



Genomic organization of the glutathione S-transferase family in insects

Robert Friedman*

Department of Biological Sciences, University of South Carolina, 715 Sumter Street, Columbia, SC 29208, USA

ARTICLE INFO

Article history:

Received 29 April 2011

Revised 28 August 2011

Accepted 30 August 2011

Available online 10 September 2011

Keywords:

Glutathione transferases

Gene duplication

Genomic organization

ABSTRACT

Cytosolic glutathione S-transferases (GSTs) are a large and diverse gene family in insects. They are classified into six major subclasses. Sigma, Omega, Zeta, and Theta have representatives across Metazoa while Delta and Epsilon are specific to Insecta and Holometabola, respectively. In this study, GSTs are assigned to a subclass by a combination of literature, phylogenetic, and genomic evidence. Moreover, it is confirmed that GSTs frequently cluster by genomic position as a result of recent gene expansions. These expansions are largely explained by the number of protein-coding genes in the genome, although life history is another contributing factor.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Glutathione S-transferases (GSTs) are a gene family of enzymes involved in detoxifying cells of natural and artificial molecules (Lumjuan et al., 2005; Ranson et al., 2000; Rogers et al., 1999; Tang and Tu, 1994). Examples include GST-D1 in *Drosophila melanogaster* and an unidentified number of GSTs in *Pediculus humanus* which are associated with resistance to insecticides (Barrios et al., 2010; Tang and Tu, 1994). GSTs mainly metabolize these substances by making them more water soluble, aiding in their excretion from the organism (Habig et al., 1974).

In insects, there are two major classes of GSTs, microsomal and cytosolic. The membrane-bound microsomal type is structurally and evolutionarily distinct from the cytosolic type (Enayati et al., 2005). The microsomal class contains few gene duplicates – such as *D. melanogaster* with only one member (Ranson et al., 2002; Toba and Aigaki, 2000).

This study is focused on the larger family of cytosolic GSTs. The cytosolic type has an active site of catalytic activity and typically forms a dimeric complex. These GSTs are characterized by two domains (Pfam domains PS50404 and PS50405; Hulo et al., 2006), a N-terminal domain of ~80 amino acid sites and a C-terminal domain of ~120 sites. The N-terminal domain contains a site which interacts with a glutathione molecule. The C-terminal domain contains a site which interacts with the substrate (review by Enayati et al. (2005)).

This highly diverse insect gene family is divided into six major subclasses: Delta, Epsilon, Sigma, Omega, Theta, and Zeta. Among the currently available genomes, the dipterans consistently have

large expansions of these genes. For instance, Ding et al. (2003) identified 37 putative cytosolic GSTs in *D. melanogaster* and 28 in *Anopheles gambiae* in contrast to the hymenopteran *Apis mellifera* with only 11 GSTs (Claudianos et al., 2006; Honeybee Genome Sequencing Consortium, 2006). Similar to the eusocial hymenopteran, the exopterogote parasite *P. humanus* has 11 GSTs which span five of the subclasses (Kirkness et al., 2010). However, this pattern is not a trend in exopterogotes as the free-living aphid *Acyrtosiphon pisum* contains 18 putative GSTs (Ramsey et al., 2010; Aphid Genomics Consortium, 2010). The current sample of insect genomes is currently insufficient to distinguish whether the GST expansion pattern is mainly driven by shared ancestry or by ecological adaptation to a niche (i.e., parasitism).

The bulk of the GST expansions are in the Delta and Epsilon subclasses. These are Insecta-specific while the other four subclasses have a broader taxonomic distribution (Low et al., 2007; Ranson et al., 2001, 2002; Sawicki et al., 2003). The overall high rate of turnover of GST genes is especially apparent from studies of the 12 drosophilid genomes (Low et al., 2007; *Drosophila* 12 Genomes Consortium, 2007). The turnover does not occur abruptly over time but instead conforms to a somewhat uniform rate in the loss and gain of genes. This is not surprising since Hahn et al. (2007) estimated that ~40% of drosophilid gene families differed in size among lineages and ~20% of gene families were lost altogether in one or more lineages.

Moreover, Severson et al. (2004) suggested that gene order across 50 kilobases has a one-half probability of being conserved between *D. melanogaster* and Culicidae. Therefore, genetic linkage of GSTs was examined across each genome (Fig. 1). In this study, a diverse set of insect genomes was examined and expands upon previous studies of GSTs which generally were restricted to a genus-level analysis. This set includes: four Diptera, two Hymenoptera, one Coleopteran, one Lepidopteran, and two outgroup species of

* Fax: +1 803 777 4002.

E-mail address: bobf@biol.sc.edu

Exopterogota. The aim was to gather evidence across Insecta of gene duplication and linkage patterns in this large and diverse gene family.

2. Methods

2.1. Sequence retrieval

Insect protein-coding sequences were available from VectorBase (Lawson et al., 2009), FlyBase (Drysdale and the FlyBase Consortium, 2008), BeeBase (Munoz-Torres et al., 2010), SilkDB (Wang et al., 2005), Baylor College of Medicine (Savard et al., 2006), AphidBase (Leggai et al., 2010) and BeetleBase (Wang et al., 2007). This study utilized the following genome assemblies from the above databases: *Aedes aegypti* version AaegL1, *A. gambiae* v. AgamP3.5, *Culex quinquefasciatus* v. Cpipj1.2, *A. mellifera* v. pre-release 2, *Tribolium castaneum* v. Tcas 2.0, *D. melanogaster* v. 5.13.58, *Nasonia vitripennis* v. 1, *Bombyx mori* v. 2.0, *A. pisum* v. ACYPI and *P. humanus* v. 1.2. *Homo sapiens* protein sequences were obtained from Ensembl (Flicek et al., 2010). A complete physical map was available for *T. castaneum*, *A. gambiae* and *D. melanogaster* genomes.

Alternatively transcribed proteins were removed wherever there was sufficient information to show that there were multiple proteins transcribed at a single locus. In addition, the genomic position of genes was used to identify gene duplication events. In other words, close proximity of highly similar genes is likely a result of tandem or segmental duplication.

Also, it was verified whether unmapped GSTs were unique loci and not redundant with those already mapped. For instance, AGAP012702 and *Anopheles*GSTD5 appear identical at the amino acid level (Fig. 3 tree), but they actually differ by five nucleotide substitutions and therefore were not predicted to be at the same locus. Likewise, AGAP012838 and *Anopheles*GSTD12 appear nearly identical at the amino acid level, yet differ by twelve nucleotide substitutions.

