GMR
The Original

ORTHOLOGY-VERIFIED COMPARATIVE ANALYSIS OF GLUTATHIONE S-TRANSFERASE EPSILON 2 (GSTE2) IN *ANOPHELES*, *AEDES* AND *CULEX* MOSQUITO VECTORS: CONSERVED GLUTATHIONE-SITE ARCHITECTURE AND LINEAGE-SPECIFIC VARIATION AT INSECTICIDE-RESISTANCE HOTSPOTS

Sridhar Gunde[1*], Nageswara Rao Apka[2], Manisha Birru[1], Srinivasa Rao Vattikonda[1]

[1]Department of Zoology, Nizam College, Osmania University, Hyderabad, Telangana, India.
[2]Department of Zoology, University College of Science, Osmania University, Hyderabad, Telangana, India.
[*]**Corresponding author**: Sridhar Gunde. E-mail address: gsridhar01@gmail.com

ABSTRACT

Glutathione S-transferase epsilon 2 (GSTe2) is one of the best-characterised metabolic insecticide-resistance enzymes in mosquitoes, with experimentally demonstrated DDT-detoxifying activity and resistance-associated substitutions in *Anopheles gambiae* (I114T) and *Anopheles funestus* (L119F). Comparative analyses of GSTe2 across vector genera are, however, vulnerable to a basic data problem: epsilon GSTs occur as tandem gene clusters, so a misassigned paralogue can silently replace the true orthologue. Here we assembled an orthology-verified GSTe2 dataset for *A. gambiae* (UniProt Q7PVS6), *A. funestus* (PDB 3ZML), *Aedes aegypti* (Q5PY77) and *Culex quinquefasciatus*, in which the *Culex* orthologue was identified from all ten members of its epsilon cluster. Sequence integrity was confirmed by CRC64 checksum against UniProtKB. The *Culex* protein B0XGJ8 (224 aa) shared 72.9–77.0% identity with the three reference GSTe2 proteins, compared with 34.4–68.0% for the other nine *Culex* epsilon GSTs, and a maximum-likelihood phylogeny (IQ-TREE 2, LG+G4) placed it within a GSTe2 clade with strong support (SH-aLRT 99.8%, UFBoot 95%). Within this clade, *Aedes* and *Culex* GSTe2 grouped together, consistent with culicine–anopheline divergence. The four orthologues were acidic (pI 5.41–5.76), 24.7–25.0 kDa, predicted stable (instability index 23.97–31.67) and mildly hydrophilic (GRAVY −0.201 to −0.006). The catalytic Ser12 and the glutathione-binding residues His53 and Ser68 were invariant across all 13 epsilon GSTs, whereas variation was concentrated in the C-terminal domain (57 of 133 positions variable among orthologues, versus 22 of 82 in the N-terminal domain). Position 114 carried Ile, Val or Leu among the orthologues, while Leu119 was conserved in all wild-type GSTe2 sequences but replaced by Phe in several *Culex* paralogues. These results provide a verified reference framework for GSTe2 comparisons and identify residue 114 and the 119–120 H-site as priorities for experimental resistance studies in *Aedes* and *Culex*.

KEYWORDS: glutathione S-transferase; GSTe2; insecticide resistance; DDT; orthology; maximum-likelihood phylogeny; mosquito vectors

1. INTRODUCTION

Mosquitoes transmit malaria, dengue, chikungunya, Zika, yellow fever and lymphatic filariasis, and insecticide-based interventions remain central to their control. The spread of insecticide resistance threatens these gains and is now monitored as a core component of vector-control programmes [1–3]. Resistance arises through target-site insensitivity, reduced cuticular penetration, behavioural avoidance and, most commonly alongside target-site changes, enhanced metabolic detoxification by cytochrome P450 monooxygenases, carboxylesterases and glutathione S-transferases (GSTs) [4–6].

GSTs catalyse the conjugation of reduced glutathione (GSH) to electrophilic substrates and can also act through peroxidase activity and ligand binding [7,8]. Cytosolic GSTs share a two-domain architecture: an N-terminal thioredoxin-like domain that forms the GSH-binding G-site and a predominantly α-helical C-terminal domain that shapes the substrate-binding H-site [7,9]. Insects possess two expanded, insect-specific classes, delta and epsilon, which account for most GST-mediated insecticide metabolism [10,11]. The epsilon class was first defined in *A. gambiae*, where its genes lie in a tandem cluster on chromosome arm 3R that is genetically linked to a major DDT-resistance locus [12,13].

Among the epsilon GSTs, GSTe2 has the strongest functional evidence. Recombinant *A. gambiae* GSTe2 shows the highest DDT dehydrochlorinase activity of the cluster [14], and its crystal structure revealed a compact H-site consistent with this activity [15]. The I114T substitution in *A. gambiae* GSTe2 lies close to the substrate-binding pocket and is associated with DDT resistance [16]. In *Aedes aegypti*, elevated epsilon GST activity was linked to DDT resistance, and gene-knockdown experiments indicated that GSTe2 contributes to resistance to DDT and pyrethroids [17,18]. In *A. funestus*, a single L119F substitution enlarges the substrate-binding cavity, increases DDT metabolism and provides a field-applicable DNA marker of resistance [19]. *Culex quinquefasciatus*, the principal vector of bancroftian filariasis in India, shows widespread DDT and pyrethroid resistance with elevated GST activity in field populations [20,21], but its GSTe2 has received far less attention.

Comparative analyses that extend GSTe2 knowledge across vector genera depend on correct orthologue assignment. This is not trivial: the *Culex* genome encodes a cluster of closely related epsilon GSTs [22], and paralogues within such clusters share 35–70% identity, enough to align plausibly while differing at functionally critical positions. Structural interpretation is similarly sensitive to template choice; for example, PDB 5D73, which has been used as a template in some analyses, is a pi-class GST from the filarial nematode *Wuchereria bancrofti* and is unsuitable for modelling mosquito GSTe2.

The aims of this study were therefore to (i) assemble a checksum-verified, orthology-confirmed GSTe2 dataset for *A. gambiae*, *A. funestus*, *Ae. aegypti* and *C. quinquefasciatus*; (ii) characterise their sequence-derived physicochemical properties; (iii) test the orthology of the *Culex* sequence by maximum-likelihood phylogeny against all members of its epsilon cluster; and (iv) map conservation and variation at experimentally defined catalytic, GSH-binding and resistance-associated residues, using experimentally determined mosquito GSTe2 structures as the structural reference (Fig. 1).

