ARTICLE

Structural genomics sheds light on protein functions and remote homologs across the insect tree of life

Weiyin Wu[1,9], Chunlai Cui[2,3,9], Yixiao Zhu[1,4,9], Jingxuan Chen[1], Qiancheng Zhuang[1], Yazhou Wang[1], Zicheng Liu[2], Han Gao[2], Guo-Zheng Ou[1], Chao Liu[1], Mei Tao[1], Yun Chen[1], Ronghui Pan[5], Guojie Zhang[6], Hua Cai[7], Jinghua Yang[1,4], Xue-xin Chen[1], Xiaofan Zhou[8]✉, Sibao Wang[2]✉ and Xing-Xing Shen[1,4,6]✉

Protein structure bridges the sequence–function relationship, enabling deep exploration of biological processes across diverse organisms. Insects, the most diverse animal lineage, accounting for over 50% of all described animal species, provide an exceptional system for exploring sequence–structure–function relationships. Here, we reconstructed a comprehensive and well-resolved phylogeny of 4854 insects, spanning all orders. Leveraging this framework, we created an atlas of 13.29 million predicted protein structures from 824 representative species, including 11.63 million newly predicted structures. Structural clustering revealed that proteins with divergent sequences but similar structures could be effectively grouped together. Structural similarity searches against proteins with well-characterized functions yielded annotations for 7.61 million insect proteins, including up to 14% of previously unannotated proteins. We further identified 750 million remote homologs between insect proteins, many of which trace back to ancient branches of the insect phylogeny. Remarkably, despite extensive sequence divergence, cGAS-like receptors (cGLRs) were structurally conserved across all 824 insects. Experimental assays demonstrated that these structurally identified cGLRs play a crucial role in antiviral defense in the yellow fever mosquito. Our findings highlight the significance of structural genomics for understanding protein function and evolution across the tree of life.

Cell Research (2026) 0:1–15; https://doi.org/10.1038/s41422-026-01220-0

INTRODUCTION

The genomics revolution, fueled by rapid advances in genome-sequencing technologies, has dramatically expanded our knowledge of molecular data and transformed our understanding of the genetic diversity and evolutionary relationships of life on Earth.[1–8] Among genomic data, protein sequences (i.e., proteomes) are rich archives for studying the molecular functions that govern cellular processes.[9,10] In comparative genomics, sequence-based homology searches have been a standard approach for the transfer of functional annotations between homologous proteins in different organisms. However, this approach often loses sensitivity when attempting to decipher the relationships between remote (or distantly related) protein homologs,[11–13] limiting the full functional characterization of proteomic data.

According to the sequence–structure–function paradigm in biology,[14,15] the three-dimensional structure of a protein, derived from its primary sequence, is a key determinant of its molecular function. Protein structures typically evolve more slowly than sequences, likely owing to constraints imposed by protein folding and preservation of function[11,16–20]; thus, protein structures and functions may be conserved over longer evolutionary timescales than protein sequences. Recent breakthrough tools in accurate structure prediction, such as AlphaFold,[21] ColabFold,[22] ESMFold,[23] and RoseTTAFold,[24] now enable structural predictions at large scale. At the same time, protein structures with well-characterized functions, available from the AlphaFold database's Swiss-Prot[25] and the Protein Data Bank (PDB),[26] together with curated domain structures from the CATH database,[27,28] collectively provide a highly precise resource for protein functional annotation. However, despite these advances in genomics, structure prediction, and functional annotation, there has not yet been a comprehensive exploration of sequence–structure–function relationships across an entire biological system within the tree of life.

In this study, we focused on the class Insecta, a highly diverse group that includes over 50% of all described living animals[29–31] and has a profound impact on ecosystems, economies, and human societies. We compiled publicly available genome and transcriptome data and used it to reconstruct a comprehensive

[1]Zhejiang Key Laboratory of Biology and Ecological Regulation of Crop Pathogens and Insects, State Key Laboratory for Vegetation Structure, Function and Construction, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang, China. [2]New Cornerstone Science Laboratory, Key Laboratory of Insect Developmental and Evolutionary Biology, State Key Laboratory of Plant Trait Design, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China. [3]Shanghai Institute of Wildlife Epidemics, School of Life Sciences, East China Normal University, Shanghai, China. [4]Hainan Institute, Zhejiang University, Yazhou Bay Science and Technology City, Sanya, Hainan, China. [5]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou, Zhejiang, China. [6]Centre for Evolutionary & Organismal Biology, Zhejiang University, Hangzhou, Zhejiang, China. [7]Sino-French Hoffmann Institute, School of Basic Medical Science, Guangzhou Medical University, Guangzhou, Guangdong, China. [8]Guangdong Laboratory for Lingnan Modern Agriculture, Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou, Guangdong, China. [9]These authors contributed equally: Weiyin Wu, Chunlai Cui, Yixiao Zhu. ✉email: xiaofan_zhou@scau.edu.cn; sbwang@cemps.ac.cn; xingxingshen@zju.edu.cn

Received: 8 September 2025 Accepted: 8 January 2026
Published online: 28 January 2026

and well-resolved phylogeny of 4854 insects representing all 28 insect orders. By sampling 824 representative insects, we generated an atlas of predicted structures for 13.29 million proteins. Analysis of structural clusters demonstrated that the use of protein structures substantially increases the sensitivity of homologous protein detection compared with the use of protein sequences. Structural similarity searches yielded functional annotations for 7.61 million proteins, primarily derived from functionally well-characterized non-insect proteins. Finally, we identified 750 million remote homologous proteins characterized by low sequence identity (< 0.25) but high structural similarity (> 0.5) across the insect tree of life. In particular, we discovered that cGAS-like receptors (cGLRs) are structurally conserved in all 824 insects, and we experimentally validated their role in antiviral defense in the yellow fever mosquito, a vector of human arboviruses.

RESULTS

A comprehensive insect tree of life

To infer a comprehensive phylogeny of the class Insecta as of October 11, 2023, we gathered genomes (1724) and transcriptomes (3130) of 4854 insects from 17 public repositories (Supplementary information, Table S1). These 4854 insects represent all 28 orders, when Blattodea including Isoptera is considered monophyletic[31,32] (Supplementary information, Fig. S1 and Table S2). Analysis of assembly completeness revealed that 3110 out of 4854 insects (~64%) had at least 80% of the 1367 full-length Benchmarking Universal Single-Copy Ortholog (BUSCO) genes[33] (Supplementary information, Fig. S2).

After constructing a multiple amino acid sequence alignment and trimming ambiguous regions for each of the 1367 BUSCO genes from 4854 insects and 10 Entognatha outgroups (Supplementary information, Fig. S3), we retained the 824 BUSCO genes with a taxon occupancy of ≥ 50% and an alignment length of ≥ 150 (Supplementary information, Fig. S4). We inferred a concatenation-based maximum likelihood (ML) phylogeny (Fig. 1; Supplementary information, Fig. S5) and a coalescent-based ASTRAL phylogeny (Supplementary information, Fig. S6). Average branch support was 99.8% for the concatenation-based phylogeny and 95.0% for the coalescent-based phylogeny (Supplementary information, Fig. S7). Comparison of the concatenation- and coalescent-based phylogenies revealed that 3731 (77%) internodes were topologically identical, whereas the remaining 1122 (23%) were topologically incongruent. Among these 1122 incongruent internodes, 842 (75%) involved relationships within families, whereas the remaining 280 (25%) involved relationships within orders.

Our concatenation- and coalescent-based phylogenies strongly supported the monophyly of each of the 28 orders and robustly resolved all intraordinal relationships with at least 95% support (Fig. 1; Supplementary information, Fig. S8). These higher-level phylogenies were consistent with those from the landmark phylogenomics study of Misof et al.[31] which used a moderate number (126) of insect genomes and transcriptomes. However, our two comprehensive phylogenies demonstrated improved resolution and robustness. For instance, our concatenation- and coalescent-based phylogenies recovered the order Raphidioptera (21 taxa) as the sister group to the orders Megaloptera (11 taxa) and Neuroptera (66 taxa) with 100% support, whereas the landmark phylogenomics study (Raphidioptera: 2 taxa; Megaloptera: 2 taxa; Neuroptera: 4 taxa) provided low support for this relationship (bootstrap values = 2%–30%) across different amino acid datasets. In addition, our comprehensive phylogenies robustly resolved the most contentious relationships among ten orders of Polyneoptera.[34–36] Overall, our results highlight the importance of comprehensive taxon sampling for improving resolution and robustness in phylogenomics. The thorough and well-resolved insect tree of life also laid a solid foundation for subsequent structural genomics investigations.

An atlas of protein structures in insects

By leveraging the new insect tree of life (Fig. 1), we created an atlas of 13.29 million predicted protein structures from 824 insect representatives spanning all 28 orders (Supplementary information, Fig. S9 and Table S3). This included 1.66 million structures (12.5%) retrieved from the AlphaFold database[21,25] and 11.63 million structures (87.5%) newly generated using ESMFold[23] (Fig. 2a). In our preliminary experiment, we randomly selected 267,694 proteins from 16 insects to compare the consistency of structure predictions between AlphaFold2 and ESMFold. We observed a strong correlation between the confidence of our ESMFold predictions and their structural similarities to AlphaFold2 predictions (Pearson's correlation coefficient $r = 0.82$, P-value < 2.2×10^{-16}) (Supplementary information, Fig. S10), consistent with a previous study.[23]

To explore the diversity of protein structures among 824 insects, we clustered the 13.29 million protein structures using the Foldseek *cluster* module,[17,37] with a structural alignment threshold of 70% coverage, TM-score of 0.4, *E*-value of 0.001, sensitivity of 7.5, and cluster reassignment value of 1 (Fig. 2a). This analysis yielded 4.22 million structural clusters, 11.3% of which were non-singleton clusters that contained 9.54 million proteins, representing 72% of the total 13.29 million proteins. This percentage of non-singleton clusters falls between those reported in the AFESM (AFDB + ESM) database (7%)[38] and the AFDB database (27%).[17] We next examined structure prediction confidence scores (predicted local distance difference test, pLDDT) and found that as cluster size increased, the proportion of cluster members with very low confidence scores (pLDDT < 50) decreased, stabilizing below 8% starting from clusters containing ≥ 10 members (Supplementary information, Fig. S11). In addition, proteins in large clusters (≥ 10 members) had lower proportions of very-low-confidence structures (pLDDT < 50: 3% vs 56%), short sequences (< 100 amino acids: 2% vs 14%), and intrinsically disordered proteins (IDPs: 12% vs 41%) compared with proteins in small clusters (< 10 members) (Fig. 2b). This observation is consistent with previous findings.[20,39,40] Therefore, our subsequent analyses focused on the 87,461 large structural clusters with at least 10 members.

To gauge the reliability of structural clustering for these 87,461 clusters, we assessed the structural similarity of each cluster using LDDT and TM-score metrics, as described in previous studies.[17,38] Our analysis showed median structural similarity values of 0.74 for LDDT and 0.63 for TM-score (Fig. 2c), indicating that the clusters tended to be structurally homogeneous. Analysis of taxonomic distributions revealed that 21,451 clusters (24.53%) were located at the branch leading to the most recent common ancestor of insects (Insecta), 51,960 (59.41%) at inter-ordinal branches (between orders), 12,981 (14.84%) at internal branches within specific orders (order-specific), and 1069 (1.22%) at external branches within specific species (species-specific) (Fig. 2d). The majority of clusters thus occurred along inter-ordinal branches of the insect tree of life. In addition, we investigated whether proteins within each structural cluster identified by the structure-based method could be reconstituted using the sequence-based method MMseqs2[41] under the same criteria. The sequence-based method failed to group all structurally similar proteins into a single cluster, with a median of 7 split clusters (Supplementary information, Fig. S12a, b). For example, structural cluster 5407 was split into 28 small sequence-based clusters, and structural cluster 7528 was split into 31 such clusters, owing to very low sequence identity (Supplementary information, Fig. S12c). This suggests that structural analysis can be more effective than sequence analysis for identifying similarity between insect proteins.