2.2. Size of proteomes

Numbers of protein-coding genes in the culicids were obtained from: Arensburg et al. (2010); *A. mellifera* from the Honeybee Genome Sequencing Consortium (2006); *T. castaneum* from the Tribolium Genome Sequencing Consortium (2008); *D. melanogaster* from the Drosophila 12 Genomes Consortium (2007); *B. mori* from the Silkworm Genome Consortium (2008); *N. vitripennis* from Werren et al. (2010); *A. pisum* from the Aphid Genomics Consortium (2010); *P. humanus* from Kirkness et al. (2010). The *A. pisum* gene count is based on the “Official Gene Set”, a combined set of NCBI RefSeq genes and non-redundant GLEAN gene predictions. The Aphid Genomics Consortium (2010) noted that this set included partial and unsupported gene models. Therefore, this species was excluded from the correlation analysis.

2.3. Classification of gene families

The protein-coding genes of Endopterogota and *P. humanus* ($n = 137,945$) were clustered into gene families of size ≥ 2 ($n = 14,743$) by BLASTCLUST (Altschul et al., 1997); default values were used, except for a change so the pairwise matching procedure required amino acid similarity at or greater than 40% across at least 60% of their sequence lengths (Hughes et al., 2005). Once all pairs of similar amino acid sequences were collated, then the software clustered them into families by a single-linkage method. This procedure is optimized for assembling gene family sets which have many informative sites, leading to robust phylogenetic tree construction (Hughes et al., 2005). The above family sets with

known glutathione S-transferase genes (GSTs) were used as a core set of GSTs. This above procedure was repeated by searching with this core set against protein sequences of the more distant vertebrate *H. sapiens* ($n = 71,913$) and the pea aphid *A. pisum* ($n = 34,821$).

To find GSTs missed by the above BLAST procedure (Altschul et al., 1997), a search was performed for GST-type domains among the set of protein-coding genes. The Prosite database (Hulo et al., 2006) and search tool (Gattiker et al., 2002; Swiss Institute of Bioinformatics) were utilized to identify genes with these two GST domains: PS50404 and PS50405; search criteria were set at default values. Perl scripts parsed the Prosite output into a computer-readable format. The GSTs identified by this procedure were added to the data set except where noted below.

In *D. melanogaster*, these genes were detected by GST domain analysis but not verified by BLAST or phylogenetic analysis: FBpp0084761 (FlyBase annotation: translation elongation factor), FBpp0085809 (FlyBase annotation: methionine-tRNA ligase), and FBpp0071969 (FlyBase annotation: unknown function). These three genes were not considered GSTs and the domain match considered a spurious result. Also, FBpp0076870 (CG4623) was excluded since it was not identified by domain analysis nor sequence similarity (Ranson et al., 2001), even though it was included as a GST by Low et al. (2007).

Unclassified GSTs were assigned to a subclass based upon phylogenetic-, linkage-, and BLAST-based evidence (Altschul et al., 1997; Table 1, Supplementary table, and Figs. 1 and 2). These assignments were annotated in the Supplementary table and verified with that in the relevant literature. GSTs were assigned to an “Unclassified” subclass (i.e., GSTu1) in cases where they were not confidently assigned to one of the six major subclasses (see column *Other name* in the Supplementary table) or in cases where the literature provided an unrefuted classification. Also, sequences too short in length for a sequence alignment were labeled as *partial* in the Supplementary table and excluded from the phylogenetic trees.

2.4. Reconciling the data set with previous studies

In *A. pisum*, Ramsey et al. (2010) identified 10 Delta-GSTs, six Sigma-GSTs, and two Theta-GSTs. Their GSTs were verified by gene name and an additional five partial GSTs were identified (Supplementary table). This procedure was supplemented by a TBLASTN search of the genomic data (Altschul et al., 1997). However, *A. pisum* sequences were excluded from the Delta/Epsilon-GST tree because these members were divergent at the amino acid sequence-level and there was insufficient evidence to consider them as members of these two subclasses.

In *B. mori*, Yu et al. (2008) identified 23 cytosolic GSTs by TBLASTN searches (Altschul et al., 1997). However, this study's GST list did not fully overlap with their list. For instance, their GSTz2 is listed as a Zeta subclass member, yet it was observed to be very divergent from other Zeta-GST members (not shown). Therefore, this gene sequence was excluded from phylogenetic analysis.

In *N. vitripennis*, Oakeshott et al. (2010) identified a Zeta-GST which was not predicted in the set of gene models, so this gene was included (located on scaffold 4). Also, their GSTS5 was reconciled as XP_001605456 instead of XP_001605470 since these two gene models are overlapping at a single gene locus.

2.5. Phylogenetic analyses

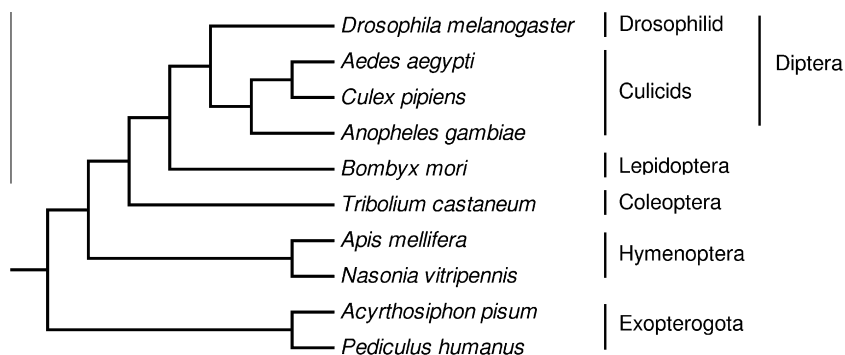
Each glutathione S-transferase (GST) subclass was globally aligned by MUSCLE (ver. 3.8.31; Edgar, 2004) so a phylogenetic analysis could generate an evolutionary tree. This analysis was

Table 1

List of glutathione S-transferase (GST) gene clusters among the Insecta genomes.