2. MATERIALS AND METHODS

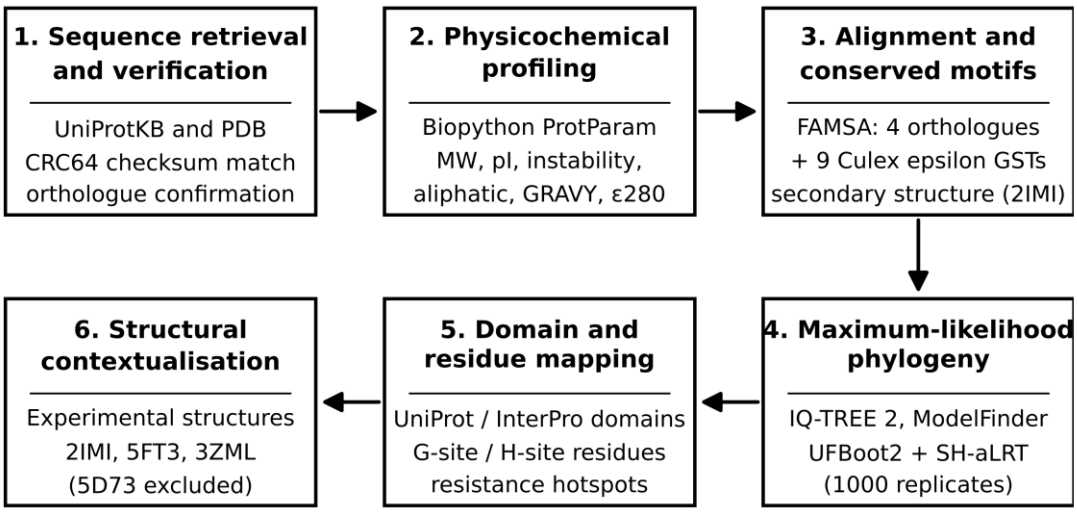


Fig. 1. Analytical workflow. Each step uses the verified sequence set from step 1; PDB 5D73 (a nematode pi-class GST) was excluded as a structural reference.

2.1. Sequence retrieval and verification

GSTe2 protein sequences were retrieved from UniProtKB [23] for *A. gambiae* (Q7PVS6; gene AGAP009194) and *Ae. aegypti* (Q5PY77). For *A. funestus*, the sequence of the experimentally determined GSTe2 structure PDB 3ZML [19,24] was used after removal of the two-residue N-terminal expression tag, yielding the 221-residue native sequence. To identify the *C. quinquefasciatus* orthologue, all ten members of the *Culex* epsilon GST cluster annotated in UniProtKB from the genome assembly [22] were retrieved (B0XGJ5–B0XGJ9 and B0XGK0–B0XGK4). Each was compared by global pairwise alignment (BLOSUM62; gap opening −10, extension −0.5) with the three reference GSTe2 sequences, and orthology was then tested phylogenetically (Section 2.4). Sequence integrity was confirmed by recomputing the CRC64 checksum of each retrieved sequence and matching it to the UniProtKB record. The verified dataset is listed in Table 1.

2.2. Physicochemical characterisation

Molecular mass, theoretical isoelectric point (pI), instability index, aliphatic index, grand average of hydropathicity (GRAVY) and molar extinction coefficient at 280 nm were calculated with the ProtParam module of Biopython 1.88 [25], which implements the algorithms of the ExPASy ProtParam tool [26]: Bjellqvist pK values for pI [27], the dipeptide instability index [28], the aliphatic index [29], Kyte-Doolittle hydropathy [30] and the Pace method for extinction coefficients [31]. These are sequence-derived descriptors; the aliphatic index in particular was used comparatively and not as evidence of measured thermostability.

2.3. Multiple sequence alignment and conservation

Two alignments were generated with FAMSA [32]: (i) the four GSTe2 orthologues, and (ii) an epsilon-cluster alignment of the four orthologues plus the nine remaining *Culex* epsilon GSTs (13 sequences). Pairwise identities were computed from alignment (i) over gap-free columns. Site-wise conservation in alignment (ii) was expressed as 1 − normalised Shannon entropy, with gaps treated as missing; a value of 1 indicates an invariant column. All positions are reported in *A. gambiae* GSTe2 (Q7PVS6) numbering.

2.4. Phylogenetic analysis

Maximum-likelihood trees were inferred with IQ-TREE 2.3.6 [33]. The substitution model was selected by ModelFinder [34] under the Bayesian information criterion, which chose LG+G4 [35] for the 13-sequence epsilon alignment and LG+I for the four-orthologue alignment. Branch support was assessed with 1000 ultrafast bootstrap replicates (UFBoot2) [36] and 1000 SH-like approximate likelihood-ratio test replicates (SH-aLRT) [37]; clades with SH-aLRT ≥ 80% and UFBoot ≥ 95% were regarded as strongly supported. The epsilon tree was midpoint-rooted for display, and a fixed random seed (12345) was used for reproducibility.

2.5. Domain annotation and residue mapping

Domain boundaries (GST N-terminal and GST C-terminal; PROSITE PS50404/PS50405) and InterPro/Pfam signatures (IPR004045, IPR004046, PF13417, PF00043) were taken from the UniProtKB records [23,38]. Residues of functional interest were mapped onto the alignments: the catalytic serine of epsilon GSTs (Ser12); the GSH-binding residues annotated from the GSH-bound *A. gambiae* GSTe2 structure PDB 2IMI (His53, Ile55, Ser68 and Arg112) [15]; the H-site residues Glu116 and Phe120 [15]; and the resistance-associated positions 114 (*A. gambiae* I114T) [16] and 119 (*A. funestus* L119F) [19].

2.6. Structural reference framework

Structural interpretation relied only on experimentally determined mosquito GSTe2 structures in the Protein Data Bank [24]: *A. gambiae* GSTe2 with GSH (2IMI) [15], *Ae. aegypti* GSTe2 (5FT3) and *A. funestus* GSTe2 with GSH (3ZML and 3ZMK) [19] (Table 6). No homology models were built; the *C. quinquefasciatus* GSTe2, which lacks an experimental structure, was interpreted through its alignment to these structures.