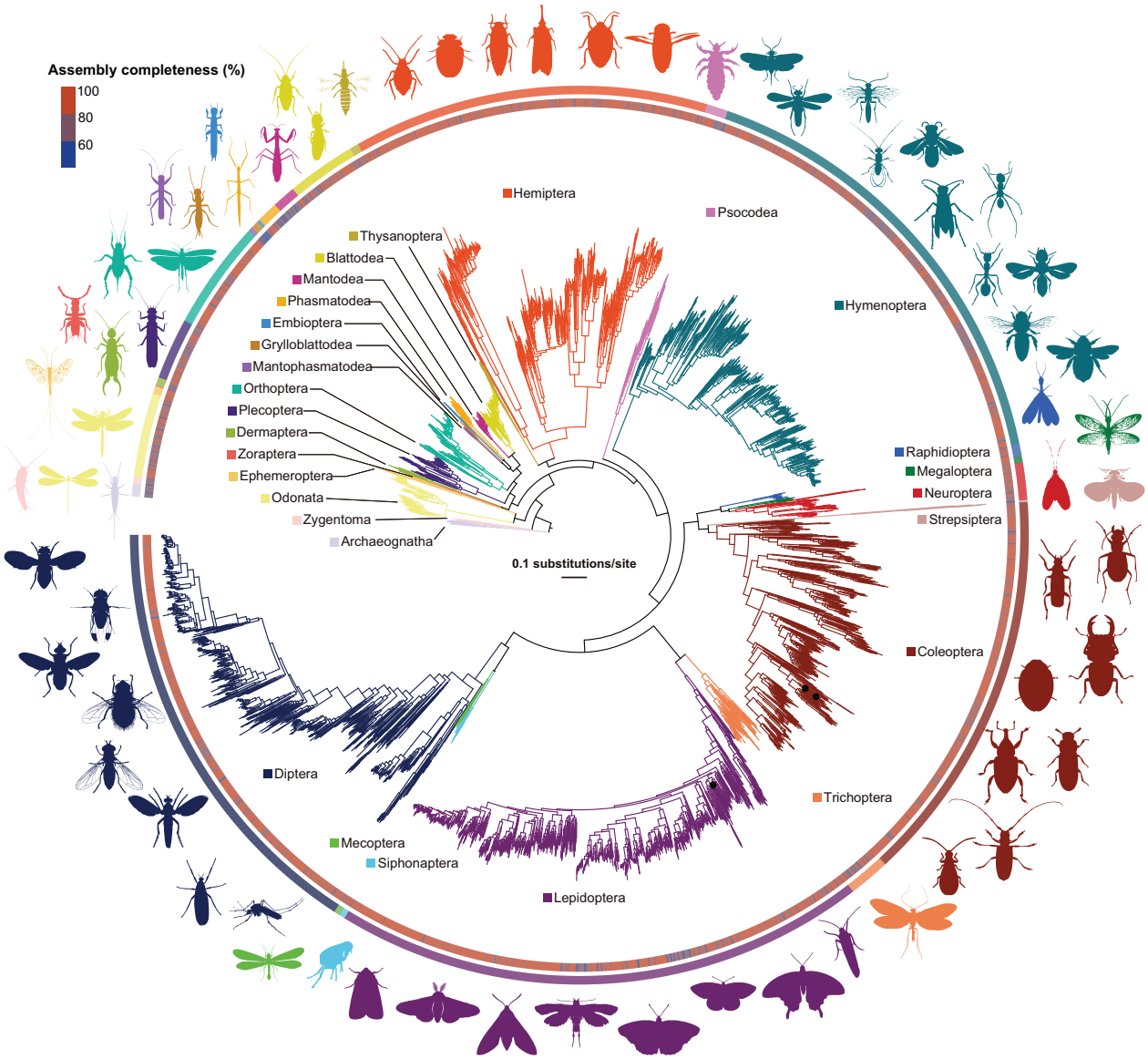


Fig. 1 A comprehensive and well-resolved phylogeny of 4854 insects. The concatenation-based ML phylogeny ($lnL = -191659866.544$) was inferred from amino acid sequences of 824 BUSCO genes (total 276,683 sites) under a single LG + G4 substitution model using IQ-TREE multicore v2.0.7. The complete phylogenetic relationships of 4854 insects, spanning all 28 orders,[31,32] are given in Supplementary information, Fig. S5. Branch support values near internal branches correspond to ultrafast bootstrap support. The only three internal branches (two within the order Coleoptera and one within the order Lepidoptera) with support values smaller than 95% are indicated with solid black dots. The branches and outer circle are colored according to their order names. The inner circle shows assembly completeness assessed with a set of 1367 conserved BUSCO genes. We also reconstructed a coalescent-based phylogeny of 4854 insects, which can be found in Supplementary information, Fig. S6. Note that the 10 Entognatha outgroups are not shown in the tree. Images representing taxa were obtained from the PhyloPic website (http://phylopic.org).

Structure-based exploration of insect functional genomics

To gain insight into the functions of the 87,461 large clusters with at least 10 members, we performed structure-based annotations using structural databases, including full-length structures with well-characterized functions from the AFDB Swiss-Prot[25] and PDB[26] databases and curated domain structures from the CATH SSG5 database,[27,28] following previous structural genomics studies.[28,38] We found that the clusters had median functional annotation consistency values (that is, the fraction of functional annotations from the highest-confidence representative shared within the cluster) of 0.89 and 0.96 for the full-length structure-based and domain structure-based approaches (Fig. 3a), respectively. This indicates that annotation of a cluster representative can reflect the overall cluster annotation. Consequently, we successfully annotated 64,356 clusters (73.6%; totaling 7.48 million proteins) via the full-length structure-based method (Fig. 3b). For the remaining clusters that were not annotated by the full-length structure-based method, we used the domain structure-based method and found that 4008 (4.6%; 0.13 million proteins) were annotated (Fig. 3b). Together, these analyses assigned functional annotations to 68,364 clusters comprising 7.61 million proteins (92% of the total 8.24 million proteins). Notably, 14.4% of these functionally annotated proteins, identified through structure-based methods, could not be annotated using sequence-based approaches in a similar manner. These proteins exhibited a wide range of functions, including involvement in cellular processes, development, response to stimuli, reproduction, the immune system, and detoxification.

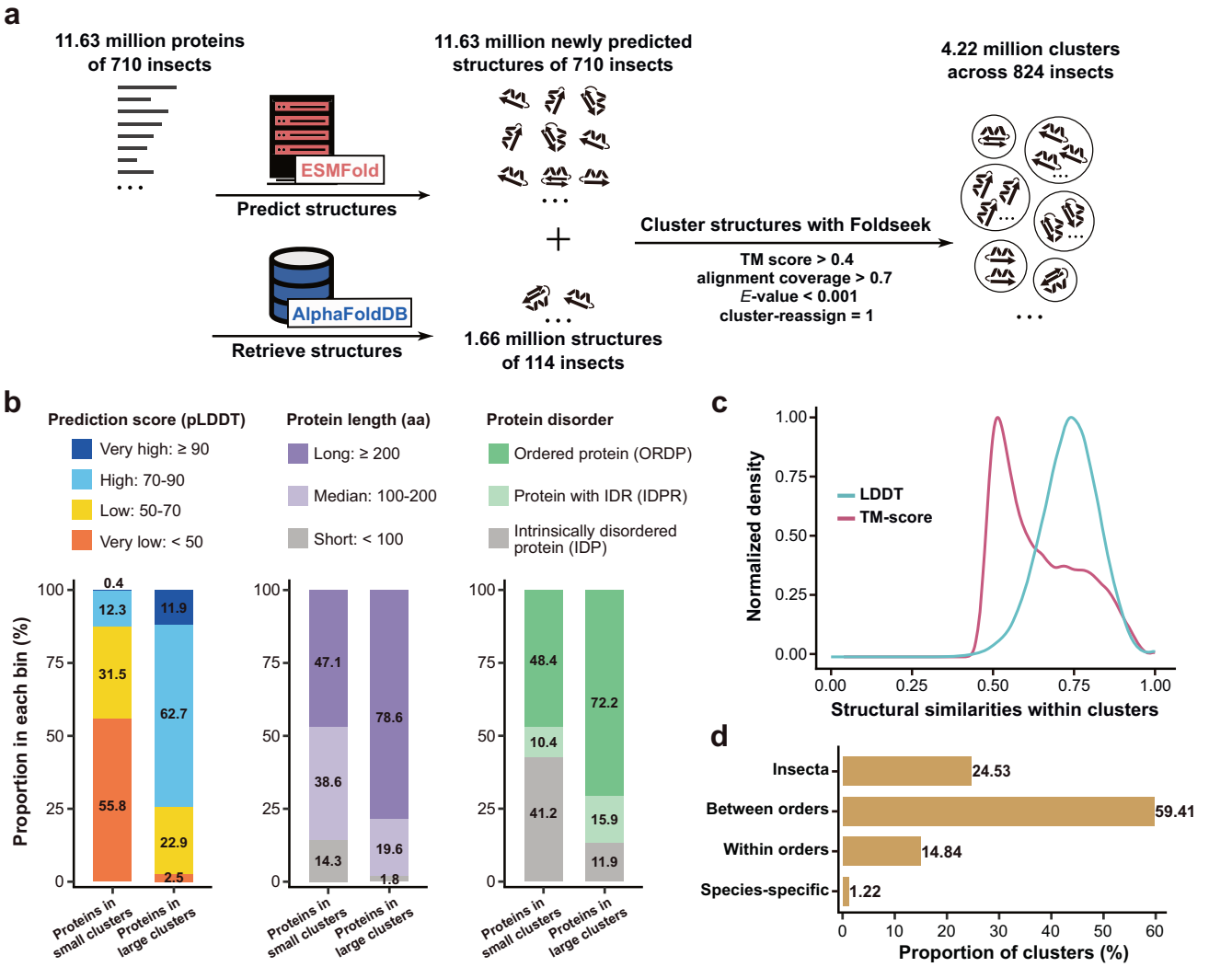
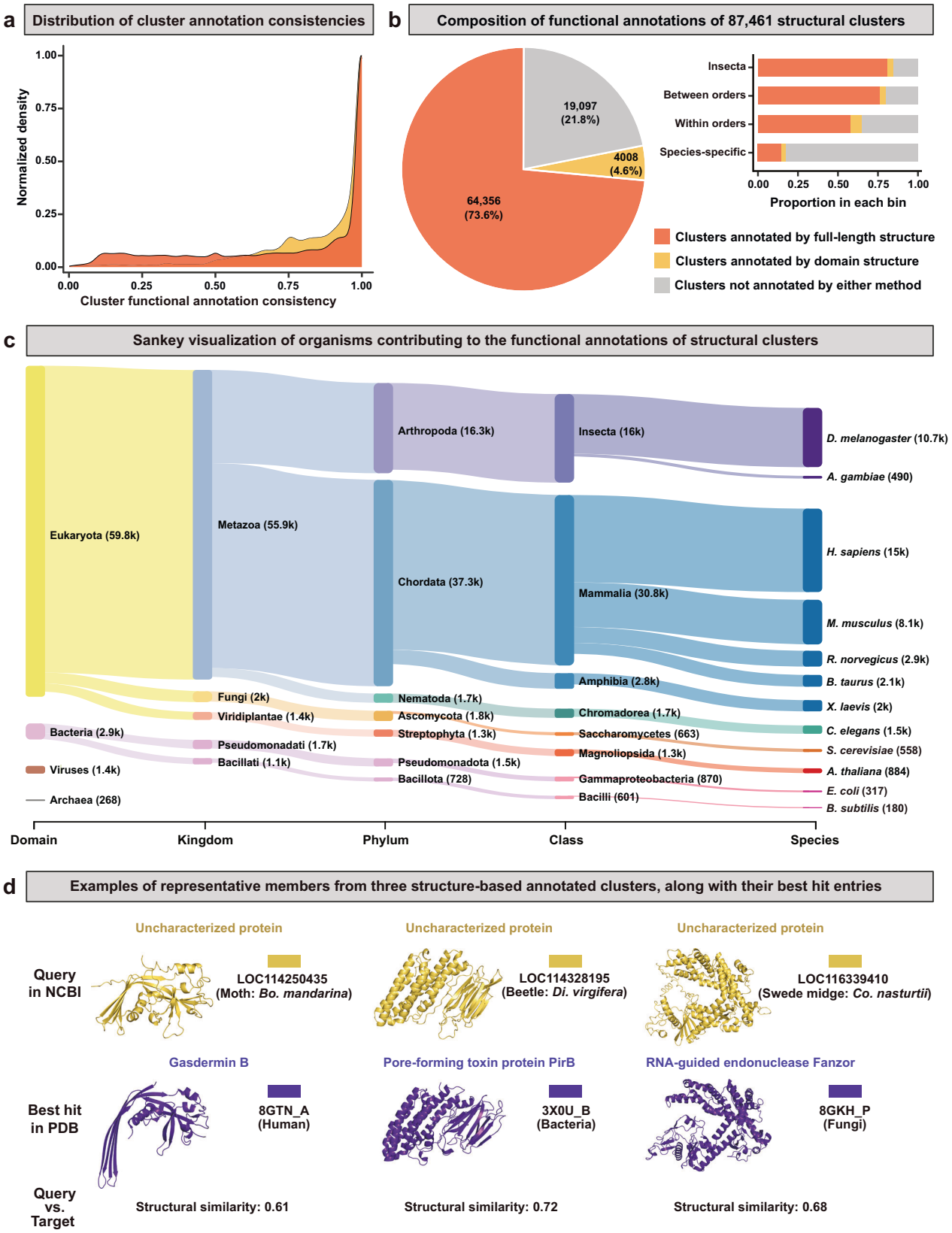


