to their evolutionarily shared functionality, i.e. mediating the molting hormone biosynthesis [29]. Ancestor for the mitochondrial P450 genes might be originated in the CYP2 clade [10]. The mitochondrial P450 genes can be categorized into two groups based on their cellular functions and sequence conservation across different taxa (orthology). The first group includes the conserved genes that are involved in molting hormone biosynthesis [30], and the second group includes the genes which are non-conserved, having species-specific gene expansion, and are possibly involved in the xenobiotic metabolism (for example, CYP12 family members) [10,12,31,32]. The evolutionarily conserved Halloween genes encode the P450 enzymes which mediate the biosynthesis of 20-hydroxyecdysone (20E) [30]; *CYP306A1* (Phantom), *CYP302A1* (Disembodied), *CYP315A1* (Shadow), *CYP314A1* (Shade), *CYP307A1* (Spook), *CYP307A2* (Spookier), *CYP307B1* (Spookiest) are the Halloween genes and are encoded by the mitochondrial and CYP2 clans [29,30]. Another enzyme, CYP301, albeit it is highly conserved across different taxa its function is yet to be elucidated [10].

The homologous genes (orthologs and paralogs), in general have conserved gene organizations including the phases of introns. However, due to less selection pressure and simpler nature of non-coding sequences of introns, they may vary in sequence and length [33]. The protein sequences of CYP2 and mitochondrial clans have very close sequence similarity than to any other gene families of the *Ag*P450s; although within and between the gene families, gene organizations of both the clans are not conserved (Supplementary figures **2**b and 2c). The inferred gene structures of *Ag*P450s were remarkably divergent within the mitochondrial, CYP2, and to some extent among certain subfamilies of the CYP4 clans. In a sharp contrast, the gene structures of CYP3 clan (introns of CYP6 and CYP9 gene families) are highly conserved (Supplementary figure 2). This may be due to a lack of expansion of the gene families through gene duplications because of their evolutionarily conserved physiological functions thereby evolutionary restrictions on their expansions [10,25]. Most of the eukaryotic P450 genes are with discontinuous gene organizations (i.e. genes are interrupted by introns) that are generally conserved within the gene family but differ dramatically between gene families in most of the cases [34]. Of the 105 total putatively active P450 genes of *An. gambiae*, 4 are intron less, while *D. melanogaster* consists of 5 intron less genes [8]. The highest numbers of introns are found in the CYP4 gene family. This is not surprising given the fact that this family of genes spread all over the genome and lacks large tandem gene organizations (Online Resource 2). In this gene family, the intron positions are highly variable and corresponding total gene lengths are also varying abruptly. In CYP6 and CYP9 gene families' intron positions and intron phases are highly conserved along with the corresponding total gene lengths. Most of the genes of CYP6 gene family members are having single, phase '1' intron at 550th position on consensus P450 sequence with a conserved glutamic acid (Glu/E) amino acid at the insertion site. In *Drosophila*, almost all the predicted CYP6 family members' gene structures are also possess the single, phase 1 conserved intron [8]. Gene structures for meprin and TRAF homology (MATH) domains from *Arabidopsis thaliana* and nematode *Caenorhabditis elegans* are also surprisingly similar. A total of 87 out of 98 proteins with known gene structures contain a phase 2 intron at the same position [35]. Remarkable conservation of intron position and phase is probably due to the deleterious effect of intron loss and is still to find if they have any functional role [36]. These observations suggest that loss of conserved introns may be more costly for these genes than their retention. In order to assess the role of conserved introns in the gene family evolution and/or functional-regulatory influences if any, the CYP6 and CYP12 gene families' introns were analyzed. For this, CYP6 and CYP12 genes introns were retrieved and MSA was performed individually for each of the gene family. Firstly, the aligned introns were used to identify the regions of high conservation between any two introns. However, not a single pair of introns is having reasonable sequence conservation among the analyzed genes from the CYP6 genes. The highest nucleotide identity was observed between two genes, namely CYP6Z2 and Z3 genes. These two genes might be the result of recent duplication event as two genes are 93.71% identical at amino acids level. Similar analyses were also performed on the conserved seven intronic positions of the CYP12 gene family. All except one intron from CYP12F3 have <100 bp intron. Interestingly, comparison