2.7. Functional association analysis

Predicted functional partners of GSTe2 were retrieved from the STRING database [39] for *Ae. aegypti* and *A. gambiae*, using GSTe2 as the single query protein for each organism. Networks were built from all active evidence channels (experiments, curated databases, co-expression, gene neighbourhood, gene fusion, co-occurrence and text mining) at the default medium-confidence threshold (combined score ≥ 0.400), with the first shell limited to ten interactors. Edges were interpreted as functional associations; they were not taken as evidence of direct physical protein–protein interaction unless the supporting evidence channel was experimental. STRING identifies proteins by its own organism-specific identifiers, so this analysis is independent of the sequence-level dataset in Table 1.

2.8. Reproducibility and quality control

All analyses used the checksum-verified sequences in Table 1, a single residue-numbering scheme (*A. gambiae* Q7PVS6) and fixed random seeds; tool versions and parameters are listed in Supplementary Table S4. No homology models were built, so template-dependent model-quality metrics (e.g., GMQE, QMEAN, Ramachandran statistics or RMSD) are not reported. Any future model of *Culex* GSTe2 should be built on a mosquito GSTe2 template from Table 6, validated by standard stereochemical checks, and reported with the alignment method, atom selection and number of aligned residues for every RMSD value.

3. RESULTS

3.1. Identification and verification of the GSTe2 orthologues

The recomputed CRC64 checksums of Q7PVS6 (FBC47191346A291E), Q5PY77 (3DADF5F337941292) and B0XGJ8 (C419701E4D160144) matched the UniProtKB records exactly. Among the ten *Culex* epsilon GSTs, B0XGJ8 (224 aa; VectorBase CPIJ018627) was the clear best match to all three references, with 76.5%, 77.0% and 72.9% pairwise identity to *A. gambiae*, *Ae. aegypti* and *A. funestus* GSTe2, respectively (global alignment; Supplementary Table S1). The next most similar paralogue, B0XGK0, shared 65.3–68.0% identity, and the remaining eight shared 34.4–59.0% (Supplementary Table S1). B0XGJ8 was therefore assigned as the *C. quinquefasciatus* GSTe2 orthologue and confirmed phylogenetically (Section 3.5). The final dataset is summarised in Table 1.

Table 1. Verified GSTe2 dataset.

Species	Identifier	Source	Length (aa)	Structures	Verification
Anopheles gambiae	Q7PVS6 (AGAP009194)	UniProtKB	221	2IMI, 2IL3, 2IMK	CRC64 match
Anopheles funestus	3ZML chain A	PDB	221*	3ZML, 3ZMK	Experimental structure
Aedes aegypti	Q5PY77	UniProtKB	222	5FT3	CRC64 match
Culex quinquefasciatus	B0XGJ8 (CPIJ018627)	UniProtKB	224	-	CRC64 match; ML orthology

*After removal of the two-residue N-terminal expression tag. The nine additional *Culex* epsilon GSTs used for orthology testing are listed in Supplementary Table S1.

3.2. Physicochemical properties

The four GSTe2 orthologues were closely matched in size (221–224 residues; 24.69–25.03 kDa) and were all acidic, with theoretical pI values of 5.41–5.76 and an excess of acidic (27–32) over basic (23–25) residues (Table 2). All four had instability indices below 40 (23.97–31.67), classifying them as stable in vitro. Aliphatic indices were high (96.65–104.11), and GRAVY values were negative but close to zero (−0.201 to −0.006), consistent with soluble cytosolic proteins carrying substantial hydrophobic surface in the substrate-binding region. The two *Anopheles* proteins were the most acidic and most hydrophilic, whereas the *Aedes* and *Culex* proteins were nearly identical in pI (5.76 and 5.75) and GRAVY (−0.008 and −0.006).

Table 2. Sequence-derived physicochemical properties of the GSTe2 orthologues.

Parameter	*A. gambiae*	*A. funestus*	*Ae. aegypti*	*C. quinquefasciatus*
Length (aa)	221	221	222	224
Molecular mass (Da)	24,849.33	25,026.51	24,694.35	24,823.43
Theoretical pI	5.54	5.41	5.76	5.75
Acidic residues (Asp+Glu)	30	32	27	27
Basic residues (Arg+Lys)	24	25	23	23
Instability index	29.02	31.67	23.97	30.56
Aliphatic index	98.37	96.65	99.28	104.11
GRAVY	−0.125	−0.201	−0.008	−0.006
ε280 ($M^{-1}cm^{-1}$), reduced	18,910	20,400	20,400	21,890
ε280 ($M^{-1}cm^{-1}$), cystine	19,035	20,525	20,650	22,015

Calculated from the verified sequences in Table 1. Instability index < 40 indicates a predicted stable protein.

3.3. Multiple sequence alignment, secondary structure and conserved motifs

The four orthologues aligned over 224 columns with only three gap-containing columns, all at the C-terminus (Fig. 2). Of the 221 gap-free columns, 140 (63.3%) were identical in all four proteins. Pairwise identity was highest between the two *Anopheles* species (91.9%) and between *Aedes* and *Culex* (77.0%), and ranged from 70.1% to 75.6% between anopheline and culicine sequences (Table 3). Mapping the experimental secondary structure of *A. gambiae* GSTe2 (PDB 2IMI) onto the alignment showed the canonical cytosolic-GST topology, a mixed α/β N-terminal domain (βαβαββα) followed by an all-α C-terminal domain; the independently determined *Ae. aegypti* structure (5FT3) gave a near-identical element pattern, confirming that the fold is conserved across the two subfamilies (Fig. 2).



Fig. 2. Multiple sequence alignment of the four verified GSTe2 orthologues (FAMSA), annotated with the experimental secondary structure of *A. gambiae* GSTe2 (PDB 2IMI; purple cylinders, α-helices; orange arrows, β-strands). Dark shading marks columns identical in all four sequences; light shading marks residues shared by three. Symbols indicate, in *A. gambiae* numbering, the catalytic Ser12 (triangle), GSH-binding residues from PDB 2IMI (circle), resistance-associated positions 114 and 119 (star) and H-site residues Glu116 and Phe120 (diamond).

Table 3. Pairwise amino-acid identity (%) among the GSTe2 orthologues, over gap-free columns of the FAMSA alignment.