Species (total # GSTs)	Cluster name	# GSTs in cluster	Chromosome/scaffold	Start position	End position
Endopterothotes					
<i>Aedes aegypti</i> (29) 15,419 protein-coding genes	Aa-1 (E)	8	1.291	894246	947890
	Aa-2 (D)	4	1.22	2894345	3015437
	Aa-3 (T)	3	1.362	1358886	1366897
	Aa-4 (D)	2	1.488	208815	210903
	Aa-5 (D)	2	1.22	1422025	1424050
<i>Anopheles gambiae</i> (32) 12,457	Ag-1 (E)	8	3R	28591663	28602188
	Ag-2 (D)	6	2R	55501957	55513812
	Ag-3 (D)	6	2R	50990112	51218384
	Ag-4 (D)	2	Unknown	33265988	33267405
<i>Tribolium castaneum</i> (33) 16,404	Tc-1 (E)	12	2	6358378	6386815
	Tc-2 (E)	6	3	2254934	2270560
	Tc-3 (S)	3	3	122680	127150
	Tc-4 (O)	2	3	14082643	14086376
<i>Culex quinquefasciatus</i> (39) 18,883	Cp-1 (D)	10	3.36	411773	448594
	Cp-2 (E)	10	3.1224	29909	46551
	Cp-3 (T)	4	3.500	205309	211831
	Cp-4 (S)	2	3.109	396925	412644
<i>Drosophila melanogaster</i> (37) 13,733	Dm-1 (D)	11	3R	8190635	8211256
	Dm-2 (E)	11	2R	14284248	14392330
	Dm-3 (O)	4	3L	8512054	8517902
	Dm-4 (Z)	2	3R	5281310	5284539
	Dm-5 (T)	2	2R	5493592	5495730
<i>Bombyx mori</i> (21) 14,623	Bm-1 (O)	3	nscaf 3031	2572527	2587558
	Bm-2 (D)	2	nscaf 2853	6789865	6799251
	Bm-3 (D)	2	nscaf 2279	197459	214701
	Bm-4 (S)	2	nscaf 2931	1432778	1456301
	Bm-5 (E)	2	nscaf 2986	6071116	6081727
<i>Nasonia vitripennis</i> (18) 17,279	Nv-1 (S)	2	NW_001816681	1057805	1060453
	Nv-2 (D)	2	NW_001820527	2557631	2563551
	Nv-3 (T)	2	NW_001820638	260237	264692
<i>Apis mellifera</i> (11) 10,157	-	-	-	-	-
Exopterothotes					
<i>Acyrtosiphon pisum</i> (24) 34,604	Ap-1 (D)	3	EQ122826	539	18433
	Ap-2 (S)	2	EQ111817	57033	69411
	Ap-3 (D)	2	EQ117811	752	44813
	Ap-4 (D)	2	EQ121193	1507	13636
	Ap-5 (D)	2	EQ123931	3108	26319
<i>Pediculus humanus</i> (11) 10,773	Ph-1 (S)	3	DS235271.1	373809	387912
	Ph-2 (D)	2	DS235150.1	52977	57929

Abbreviations for subclass in "Cluster name" column: S = Sigma; O = Omega; Z = Zeta; T = Theta; D = Delta; E = Epsilon.

**Fig. 1.** Rooted cladogram of 10 insect species (Gaunt and Miles, 2002; Tamura et al., 2004; Honeybee Genome Sequencing Consortium, 2006; Savard et al., 2006; Drosophila Genomes Consortium, 2007; Wiegmann et al., 2009). A cladogram is not drawn to a scale.

based upon an appropriate model of sequence evolution which was determined by PROTTEST (ver. 2.4; Abascal et al., 2005). The result showed that LG + I + G (Le and Gascuel, 2008) as the best model that fits to the sequence data set and was supported by all software-generated confidence measurements. The I + G is an abbreviation for invariant sites plus the gamma distribution of

substitution rates among sites. However, the LG + G (four categories of the rate G) was chosen and the parameter for invariant sites was removed because I and G are correlated (Yang, 2006).

RaxML (ver. 7.2.6; Stamatakis, 2006) was used to construct maximum likelihood phylogenies (LG + G model) from the amino acid sequence alignments. The support values were calculated by

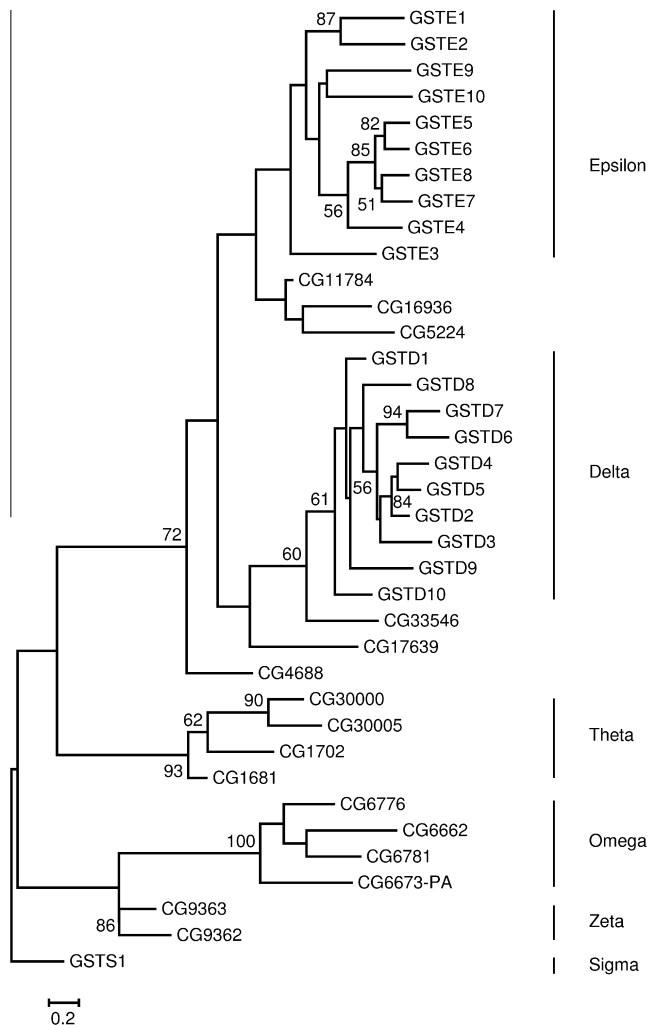


Fig. 2. Rooted phylogenetic tree of glutathione S-transferases in *Drosophila melanogaster*. The six subclasses are labeled on the right-hand side. Bootstrap values are placed at the nodes where the value is ≥ 50 . Scale bar shows the inferred amino acid distance.

a “rapid bootstrapping” procedure (Stamatakis et al., 2008). One thousand bootstrap sequence pseudoreplicates were generated to compute a support value at each node (Felsenstein, 1985); the exception was the condensed topology where five hundred bootstrap pseudoreplicates were generated (Supplementary Fig. 1; Felsenstein, 1989). For the Fig. 5A phylogeny, the 5' portion of the gene sequence alignment was deleted because it caused long branch attraction which led to the highly significant clustering of the *B. mori* and *A. aegypti* GSTs. This edit led to a topology more consistent with the expected species tree.

Finally, four of the six subclasses, Theta, Zeta, Omega, and Sigma, were rooted with a GST from the outgroup *H. sapiens*. This outgroup helped to establish the phylogenetic-based evidence of an unassigned GST belonging to a major subclass. However, this is a somewhat liberal procedure as it is not known whether a particular insect GST evolved faster than the human homolog or instead is older in origin than the vertebrate/insect divergence in Metazoa.

3. Results

The results include phylogenetic trees of the glutathione S-transferases (GSTs) in Insecta and a table of their linkage patterns. Fig. 2 has an overview of the six GST subclasses in *D. melanogaster*:

Sigma, Theta, Omega, Zeta, Delta and Epsilon. The subclasses are defined by the vast biological literature which is supplemented by phylogenetic-, linkage-, and BLAST-based evidence (Altschul et al., 1997; Corona and Robinson, 2006; Ranson et al., 2001). Trees are shown of each of these subclasses for the class Insecta (Figs. 3–6) along with any applicable outgroup, such as provided by the vertebrate.

