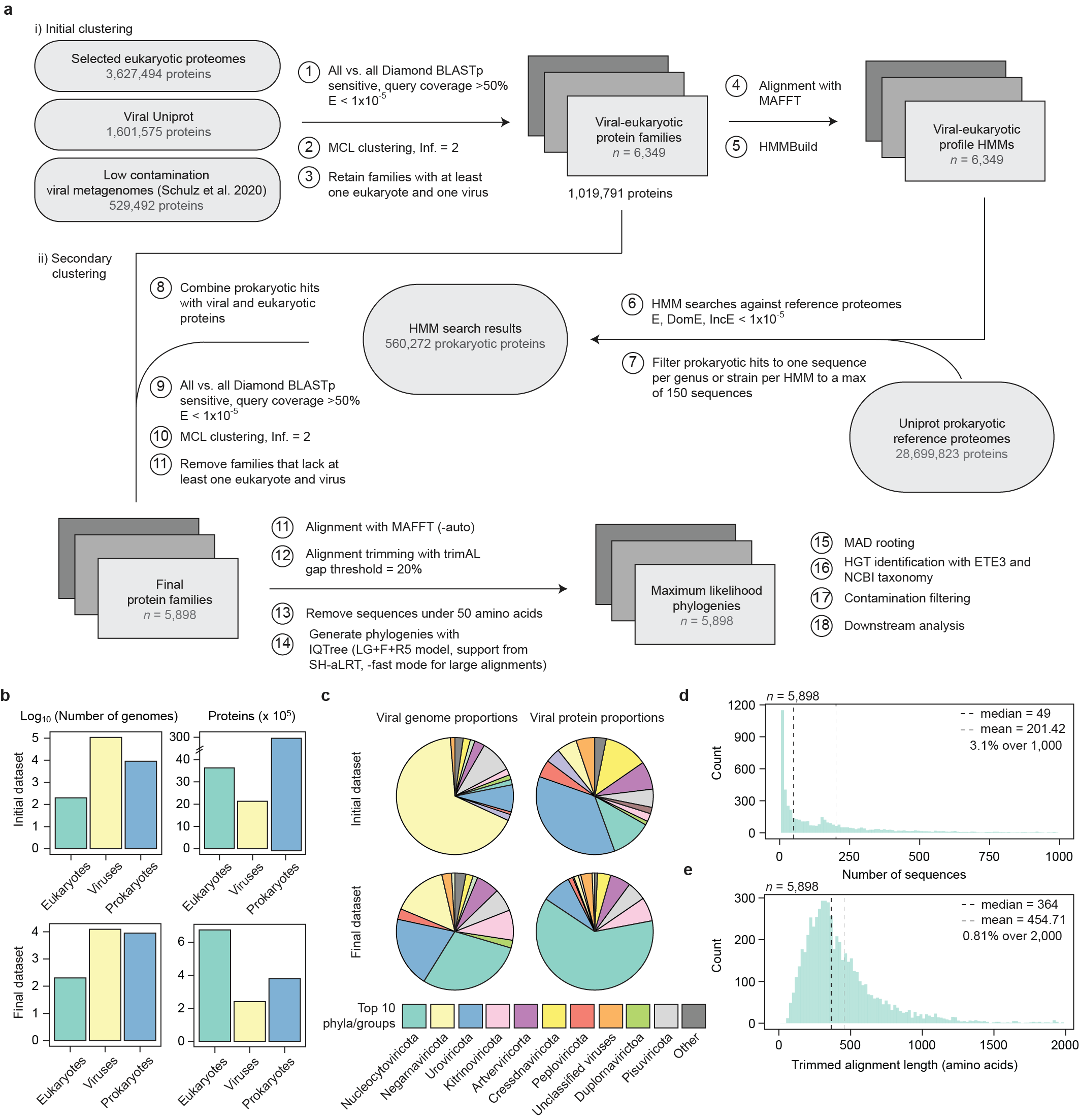
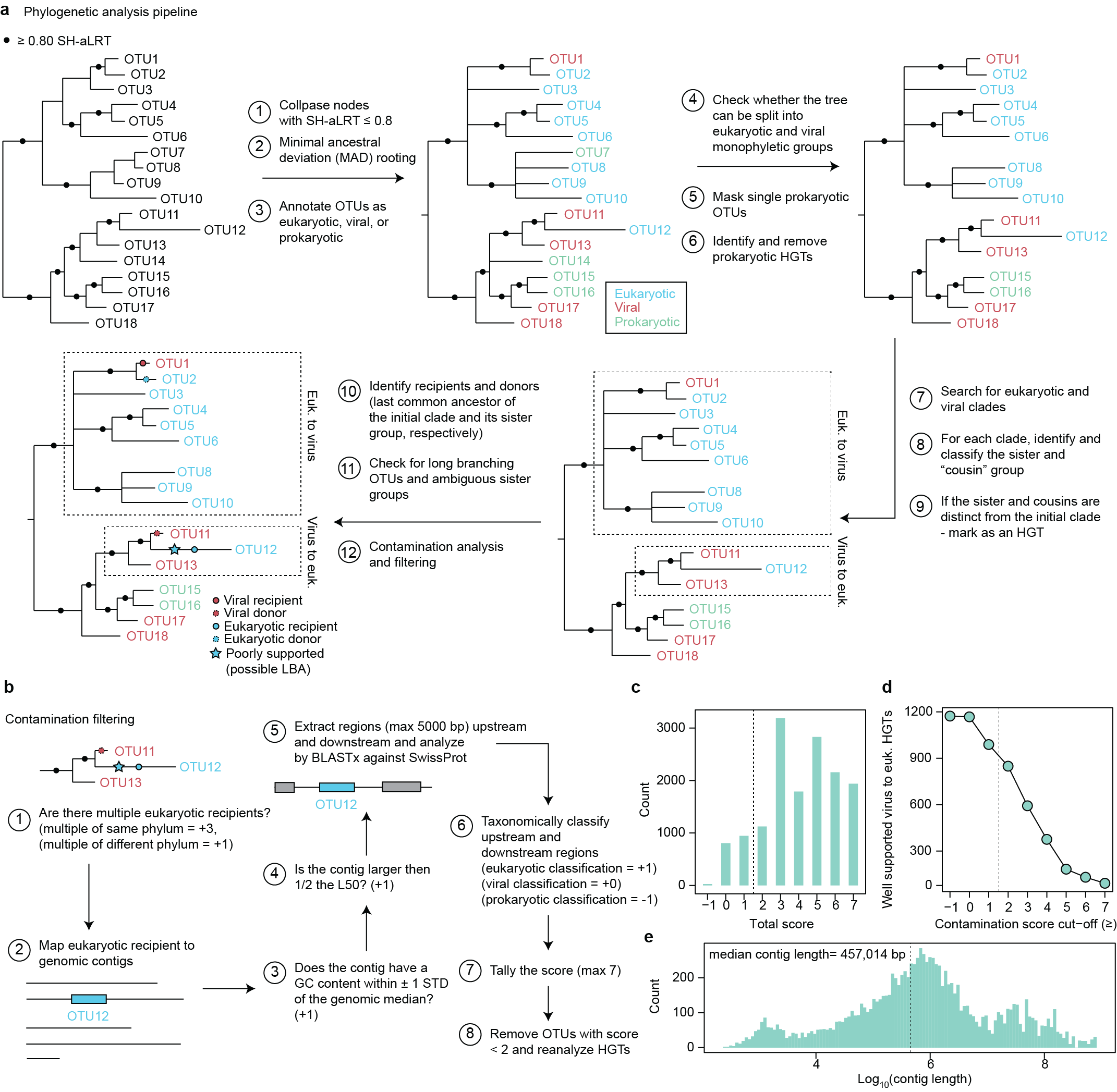
**Extended Data**