	A. gambiae	*A. funestus*	*Ae. aegypti*	*C. quinquefasciatus*
A. gambiae	100	91.9	71.0	75.6
A. funestus	91.9	100	70.1	71.9
Ae. aegypti	71.0	70.1	100	77.0
C. quinquefasciatus	75.6	71.9	77.0	100

Eleven blocks of five or more consecutive residues were identical across all four orthologues; the eight most informative are listed in Table 4 with the secondary-structure element they occupy. These conserved motifs cluster in and around the G-site (the Ser12 loop, the His53 and Ser68 segments and the Arg112-bearing helix), consistent with their proposed role in glutathione recognition rather than being presented as experimentally proven catalytic motifs.

Table 4. Conserved sequence motifs (identical in all four GSTe2 orthologues), with the PDB 2IMI secondary-structure element each occupies.

Residues	Motif sequence	Secondary-structure element	Functional context
12–20	SPPCRAVEL	helix α1	Catalytic Ser12; G-site thiol loop
50–54	NPQHT	β3–α3 loop	Adjacent to GSH-binding His53
67–76	ESHAIMIYLV	helix α3	Contains GSH-binding Ser68
78–83	KYGKDD	α3–β4 loop	G-site
99–110	ALHFESGVLFAR	helix α4	Leads into GSH-binding Arg112
140–145	LLEDTL	helix α5	C-terminal domain
156–162	TIADFSC	helix α6	Conserved C-domain motif
193–197	LPYYE	helix α8	Conserved C-terminal motif

Secondary-structure elements from PDB 2IMI; residue numbering as in *A. gambiae* Q7PVS6. Functional context is interpretive and literature-based, not an experimental assignment.

3.4. Domain architecture and functional residues

All three UniProtKB-annotated orthologues shared the canonical cytosolic GST architecture, with a GST N-terminal domain spanning residues 2–83 and a GST C-terminal domain beginning at residue 89 (ending at 221, 210 and 211 in *A. gambiae*, *Ae. aegypti* and *C. quinquefasciatus*, respectively), and identical InterPro and Pfam signatures (IPR004045/PF13417 and IPR004046/PF00043) (Fig. 3). The conservation of mapped functional residues is summarised in Table 5.

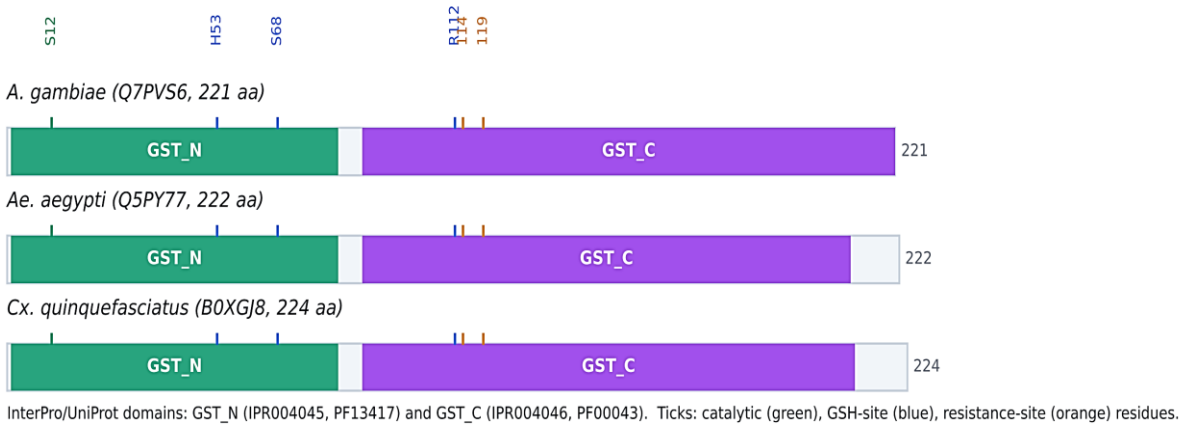


Fig. 3. Domain architecture of the three UniProtKB-annotated GSTe2 orthologues, drawn to scale. Green, GST N-terminal domain (IPR004045/PF13417); purple, GST C-terminal domain (IPR004046/PF00043). Ticks mark the catalytic serine (green), GSH-binding residues (blue) and resistance-associated positions (orange) in *A. gambiae* numbering.

The G-site was strongly constrained. Ser12, the catalytic serine of epsilon-class GSTs, was conserved in all four orthologues and in 11 of the 13 epsilon GSTs (two *Culex* paralogues carry Thr). His53 and Ser68, which contact GSH in the *A. gambiae* structure, were invariant in all 13 sequences, and Arg112 was conserved in all four orthologues and 11 of 13 epsilon GSTs. In contrast, the H-site was variable. Position 114 carried Ile in both *Anopheles* species, Val in *Aedes* and Leu in *Culex*. Leu119 was conserved in all four GSTe2 orthologues, including the deposited 3ZML structure; however, Phe occurred naturally at this position in five of the nine *Culex* paralogues. Phe120 was invariant among the orthologues but replaced by Tyr, Leu, Ser or His in eight of nine paralogues. Consistent with this, Leu119 and Phe120 were among 21 positions identical in all four orthologues but present in no more than two of the nine *Culex* paralogues (Supplementary Table S2), identifying them as GSTe2-specific features of the H-site.

Table 5. Catalytic, glutathione-binding and resistance-associated residues, classified by structural region (*A. gambiae* numbering).

Position	Structural region	Role / evidence	Ag	Af	Aa	Cq	n=13
Ser12	helix α1 (G-site)	Catalytic serine (epsilon class)	S	S	S	S	11
His53	G-site (2IMI)	GSH binding (experimental)	H	H	H	H	13
Ile55	G-site (2IMI)	GSH binding (experimental)	I	I	V	I	6
Ser68	helix α3 (G-site)	GSH binding (experimental)	S	S	S	S	13
Arg112	G-site (2IMI)	GSH binding (experimental)	R	R	R	R	11
Ile114	H-site	*A. gambiae I114T resistance site*	I	I	V	L	4
Glu116	H-site	Substrate-binding residue	E	E	E	E	9
Leu119	H-site	*A. funestus* L119F resistance site	L	L	L	L	6
Phe120	H-site	Substrate-binding residue	F	F	F	F	4

Ag, *A. gambiae*; Af, *A. funestus*; Aa, *Ae. aegypti*; Cq, *C. quinquefasciatus*. Final column: number of the 13 epsilon GSTs carrying the *A. gambiae* residue. "Experimental" denotes residues annotated from the GSH-bound structure PDB 2IMI; the catalytic serine is assigned from epsilon-class literature, not re-demonstrated here.