Fig. 2 The insect protein-structure universe and statistics. a Workflow for generating and clustering a structural atlas of 13.29 million proteins from 824 insects representing all 28 orders, including 11.63 million newly predicted structures. **b** Proportions of proteins classified on the basis of three properties in small clusters (< 10 members) and large clusters (≥ 10 members). The cluster size threshold of 10 was selected because the proportion of cluster members with very low confidence scores stabilized starting from clusters with ≥ 10 members (Supplementary information, Fig. S11). The three properties were prediction confidence score, protein length, and protein disorder level. **c** Distribution of structural similarities within 87,461 large structural clusters (≥ 10 members). Following previous studies,[17,38] structural similarities within each cluster were assessed by calculating the average of all member-to-representative LDDT and TM-scores. **d** Summary of high-level taxonomic classifications of 87,461 structural clusters across the tree of life for 824 insects.

We next mapped the taxonomic distributions of the 68,364 annotated clusters and 19,097 unannotated clusters onto the phylogeny of 824 insect species. Our analysis showed that species-specific clusters had a substantially higher proportion of unannotated clusters compared with those involving two or more species (Fig. 3b). To identify the organisms that contributed to the annotations of our structural clusters, we extracted taxonomic information from the best hits of the cluster representatives. The cluster functional annotations were derived from a wide range of species, including animals, fungi, plants, bacteria, archaea, and viruses. Most of these species were non-insect model organisms such as *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Escherichia coli* (Fig. 3c). Gene Ontology (GO) analysis of the cluster representatives revealed diverse functions, including many not previously reported in insects (Supplementary information, Fig. S13a). Representative examples are shown in Fig. 3d: Cluster 10381, including 138 proteins from 138 insects covering five different orders, was structurally homologous to human Gasdermin B (PDB: 8GTN_A), an executor of inflammasome-dependent pyroptotic cell death[42–46]; Cluster 20226, containing 57 proteins from 19 beetles, was structurally homologous to the bacterial pore-forming toxin protein PirB (PDB: 3X0U_B) from a family of insecticidal toxins[47–49]; and Cluster 18826, including 63 proteins from *Contarinia nasturtii*, was structurally homologous to fungal Fanzor (PDB: 8GKH_P), a eukaryotic programmable RNA-guided DNA endonuclease for genome editing.[50–54]

We predicted the functions of the 19,097 large clusters (21.8% of the total) that were not annotated by either the full-length or the domain structure-based method using the structure-based functional prediction tool DeepFRI.[55] We found that these proteins exhibited distinct functions compared with proteins successfully annotated with well-characterized proteins (Supplementary information, Fig. S13a, b). For example, some of these unannotated clusters were predicted to participate in membrane transport activities, a functional category that was underrepresented in clusters annotated with functionally well-characterized proteins.



Lastly, for the remaining small structural clusters with fewer than 10 members, we also performed structure-based functional annotations of their representatives with high-confidence predictions (pLDDT > 70) and found that 6.2% could be functionally annotated. To enhance access to all annotated functions of the insect proteins via structure or sequence queries, we set up The Insect Protein Structure (TIPS) database (http://tips.shenxlab.com/).

Fig. 3 Insect functional genomics. We performed structural similarity searches against structural databases, including full-length structures with well-characterized functions from the AFDB Swiss-Prot[25] and PDB[26] databases and curated domain structures from the CATH SSG5 database.[27,28] **a** Distribution of cluster functional-annotation consistencies. Consistency within each cluster was calculated as the fraction of functional annotations from the cluster representative that were also present in the list of functional annotations of all cluster members. **b** The pie chart shows the composition of annotations for the 87,461 large clusters. Note that if the full-length structure-based annotation method was not applicable, we then used the domain structure-based annotation method. This prioritization was made because the full-length structure-based annotation method incorporates the entire protein structure. The stacked bars show the relationship between cluster annotations and their taxonomic levels. All 7.61 million proteins from the 68,364 clusters annotated by either the full-length structure-based method or the domain structure-based method are accessible via structures and sequences through The Insect Protein Structure (TIPS) database (http://tips.shenxlab.com/). **c** Sankey visualization of organisms contributing to the functional annotations of structural clusters. The numbers in parentheses represent the numbers of annotated clusters. **d** Three examples of cluster representatives (in yellow) functionally annotated by the structure-based method, together with their best hits (in blue). These representatives were selected on the basis of their various timescales and functional significances.

Massive remote homologs between insect proteins

Given that protein structures are generally conserved over longer evolutionary timescales than are protein sequences,[17–19] we identified remote (or distantly related) homologs between insect proteins within each cluster (Supplementary information, Fig. S14). We defined a pair of insect proteins as remote homologs if their protein sequence identity was < 0.25 and their structural similarity was > 0.5 (Fig. 4a). These thresholds were chosen because sequence-based homology detection typically fails below 0.25 amino acid identity,[11–13] and a structural similarity above 0.5 indicates the presence of shared structural folds.[56,57] We identified 750 million remote homologs with highly similar structures but markedly divergent sequences from 12,308 distinct clusters. To assess their prevalence, we calculated the percentage of unique cluster-member pairs that exhibited remote homology, excluding self-comparisons. The prevalence ranged from 10% to 79%, with a median of 18% (Fig. 4b).

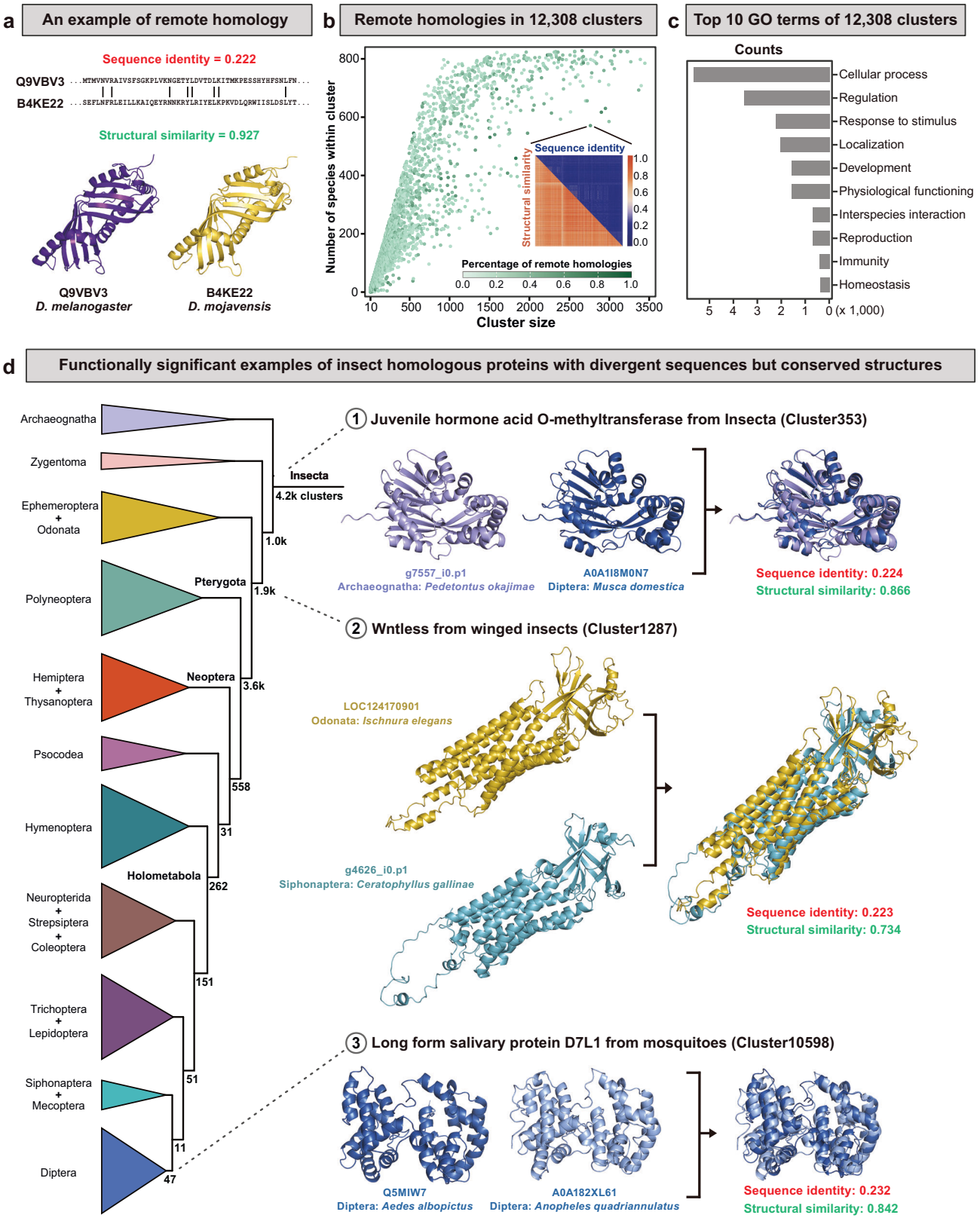
Examination of the functional annotations for these 12,308 cluster representatives showed that most were associated with cellular-, regulation-, response to stimulus-, localization-, and development-related biological processes (Fig. 4c). Evolutionary analysis revealed that among these 12,308 clusters containing remote homologies (left panel in Fig. 4d; Supplementary information, Fig. S15), 4191 (34.1%) occurred at the branch leading to the most recent common ancestor of 824 insects (Insecta), 7709 (62.6%) at the inter-ordinal branches (between orders), and 408 (3.3%) at internal branches within specific orders (order-specific). These results suggest that the majority of clusters that contain remote homologies arose during the early diversification of insects and were established deep within the insect tree of life.

Finally, we present three functionally significant examples of structural alignments that reveal connections among remote homologous proteins across different evolutionary timescales (right panel in Fig. 4d). In the first example, for the class Insecta, a jumping bristletail protein (Archaeognatha: *Pedetontus okajimae*, protein ID: g7557_i0.p1) and a housefly protein (Diptera: *Musca domestica*, UniProt: A0A1I8M0N7) cluster with high structural similarity (0.866) despite low sequence identity (0.224) in Cluster 353. Both exhibit structural homology to juvenile hormone (JH) acid *O*-methyltransferase (UniProt: Q9VJK8) from the dipteran *D. melanogaster*, which catalyzes the final step of JH biosynthesis and is essential for regulating insect growth and metamorphosis.[58,59] The second example involves the subclass Pterygota (winged insects), in which a protein from an odonate (Odonata: *Ischnura elegans*, NCBI: LOC124170901) clusters with one from a hen flea (Siphonaptera: *Ceratophyllus gallinae*, protein ID: g4626_10_p1) in Cluster 1287, displaying high structural similarity (0.734) but low sequence identity (0.223). These proteins are structural homologs of Wntless (UniProt: B4J2W3) from the dipteran *Drosophila grimshawi*, which is a cargo for transporting Wnt signaling molecules and is critical for embryonic development.[60,61] The third example involves salivary proteins in the order Diptera, in which a protein from the mosquito *Aedes albopictus* (UniProt: Q5MIW7) and another from the mosquito *Anopheles quadrimaculatus* (UniProt: A0A182XL61) have a structural similarity of 0.842 and a sequence identity of 0.232 in Cluster 10598. Both are structural homologs of the long-form salivary protein D7L1 (PDB: 6v4c_A) from the mosquito *Culex quinquefasciatus*, which plays a key role in blood feeding by inhibiting host hemostasis.[62,63] Together, these three cases exemplify how structural homology can reveal functional relationships that remain hidden at the sequence level, underscoring their importance for understanding protein function and evolution.