3.1. Genomic clustering of GSTs

Table 1 enumerates the GST clusters by genomic position. As compared to the other insects, larger clusters are observed among the dipterans and the sole coleopteran. Moreover, the clusters are generally distributed among different chromosomes as opposed to localizing on a single chromosome (Table 1). These clusters likely formed by tandem and segmental gene duplication since they tend to contain members of a single GST subclass with high sequence similarity among members (see figures and table). Unfortunately, the genomic map in *A. mellifera* is insufficient to identify its clusters.

Wherever possible, unclassified GSTs were assigned to a subclass by phylogenetic-, linkage-, and BLAST-based evidence (Altschul et al., 1997). For instance, AGAP009190 is unassigned and named GSTU4 by the genome database, yet located in linkage with an Epsilon-GST (GSTE6). In addition, there are six other Epsilon-GSTs arrayed in a tight cluster with AGAP009190. Therefore, this linkage pattern (along with sequence similarity within this cluster) is evidence that this unassigned gene is an Epsilon-GST. The alternative is that AGAP009190 is not an Epsilon-GST yet evolution resulted in a recent genome rearrangement leading to its co-localization with the other Epsilon-GSTs. However, one reason for discounting this possibility is that the majority of clusters contain members of a single GST subclass with high sequence similarity. Therefore, AGAP009190 is labeled as an Epsilon-GST (also, see Ding et al., 2003; Fig. 3).

Furthermore, *D. melanogaster* CG5224, CG16936, and CG11784 are assigned to the Epsilon subclass by phylogenetic- along with any synteny-based evidence (see Supplementary table; Claudianos et al., 2006). Likewise, CG17639 is assigned to the Delta subclass and CG33546 to the Delta/Epsilon superclass (Supplementary table). Where there is no evidence for a subclass assignment, but there is a literature assignment, then the latter source is used to classify the gene. For instance, Lumjuan and others (2007) assigned unclassified GSTs to a superclass (Delta, Epsilon, several Unclassified members, and the minor subclasses Xi, Iota) and, in this study, these assignments are used so a common vocabulary exists to document the GSTs.

3.2. Correlating GST family size with proteome size

Among the culicids, Arensburger et al. (2010) identified a higher gene count in *C. quinquefasciatus* as compared to *A. aegypti* and *A. gambiae*. This observation was credited to large gene family sizes. However, the Delta/Epsilon-GST tree does not support this pattern. Although there are a few more copies in *C. quinquefasciatus* (24) than in *A. aegypti* (20), *A. gambiae* has the most (25). Instead, it is expected that the *Culex* species should have more GST members since it has a larger proteome. However, with the power of a large number of insect taxa, a positive trend is observed between the total number of genes in the genome and the number of GSTs present ($R = 0.60$, $p < 0.05$, one-tailed test).

3.3. Phylogenies of the Delta and Epsilon subclasses

Fig. 2 shows a close relationship between the Delta- and Epsilon-GST subclasses as evidenced by their sharing a common branch

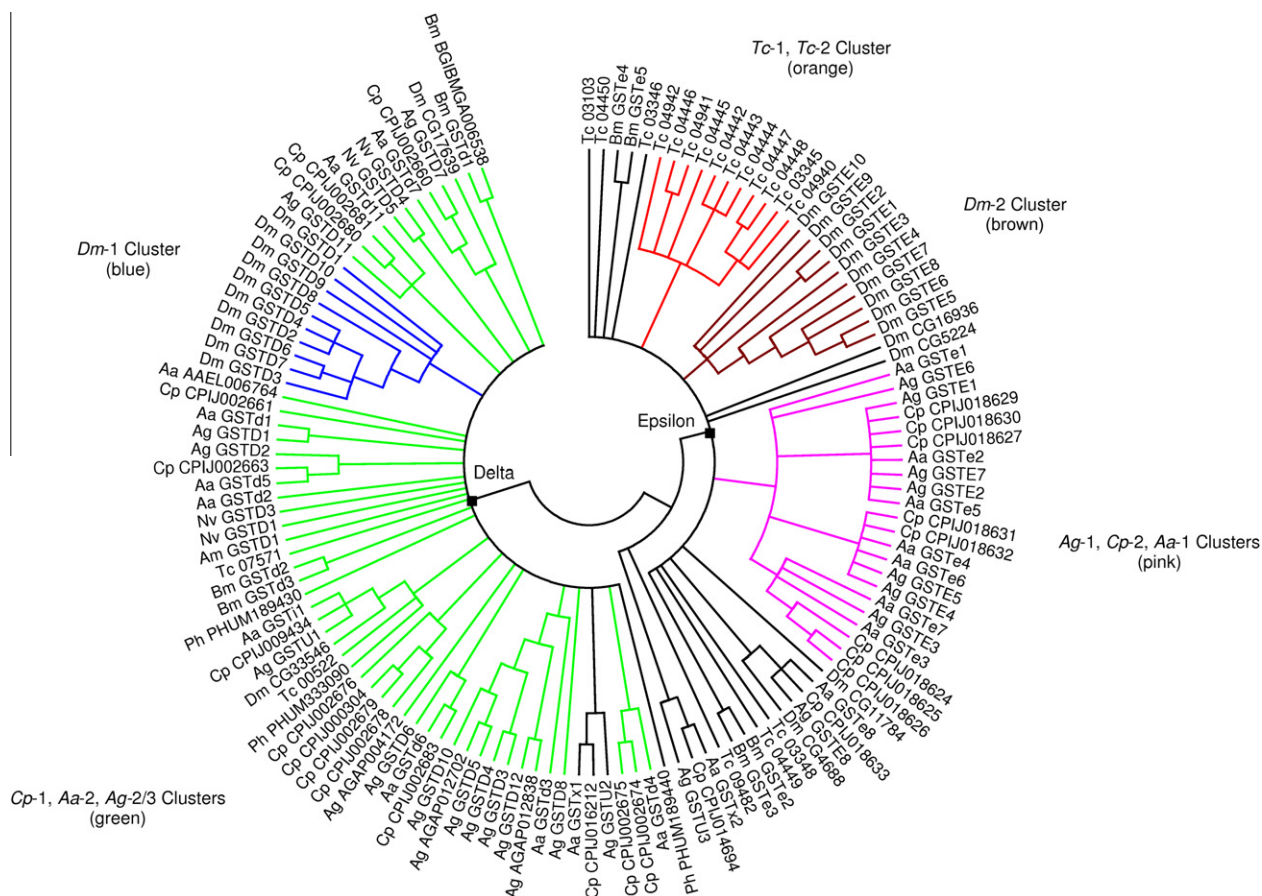


Fig. 3. Unrooted cladogram of the Delta/Epsilon-GST superfamily. The topology is based on a 75% condensed tree obtained by bootstrap analysis. The names of the "Clusters" correspond to those in Table 1 and branches are colored by "Cluster". Species abbreviations occur before the gene name and the cluster names are as follows: *Aa* = *Aedes aegypti*, *Ag* = *Anopheles gambiae*, *Cp* = *Culex quinquefasciatus*, *Dm* = *Drosophila melanogaster*, *Bm* = *Bombyx mori*, *Tc* = *Tribolium castaneum*, *Am* = *Apis mellifera*, *Nv* = *Nasonia vitripennis*, *Ap* = *Acyrtosiphon pisum*, *Ph* = *Pediculus humanus*.

