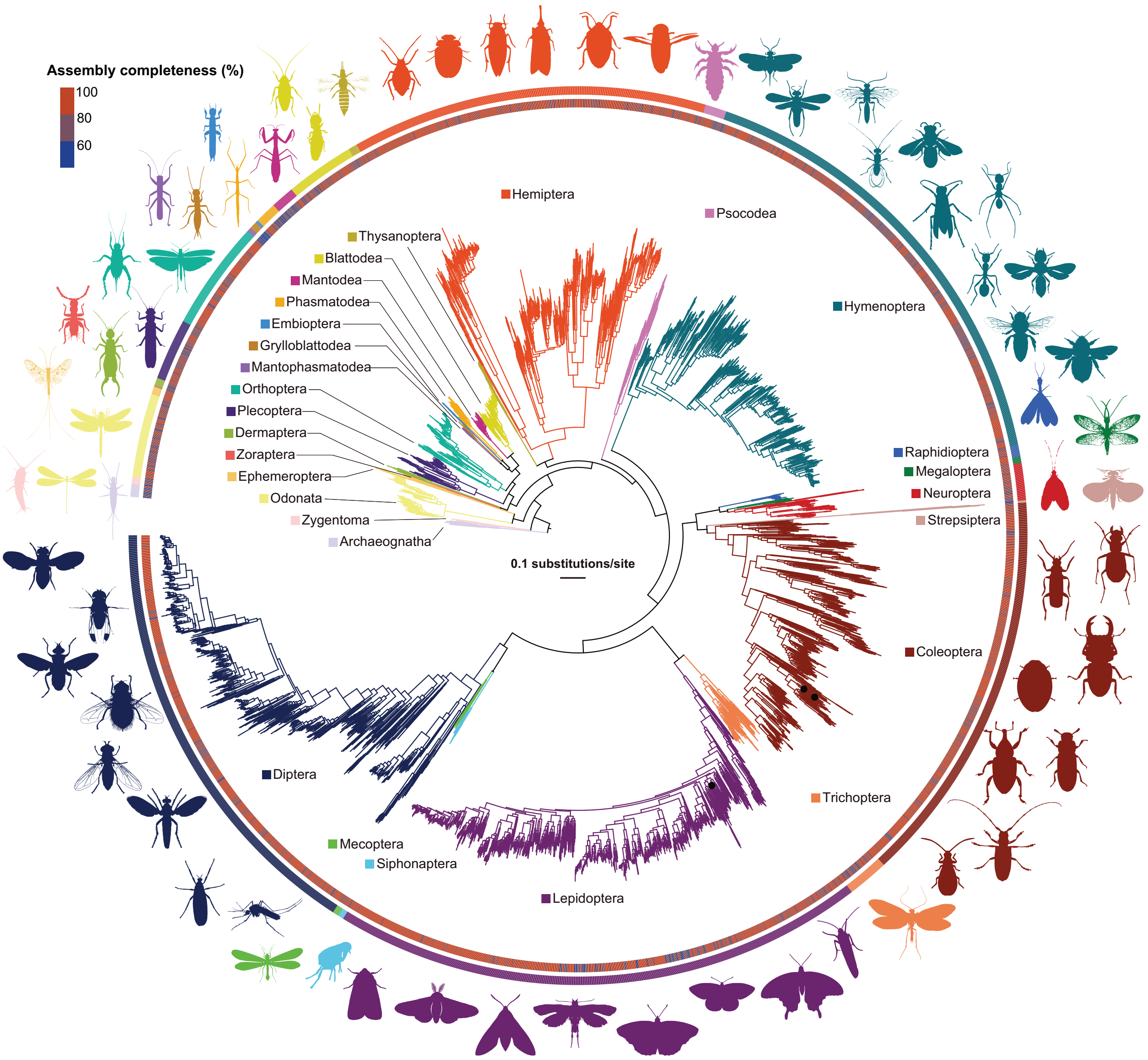

Fig. 1 A comprehensive and well-resolved phylogeny of 4854 insects. The concatenation-based ML phylogeny ($lnL = -191659866.544$) was inferred from amino acid sequences of 824 BUSCO genes (total 276,683 sites) under a single LG + G4 substitution model using IQ-TREE multicore v2.0.7. The complete phylogenetic 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.