Structurally conserved but sequence-divergent cGLRs pervade the insect tree of life

In the set of 12,308 structural clusters that contained remote homologies (Fig. 4), Cluster 142 was the most prevalent, accounting for 41% of remote homologies and including all 824 insect species (Supplementary information, Fig. S16). This cluster contained 3056 proteins, with 1–18 copies per species across the 824 insects (Fig. 5a; Supplementary information, Table S4), including two functionally well-characterized cGLRs from *D. melanogaster* that serve as sensors in antiviral innate immunity.[64,65] Compared with fruit fly cGLRs, those in other insects exhibited remarkable sequence divergence, with an average identity of 0.24, but maintained a high structural similarity of 0.73 (Fig. 5a). A recent study by Li et al.[66] used an amino acid position-specific approach to identify over 3000 putative cGLRs across 583 animal species, including 413 putative cGLRs from 193 of the insects we examined. Our analysis identified 944 putative cGLRs in the same 193 insects, including 404 (98%) from Li et al. and 540 unique to our study (Supplementary information, Fig. S17). Structural alignment revealed that the 9 cGLRs exclusive to Li et al. showed lower structural similarity (average 0.56) than the 540 unique cGLRs from our study (0.73) and the 404 overlapping cGLRs (0.77).

We characterized the molecular functions of two putative cGLRs from the yellow fever mosquito (*Aedes aegypti*) that had not been reported previously, including in Li et al.'s study. Knockdown (70%–80% efficiency) of *Aa-cGLR1* or *Aa-cGLR2* significantly increased the prevalence (Fig. 5b) and intensity (Supplementary information, Fig. S18 and Table S5) of dengue and Zika virus infections. Overexpression of these cGLRs in C6/36 cells significantly reduced viral infections (Fig. 5c). Transcriptome analysis revealed that 238 genes downregulated in *Aa-cGLR1* knockdowns and 247 in *Aa-cGLR2* knockdowns were mainly enriched in pathways associated with the Toll and Imd signaling pathways, as well as in metabolic pathways (Supplementary information, Fig. S19 and Table S6). The enrichment of metabolic pathways is consistent with recent reports on the non-canonical functions of cGAS in metabolic regulation.[67–69] Notably, knockdown resulted in a 2.5- to 5.0-fold decrease in expression of the immune deficiency gene *imd* and the NF-κB transcription factor gene *Relish/REL2*, both of which are crucial for antiviral immunity.[70–72] In addition,



both *Aa-cGLR1* and *Aa-cGLR2* were upregulated following infection with either DENV2 or ZIKV compared with uninfected controls (Supplementary information, Fig. S20).

We next individually expressed *Aa*-cGLR1 and *Aa*-cGLR2, as well as three positive controls (human cGAS, fruit fly cGLR1, and cGLR2) in HEK293T cells, following the method described in Holleufer et al.[64] Co-transfection with STING showed that human cGAS and fruit fly cGLR2 increased IFNB1 activity, which is essential for the innate immune response, whereas fruit fly cGLR1, *Aa*-cGLR1, and *Aa*-cGLR2 did not (Fig. 5d). When transfected with STING and

Fig. 4 Structural alignments reveal massive numbers of remote homologous proteins across the insect tree of life. a An example of a pair of remote homologous proteins (or distantly related proteins) characterized by a sequence identity of < 0.25 and a structural similarity > 0.5. These thresholds were chosen because sequence homology detection typically fails below 0.25 identity,[11–13] whereas a structural similarity above 0.5 suggests the presence of shared folds.[56,57] **b** Dot plot of 12,308 distinct clusters containing remote homologous proteins. A total of 750 million instances of remote homologies from 12,308 clusters were identified. The *x*-axis depicts the cluster size, and the *y*-axis indicates the number of insect species in each cluster. Each dot corresponds to a cluster, with its color indicating the percentage of remote homologies in the cluster. **c** Top ten Gene Ontology (GO) terms of 12,308 clusters containing remote homologous proteins, focusing on high-level categories of biological processes. Note that these GO terms are categorized at the same level, without overlapping. **d** Schematic representation of our phylogenetic tree, showing major clades and internal nodes, together with the numbers of clusters containing remote homologous proteins. The right panel shows three functionally significant examples of insect remote homologous proteins with divergent sequences but conserved structures across different evolutionary timescales.

poly(I:C), a dsRNA analog, all five candidates upregulated IFNB1 (Fig. 5d; Supplementary information, Table S7). However, comparison of IFNB1 activity with and without poly(I:C) showed a significant increase for fruit fly cGLR1, *Aa*-cGLR1, and *Aa*-cGLR2 but not for human cGAS or fruit fly cGLR2 (Fig. 5d). These results indicate that mosquito cGLRs, like fruit fly cGLR1, can sense dsRNA. Mutagenesis analysis identified four conserved residues (F80, E81, Q175, and R253) in the *Aa*-cGLRs whose mutations significantly reduced IFNB1 activity (Fig. 5e; Supplementary information, Fig. S21a). Surface electrostatic modeling showed that the structures of *Aa*-cGLR1 and *Aa*-cGLR2 might feature a putative positively charged ligand-binding surface (Supplementary information, Fig. S21b). Liquid chromatography mass spectrometry (LC-MS) analysis revealed that *Aa*-cGLR1 produced the cyclic dinucleotide (CDN) 2′3′-cGAMP, but the product of *Aa*-cGLR2 remains unidentified, despite screening for six typical CDNs (Supplementary information, Fig. S21c). Furthermore, injection of 2′3′-cGAMP into yellow fever mosquitoes enhanced their antiviral defense against dengue and Zika infections (Supplementary information, Fig. S22). Taken together, these results demonstrate that mosquito cGLRs play a crucial role in antiviral defense.

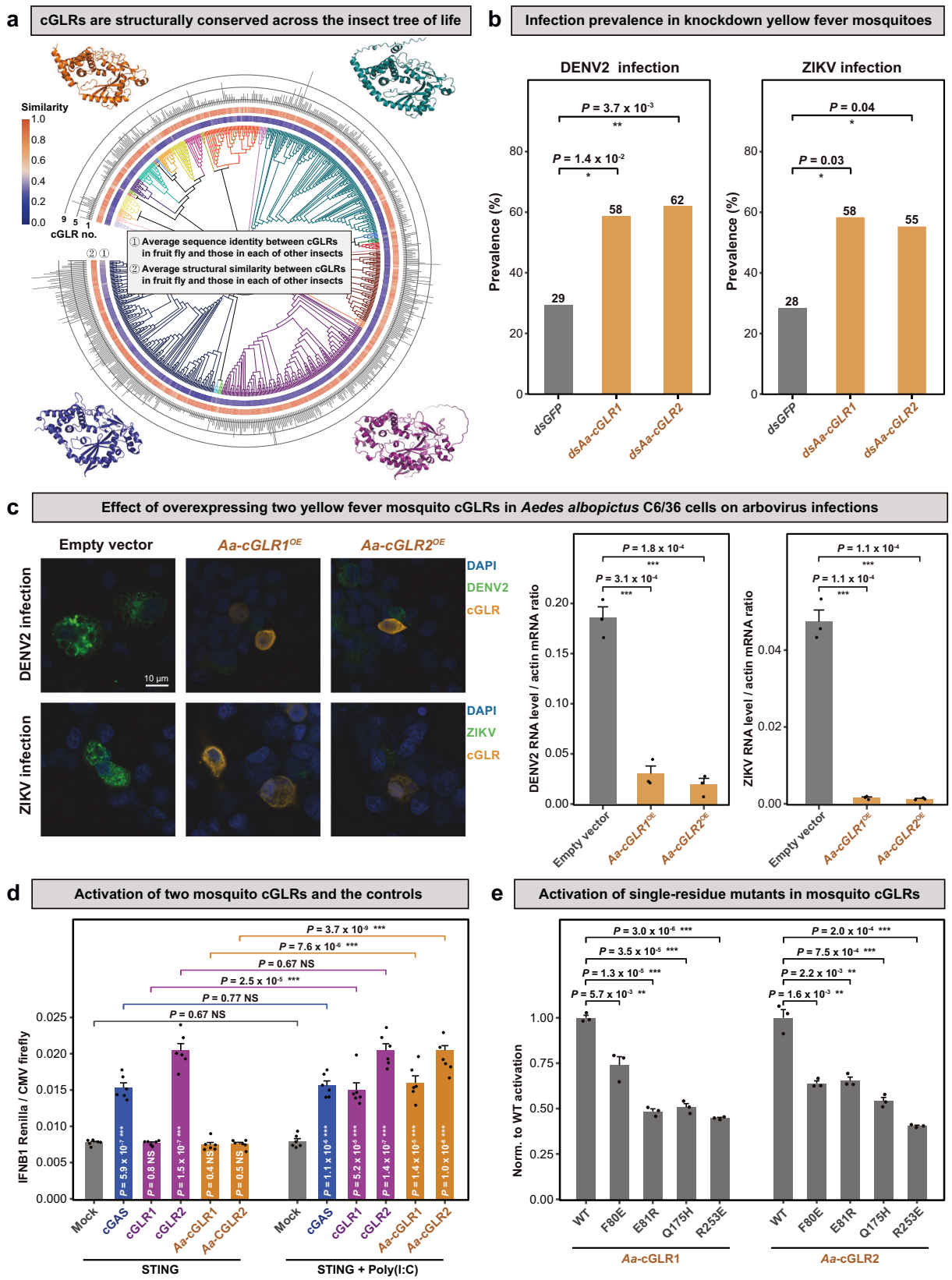
DISCUSSION

Advances in genomics and accurate structure predictions have ushered in a new era, offering unprecedented opportunities to deepen our understanding of protein function and evolution. In this study, we harnessed publicly available genome and transcriptome data to reconstruct a comprehensive and well-resolved phylogeny of 4854 insects representing all 28 insect orders (Fig. 1).[31,73] From this comprehensive framework, we selected 824 representatives and created an atlas of predicted structures for 13.29 million proteins (Fig. 2), including 11.63 million new proteins, to explore the relationships among sequence, structure, and function in the insect tree of life.

A robust phylogenomics framework is essential for elucidating the tempo and mode of insect evolution. Numerous phylogenomics studies have provided valuable insights into the evolutionary relationships among insects,[2,74–81] but most of their taxon sampling strategies focused on either all orders with a moderate number of insects or on insects from specific orders.[2,36,82–90] In their landmark 2014 study, Misof et al.[31] sampled all 28 insect orders using 126 transcriptomes and genomes, providing a holistic view of insect evolution. However, the past decade has seen significant advances in high-throughput sequencing technologies, greatly increasing the availability of genome and transcriptome data. Using this wealth of data, we reconstructed a comprehensive phylogeny of 4854 insects spanning all 28 orders, providing a robust evolutionary framework. Our phylogeny not only significantly expands species sampling but also improves resolution relative to the landmark phylogeny of Misof et al.[31] Recent phylogenomics studies have resolved many branches in the tree of life, but the resolution of problematic internodes — often influenced by phylogenetic methods, gene selection, and taxon choices, remains challenging, especially for contentious branches.[91,92] We observed that 23% of internodes exhibited topological conflicts between concatenation- and coalescent-based phylogenies. This may be attributed to gene-tree variation due to various biological processes across loci (e.g., deep coalescence, gene duplication and loss, and ILS) and gene-tree estimation error due to inadequacy of the multispecies coalescent model (e.g., recombination within a locus).[93–95] Through large-scale taxon sampling and advances in genome sequencing, comprehensive phylogenomic sampling shows great potential for reducing the number of contentious branches in the tree of life. Although the use of protein structures rather than protein sequences to detect homologs represents a significant advance, the use of protein structures for reconstruction of phylogenetic trees is still in its early stages.[96] Nevertheless, future advances in evolutionary models of protein structures and structural ortholog inference may enable structure-based methods to surpass sequence-based methods in phylogenomics.