not shared with the other subclasses (moderate bootstrap support and close sequence similarity). Furthermore, the approximate origination time of these groups is determined by examining their taxonomic distribution (Fig. 3; Supplementary Fig. 2). Given the Delta subclass contains earlier diverging insects, it is considered older in origin than Epsilon. For instance, the Delta subclass contains the earlier diverging exopterogote *P. humanus* whereas the Epsilon subclass has no members among the exopterogotes. Therefore, the Delta subclass is at least as old as the Exopterogota lineage.

The Epsilon-GSTs are confined to the dipterans (*Culex*, *Drosophila*, *Aedes*, *Anopheles*), a coleopteran, and a lepidopteran (Fig. 3). However, Epsilon does not have a GST from Hymenoptera (*Nasonia*, *Apis*). Given this is not a result of gene loss along the hymenopteran lineage, this lends support to the latest findings that the hymenopterans have an earlier origin than the coleopterans. It also suggests that the Epsilon-GSTs evolved from the Delta subclass and between the times when Hymenoptera and Coleoptera originated as a lineage (Figs. 1–3).

The Delta and Epsilon subclasses have a general pattern of vast and recent diversification as supported by individual clades reflecting the speciation events (Figs. 1 and 3). It could be surmised that these are recent species-level events, but Low and others (2007) showed that the duplications are often older, at least among the drosophilids. Tandem and segmental gene duplication are the likely causes of this pattern of vast gene diversification. Evidence supporting this mechanism are the phylogenetic branching conforming to the species tree along with the genomic clustering of GSTs (Table 1; Fig. 3).

The diversification is most pronounced in Diptera and in the coleopteran *T. castaneum* (Fig. 3, see Tc-1 cluster). For Diptera, this pattern was created in a fairly uniform way over evolutionary time (Low et al., 2007). However, the pattern of diversification is unknown for the coleopteran. Interestingly, many of the culicid (*Aedes*, *Anopheles*, *Culex*) GSTs are clustering together and are branching by the culicid speciation pattern. This is evidence that the culicid GST duplications occurred early in the culicid lineage, but not earlier than the drosophilid and culicid common ancestor.

In addition, there is synteny evidence supporting the time of diversification in the Delta-GST subtree (Fig. 3; Supplementary table). For example, there is a clade of *D. melanogaster* GSTs which are in linkage (*Dm*-2 cluster, defined in Fig. 3 legend) and a clade of *A. gambiae* GSTs which are also in linkage (*Ag*-1 cluster). The *Dm*-2 cluster is localized to chromosome 2R while *Ag*-1 is on chromosome 3R. Severson et al. (2004) showed evidence that these two regions are syntenic at the chromosomal level, supporting that these clades share a common GST ancestor (in an ancient dipteran or earlier insect ancestor) even though their gene expansions occurred independently. If the majority of these duplications had not occurred independently in these two lineages, then they would have occurred in their ancient common ancestor and not along each of their lineages (Fig. 3).

3.4. Phylogenies of the Theta, Zeta, Omega, and Sigma subclasses

The Theta-GST phylogeny (Fig. 4) contains a member in all sampled insect species and the root is provided by a vertebrate GST. In

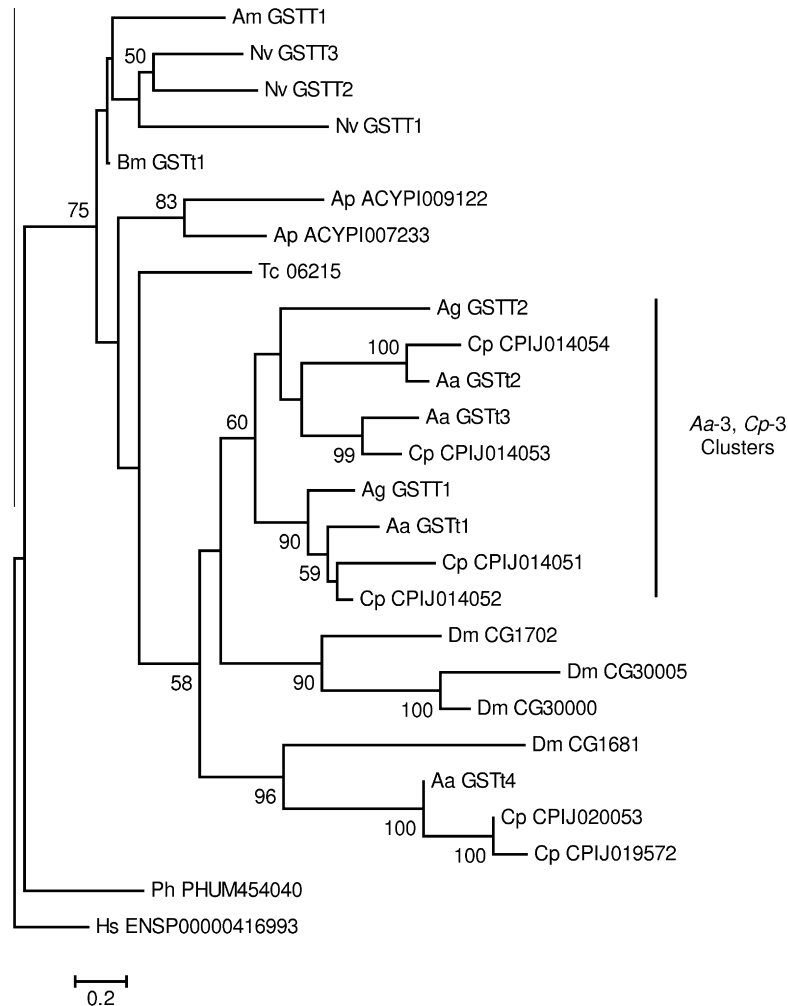


Fig. 4. Rooted phylogenetic tree of Theta-GST subclass. Scale bar shows the inferred amino acid distance. Bootstrap values are placed at the nodes where the value is ≥ 50 . The names of the “Clusters” correspond to those in Table 1. Species abbreviations occur before the gene name as described in Fig. 3.

this tree, there are four *D. melanogaster* genes where three of these cluster together with a bootstrap support value of 90. They are also neighbors as defined by position in the genome, likely an effect of tandem gene duplication. Similarly, there is a cluster of culicid genes (labeled in the figure) whose diversification occurred in an ancestral culicid. This timing of diversification is supported by the bootstrap-supported topology which shows the GST branching pattern (Fig. 4) as mirroring the culicid speciation pattern (Fig. 1). Lastly, among the hymenopterans, there is one GST in *A. mellifera* yet three copies in *N. vitripennis* (Oakeshott et al., 2010).