3.5. Phylogenetic confirmation of orthology

In the maximum-likelihood tree of the 13 epsilon GSTs (LG+G4; log-likelihood −3604.88), the four GSTe2 sequences formed a single clade with strong support (SH-aLRT 99.8%, UFBoot 95%), separating *Culex* B0XGJ8 from all nine other *Culex* epsilon GSTs (Fig. 4). The closest *Culex* paralogues, B0XGK0 and B0XGJ9, formed a separate, well-supported clade (100/97). Within the GSTe2 clade, the two *Anopheles* sequences grouped together, as did the *Aedes* and *Culex* sequences (SH-aLRT 92.5%, UFBoot 54%); in the four-taxon orthologue analysis this anopheline–culicine split received 100% UFBoot support. This topology matches the accepted divergence of the subfamilies Anophelinae and Culicinae. Other relationships among the *Culex* paralogues were resolved with variable support and are not interpreted further.

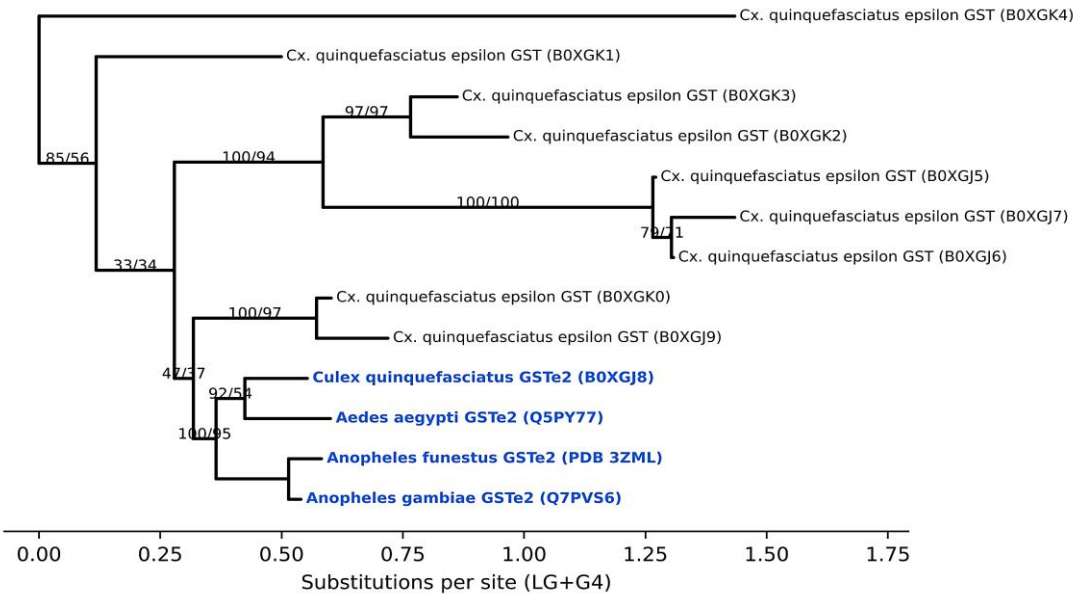


Fig. 4. Maximum-likelihood phylogeny of the four GSTe2 orthologues (blue) and nine additional *C. quinquefasciatus* epsilon GSTs (IQ-TREE 2, LG+G4, midpoint-rooted). Node labels give SH-aLRT (%)/UFBoot (%) support from 1000 replicates each. Scale: substitutions per site.

3.6. Residue conservation across the epsilon class

Variation was distributed unevenly between the two domains (Fig. 5). Of the 81 variable positions that map to the *A. gambiae* sequence, 22 lay in the N-terminal domain (residues 2–83; 26.8% of positions), two in the inter-domain linker and 57 in the C-terminal domain (residues 89–221; 42.9% of positions). Across the 13-sequence epsilon alignment, mean conservation was correspondingly higher in the N-terminal domain (0.75) than in the C-terminal domain (0.65), and the most conserved stretch overall surrounded the GSH-binding Ser68. A pronounced drop in conservation began near residue 112 and extended through the H-site segment containing positions 114–120, the same region that carries the resistance-associated substitutions.

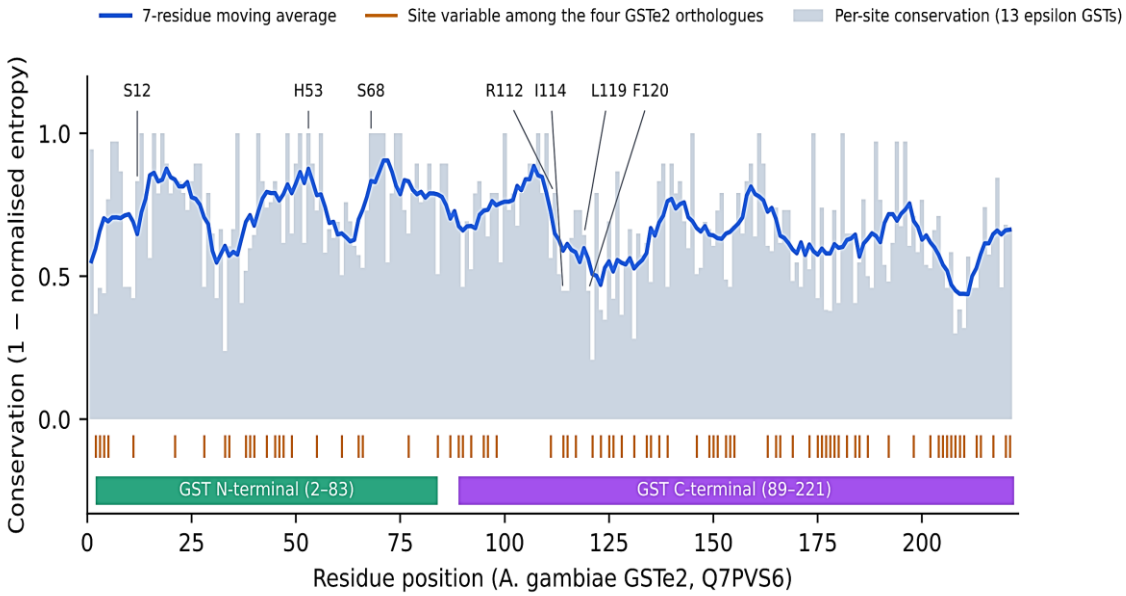


Fig. 5. Site-wise conservation (1 − normalised Shannon entropy) across the 13-sequence epsilon GST alignment, in *A. gambiae* GSTe2 numbering, with a 7-residue moving average. Orange ticks mark positions that vary among the four GSTe2 orthologues. Domain boundaries are from UniProtKB (PROSITE PS50404/PS50405).

