

SUPPORTING INFORMATION

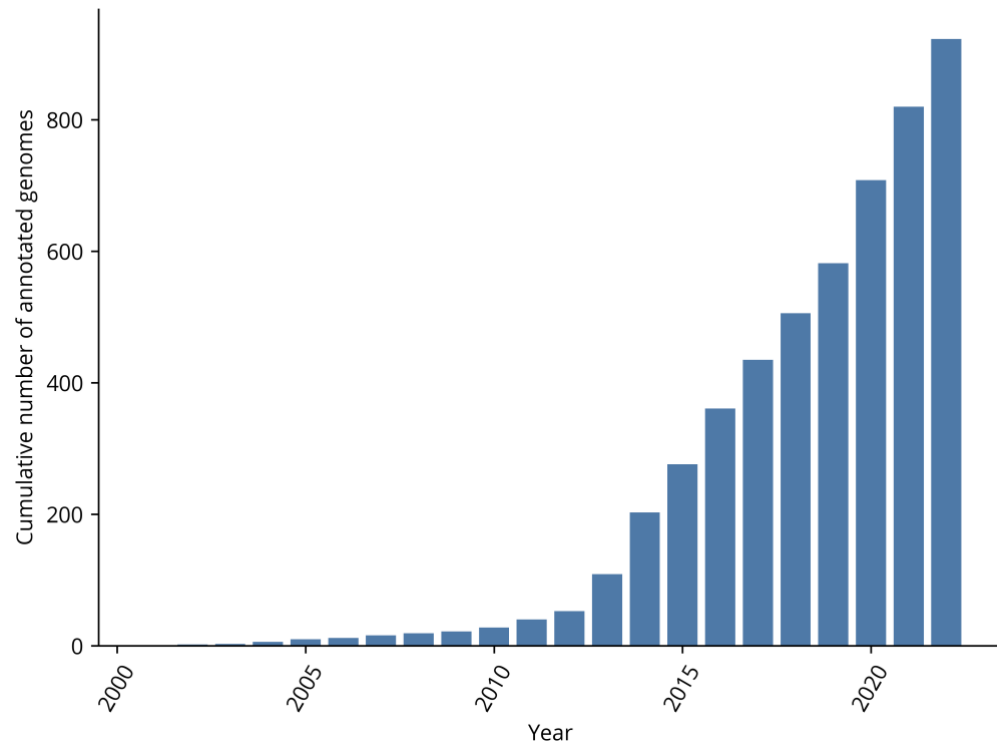


Figure S1: Cumulative number of different eukaryotic genomes annotated by NCBI.