The Zeta-GST subclass (Fig. 5A) shows single members among the insects except for two species. The exceptions are *D. melanogaster* with two duplicated members and *A. pisum* which does not have any members. Overall the pattern shows a gene family with little gain and loss of members.

The Omega subclass shows four GSTs from *D. melanogaster* which are also linked along the chromosome (Fig. 5B; Table 1). This is evidence of tandem and/or segmental duplication events. If one accepts the rule of parsimony, then one would favor the duplication events as caused by segmental duplication since it requires a fewer number of events. As in the Zeta-GSTs, this subclass mostly has single members from each species. The exceptions are: *D. melanogaster*, *B. mori* and *T. castaneum*, each with three members, and *N. vitripennis* with two members.

The sixth subclass, Sigma (Fig. 6), contains a member of all species from this study. There is a single GST in *D. melanogaster*.

However, there are gene duplicates in other species. *N. vitripennis* and *A. mellifera* have eight and four GSTs, respectively (Oakeshott et al., 2010); note that two of the GSTs in each of the hymenopterans do not appear in Fig. 6 because they are too short to contribute to a sequence alignment (see Supplementary table for assignments). *A. pisum* and *T. castaneum* show gene expansions in this subclass where each has five members which mainly cluster by species (Fig. 6). The *A. pisum* genome is not sufficiently assembled to confidently map its genes, but *T. castaneum* is well mapped and allows the identification of a cluster of four GSTs (labeled as Tc-3).

Each of these four subclasses has a homolog in vertebrate animals. Therefore, the origins of each of the Sigma, Theta, Omega, and Zeta subclasses predate the vertebrate and insect divergence (see Figs. 4–6). It was even proposed that these divergences occurred much further back in time. Enayati et al. (2005) pushed the origin of the Zeta and Theta subclasses to before the origins of multicellular life. However, the other two subclasses, Delta and Epsilon, have no matches to non-insect genes. Therefore, they are considered to be innovations along the relatively more recent Insecta lineage.

4. Discussion

The glutathione S-transferase (GST) trees show the birth-and-death process of genes (Nei and Rooney, 2005). This is particularly pronounced among the Delta and Epsilon subclasses (Fig. 3) where

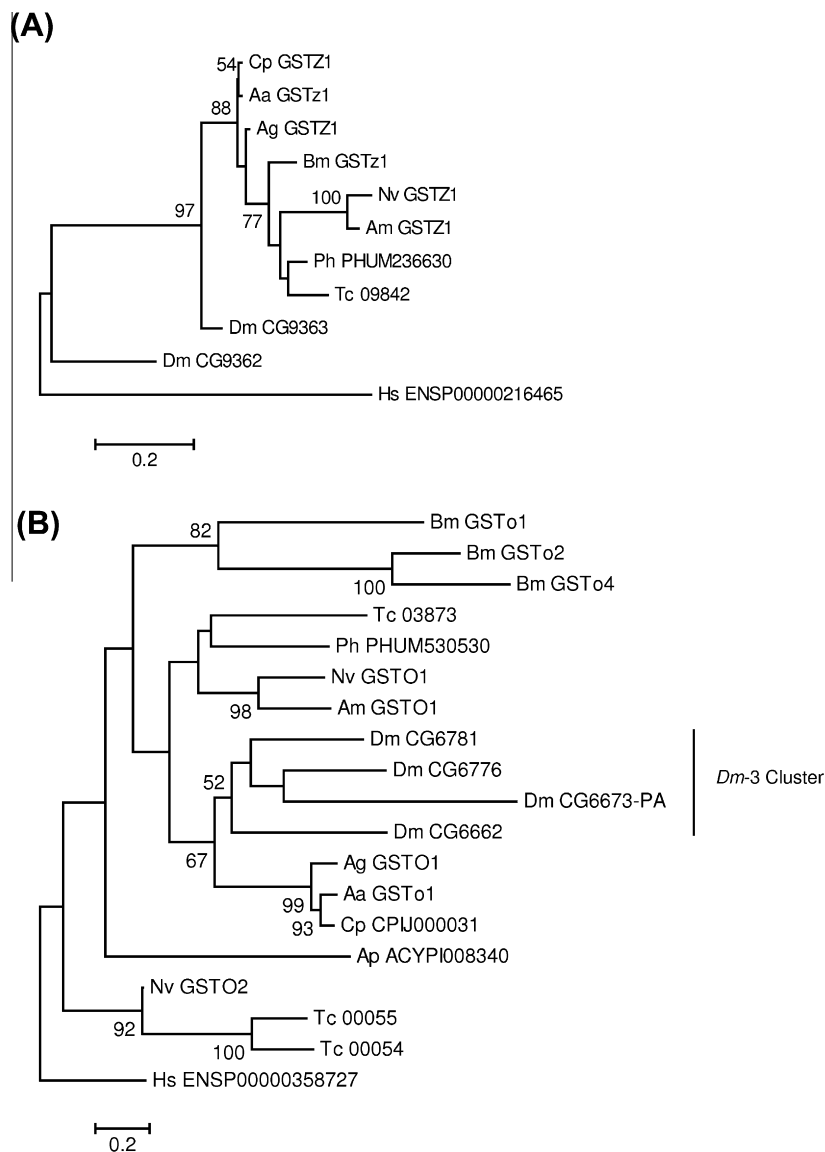


Fig. 5. (A) Rooted phylogenetic tree of Zeta-GST subclass. (B) Rooted phylogenetic tree of Omega subclass of GSTs. See Fig. 4 legend for description of annotations.

the turnover of genes in this family is consistent with their role in detoxifying cells from a varying set of chemical compounds. The other pattern of interest is the genomic clustering of these genes (Table 1). This clustering could be evidence of recent duplication or nature selecting for genomic proximity and efficiency in producing the GST enzymes. However, it is most likely that the clustering resulted from recent tandem and segmental duplication events and this information is utilized to assign GSTs to the subclasses.

The highest number of GST duplicates are in the dipterans and the sole coleopteran while the two hymenopterans and two exopterogotes have less copies. Within Diptera, the phylogenies show that the duplication events frequently occurred along the drosophilid lineage (Low et al., 2007) or along the culicid lineage prior to the divergence of the three culicids (Lumjuan et al., 2005; Fig. 3). These independently occurring duplication events correspond to separate gene family radiations (Ranson et al., 2002; Claudianos et al., 2006; Low et al., 2007). In addition, this study has shown that much of the variation in GST family size is explained by number of protein-coding genes in the genome; however, there are other contributing causes to the correlation, such as parasitic lifestyle (*P. humanus*) and eusociality (*A. mellifera*).