****

**Extended Data Fig. 1. Dataset assembly and statistics. a.** A schematic representation of the dataset assembly pipeline. **b.** The numbers of eukaryotic, viral, and prokaryotic genomes and proteins examined and included in the final dataset. **c.** The representation of viral phyla and other groups (i.e., those lacking phyla classifications) in the initial and final datasets. **d, e.** Summary statistics for the final clustered protein families including the number of sequences present (**d**) and the trimmed alignment length (**e**). See Materials and Methods for additional information.

****

**Extended Data Fig. 2. Phylogenetic pipeline and contamination analysis overview. a, b.** Schematic representation of the HGT identification and phylogenetic analysis pipeline (**d**) and the contamination scoring protocol (**b**). **c.** The distribution of contamination scores across eukaryotic recipient sequences. **d.** The number of well-supported HGTs that are identified using different contamination score thresholds. Dashed lines denote the defined scoring threshold (≥ 2). **e.** The distribution of eukaryotic contig lengths that contain viral HGTs. See Materials and Methods for additional information.

****

**Extended Data Fig. 3. Gene origin identification and transduction examples. a.** A schematic illustrating how viral gene origins were approximated by moving up through the phylogeny from the donor towards the root until a cellular lineage was encountered. The pie chart reflects the proportion of well supported virus-to-eukaryotic HGTs that were assigned a given origin. **b, c.** Example phylogenies illustrating cases of eukaryote-to-eukaryote and prokaryote-to-eukaryote transduction. Phylogenies were generated in IQ-Tree using the LG+R7 (**b**) or LG+R9 (**c**) substitution models as selected using ModelFinder and statistical support was assessed using SH-aLRT (*n* = 1,000) 48,49.

**Extended Data Fig. 4. Phylogenies for glycosyltransferases denoted in Figure 4.** Phylogenetic analyses were conducted in IQ-Tree with statistical support assessed using SH-aLRT (*n* = 1,000). Substitution models were selected using ModelFinder and included LG+R10 (**a**), LG+F+R10 (**b**), LG+R9 (**c**), LG+F+R8 (**d, e**), LG+R6 (**f**), LG+F+R7 (**g**), LG+F+G4 (**h**), LG+F+R5 (**i**, **j**), and LG+I+G4 (**k**).