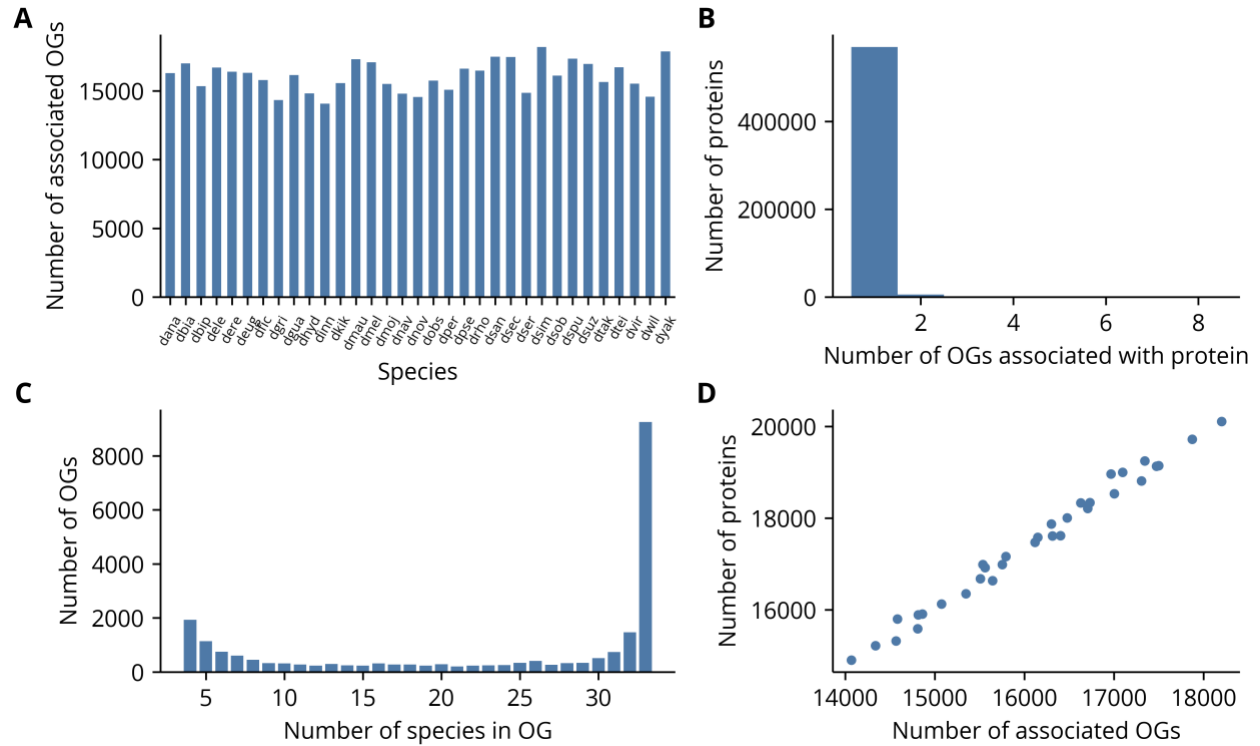


Figure S2: **Statistics of orthologous groups.** (A) Each species is equally represented in orthologous groups (OGs). (B) Nearly all proteins are associated with a single orthologous group. (C) A plurality of orthologous groups contain all species. (D) The number of orthologous groups associated with a species is strongly correlated with the number of unique annotated proteins, which suggests the annotation pipeline generally identifies conserved genes.