The wealth of insect genome and transcriptome data not only aids in reconstructing the insect tree of life but also serves as a valuable resource for predicting protein structures. Recent advances in accurate structure prediction[21–24] have enabled diverse applications. These include establishment of structural databases such as the AlphaFold Protein Structure Database (AFDB),[25] the ESM Metagenomic Atlas (ESMatlas),[23] AFESM (AFDB + ESMatlas),[38] and the Encyclopedia of Domains (TED),[28] identification of novel protein families,[17,19,20,97,98] exploration of new genome editing tools,[99] and discovery of fungal effectors.[16] In particular, AlphaFold2 has substantially expanded the protein structural space, contributing over 200 million predicted structures to AFDB, including more than 0.5 million structures for expert-reviewed proteins in Swiss-Prot.[25] AFDB Swiss-Prot, together with PDB[26] and CATH,[27,28] provides a highly precise resource for protein functional annotations. In this study, we created an atlas of over 13 million predicted protein structures from 824 representative insects spanning all 28 orders, then clustered this structural universe. By structurally aligning our predicted structures with those of proteins whose functions are well characterized, we generated functional annotations for 7.61 million proteins (Fig. 3), 14% of which were not annotated by similar sequence-based approaches. These annotations, mainly derived from well-characterized non-insect proteins, reveal diverse functions, including many not previously reported in insects, such as the pore-forming toxin protein PirB[47–49] and the eukaryotic genome-editing endonuclease Fanzor.[50–54] Therefore, we propose that structure-based approaches may be considered a crucial strategy for functional protein annotations in the future.

In addition to their utility in functional annotation, analyses of protein structure are also effective for detecting remote homologous proteins with dissimilar sequences but similar functions.[17–19,100–102] Using a sequence identity of < 0.25[11–13] and a structural similarity of > 0.5,[56,57] we identified 750 million remote homologous proteins from 12,308 distinct clusters and found that many clusters could be traced back to ancient origins in the insect tree of life (Fig. 4). These families of remote homologous proteins, whose members share conserved structures and have similar



functional annotations, are involved in a variety of biological processes, such as immune responses, cell differentiation, and circadian rhythms. Notably, all 824 insects examined encoded putative cGLRs, critical components of antiviral immunity.[66,103–106]

Despite substantial sequence divergence, these putative cGLRs formed a single structural cluster that included two functionally characterized fruit fly cGLRs.[64,65] Experimental analysis revealed that cGLRs play a crucial role in the antiviral defense of *Aedes*

Fig. 5 A notable case of remote homologs: cGLRs are structurally conserved but show marked sequence divergence across the insect tree of life, with functional characterization in the yellow fever mosquito. Our study identified 12,308 structural clusters of remote homologs with highly similar structures but markedly divergent sequences. Among these, Cluster 142 emerged as the most prevalent, containing sequences from all 824 insect species examined. This cluster comprises 3056 proteins (1–18 per species), including two well-characterized cGLRs (cGLR1 and cGLR2) from *D. melanogaster*.[64,65] **a** Distribution of 3056 putative cGLRs across the tree of life for 824 insects. Branches are colored according to their order names as depicted in Fig. 1. The inner circle shows the average sequence identity between fruit fly cGLRs and those in each of the remaining 823 insects. The outer circle shows the average structural similarity between fruit fly cGLRs and those in each of the remaining 823 insects. Gray bars indicate the number of putative cGLRs identified for each species. Representative structures are displayed outside the circles. **b** Effect of knocking down two cGLRs (*Aa*-cGLR1 and *Aa*-cGLR2) on the prevalence of dengue and Zika viral infection (%) in the yellow fever mosquito. These two mosquito cGLRs have not been reported previously. *P* values were calculated using a two-tailed Fisher's exact test. The viral infection intensity in knockdown yellow fever mosquitoes is shown in Supplementary information, Fig. S18. **c** Effect of overexpressing *Aedes aegypti* cGLRs in *Aedes albopictus* C6/36 cells on dengue and Zika virus infections. The left panels display confocal images of immunostained cells. The nuclei were stained with DAPI (blue). Viruses were stained with dengue virus antibody (D1-11) or Zika virus envelope protein antibody (green). Yellow fever mosquito cGLRs were stained with anti-V5-tag antibody (yellow). The right panels present bar graphs quantifying virus infection intensity in the overexpressing cells. Data are presented as mean ± SD. *P* values were calculated using a two-tailed *t*-test. **d** Two yellow fever mosquito cGLRs sense poly(I:C), a dsRNA analog. *IFNB1* (a critical component of the innate immune response to infection) reporter activity in HEK293T cells transfected with each of three positive controls (human cGAS and fruit fly cGLR1 and cGLR2) and two mosquito cGLRs, with STING (left panel), or together with poly(I:C) (a dsRNA analog) (right panel). Data are presented as mean ± SD. *P* values were calculated using a two-tailed *t*-test. **e** Analysis of the activation of mosquito single-residue cGLR mutants in HEK293T cells transfected with human STING and poly(I:C). The activation of each single-residue cGLR mutant was normalized to the mean value of the wild-type (WT) activation. Data are presented as mean ± SD. *P* values were calculated using a two-tailed *t*-test.

aegypti, a vector of human arboviruses (Fig. 5). Comparisons of mosquito cGLRs and vertebrate cGASs revealed that mosquito cGLR structures are markedly similar to vertebrate cGAS structures, including experimentally determined structures of human, mouse, and pig cGASs, despite sharing very low sequence similarity (Supplementary information, Fig. S23). This deep structural conservation implies that these proteins may have an evolutionarily ancient function. Lastly, although structural and sequence analyses revealed that mosquitoes lack a canonical STING, a recent study has shown that cGLRs can exist without STING in some metazoans.[107] We propose that CDNs produced by *Aedes* mosquito cGLR may signal through alternative proteins. For instance, in bacterial CBASS systems,[108] which share evolutionary ancestry with the cGAS-STING pathway, CD-NTase (cGAS/DncV-like nucleotidyltransferase) synthesizes CDNs that bind and activate the Cap effector (CD-NTase-associated protein), triggering an anti-phage response without STING involvement. Similarly, mammalian proteins such as RECON[109] and ERAdP[110] can sense bacterial CDNs and initiate anti-bacterial immunity independently of STING. Therefore, identifying the protein(s) that interact with cGLR-generated CDNs in mosquitoes would be an intriguing topic for future research.

Although cGLRs and Mab21/Mab21-like proteins share some degree of sequence homology,[66,107] they differ in functional characteristics. Early studies reported that Mab21 protein in *Caenorhabditis elegans* and Mab21-like proteins in mice are involved in developmental regulation.[111,112] However, the functional roles of Mab21/Mab21-like proteins in insects remain largely uncharacterized. By contrast, as core components of the innate immune pathway, cGLRs possess the ability to mediate antiviral responses.[64,66] In this study, we identified two *Aedes aegypti* Mab21-like proteins, and multiple lines of evidence supported their classification as cGLRs (*Aa*-cGLRs). These proteins not only showed high structural similarity to two well-characterized *Drosophila* cGLRs but also were functionally validated through several findings: (1) they exhibited antiviral activity; (2) upon stimulation with the dsRNA analog poly(I:C), they activated expression of IFN-β1 in a STING-dependent manner; (3) LC-MS analysis revealed that mosquito cGLR1 was capable of producing 2′3′-cGAMP; (4) injection of chemically synthesized 2′3′-cGAMP into mosquitoes enhanced their resistance to viral infection. These functional characteristics align with the criteria used to identify *Drosophila* cGLRs in a previous study,[64] further supporting the immune functions of mosquito Mab21-like proteins. This discovery not only broadens our understanding of cGLRs but also provides valuable insight into antiviral innate immunity in mosquitoes. However, the innate immune functions of Mab21/Mab21-like proteins in other insects remain to be investigated and confirmed experimentally. In addition, no detectable CDN signals were observed when cGLR2 was heterologously transfected into HEK293T cells. This result may stem from a combination of factors, including insufficient sensitivity of CDN detection, low protein expression efficiency in the heterologous expression system, presence of endogenous 2′3′-cGAMP-degrading enzymes,[113,114] potential export of 2′3′-cGAMP,[115] or generation of uncharacterized CDN signaling molecules.

Despite its merits, our insect structural genomics study encountered challenges with proteins that yielded very low-confidence structural predictions. This issue is not unique to our insect structural data; existing large-scale structure databases such as AFDB, ESM Atlas, and BFVD also contain some proportion of very low-quality structures. For example, 27% of the predicted structures in the human proteome have very low confidence.[116] There are multiple underlying causes, potentially including uneven representation in training data for state-of-the-art prediction tools,[117] limited information within short sequences,[40] and/or the presence of intrinsically disordered regions in proteins.[118]

MATERIALS AND METHODS

Taxon sampling

Genome collection. To compile a dataset with extensive taxonomic sampling as of October 11, 2023, we initially gathered publicly available insect genome information, including species names, assembly accession numbers, assembly release dates, and assembly levels, from 17 public repositories (Supplementary information, Table S1), including FlyBase,[119] NCBI,[120] BIPAA, InsectBase2,[121] i5k workspace,[122] Ensembl,[123] UCSC Genome Browser,[124] and NGDC.[125] Next, for species with multiple sequenced genomes, we retrieved only the genome that had the highest assembly level and most recent release date. After filtering out assemblies with completeness < 30%, we retained 1724 genomes (Supplementary information, Table S2).

Transcriptome collection. To obtain transcriptome data, we used "Insecta" as the search term in NCBI's Sequence Read Archive (SRA) Browser (https://www.ncbi.nlm.nih.gov/sra) to obtain basic information on species names, SRA accession numbers, SRA sizes, sequencing strategies, and release dates. For species without publicly available genomes, we selected the dataset with the largest SRA size and the most recent release date. Paired-end sequences were chosen over single-end reads when available. For each species, we retrieved and decompressed the raw reads using the prefetch and fastq-dump programs of the SRA Toolkit v2.10.7 (https://github.com/ncbi/sra-tools). We then processed the raw reads by trimming adapters with TrimGalore v0.6.10 (https://github.com/FelixKrueger/

TrimGalore). Clean reads were used for de novo transcriptome assembly with default parameters of Trinity v2.11.1.[126] Putative proteins were identified using TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) with a minimum open reading frame of 15 amino acids, and the longest putative protein per gene was retained. After filtering out assemblies with completeness < 30%, we retained 3130 transcriptomes (Supplementary information, Table S2).

Assembly quality assessment. To evaluate the quality of publicly available genomes and our newly assembled transcriptomes, we used BUSCO v5.2.2.[33] For genome assemblies, we used the genome mode option "-m genome", and for assembled transcriptomes, we used the transcriptome mode option "-m transcriptome". Each assembly's completeness was assessed on the basis of the presence or absence of a set of 1367 conserved, full-length BUSCO nuclear genes from 75 insect genomes in the OrthoDB v10 database.[127] We considered both single-copy full-length genes and duplicated full-length genes as complete genes for the assembly assessment.

Phylogenetic analyses

Given that BUSCO assignments do not depend on genome annotations and have been widely used in studies involving insects,[128] plants,[129] and fungi,[130] we began construction of the phylogenomic data matrix with a set of 1367 single-copy full-length BUSCO genes from 4854 insects and 10 Entognatha outgroups based on the BUSCO output folders "single_copy_busco_sequences". We aligned the amino acid sequences for each BUSCO gene using MAFFT v7.505[131] with the options "--thread 8 --auto" and trimmed the amino acid alignments using trimAl v1.4.rev15[132] with the options "-automated1 -colnumbering". We excluded 543 BUSCO gene alignments with a taxon occupancy (i.e., the percentage of taxa whose sequences were present in the trimmed amino acid alignment) < 50% and a trimmed alignment length < 150 amino acids. These filters resulted in a data matrix that contained 4864 taxa, 824 genes, and 276,683 amino acid sites.