Further, with the addition of the exopterogotes, this study has shown that the Delta subclass is Insecta-specific while the Epsilon subclass is specific to a clade of endopterogotes (Coleoptera/Lepidoptera/Diptera). Therefore, these two highly diverse subclasses are expected to have an ancient role unique to insects. In other words, these genes were removing chemical compounds from natural sources and not from artificial sources. This is asserted because the gene duplications are not recent as they are not highly similar at the amino acid level (Li, 1997), at least not recent in the time frame of human society. Therefore, it must be artifactual that the GSTs are functional in the pathway to remove products or byproducts of artificial insecticide use.

The other four subclasses have a more ancient origin in Metazoa or earlier. A very early scenario was introduced by Da Fonseca et al. (2010) where they identified bacterial homologs with GSTs from the four subclasses. However, it should be noted this result was based upon confidence in a Bayesian phylogeny and its posterior probability support values (Suzuki et al., 2002). Also, these four subclasses show much less extensive cases of gene expansions, although some species show an equal or higher count of GSTs among these subclasses as compared to Delta and Epsilon. For

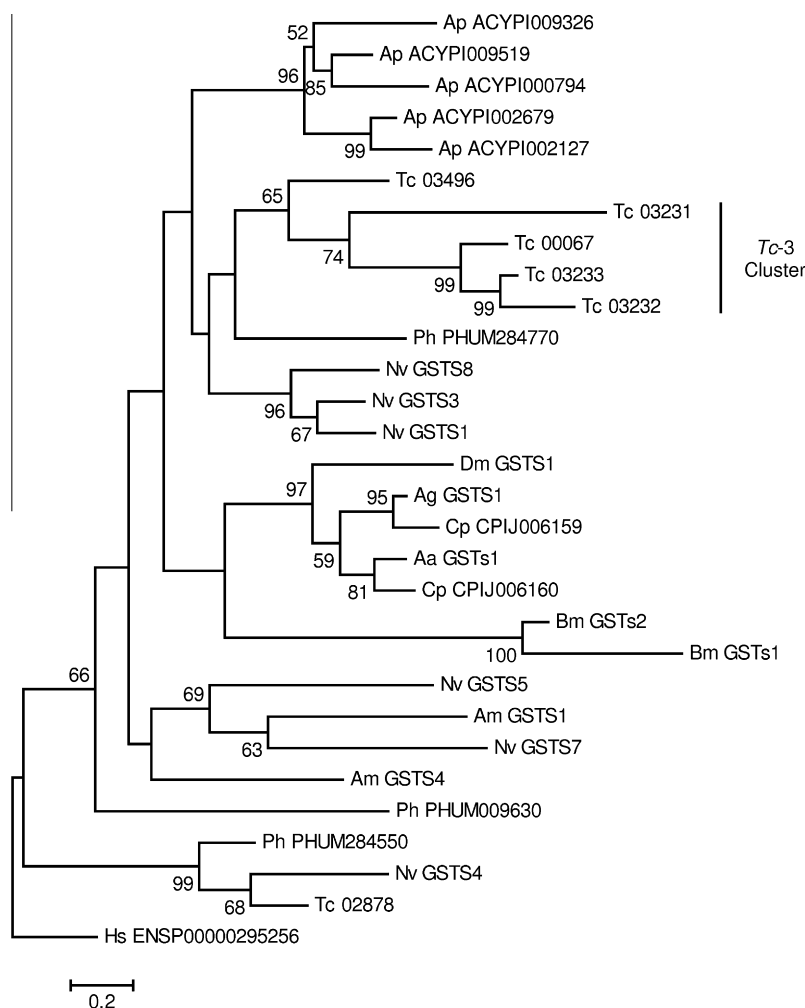


Fig. 6. Rooted phylogenetic tree of Sigma-GST subclass. See Fig. 4 legend for description of annotations.

example, *P. humanus* has three GSTs in both the Sigma and Delta subclasses (Supplementary table).

In conclusion, this study provides a comprehensive survey of GSTs among the insects. Also, their pattern of forming genomic clusters is confirmed and it is determined that these clusters are not orthologous among genera but rather formed independently. Clusters are formed and maintained across a few million years, but then are slowly erased by the effects of gene loss. The turnover of genes by the birth-and-death process has the result of destroying previously established genomic clusters and creating new ones.

Acknowledgments

Three anonymous reviewers provided invaluable suggestions on nomenclature and phylogenetic methods to reliably classify genes into families.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.08.027. 1. Supplementary table of GST genes: http://www.biol.sc.edu/~bobf/supplementary_table.pdf. 2. Supplementary Fig. 1 is a condensed topology of the full length GSTs: http://www.biol.sc.edu/~bobf/supplementary_figure1.pdf. 3. Supplementary Fig. 2 is a rectangular

phylogeny of the Delta/Epsilon-GST superfamily: http://www.biol.sc.edu/~bobf/supplementary_figure2.pdf.

References

- Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.
- Aphid Genomics Consortium, 2010. Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol.* 8, e1000313.
- Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., et al., 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* 330, 86–88.
- Barrios, S., Zerba, E., Picollo, M.I., Audino, P.G., 2010. Activity of increased specific and non-specific esterases and glutathione transferases associated with resistance to permethrin in *Pediculus humanus capitis* (phthiraptera: pediculidae) from Argentina. *Parasitol. Res.* 106, 415–421.
- Claudianos, C., Ranson, H., Johnson, R.M., Biswas, S., Schuler, M.A., Berenbaum, M.R., Feyereisen, R., Oakeshott, J.G., 2006. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* 15, 615–636.
- Corona, M., Robinson, G.E., 2006. Genes of the antioxidant system of the honey bee: annotation and phylogeny. *Insect Mol. Biol.* 15, 687–701.
- da Fonseca, R.R., Johnson, W.E., O'Brien, S.J., Vasconcelos, V., Antunes, A., 2010. Molecular evolution and the role of oxidative stress in the expansion and functional diversification of cytosolic glutathione transferases. *BMC Evol. Biol.* 10, 281.
- Ding, Y., Ortel, F., Rossiter, L.C., Hemingway, J., Ranson, H., 2003. The *Anopheles gambiae* glutathione transferase supergene family: annotation, phylogeny and expression profiles. *BMC Genom.* 4, 35.