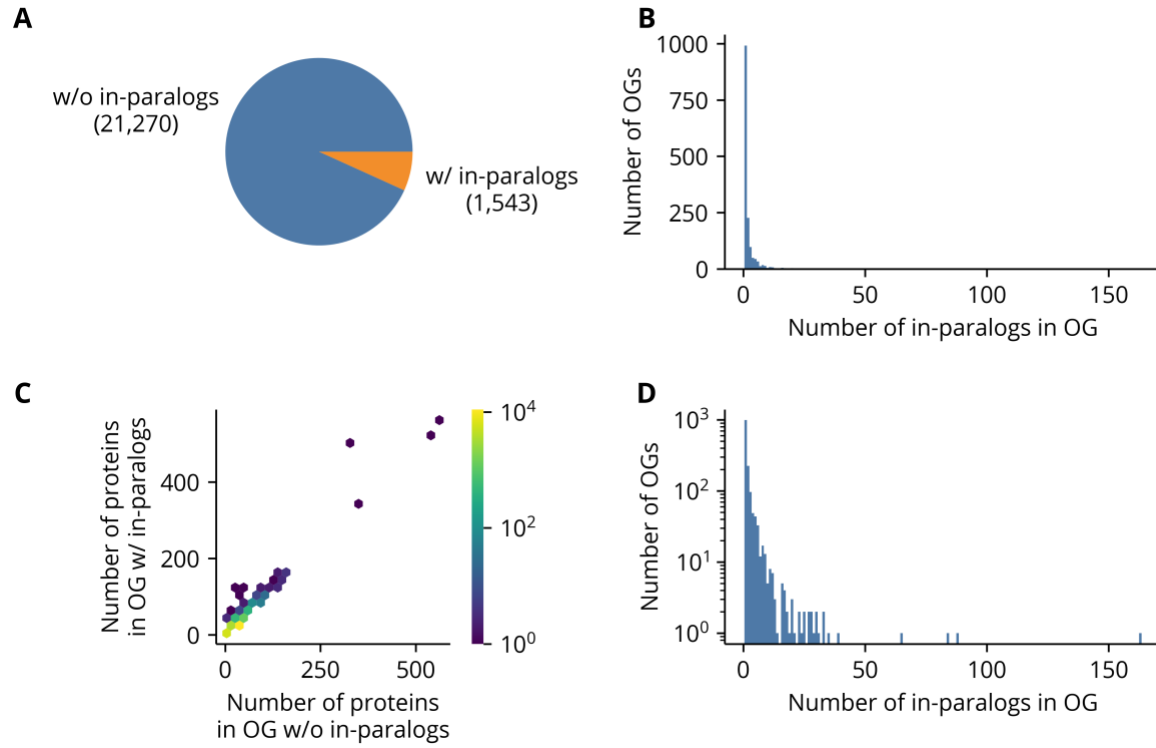


Figure S3: **Addition of paralogs to orthologous groups.** (A) Most orthologous groups (OGs) have no in-paralogs. (B, D) Of the groups with paralogs, most have fewer than five. (C) The in-paralogs are generally only a small fraction of the sequences in an orthologous group.