of the nucleotide sequences revealed that all the introns of CYP12 gene family might have a common origin. For example, first and sixth introns of the CYP12F4 gene, and seventh intron of the CYP12F3 and fifth intron of the CYP12F1 have shared 55.7% and 55.2% nucleotide identities, respectively. This study identifies the need of systematic functional evaluation of the single phase '1' intron which is highly conserved in both *An. gambiae* and *D. melanogaster* CYP6 gene family. In a further attempt to understand the reason for high conservation of intron positions and phases in CYP6 and CYP12 gene families, intronic sequences were searched for microsatellites, regulatory sequences like mi-RNAs, and signatures of gene conversions. However, analyses failed to identify any significant information that can explain why introns are conserved in these important gene families for the cause of insecticide resistance. One of the possible reasons for high conservation of the CYP6 families' single, phase 1 intron is "forcible conservation". This is so because the conserved intron is found to be located within the heme-ligand signature sequence which is a hallmark of the P450 sequences and is involved in substrate recognition. Any disturbance in the intron either by accumulating more mutations or its excision may cause a loss of function of these one of the functionally important genes.

The CYP4 and CYP325 gene families together consist of 157/296 introns. The position and phase of introns within the gene families are conserved (Supplementary figure 2c) but the intron lengths are highly variable (Online Resource 3). The CYP307 gene family has two genes with two intron positions with conserved intron phase and position. But the first intron of *CYP307A1* has expanded exceptionally after duplication (Online Resource 3). The small introns (<100 bp) have dominated the gene organization of *Ag*P450s (Figure 3b). In *Drosophila* and *Arabidopsis,* introns are much shorter than that of human introns and the most abundant size is 50-100 nt [37]. It has been noticed that the members of subfamily or family with conserved intron position and phase does not necessarily have conserved intronic length. There are 21 introns whose length is more than 1000 bp. There may be selection on longer introns to maintain pre-mRNA secondary structure that contribute to the formation of RNA secondary structure involved in gene expression [38]. Similar to *D. melanogaster*, the *Ag*P450s can be primarily explained as more intron loss and few intron gains. The amount of intron loss is not surprising considering the extensive loss of introns seems to have occurred in insects, probably as a result of evolutionary selection for compact insect genomes. Dollo parsimony allows only one insertion of an intron at each site. Even if more than one insertion has taken place, the model accommodate only one insertion event at an ancestral time point. Model ignores the intron that could have lost in all the extant lineages, and does not consider the possibility of parallel intron gains or successive gain-loss-gain events [23]. For this reason, we have also compared the intron losses and gains using probabilistic rate models that can be used directly to estimate the ancestral intron presence, as well as intron gains and losses. Finally, the analysis suggests that the exon-intron structure of the cytochrome P450 superfamily is highly dynamic and this superfamily must have undergone numerous intron losses and gains during its intriguing evolution process.

Conclusions

The phylogenetic analysis of *Ag*P450 supergene family revealed four unambiguously distinguishable clades, namely mitochondrial, CYP2, CYP3, and CYP4. Our analysis suggests cautious use of the Dollo parsimony analysis that uses intronic positions as phylogenetic informative characters to find out evolutionary relationship among the members of gene families that are evolving through gene duplications. The gene organization of the *Ag*P450 supergene family revealed that the exon-intron organizations are highly conserved within subfamilies and gene families, and between the gene families for some intronic positions. Furthermore, gene organizations and gene clusters formation have closely followed the phylogenetic relationship that is inferred using the amino acid sequences. Lack of clear orthologs of CYP4 and CYP6 families' genes and their occurrence in a large gene clusters with conserved gene organizations and gene orientations further supports the species-specific expansion of the *Ag*P450 supergene family through extensive gene duplications. The generated gene organization data of *Ag*P450 could be used for isolating and annotating more similar function genes in related organisms. The description of a protein family by its conserved regions focuses on the gene family's characteristics and distinctive sequence features. Thus, the generated information can very well be employed for designing the degenerate primers to isolate the P450 gene sequences from the insect species for which genome sequence information is yet to be generated.

Acknowledgements

BPNR was supported by Council of Scientific and Industrial Research (CSIR)-SRF fellowship.