We inferred the concatenation-based ML tree using IQ-TREE multicore v2.0.7[133,134] on a single computing node with 256 CPU cores and 2 TB RAM under a single "LG + G4" model with the options "--runs 1 -T 240 -m LG + G4 --ufboot 1000", as 438 out of 824 genes favored "LG + G4" as the best-fitting model. We ran five independent tree searches using five different seeds to obtain the best-scoring concatenation-based ML tree. We inferred the coalescent-based species phylogeny with ASTRAL-III v4.10.2[135,136] using the set of 824 individual ML single-gene trees. Finally, we visualized the phylogenetic trees using iTOL v5.[137]

Protein structure prediction and structural clustering

Owing to the substantial computational burden of predicting structures for all 4854 insect species, we selected a subset of insects for structure predictions on the basis of our concatenated phylogenetic tree (Fig. 1). Our selection process involved three criteria: First, we ensured taxonomic diversity by representing a wide range of clades and avoiding bias toward specific groups. Second, we excluded species with long branch lengths. Third, we selected species with high-quality data, as assessed by BUSCO completeness. Consequently, we retained 824 representative insects that were broadly distributed across the insect tree of life (Supplementary information, Table S3). Initially, we retrieved 1.66 million predicted protein structures for 114 insects from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/),[21,25] with each species having > 10,000 structures. For the 11.78 million protein sequences of the remaining 710 insects, we predicted their structures using a heterogeneous GPU cluster at the Center for Engineering and Scientific Computation (CESC) at Zhejiang University. This cluster comprised an NVIDIA Tesla H800 (80 GB RAM), H100 (80 GB RAM), and V100 (32 GB RAM). To ensure computational tractability, we restricted our analysis to 11.63 million protein sequences (~98.7% of the 11.78 million proteins) with lengths ranging from 16 to 2000 amino acids.

The structure of each protein sequence was predicted using ESMFold,[23] a fast and comparably accurate structural prediction method. Specifically, we used Hugging Face transformers v4.35.2 to run the esm.pretrained.esmfold_v1 model for prediction, with the remaining parameters left at their default values. As a result, we obtained 13.29 million predicted protein structures, including 11.63 million newly generated in this study and 1.66 million publicly available from the AlphaFold2 database, from 824 insects across the insect tree of life.

To cluster this structural universe, we used Foldseek *cluster* v9.427df8a[17,37] with the following thresholds: minimum TM-score of 0.4, minimum query and target coverage of 70%, *E*-value = 0.001, and sensitivity of 7.5. The parameters used were --threads 250 -s 7.5 --cluster-mode 0 -c 0.7 --tmscore-threshold 0.4 -e 0.001 --cluster-reassign 1. In addition, we clustered 527,789 structures with well-characterized functions from the Swiss-Prot database using the same methodology. This analysis revealed high functional consistency within the structural clusters, with a median value of 91%. These results indicate that structurally similar proteins are very likely to exhibit functional similarity.

Examination of three protein properties

The prediction confidence value for a given predicted protein structure was calculated as the average of the pLDDT scores for all residues, and its length was determined by the total number of residues. To assess the level of disorder in a protein, we first used flDPnn to predict disordered residues, identifying those with a predicted disorder score > 0.3.[138] We then classified the proteins into three levels based on the method described by Deiana et al.[139]:

(i). **Intrinsically disordered protein**: a protein in which more than 30% of residues are predicted to be disordered.
(ii). **Protein with intrinsically disordered regions**: a protein with fewer than 30% disordered residues overall but which contains at least one segment of more than 30 consecutive disordered residues.
(iii). **Ordered protein**: a protein with fewer than 30% disordered residues and no segments of more than 30 consecutive disordered residues.

Analysis of structural similarity within a cluster

We used two structural similarity metrics — LDDT and TM-score — for each structural cluster, following previous studies.[17,38] For a given cluster, we first performed pairwise alignments between cluster members using the Foldseek *structurealign* module with the options "-a -e INF --threads 120". We then used the Foldseek *convertalis* module to customize the output format with the parameters "--format-output query,target,evalue,lddt,alntmscore". Lastly, we calculated the average LDDT and TM-score of all member-to-representative alignments per cluster.

Functional annotations

To obtain functional annotations for each structural cluster, we performed structural similarity searches against structural databases with highly precise functions: AFDB Swiss-Prot[25] and PDB[26] for full-length structure-based annotation, and CATH[27,28] for domain structure-based annotation. If the full-length structure-based annotation method was not applicable, we then used the domain structure-based annotation method. This is because the former takes into account the entire protein structure.

(i). Full-length structure-based annotation: we searched each cluster member against full-length structures in AFDB Swiss-Prot and PDB using the Foldseek *easy-search* module with the parameters "--max-seqs 10000 -s 9.5 -e 0.001 -c 0.4 --alignment-type 2 --cov-mode 0".
(ii). Domain structure-based annotation: using the TED pipeline as described by Lau et al.[28] we identified protein domain boundaries using Chainsaw,[140] Merizo,[141] and UniDoc,[142] retaining consensus domains (≥ 2 predictor agreement). Consensus domains were assigned to 31,574 structures in existing CATH Structural Similarity Groups at 5 Å (SSG5), where each SSG5 is a cluster representative for CATH domains that superpose within 5 Å, using the Foldseek *easy-search* module with the parameters "-s 10 --cov-mode 5 --alignment-type 2 -e 0.108662 -c 0.366757 -a". Assignments to superfamily (H-level) or fold (T-level) were determined using cutoffs from Lau et al. (*E*-values 0.019000 and 0.108662, coverages 0.366757 and 0.786333, and TM-scores 0.560000 and 0.416331, respectively). We found that 59.6% of the representatives from all large clusters with at least 10 members exhibited a single structural domain, whereas the remaining 40.4% consisted of two or more structural domains.

We investigated whether sequence-based methods could capture our proteins annotated by structure-based methods. For proteins annotated by the full-length structure-based approach, we performed sequence-based

searches against full-length protein sequences from the AFDB Swiss-Prot and PDB databases using the MMseqs2 v17.b804f *easy-search* module.[41] This was performed in the same manner as the full-length structure-based annotation. For proteins annotated by the domain structure-based approach, we performed sequence-based searches against sequence domains in Gene3D assigned through structures,[143] using InterProScan v5.66-98.0[144] with the parameters "-dp -appl Gene3D -f TSV -T tmp".

Assessment of cluster functional-annotation consistency

For structural clusters annotated by the full-length structure-based annotation method, we calculated the fraction of functional annotations from the cluster representative that were also present in the functional annotation lists of all cluster members. For structural clusters annotated by the domain structure-based annotation method, following the strategy outlined in a previous study,[17] we determined the fraction of correctly matched CATH domains for all member-to-representative pairs. True positives were defined as pairs of CATH domains in the same clan. Functional-annotation consistency was then calculated as the proportion of true positives within the cluster.

Identification of remote homologous proteins

For each structural cluster, we performed an all-vs-all sequence alignment of all its members using MMseqs *align* with the options "--threads 10 --alignment-mode 3 -e inf --comp-bias-corr 0". In addition, we performed an all-vs-all structural alignment of its members using Foldseek *structure-align* with the same options. The *Convertalis* module in MMseqs and Foldseek was used to customize the output format. To identify remote homologous protein pairs, we used a sequence identity threshold of < 0.25 and a structural similarity threshold of > 0.5. These thresholds were chosen because sequence-based homology detection typically fails below an amino acid identity of 0.25,[11–13] and a structural similarity score > 0.5 is indicative of shared structural folds.[56,57]

Functional validation of two yellow fever mosquito cGLRs

Gene silencing in mosquitoes. DNA templates were PCR amplified with T7 promoter-flanked primers for mosquito *LOC5570128* (*Aa-cGLR1*) and *LOC5570126* (*Aa-cGLR2*) (Supplementary information, Table S8) and used to synthesize dsRNA in vitro with the MEGAscript RNAi kit (Thermo Fisher Scientific). For mosquito microinjection, 138 nL of dsRNA solution (3 μg/μL) was injected into the hemocoel of 3-day-old female mosquitoes using a Nanoject III microinjector (Drummond). The injected mosquitoes were allowed to recover for 2–3 d before performing a blood meal. The gene-silencing efficiency was assessed by qPCR. Knockdown efficiency was 79% for *Aa-cGLR1* and 70% for *Aa-cGLR2*.

Virus passage and infection assay in mosquitoes. DENV2 (New Guinea C strain, AF038403.1) and ZIKV (ZJ03 strain) were passaged in C6/36 cells. The supernatant was harvested, filtered through a 0.22-μm filter, separated into 0.5-mL aliquots, and frozen at −80 °C. For virus infection assays in mosquitoes, female *Ae. aegypti* mosquitoes cultured in paper cups covered with mesh were starved for 12–24 h before blood feeding to ensure engorgement. An infectious blood meal was prepared by mixing heat-inactivated defibrinated sheep blood (Yuanye Biotech) with a virus suspension at a ratio of 1:1. Mosquitoes were fed the blood meal using Parafilm membrane–covered glass feeders that were warmed by 37 °C circulating water. Fully engorged female mosquitoes were transferred to new paper cups and maintained under standard conditions for further investigation. The viral load in the whole body was determined at 10 d post infection by qPCR.

Transfection of C6/36 cells. To overexpress *Ae. aegypti* cGLRs in C6/36 cells, 24-well tissue culture plates were seeded with 1×10^5 C6/36 cells per well. After 24 h, each well was transfected with 400 ng of plasmids expressing cGLR under the control of the poly-ubiquitin promoter or an empty vector. The DNA and 1 μL Attractene Transfection Reagent (Qiagen) were dissolved and mixed in 60 μL of RPMI-1640 medium. The mixture of DNA and Attractene was incubated for 15 min, then added dropwise to cells. At 24 h post transfection, DENV2 or ZIKV was inoculated into treated cells at MOI = 1.

Quantification and immunostaining of virus in C6/36 cells. At 48 h post virus infection, C6/36 cells were collected, and virus infection loads were quantified by qPCR. To detect virus infection by immunostaining, the virus-infected cells were fixed with 4% (m/vol) paraformaldehyde (Sigma) for 10 min at room temperature, then blocked in Immunol Staining Blocking Buffer (Beyotime) for 60 min. The cells were then incubated overnight with Dengue virus antibody (Santa Cruz Biotechnology, 500-fold dilution), Zika virus envelope protein antibody (GeneTex, 400-fold dilution), or anti-V5 tag antibody (Abcam, 500-fold dilution) in blocking buffer. The cells were washed with 1× PBST three times, incubated with Alexa Fluor 488 or 555 anti-mouse or anti-rabbit IgG (Beyotime, 200-fold dilution) for 1 h at room temperature, and washed with 1× PBST three more times. For confocal observation, cells were mounted with VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories) for 5 min, and fluorescence signals were visualized with a Nikon AXR confocal microscope system.

RNA sequencing. Total RNA was isolated from *Aa-cGLR1*- and *Aa-cGLR2*-knockdown and control female *Ae. aegypti* adults using RNAiso Plus (TaKaRa) and treated with DNase I (TaKaRa). Each treatment had three replicates. RNA-seq libraries were constructed and sequenced on the Illumina HiSeq 2000 platform to obtain paired-end reads. Trimmomatic v0.39[145] was used to remove low-quality reads and adapter sequences. Clean reads were mapped to the reference genome using STAR v2.7.10a,[146] and featureCounts v2.0.1[147] was used to count reads per gene. DESeq2 v1.30.1[148] was used to identify differentially expressed genes, defined as those with an adjusted *P*-value ≤ 0.05 and at least a 1.5-fold expression change. KEGG enrichment analysis was performed using ShinyGO 0.80.[149]

Transfection of HEK293T cells. To test the induction of human STING by cGLRs, 24-well tissue culture plates were seeded with 1×10^5 HEK293T cells per well. After 24 h, each well was transfected with 400 ng of dual-luciferase plasmid expressing firefly luciferase under the constitutive CMV promoter and *Renilla* luciferase under the *IFNB1* promoter, 100 ng pcDNA3.1 plasmid expressing human STING, and 300 ng pcDNA3.1 expressing human cGAS, fruit fly cGLR1, fruit fly cGLR2, mosquito *Aa*-cGLR1, mosquito *Aa*-cGLR2, or empty pcDNA3.1 plasmid to reach a total of 800 ng plasmid per well. The DNA was dissolved in 50 μL of DMEM, and 2 μL LipoFiter (Hanbio) was dissolved in another 50 μL of DMEM and incubated for 5 min. The DNA and LipoFiter were then mixed and incubated for another 20 min before dropwise addition to cells. Three hours later, cells were transfected with 300 ng of poly(I:C) (APExBIO) per well using the LipoFiter reagent.