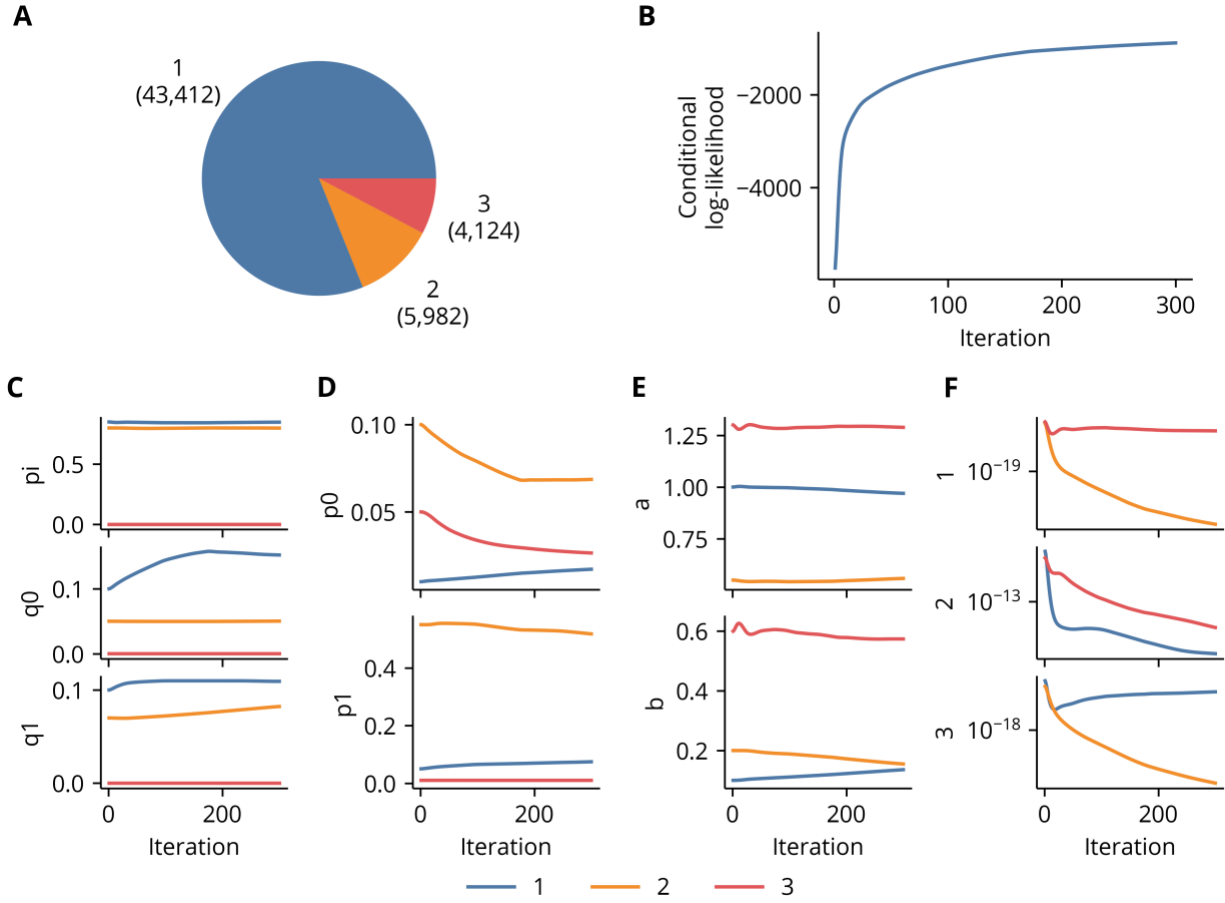


Figure S4: **Insertion phylo-HMM data and training details.** (A) Most columns in the training data were labeled as state 1A or 1B. (B) The model loss stabilized by the final training iteration. (C-F) The values of parameters in the phylogenetic process, the jump process, the pattern stickiness model, and the transition matrix, respectively, at each training iteration. The transition matrix plots are the transition rates to the state indicated on the vertical axis and given in log scale. Self transitions are excluded.