3.7. Structural reference framework

Experimental structures were available for three of the four orthologues (Table 6). The *A. gambiae* (2IMI, 1.40 Å) and *A. funestus* (3ZML, 1.64 Å; 3ZMK, 2.20 Å) structures contain bound GSH and define the G-site and H-site residues mapped above; the *Ae. aegypti* structure (5FT3, 1.43 Å) contains no bound ligand. The sequence of 3ZML carries Leu at position 119 and therefore represents the L119 form of *A. funestus* GSTe2. Because the *Culex* GSTe2 shares 72.9–77.0% identity with these three proteins, with no internal insertions or deletions relative to them, its fold and G-site geometry can be inferred with confidence from them, whereas the conformation of its variable H-site cannot. PDB 5D73, a pi-class GST from *W. bancrofti*, shares neither class nor phylum with mosquito GSTe2 and was not used.

Table 6. Experimental structures used as the structural reference framework.

PDB	Organism	Protein	Method	Resolution	Ligand	Use
2IMI	*A. gambiae*	GSTe2	X-ray	1.40 Å	GSH	G-site/H-site reference
3ZML	*A. funestus*	GSTe2 (L119)	X-ray	1.64 Å	GSH	Sequence/resistance-site reference
3ZMK	*A. funestus*	GSTe2	X-ray	2.20 Å	GSH	Supporting reference
5FT3	*Ae. aegypti*	GSTe2	X-ray	1.43 Å	-	Culicine fold reference
5D73	*W. bancrofti*	Pi-class GST	X-ray	2.33 Å	-	Excluded (not GSTe2)

Metadata from the Protein Data Bank (accessed October 2026).

3.8. Functional association networks

The STRING networks differed in composition between the two species (Fig. 6). In *Ae. aegypti* (Fig. 6A), nine of the ten predicted partners of GSTe2 were glutathione- or redox-related enzymes, and the tenth was an uncharacterised protein: GSTs of the delta (GSTD3), theta (GSTt1, GSTT2, GSTT3, GSTT4), sigma (GSTS1) and omega (GSTO1) classes, the glutathione peroxidase GPXH1, the thioredoxin peroxidase TPX5 and the uncharacterised protein AAEL005801. The network was densely interconnected, with most edges supported by several evidence channels. In *A. gambiae* (Fig. 6B), GSTe2 again associated with a densely connected GST module (GSTe8, GSTD3, gstT1, gstT2, GstS1, GSTo1 and Q7PRA2), but it was additionally linked to the cytochrome P450s CYP6Z1 and CYP9K1 and to acetylcholinesterase (Ace). These three links formed a separate, sparsely connected module joined to GSTe2 by single-channel edges, consistent with text-mining evidence. Throughout, edges denote predicted functional associations and do not by themselves indicate direct physical interaction.

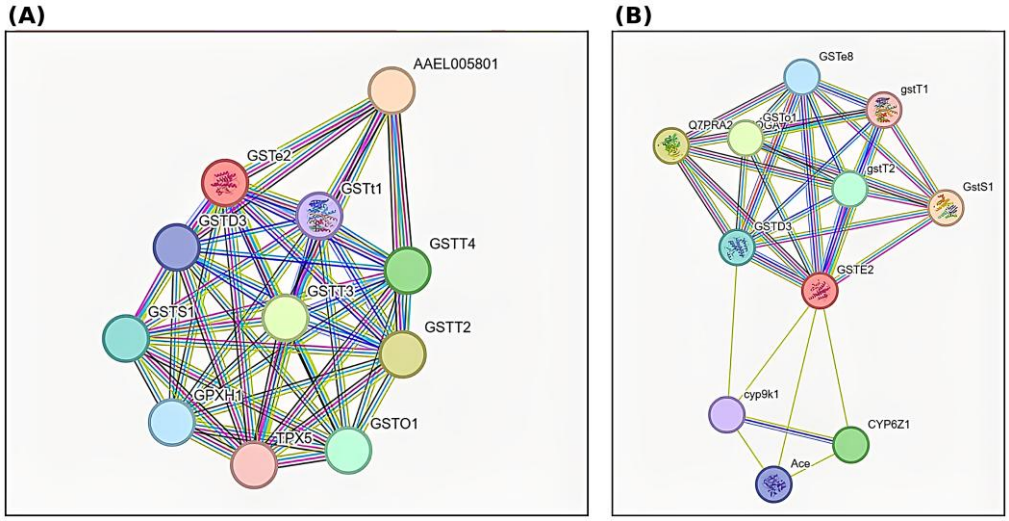


Fig. 6. STRING functional-association networks centred on GSTe2 in (A) *Ae. aegypti* and (B) *A. gambiae*. Nodes are proteins; edge colours denote the evidence channel supporting each association. Edges represent predicted functional associations and do not by themselves indicate direct physical interaction.

4. DISCUSSION

4.1. Orthology verification is a prerequisite for comparative GSTe2 analysis

Besides GSTe2, the *Culex* epsilon cluster contains nine genes whose products share up to 68% identity with it. Without explicit orthology testing, any of these could be mistaken for GSTe2, and because the paralogues differ precisely at H-site positions such as 114, 119 and 120 (Table 5), a misassignment would directly distort conclusions about substrate binding and resistance. Combining checksum verification with a maximum-likelihood test against the complete cluster resolved B0XGJ8 as the *Culex* GSTe2 with strong support. We recommend that comparative studies of epsilon GSTs report the accession, length and orthology evidence for every sequence used, because similarly sized paralogues can otherwise be substituted without detection.