Mutation of single residues in cGLR proteins. To introduce a single F80, E81, Q175, or R253 mutation into *Aa*-cGLR1 and *Aa*-cGLR2, mutagenesis PCR was performed with pcDNA3.1 plasmids expressing *Aa*-cGLR1 and *Aa*-cGLR2 as templates. In brief, PCR primers containing mutant DNA sequences were used to amplify *Aa*-cGLR1 or *Aa*-cGLR2 with Phanta Max Super-Fidelity DNA Polymerase (Vazyme). The PCR products were digested with *Dpn*I (NEB), purified with the Cycle-Pure Kit (Omega), and ligated with the *OK Clon* DNA Ligation Kit II (Accurate Biology). The resulting plasmids were verified by Sanger sequencing and used in the HEK293T transfection test.

Measurement of luciferase activity in transfected cells. At 48 h post transfection, HEK293T cells were lysed in 100 μL of 1× passive lysis buffer (Promega) per well. Firefly and *Renilla* luciferase activity were sequentially measured with 10 μL of lysate using the dual-luciferase reporter assay system (Promega).

Identification of CDNs using LC-MS

CDNs were identified by LC-MS as described previously.[64,150] In brief, for CDN extraction from HEK293T cells ectopically expressing yellow fever mosquito cGLRs in the presence of poly(I:C), the cells were collected in a 1.5-mL Eppendorf tube, to which 1 mL of a precooled extraction reagent (2/2/1 (v/v/v) methanol, acetonitrile, and water mixture) was added. After centrifugation at 20,000× *g* for 15 min at 4 °C, the supernatant was transferred to a new Eppendorf tube for evaporation. The residue was then reconstituted with 1 mL of 20 mM ammonium carbonate and loaded onto P-SAX SPE columns. Eluents were concentrated by evaporation, resuspended in 200 μL of 0.1% formic acid in water, and prepared for LC-MS/MS analysis.

CDN production was analyzed by high-resolution LC-MS using an UltiMate 3000 system (Thermo Fisher Scientific) linked to a Q Exactive HFX Quadrupole-Orbitrap hybrid mass spectrometer (Thermo Fisher Scientific). A 20-μL sample was injected into a Poroshell 120 AQ-C18 column (2.7 μm,

2.1 × 150 mm; Agilent Technologies) maintained at 40 °C. Mobile phase A was 5 mM ammonium carbonate in water, and mobile phase B was acetonitrile with 0.1% formic acid. The HPLC gradient used was as follows: 0%–14% B in 6.0 min, 14%–25% B in 7.4 min, 25%–95% B in 8.0 min, 100% B in 13.0 min, 100%–0% B in 13.1 min, and 0% B in 17.0 min, with a flow rate of 0.300 mL/min. Mass spectra were recorded using positive-ion full-scan mode with m/z from 100 to 1500. Accurate mass measurement was accomplished by the Orbitrap-MS with a mass resolution of 70,000. Source parameters included a capillary temperature of 350 °C, maximum injection time of 100 ms, AGC target of 1E6, and S-lens RF level of 55. Target ions were isolated using high-energy collision dissociation fragmentation, and progeny ions were detected with dd-MS2 mode. The parent ion was isolated with an isolation window of 1 m/z unit and fragmented (resolution = 17,500; NCE = 20, maximum injection time: 50 ms; loop count: 5; topN: 5). CDNs were identified by targeted mass analysis for exact masses and formulae of the targeted CDNs. Xcalibur 4.4 (Thermo Fisher Scientific, CA) software was used for equipment control and data acquisition.

Injection of 2′3′-cGAMP into mosquitoes

To test the effects of 2′3′-cGAMP (Biolog) in mosquitoes, 69 nL of 2′3′-cGAMP solution (1 mg/mL) was injected into the hemocoel of 3-day-old female mosquitoes using a Nanoject III microinjector (Drummond). Mosquitoes injected with 1× phosphate-buffered saline were used as controls. The injected mosquitoes were allowed to recover for 3 d before the DENV or ZIKV infection assay was performed. The viral load in the whole body was determined at 10 d post infection by qPCR.

Statistical analyses

All statistical analyses and plots were performed in R v3.6.3 (R core team 2021).

DATA AVAILABILITY

All gene alignments and gene trees are available on the figshare repository (https://doi.org/10.25452/figshare.plus.25906339). Raw RNA sequencing data has been deposited in GenBank under Bioproject ID: PRJNA1173893. Protein structures are freely available on TIPS database (http://tips.shenxlab.com/). The web server offers options for searching, visualizing, and downloading protein structures, as well as accessing the comprehensive insect tree of life.

REFERENCES

1. Lewin, H. A. et al. The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. USA* **119**, e2115635118 (2022).
2. Thomas, G. W. C. et al. Gene content evolution in the arthropods. *Genome Biol.* **21**, 1–14 (2020).
3. Marks, R. A., Hotaling, S., Frandsen, P. B. & VanBuren, R. Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* **7**, 1571–1578 (2021).
4. Opulente, D. A. et al. Genomic factors shape carbon and nitrogen metabolic niche breadth across Saccharomycotina yeasts. *Science* **384**, eadj4503 (2024).
5. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 1–6 (2016).
6. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
7. Perez-Sepulveda, B. M. et al. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biol.* **22**, 349 (2021).
8. Shen, X.-X. et al. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545.e20 (2018).
9. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17**, 121–127 (1999).
10. Anderson, N. L. & Anderson, N. G. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19**, 1853–1861 (1998).
11. Hamamsy, T. et al. Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.* **42**, 975–985 (2024).
12. Kilinc, M., Jia, K. & Jernigan, R. L. Improved global protein homolog detection with major gains in function identification. *Proc. Natl. Acad. Sci. USA* **120**, e2211823120 (2023).
13. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* **12**, 85–94 (1999).
14. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**, 823–826 (1986).
15. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422**, 216–225 (2003).
16. Seong, K. & Krasileva, K. V. Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. *Nat. Microbiol.* **8**, 174–187 (2023).
17. Barrio-Hernandez, I. et al. Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).
18. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins Struct. Funct. Bioinform.* **77**, 499–508 (2009).
19. Nomburg, J. et al. Birth of protein folds and functions in the virome. *Nature* **633**, 710–717 (2024).
20. Kim, R. S., Levy Karin, E., Mirdita, M., Chikhi, R. & Steinegger, M. BFVD—a large repository of predicted viral protein structures. *Nucleic Acids Res.* **53**, D340–D347 (2025).
21. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
22. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
23. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
24. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
25. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
26. Burley, S. K. et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
27. Sillitoe, I. et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
28. Lau, A. M. et al. Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**, eadq4946 (2024).
29. Stork, N. E. How many species of insects and other terrestrial arthropods are there on earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).
30. May, R. M. How many species are there on earth?. *Science* **241**, 1441–1449 (1988).
31. Misof, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
32. Rainford, J. L., Hofreiter, M., Nicholson, D. B. & Mayhew, P. J. Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. *PLoS One* **9**, e109085 (2014).
33. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
34. Whitfield, J. B. & Kjer, K. M. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu. Rev. Entomol.* **53**, 449–472 (2008).
35. Sharma, P. P. Integrating morphology and phylogenomics supports a terrestrial origin of insect flight. *Proc. Natl. Acad. Sci. USA* **116**, 2796–2798 (2019).
36. Wipfler, B. et al. Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects. *Proc. Natl. Acad. Sci. USA* **116**, 3024–3029 (2019).
37. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
38. Yeo, J. et al. Metagenomic-scale analysis of the predicted protein structure universe. *bioRxiv* https://doi.org/10.1101/2025.04.23.650224 (2025).
39. Akdel, M. et al. A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
40. Monzon, V., Haft, D. H. & Bateman, A. Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinforma. Adv.* **2**, vbab043 (2022).
41. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
42. Zhong, X. et al. Structural mechanisms for regulation of GSDMB pore-forming activity. *Nature* **616**, 598–605 (2023).
43. Johnson, A. G. et al. Structure and assembly of a bacterial gasdermin pore. *Nature* **628**, 657–663 (2024).
44. Johnson, A. G. et al. Bacterial gasdermins reveal an ancient mechanism of cell death. *Science* **375**, 221–225 (2022).
45. Wang, C. et al. Structural basis for GSDMB pore formation and its targeting by IpaH7.8. *Nature* **616**, 590–597 (2023).
46. Devant, P. & Kagan, J. C. Molecular mechanisms of gasdermin D pore-forming activity. *Nat. Immunol.* **24**, 1064–1075 (2023).
47. Prashar, A. et al. Crystal structures of PirA and PirB toxins from Photorhabdus akhurstii subsp. akhurstii K-1. *Insect Biochem. Mol. Biol.* **162**, 104014 (2023).