- Drosophila* 12 Genomes Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218.
- Drysdale, R., FlyBase Consortium, 2008. FlyBase: a database for the *Drosophila* research community. *Methods Mol. Biol.* 420, 45–49.
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32, 1792–1797.
- Enayati, A.A., Ranson, H., Hemingway, J., 2005. Insect glutathione transferases and insecticide resistance. *Insect Mol. Biol.* 14, 3–8.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.
- Felsenstein, J., 1989. PHYLIP – phylogeny inference package (version 3.2). *Cladistics* 5, 164–166.
- Flück, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., et al., 2010. Ensembl 2011. *Nucl. Acids Res.* 39 (Suppl. 1), D800–D806.
- Gattiker, A., Gasteiger, E., Bairoch, A., 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinformatics* 1, 107–108.
- Gaunt, M.W., Miles, M.A., 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol. Biol. Evol.* 19, 748–761.
- Habig, W.H., Pabst, M.J., Jakoby, W.B., 1974. Glutathione S-transferases. The first enzymatic step in mercapturic acid formation. *J. Biol. Chem.* 249, 7130–7139.
- Hahn, M.W., Han, M.V., Han, S.-G., 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.* 3, e197.
- Honeybee Genome Sequence Consortium, 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443, 931–949.
- Hughes, A.L., Ekollu, V., Friedman, R., Rose, J.R., 2005. Gene family content-based phylogeny of prokaryotes: the effect of criteria for inferring homology. *Syst. Biol.* 54, 268–276.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Castro, E.D., Langendijk-Genevaux, P.S., Pagni, M., Sigrist, C.J.A., 2006. The PROSITE database. *Nucl. Acids Res.* 34, D227–D230.
- Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., et al., 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl. Acad. Sci. USA* 107, 12168–12173.
- Lawson, D., Arensburg, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., et al., 2009. VectorBase: a data resource for invertebrate vector genomics. *Nucl. Acids Res.* 37, D583–D587.
- Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320.
- Legeai, F., Shigenobu, S., Gauthier, J.-P., Colbourne, J., Risper, C., et al., 2010. AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol. Biol.* 19 (Suppl. 2), 5–12.
- Li, W.-H., 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Low, W.Y., Ng, H.L., Morton, C.J., Parker, M.W., Batterham, P., Robin, C., 2007. Molecular evolution of glutathione S-transferases in the genus *Drosophila*. *Genetics* 177, 1363–1375.
- Lumjuan, N., McCarroll, L., Prapanthadara, L., Hemingway, J., Ranson, H., 2005. Elevated activity of an Epsilon class glutathione transferase confers DDT resistance in the dengue vector, *Aedes aegypti*. *Insect Biochem. Mol. Biol.* 35, 861–871.
- Lumjuan, N., Stevenson, B.J., Prapanthadara, L., Somboon, P., Brophy, P.M., et al., 2007. The *Aedes aegypti* glutathione transferase family. *Insect Biochem. Mol. Biol.* 37, 1026–1035.
- Munoz-Torres, M.C., Reese, J.T., Childers, C.P., Bennett, A.K., Sundaram, J.P., Childs, K.L., Anzola, J.M., Milshina, N., Elsik, C.G., 2010. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucl. Acids Res.* 39, D658–D662.
- Nei, M., Rooney, A.P., 2005. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* 39, 121–152.
- Oakeshott, J.G., Johnson, R.M., Berenbaum, M.R., Ranson, H., Cristino, A.S., Claudianos, C., 2010. Metabolic enzymes associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*. *Insect Mol. Biol.* 19 (Suppl. 1), 147–163.
- Ramsey, J.S., Rider, D.S., Walsh, T.K., Vos, M.D., Gordon, K.H.J., Ponnala, L., Macmil, S.L., Roe, B.A., Jander, G., 2010. Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*. *Insect Mol. Biol.* 19 (Suppl. 2), 155–164.
- Ranson, H., Jensen, B., Wang, X., Prapanthadara, L., Hemingway, J., Collins, F.H., 2000. Genetic mapping of two loci affecting DDT resistance in the malaria vector *Anopheles gambiae*. *Insect Mol. Biol.* 9, 499–507.
- Ranson, H., Rossiter, L., Ortelli, F., Jensen, B., Wang, X., Roth, C.W., Collins, F.H., Hemingway, J., 2001. Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector *Anopheles gambiae*. *Biochem. J.* 359, 295–304.
- Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M.V., Unger, M.F., Collins, F.H., Feyereisen, R., 2002. Evolution of supergene families associated with insecticide resistance. *Science* 298, 179–181.
- Rogers, M.E., Jani, M.K., Vogt, R.C., 1999. An olfactory-specific glutathione-S-transferase in the sphinx moth *Manduca sexta*. *J. Exp. Biol.* 202, 1625–1637.
- Savard, J., Tautz, D., Richards, S., Weinstock, G.M., Gibbs, R.A., Werren, J.H., Tettelin, H., Lercher, M.J., 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* 16, 1334–1338.
- Sawicki, R., Singh, S.P., Mondal, A.K., Benes, H., Zimniak, P., 2003. Cloning, expression and biochemical characterization of one Epsilon-class (GST-3) and ten Delta-class (GST-1) glutathione S-transferases from *Drosophila melanogaster*, and identification of additional nine members of the Epsilon class. *Biochem. J.* 370, 661–669.
- Severson, D.W., DeBruyn, B., Lovin, D.D., Brown, S.E., Knudson, D.L., Morlais, I., 2004. Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*. *J. Hered.* 95, 103–113.
- Silkworm Genome Consortium, 2008. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1036–1045.
- Stamatakis, A., 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690.
- Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAXML web servers. *Syst. Biol.* 57, 758–771.
- Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Natl. Acad. Sci. USA* 99, 16138–16143.
- Tamura, K., Subramanian, S., Kumar, S., 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* 21, 36–44.
- Tang, A.H., Tu, C.P., 1994. Biochemical characterization of *Drosophila* glutathione S-transferases D1 and D21. *J. Biol. Chem.* 269, 27876–27884.
- Toba, G., Aigaki, T., 2000. Disruption of the microsomal glutathione S-transferase-like gene reduces life span of *Drosophila melanogaster*. *Gene* 253, 179–187.
- Tribolium Genome Consortium, 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452, 949–955.
- Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., et al., 2005. SilkDB: a knowledgebase for silkworm biology and genomics. *Nucl. Acids Res.* 33, D399–D402.
- Wang, L., Wang, S., Li, Y., Paradesi, M.S.R., Brown, S.J., 2007. BeetleBase: the model organism database for *Tribolium castaneum*. *Nucl. Acids Res.* 35, D476–D479.
- Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., et al., 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327, 343–348.
- Wiegmann, B.M., Trautwein, M.D., Kim, J.-W., Cassel, B.K., Bertone, M.A., et al., 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol.* 7, 34.
- Yang, Z., 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford, UK.
- Yu, Q., Lu, C., Li, B., Fang, S., Zuo, W., et al., 2008. Identification, genomic organization and expression pattern of glutathione S-transferase in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1158–1164.