4.2. A conserved G-site and a variable H-site

The conservation pattern mirrors the functional architecture of cytosolic GSTs [7,9]. The G-site, which binds the same co-substrate in every GST, was almost invariant: Ser12, His53 and Ser68 were conserved across all orthologues and nearly all paralogues. The H-site, which determines substrate range, absorbed most of the variation: the C-terminal domain carried a higher proportion of variable positions than the N-terminal domain, and the steepest decline in conservation coincided with the segment containing residues 112–120. In *A. gambiae* GSTe2 this region forms part of a compact substrate pocket associated with the enzyme's high DDT-dehydrochlorinase activity [15]. Our data indicate that this pocket is where GSTe2 orthologues have diverged among genera, while the catalytic machinery has been preserved.

4.3. Natural variation at characterised resistance positions

Two observations relate to resistance biology. First, position 114 differs among the four vectors (Ile in both *Anopheles* species, Val in *Aedes*, Leu in *Culex*); in *A. gambiae*, replacement of Ile114 by Thr accompanies DDT resistance [16]. The natural occurrence of three different residues at this position suggests that the local environment tolerates substitution and that the functional consequence depends on the specific change rather than on simple variability. Second, Leu119 was the wild-type residue in all four GSTe2 orthologues, including the deposited 3ZML structure, yet Phe, the residue of the resistant *A. funestus* allele [19], occurred naturally at this position in several *Culex* epsilon paralogues. This shows that the Phe119 side chain is compatible with a folded, functional epsilon GST in *Culex*, and makes residue 119 of *Culex* GSTe2 a logical target for resistance surveillance. These are comparative, hypothesis-generating observations; attributing a resistance phenotype to any residue in *Aedes* or *Culex* GSTe2 requires direct biochemical and genetic testing.

4.4. GSTe2 within a detoxification context

GSTe2 does not act in isolation: the epsilon GSTs form a co-localised gene cluster whose members are co-expressed and collectively up-regulated in resistant *A. gambiae*, and metabolic resistance in the field is typically polygenic, involving GSTs together with cytochrome P450s and esterases [5,12,13]. The nine additional *Culex* epsilon paralogues analysed here are the counterparts of that cluster in *C. quinquefasciatus* and are candidates for the same kind of coordinated regulation. Any such relationship is a functional association inferred from genomic co-localisation and co-expression, not a direct physical protein–protein interaction; establishing physical partners would require co-immunoprecipitation, yeast two-hybrid or other direct interaction assays. We therefore frame the broader detoxification context as a set of candidate functional associations to be tested, rather than as an interaction network.

4.5. Physicochemical conservation

The four orthologues were closely matched in mass, charge, predicted stability and hydropathy, as expected for proteins under strong structural constraint. All were acidic and predicted stable, consistent with soluble cytosolic enzymes. The near-identical pI and GRAVY of the *Aedes* and *Culex* proteins parallel their higher pairwise identity and shared position in the phylogeny, whereas the more acidic, more hydrophilic *Anopheles* proteins form the other orthologue pair.

4.6. Conclusions and future work

This study establishes an orthology-verified, checksum-controlled framework for comparing GSTe2 across the three major mosquito vector genera. By resolving the true *C. quinquefasciatus* orthologue (B0XGJ8) from a ten-member epsilon cluster, and by anchoring all structural interpretation to experimentally determined mosquito GSTe2 structures while

explicitly excluding the misassigned *Wuchereria bancrofti* template PDB 5D73, it removes two recurrent sources of error that can silently distort comparative and structural analyses of this enzyme. Within this verified framework, mosquito GSTe2 emerges as a two-speed protein: a highly conserved N-terminal glutathione-binding core, in which the G-site residues His53 and Ser68 are invariant across all thirteen epsilon GSTs examined and the catalytic Ser12 and Arg112 are conserved in eleven of them, coupled to a markedly more variable C-terminal substrate-binding domain that concentrates the differences between vectors, including at the experimentally defined resistance positions 114 and 119.

This pattern is biologically coherent. The machinery that every GST must preserve in order to bind glutathione is held constant, whereas the H-site that defines substrate range, and therefore the capacity to metabolise particular insecticides, is where lineage-specific divergence accumulates. The natural occurrence of three different residues at position 114, and of the *A. funestus* resistance residue phenylalanine at position 119 in several *Culex* paralogues, indicates that these functionally pivotal positions are evolutionarily labile, and identifies them as priority candidates for resistance surveillance in *Aedes* and *Culex*, where GSTe2 has received far less attention than in *Anopheles*. We stress, however, that these are comparative, hypothesis-generating observations: sequence conservation and structural similarity establish constraint and plausibility, not catalytic activity or a resistance phenotype.

Translating this framework into mechanistic understanding will require targeted experiments. The most informative next steps are recombinant expression of *Aedes* and *Culex* GSTe2 with direct DDT and pyrethroid metabolism assays; site-directed mutagenesis at positions 114 and 119 to isolate their functional contribution; and population-level genotyping of these positions alongside measured resistance phenotypes in field populations. Where an experimental structure is still lacking, most notably for *Culex* GSTe2, a homology model built on a verified mosquito template (2IMI, 5FT3 or 3ZML) and validated by standard stereochemical criteria would allow the variable H-site to be examined in three dimensions. Together, such work would connect the conserved-core and variable-H-site architecture described here to the metabolic resistance phenotypes observed in the field, and would help determine whether GSTe2 can serve as a cross-vector molecular marker for monitoring insecticide resistance. More broadly, the orthology-verification and checksum-control steps applied here offer a template for rigorous comparative analysis of other detoxification gene families that, like the epsilon GSTs, are organised as closely related clusters prone to paralogue misassignment.