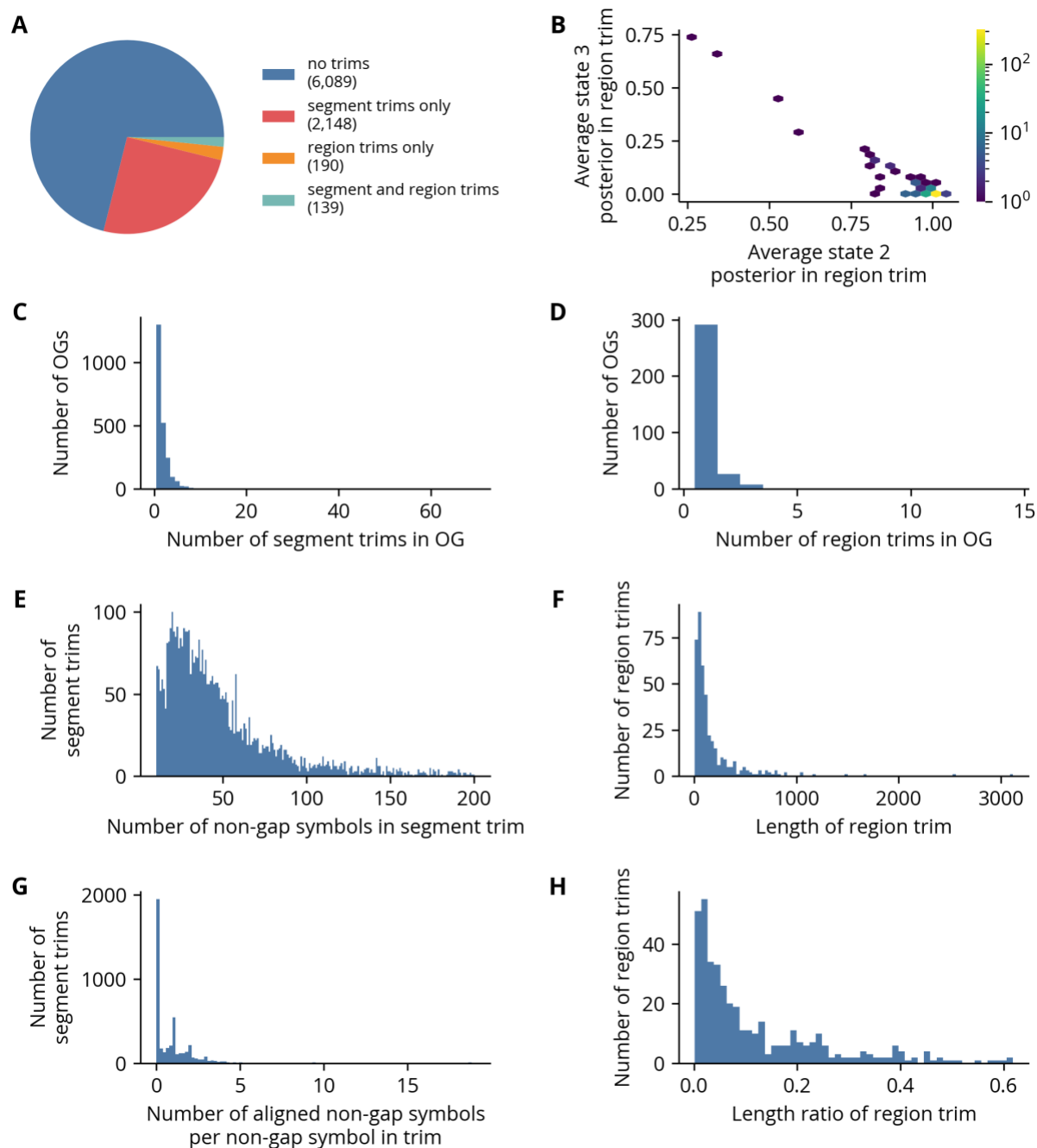


Figure S5: Insertion phylo-HMM trimming details. (A) Most alignments were not trimmed. Of the alignments with trims, most were trimmed only at the level of sequences. (B) Most trimmed regions were inferred primarily as state 2. (C) Most alignments with sequence trims have fewer than 10 segments removed. (D) Most alignments with region trims have fewer than five regions removed. (E, G) The number of non-gap symbols in sequence trims can vary considerably, but for nearly all sequence trims each non-gap symbol in the removed segment is aligned to fewer than five non-gap symbols on average. Only the lower 95% of the distribution of the number of non-gap symbols in the sequence trims is shown. (F, H) The length of region trims can also vary considerably, but generally each region trims accounts for fewer than 10% of the columns in the original alignment.