References

1. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129-149.
2. Hemingway J, Field L, Vontas J (2002) An overview of insecticide resistance. Science 298: 96-97.
3. Califano A (2001) Advances in sequence analysis. Curr Opin Struct Biol 11: 330-333.
4. Yadav JS, Doddapaneni H, Subramanian V (2006) P450ome of the white rot fungus Phanerochaete chrysosporium: structure, evolution and regulation of expression of genomic P450 clusters. Biochem Soc Trans 34: 1165-1169.
5. Scott JG, Wen Z (2001) Cytochromes P450 of insects: the tip of the iceberg. Pest Manag Sci 57: 958-967.
6. Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, et al. (1993) The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes and nomenclature. DNA Cell Biol 12: 1-51.
7. Werck-Reichhart D, Feyereisen R (2000) Cytochromes P450: a success story. Genome Biol 1: 3003.1-3003.9.
8. Tijet N, Helvig C, Feyereisen R (2001) The cytochrome P450 gene superfamily in Drosophila melanogaster: annotation, intron-exon organization and phylogeny. Gene 262: 189-198.
9. Ranson H, Nikou D, Hutchinson M, Wang X, Roth CW, et al. (2002) Molecular analysis of multiple cytochrome P450 genes from the malaria vector, Anopheles gambiae. Insect Mol Biol 11: 409-418.
10. Feyereisen R (2006) Evolution of insect P450. Biochem Soc Trans 34: 1252-1255.
11. Amenya DA, Koekemoer LL, Vaughan A, Morgan JC, Brooke BD, et al. (2005) Isolation and sequence analysis of P450 genes from a pyrethroid resistant colony of the major malaria vector Anopheles funestus. DNA Seq 16: 437-445.
12. David JP, Strode C, Vontas J, Nikou D, Vaughan A, et al. (2005) The Anopheles gambiae detoxification chip: A highly specific microarray to study metabolic-based insecticide resistance in malaria vectors. Proc Natl Acad Sci U S A 102: 4080-4084.
13. Feyereisen R (1999) Insect P450 enzymes. Annu Rev Entomol 44: 507-533.
14. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
15. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

16. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599.

17. Charif D, Thioulouse J, Lobry JR, Perrière G (2005) Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21: 545-547.

18. Pietrokovski S, Henikoff JG, Henikoff S (1996) The Blocks database-a system for protein classification. Nucleic Acids Res 24: 197-200.

19. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 163: GC17-GC26.

20. Farris JS (1977) Phylogenetic analysis under Dollo's Law. Syst Zool 26: 77-88.

21. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13: 1512–1517.

22. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform 6: 118-134.

23. Csuros M (2005) Likely scenarios of intron evolution. Lecture Notes in Computer Science 3678: 47-60.

24. Chung H, Sztal T, Pasricha S, Sridhar M, Batterham P, et al. (2009) Characterization of Drosophila melanogaster cytochrome P450 genes. Proc Natl Acad Sci U S A 106: 5731-5736.

25. Feyereisen R (2005) Insect cytochrome P450. Comprehensive Molecular Insect Science 4: 1–77.

26. Claudianos C, Ranson H, Johnson RM, Biswas S, Schuler MA, et al. (2006) A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. Insect Mol Biol 15: 615-636.

27. Strode C, Wondji CS, David JP, Hawkes NJ, Lumjuan N, et al. (2008) Genomic analysis of detoxification genes in the mosquito Aedes aegypti. Insect Biochem Mol Biol 38: 113-123.

28. Ranson H, Claudianos C, Ortelli F, Abgrall C, Hemingway J, et al. (2002) Evolution of supergene families associated with insecticide resistance. Science 298: 179-181.

29. Rewitz KF, Rybczynski R, Warren JT, Gilbert LI (2006) The Halloween genes code for cytochrome P450 enzymes mediating synthesis of the insect moulting hormone. Biochem Soc Trans 34: 1256-1260.

30. Rewitz KF, O'Connor MB, Gilbert LI (2007) Molecular evolution of the insect Halloween family of cytochrome P450s: phylogeny, gene organization and functional conservation. Insect Biochem Mol Biol 37: 741-753.

31. Guzov VM, Unnithan GC, Chernogolov AA, Feyereisen R (1998) CYP12A1, a mitochondrial cytochrome P450 from the house fly. Arch Biochem Biophys 359: 231-240.

32. Strode C, Steen K, Ortelli F, Ranson H (2006) Differential expression of the detoxification genes in the different life stages of the malaria vector Anopheles gambiae. Insect Mol Biol 15: 523-530.

33. de Souza SJ, Long M, Gilbert W (1996) Introns and gene evolution. Genes Cells 1: 493-505.

34. Gotoh O (1993) Structure of P450 genes. Tokyo: Kodansha.

35. Betts MJ, Guigo R, Agarwal P, Russell RB (2001) Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? The EMBO Journal 20: 5354-5360.

36. Fedorova L, Fedorov A (2003) Introns in gene evolution. Genetica 118: 123-131.

37. Fedorov A, Roy S, Fedorova L, Gilbert W (2003) Mystery of intron gain. Genome Res 13: 2236-2241.

38. Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ (2005) Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics 170: 661-674.