REFERENCES

1. World Health Organization. World Malaria Report 2023. Geneva: World Health Organization; 2023.
2. World Health Organization. Manual for Monitoring Insecticide Resistance in Mosquito Vectors and Selecting Appropriate Interventions. Geneva: World Health Organization; 2022.
3. Hemingway J, Ranson H, Magill A, Kolaczinski J, Fornadel C, Gimnig J, et al. Averting a malaria disaster: will insecticide resistance derail malaria control? Lancet 2016;387:1785–1788. https://doi.org/10.1016/S0140-6736(15)00417-1
4. Hemingway J, Ranson H. Insecticide resistance in insect vectors of human disease. Annu Rev Entomol 2000;45:371–391. https://doi.org/10.1146/annurev.ento.45.1.371
5. Hemingway J, Hawkes NJ, McCarroll L, Ranson H. The molecular basis of insecticide resistance in mosquitoes. Insect Biochem Mol Biol 2004;34:653–665. https://doi.org/10.1016/j.ibmb.2004.03.018
6. Liu N. Insecticide resistance in mosquitoes: impact, mechanisms, and research directions. Annu Rev Entomol 2015;60:537–559. https://doi.org/10.1146/annurev-ento-010814-020828
7. Hayes JD, Flanagan JU, Jowsey IR. Glutathione transferases. Annu Rev Pharmacol Toxicol 2005;45:51–88. https://doi.org/10.1146/annurev.pharmtox.45.120403.095857
8. Oakley AJ. Glutathione transferases: new functions. Curr Opin Struct Biol 2005;15:716–723. https://doi.org/10.1016/j.sbi.2005.10.005
9. Frova C. Glutathione transferases in the genomics era: new insights and perspectives. Biomol Eng 2006;23:149–169. https://doi.org/10.1016/j.bioeng.2006.05.020
10. Enayati AA, Ranson H, Hemingway J. Insect glutathione transferases and insecticide resistance. Insect Mol Biol 2005;14:3–8. https://doi.org/10.1111/j.1365-2583.2004.00529.x
11. Ranson H, Hemingway J. Mosquito glutathione transferases. Methods Enzymol 2005;401:226–241. https://doi.org/10.1016/S0076-6879(05)01014-1
12. Ranson H, Rossiter L, Ortelli F, Jensen B, Wang X, Roth CW, et al. Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector *Anopheles gambiae*. Biochem J 2001;359:295–304. https://doi.org/10.1042/0264-6021:3590295
13. Ding Y, Ortelli F, Rossiter LC, Hemingway J, Ranson H. The *Anopheles gambiae* glutathione transferase supergene family: annotation, phylogeny and expression profiles. BMC Genomics 2003;4:35. https://doi.org/10.1186/1471-2164-4-35
14. Ortelli F, Rossiter LC, Vontas J, Ranson H, Hemingway J. Heterologous expression of four glutathione transferase genes genetically linked to a major insecticide-resistance locus from the malaria vector *Anopheles gambiae*. Biochem J 2003;373:957–963. https://doi.org/10.1042/BJ20030169
15. Wang Y, Qiu L, Ranson H, Lumjuan N, Hemingway J, Setzer WN, et al. Structure of an insect epsilon class glutathione S-transferase from the malaria vector *Anopheles gambiae* provides an explanation for the high DDT-detoxifying activity. J Struct Biol 2008;164:228–235. https://doi.org/10.1016/j.jsb.2008.08.003
16. Mitchell SN, Rigden DJ, Dowd AJ, Lu F, Wilding CS, Weetman D, et al. Metabolic and target-site mechanisms combine to confer strong DDT resistance in *Anopheles gambiae*. PLoS One 2014;9:e92662. https://doi.org/10.1371/journal.pone.0092662

17. Lumjuan N, McCarroll L, Prapanthadara L, Hemingway J, Ranson H. Elevated activity of an Epsilon class glutathione transferase confers DDT resistance in the dengue vector, *Aedes aegypti*. Insect Biochem Mol Biol 2005;35:861–871. https://doi.org/10.1016/j.ibmb.2005.03.008
18. Lumjuan N, Rajatileka S, Changsom D, Wicheer J, Leelapat P, Prapanthadara L, et al. The role of the *Aedes aegypti* Epsilon glutathione transferases in conferring resistance to DDT and pyrethroid insecticides. Insect Biochem Mol Biol 2011;41:203–209. https://doi.org/10.1016/j.ibmb.2010.12.005
19. Riveron JM, Yunta C, Ibrahim SS, Djouaka R, Irving H, Menze BD, et al. A single mutation in the GSTe2 gene allows tracking of metabolically based insecticide resistance in a major malaria vector. Genome Biol 2014;15:R27. https://doi.org/10.1186/gb-2014-15-2-r27
20. Sarkar M, Bhattacharyya IK, Borkotoki A, Goswami D, Rabha B, Baruah I, et al. Insecticide resistance and detoxifying enzyme activity in the principal bancroftian filariasis vector, *Culex quinquefasciatus*, in northeastern India. Med Vet Entomol 2009;23:122–131. https://doi.org/10.1111/j.1365-2915.2009.00805.x
21. Rai P, Bharati M, Subba A, Saha D. Insecticide resistance mapping in the vector of lymphatic filariasis, *Culex quinquefasciatus* Say from northern region of West Bengal, India. PLoS One 2019;14:e0217706. https://doi.org/10.1371/journal.pone.0217706
22. Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, et al. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science 2010;330:86–88. https://doi.org/10.1126/science.1191864
23. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 2023;51:D523–D531. https://doi.org/10.1093/nar/gkac1052
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. https://doi.org/10.1093/nar/28.1.235
25. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163
26. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005. p. 571–607. https://doi.org/10.1385/1-59259-890-0:571
27. Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, et al. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis 1993;14:1023–1031. https://doi.org/10.1002/elps.11501401163
28. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition. Protein Eng 1990;4:155–161. https://doi.org/10.1093/protein/4.2.155
29. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem 1980;88:1895–1898. https://doi.org/10.1093/oxfordjournals.jbchem.a133168
30. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132. https://doi.org/10.1016/0022-2836(82)90515-0
31. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T. How to measure and predict the molar absorption coefficient of a protein. Protein Sci 1995;4:2411–2423. https://doi.org/10.1002/pro.5560041120
32. Deorowicz S, Debudaj-Grabysz A, Gudys A. FAMSA: fast and accurate multiple sequence alignment of huge protein families. Sci Rep 2016;6:33964. https://doi.org/10.1038/srep33964
33. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 2020;37:1530–1534. https://doi.org/10.1093/molbev/msaa015
34. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 2017;14:587–589. https://doi.org/10.1038/nmeth.4285
35. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol 2008;25:1307–1320. https://doi.org/10.1093/molbev/msn067
36. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 2018;35:518–522. https://doi.org/10.1093/molbev/msx281
37. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307–321. https://doi.org/10.1093/sysbio/syq010
38. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 2021;49:D344–D354. https://doi.org/10.1093/nar/gkaa977
39. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 2023;51:D638–D646. https://doi.org/10.1093/nar/gkac1000