48. Lee, C.-T. et al. The opportunistic marine pathogen Vibrio parahaemolyticus becomes virulent by acquiring a plasmid that expresses a deadly toxin. *Proc. Natl. Acad. Sci. USA* **112**, 10798–10803 (2015).
49. Wang, H.-C. et al. A bacterial binary toxin system that kills both insects and aquatic crustaceans: Photorhabdus insect-related toxins A and B. *PLoS Pathog.* **19**, e1011330 (2023).
50. Saito, M. et al. Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620**, 660–668 (2023).
51. Jiang, K. et al. Programmable RNA-guided DNA endonucleases are widespread in eukaryotes and their viruses. *Sci. Adv.* **9**, eadk0171 (2023).
52. Altae-Tran, H. et al. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).
53. Bao, W. & Jurka, J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob. DNA* **4**, 12 (2013).
54. Yoon, P. H. et al. Eukaryotic RNA-guided endonucleases evolved from a unique clade of bacterial enzymes. *Nucleic Acids Res.* **51**, 12414–12427 (2023).
55. Gligorijević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
56. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
57. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E. & Skolnick, J. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. USA* **103**, 2605–2610 (2006).
58. Shinoda, T. & Itoyama, K. Juvenile hormone acid methyltransferase: a key regulatory enzyme for insect metamorphosis. *Proc. Natl. Acad. Sci. USA* **100**, 11986–11991 (2003).
59. Jindra, M., Palli, S. R. & Riddiford, L. M. The juvenile hormone signaling pathway in insect development. *Annu. Rev. Entomol.* **58**, 181–204 (2013).
60. Bänziger, C. et al. Wntless, a conserved membrane protein dedicated to the secretion of Wnt proteins from signaling cells. *Cell* **125**, 509–522 (2006).
61. Korkut, C. et al. Trans-synaptic transmission of vesicular Wnt signals through Evi/Wntless. *Cell* **139**, 393–404 (2009).
62. Martin-Martin, I. et al. ADP binding by the Culex quinquefasciatus mosquito D7 salivary protein enhances blood feeding on mammals. *Nat. Commun.* **11**, 2911 (2020).
63. Martin-Martin, I. et al. Aedes aegypti D7 long salivary proteins modulate blood feeding and parasite infection. *MBio* **14**, e0228923 (2023).
64. Holleufer, A. et al. Two cGAS-like receptors induce antiviral immunity in *Drosophila*. *Nature* **597**, 114–118 (2021).
65. Slavik, K. M. et al. cGAS-like receptors sense RNA and control 3′2′-cGAMP signalling in Drosophila. *Nature* **597**, 109–113 (2021).
66. Li, Y. et al. cGLRs are a diverse family of pattern recognition receptors in innate immunity. *Cell* **186**, 3261–3276.e20 (2023).
67. Wang, J. & Meng, W. cGAS: Bridging immunity and metabolic regulation. *J. Mol. Cell Biol.* mjaf018 (2025).
68. Palmer, C. S. Innate metabolic responses against viral infections. *Nat. Metab.* **4**, 1245–1259 (2022).
69. Liu, H., Wang, F., Cao, Y., Dang, Y. & Ge, B. The multifaceted functions of cGAS. *J. Mol. Cell Biol.* **14**, mjac031 (2022).
70. Cai, H. et al. 2′3′-cGAMP triggers a STING- and NF-κB–dependent broad antiviral response in Drosophila. *Sci. Signal.* **13**, eabc4537 (2020).
71. Antonova, Y., Alvarez, K. S., Kim, Y. J., Kokoza, V. & Raikhel, A. S. The role of NF-κB factor REL2 in the *Aedes aegypti* immune response. *Insect Biochem. Mol. Biol.* **39**, 303–314 (2009).
72. Martin, M., Hiroyasu, A., Guzman, R. M., Roberts, S. A. & Goodman, A. G. Analysis of *Drosophila* STING reveals an evolutionarily conserved antimicrobial function. *Cell Rep.* **23**, 3537–3550.e6 (2018).
73. Kristensen, N. P. Phylogeny of insect orders. *Annu. Rev. Entomol.* **26**, 135–157 (1981).
74. Ribeiro, T. M. & Espíndola, A. Integrated phylogenomic approaches in insect systematics. *Curr. Opin. Insect Sci.* **61**, 101150 (2024).
75. Chesters, D. The phylogeny of insects in the data-driven era. *Syst. Entomol.* **45**, 540–551 (2020).
76. Trautwein, M. D., Wiegmann, B. M., Beutel, R., Kjer, K. M. & Yeates, D. K. Advances in insect phylogeny at the dawn of the postgenomic era. *Annu. Rev. Entomol.* **57**, 449–468 (2012).
77. Behura, S. K. Insect phylogenomics. *Insect Mol. Biol.* **24**, 403–411 (2015).
78. Yeates, D. K., Meusemann, K., Trautwein, M., Wiegmann, B. & Zwick, A. Power, resolution and bias: recent advances in insect phylogeny driven by the genomic revolution. *Curr. Opin. Insect Sci.* **13**, 16–23 (2016).
79. Giribet, G. & Edgecombe, G. D. The phylogeny and evolutionary history of arthropods. *Curr. Biol.* **29**, R592–R602 (2019).
80. Johnson, K. P. Putting the genome in insect phylogenomics. *Curr. Opin. Insect Sci.* **36**, 111–117 (2019).
81. Tihelka, E. et al. The evolution of insect biodiversity. *Curr. Biol.* **31**, R1299–R1311 (2021).
82. Kohli, M. et al. Evolutionary history and divergence times of Odonata (dragonflies and damselflies) revealed through transcriptomics. *iScience* **24**, 103324 (2021).
83. Kawahara, A. Y. et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. USA.* **116**, 22657–22663 (2019).
84. Johnson, K. P. et al. Phylogenomics and the evolution of hemipteroid insects. *Proc. Natl. Acad. Sci. USA* **115**, 12775–12780 (2018).
85. Peters, R. S. et al. Evolutionary history of the hymenoptera. *Curr. Biol.* **27**, 1013–1018 (2017).
86. Blaimer, B. B. et al. Key innovations and the diversification of Hymenoptera. *Nat. Commun.* **14**, 1212 (2023).
87. Almeida, E. A. B. et al. The evolutionary history of bees in time and space. *Curr. Biol.* **33**, 3409–3422.e6 (2023).
88. de Moya, R. S. et al. Phylogenomics of parasitic and nonparasitic lice (Insecta: Psocodea): combining sequence data and exploring compositional bias solutions in next generation data sets. *Syst. Biol.* **70**, 719–738 (2021).
89. Kawahara, A. Y. et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat. Ecol. Evol.* **7**, 903–913 (2023).
90. McKenna, D. D. et al. The evolution and genomic basis of beetle diversity. *Proc. Natl. Acad. Sci. USA* **116**, 24729–24737 (2019).
91. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**, 361–375 (2005).
92. Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* **36**, 541–562 (2005).
93. Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X. & Rokas, A. Incongruence in the phylogenomics era. *Nat. Rev. Genet.* **24**, 834–850 (2023).
94. Shen, X.-X., Steenwyk, J. L. & Rokas, A. Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. *Syst. Biol.* **70**, 997–1014 (2021).
95. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* **1**, 0126 (2017).
96. Mutti, G., Ocaña-Pallarès, E. & Gabaldón, T. Newly developed structure-based methods do not outperform standard sequence-based methods for large-scale phylogenomics. *Mol. Biol. Evol.* **42**, msaf149 (2025).
97. Durairaj, J. et al. Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).
98. Mifsud, J. C. O. et al. Mapping glycoprotein structure reveals *Flaviviridae* evolutionary history. *Nature* **633**, 695–703 (2024).
99. Huang, J. et al. Discovery of deaminase functions by structure-based protein clustering. *Cell* **186**, 3182–3195.e14 (2023).
100. Himmel, N. J., Moi, D. & Benton, R. Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life. *Curr. Biol.* **33**, 5023–5033.e4 (2023).
101. Liu, W. et al. PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nat. Commun.* **15**, 2775 (2024).
102. Hong, L. et al. Fast, sensitive detection of protein homologs using deep dense retrieval. *Nat. Biotechnol.* **43**, 983–995 (2025).
103. Jenson, J. M. & Chen, Z. J. cGAS goes viral: a conserved immune defense system from bacteria to humans. *Mol. Cell* **84**, 120–130 (2024).
104. Wein, T. & Sorek, R. Bacterial origins of human cell-autonomous innate immune mechanisms. *Nat. Rev. Immunol.* **22**, 629–638 (2022).
105. Hobbs, S. J. & Kranzusch, P. J. Nucleotide immune signaling in CBASS, Pycsar, thoeris, and CRISPR antiphage defense. *Annu. Rev. Microbiol.* **78**, 255–276 (2024).
106. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type i interferon pathway. *Science* **339**, 786–791 (2013).
107. Culbertson, E. M. & Levin, T. C. Eukaryotic CD-NTase, STING, and viperin proteins evolved via domain shuffling, horizontal transfer, and ancient inheritance from prokaryotes. *PLoS Biol* **21**, e3002436 (2023).
108. Millman, A., Melamed, S., Amitai, G. & Sorek, R. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat. Microbiol.* **5**, 1608–1615 (2020).
109. McFarland, A. P. et al. Sensing of bacterial cyclic dinucleotides by the oxidoreductase RECON promotes NF-κB activation and shapes a proinflammatory antibacterial state. *Immunity* **46**, 433–445 (2017).
110. Xia, P. et al. The ER membrane adaptor ERAdP senses the bacterial second messenger c-di-AMP and initiates anti-bacterial immunity. *Nat. Immunol.* **19**, 141–150 (2018).

111. Chow, K. L., Hall, D. H. & Emmons, S. W. The mab-21 gene of *Caenorhabditis elegans* encodes a novel protein required for choice of alternate cell fates. *Development* **121**, 3615–3626 (1995).
112. Yamada, R. et al. Cell-autonomous involvement of Mab21l1 is essential for lens placode development. *Development* **130**, 1759–1770 (2003).
113. Li, L. et al. Hydrolysis of 2′3′-cGAMP by ENPP1 and design of nonhydrolyzable analogs. *Nat. Chem. Biol.* **10**, 1043–1048 (2014).
114. Hou, Y. et al. SMPDL3A is a cGAMP-degrading enzyme induced by LXR-mediated lipid metabolism to restrict cGAS-STING DNA sensing. *Immunity* **56**, 2492–2507.e10 (2023).
115. Maltbaek, J. H., Cambier, S., Snyder, J. M. & Stetson, D. B. ABCC1 transporter exports the immunostimulatory cyclic dinucleotide cGAMP. *Immunity* **55**, 1799–1812.e4 (2022).
116. Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. & Valencia, A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput. Biol.* **18**, e1009818 (2022).
117. Derry, A., Carpenter, K. A. & Altman, R. B. Training data composition affects performance of protein structure analysis algorithms. *Pac. Symp. Biocomput.* **27**, 10–21 (2022).
118. Necci, M. et al. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **18**, 472–481 (2021).
119. Gramates, L. S. et al. FlyBase: a guided tour of highlighted features. *Genetics* **220**, iyac035 (2022).
120. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.* **51**, D29–D38 (2023).
121. Mei, Y. et al. InsectBase 2.0: a comprehensive gene resource for insects. *Nucleic Acids Res.* **50**, D1040–D1045 (2022).
122. Poelchau, M. et al. The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.* **43**, D714–D719 (2015).
123. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
124. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
125. Bai, X. et al. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.* **52**, D18–D32 (2024).
126. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
127. Kriventseva, E. V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
128. Li, Y. et al. HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **185**, 2975–2987.e10 (2022).
129. Zhao, T. et al. Whole-genome microsynteny-based phylogeny of angiosperms. *Nat. Commun.* **12**, 3498 (2021).
130. Steenwyk, J. L., Shen, X.-X., Lind, A. L., Goldman, G. H. & Rokas, A. A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. *MBio* **10**, 1–25 (2019).
131. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
132. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
133. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
134. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
135. Yin, J., Zhang, C. & Mirarab, S. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics* **35**, 3961–3969 (2019).
136. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
137. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
138. Hu, G. et al. flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **12**, 4438 (2021).
139. Deiana, A., Forcelloni, S., Porrello, A. & Giansanti, A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One* **14**, e0217889 (2019).
140. Wells, J. et al. Chainsaw: protein domain segmentation with fully convolutional neural networks. *Bioinformatics* **40**, btae296 (2024).
141. Lau, A. M., Kandathil, S. M. & Jones, D. T. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. *Nat. Commun.* **14**, 8445 (2023).
142. Zhu, K., Su, H., Peng, Z. & Yang, J. A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinformatics* **39**, btad070 (2023).
143. Lees, J. et al. Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* **40**, D465–D471 (2012).
144. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
145. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
146. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
147. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
148. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
149. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
150. Cai, H. et al. The virus-induced cyclic dinucleotide 2′3′-c-di-GMP mediates STING-dependent antiviral immunity in *Drosophila*. *Immunity* **56**, 1991–2005.e9 (2023).

ACKNOWLEDGEMENTS

We thank Antonis Rokas for insightful comments and Xin Qiao for help with the cell culture and transfection. This work was conducted in part using the resources of the Information Technology Center and State Key Lab of CAD&CG at Zhejiang University. This work was supported by the Scientific Research Innovation Capability Support Project for Young Faculty (ZYGXQNJSKYCXNLZCXM-A12 to X.-X.S.), the Key Program of National Natural Science Foundation of China (32530086 to X.-X.S.), the National Science Foundation for Distinguished Young Scholars of Zhejiang Province (LR23C140001 to X.-X.S.), Shanghai Municipal Science and Technology Major Project (S.W.), the New Cornerstone Science Foundation (NCI202328 to S.W.), the National Natural Science Foundation of China (32230015 and 32021001 to S.W., 32200395 to C.C.), Zhejiang Provincial Natural Science Foundation of China (LZ23C020002 to R.P.), the Key-Area Research and Development Program of Guangdong Province (2018B020205003 and 2020B0202090001 to X.Z.), and the Key International Joint Research Program of National Natural Science Foundation of China (31920103005 to X.-X.C.).

AUTHOR CONTRIBUTIONS

X.-X.S. and S.W. conceived and designed the study. X.-X.S., W.W., C.C., Y.Z., J.C., Q.Z., Y.W., H.C., Z.L., H.G., G.-Z.O., C.L., and M.T. performed computational analyses and experiments. X.-X.S., S.W., W.W., C.C., Y.Z., X.Z., Y.C., R.P., J.Y., H.C., G.Z., and X.-X.C. interpreted results. X.-X.S. wrote the manuscript with input from all authors. X.-X.S., S.W., W.W., C.C., X.Z., and H.C. edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41422-026-01220-0.

Correspondence and requests for materials should be addressed to Xiaofan Zhou, Sibao Wang or Xing-Xing Shen.