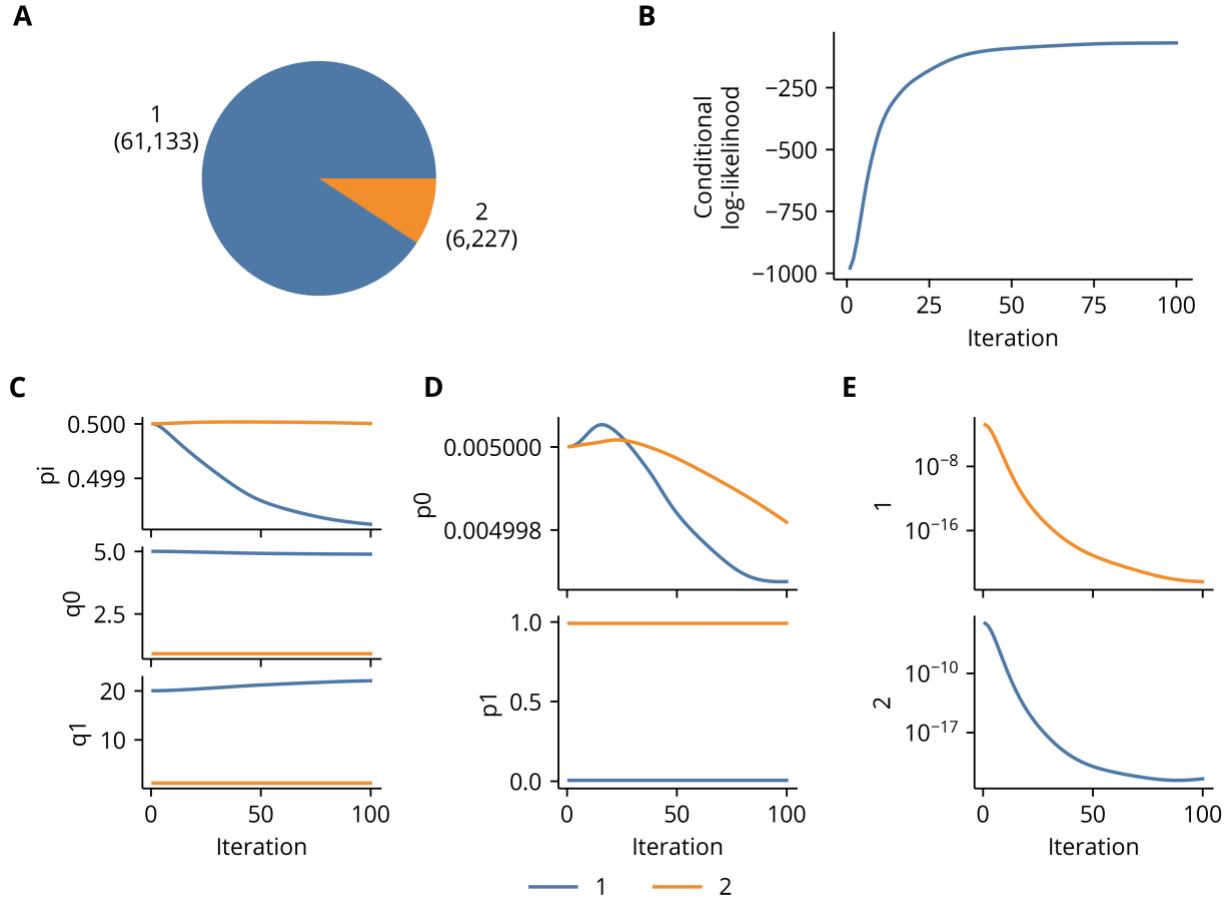


Figure S6: **Missing phylo-HMM data and training details.** (A) Most columns in the training data were labeled as state 1, which is referred to as the “not missing” state in the main text. (B) The model loss stabilized by the final training iteration. (C-F) The values of parameters in the phylogenetic process, the jump process, and the transition matrix, respectively, at each training iteration. The transition matrix plots are the transition rates to the state indicated on the vertical axis and given in log scale. Self transitions are excluded.

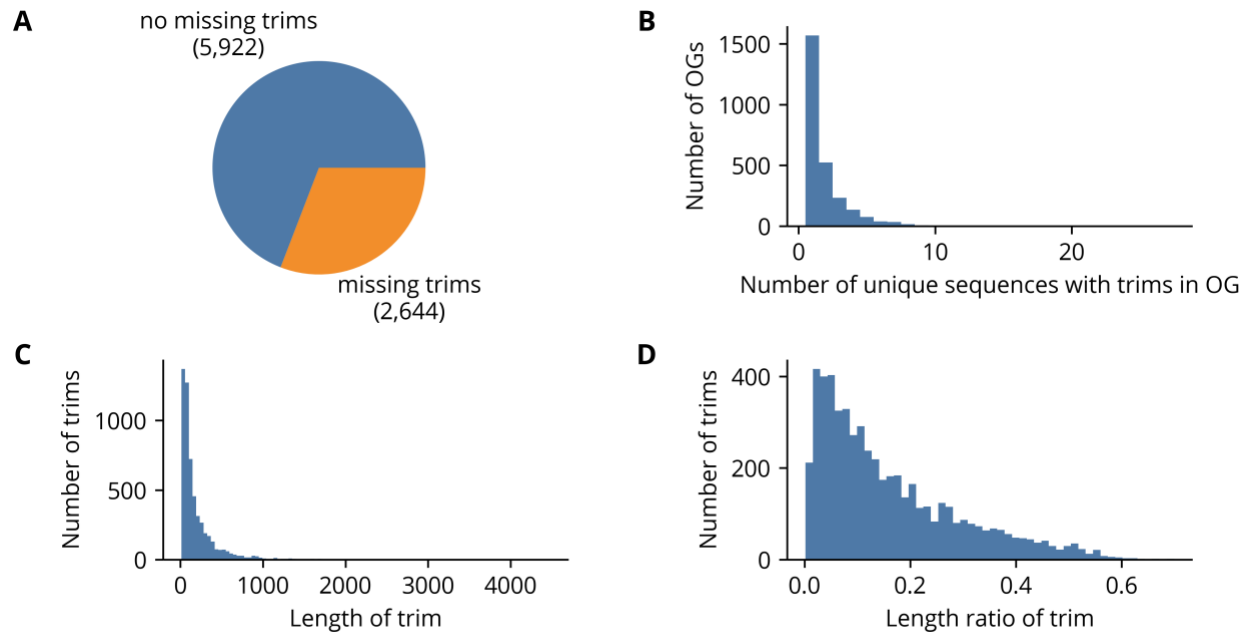


Figure S7: **Missing phylo-HMM trimming details.** (A) Most alignments have no sequences with “missing” segments. (B) Of the alignments with sequences trimmed of missing segments, a majority have only one trimmed sequence. (C-D) The length of missing segments can vary considerably, both in terms of the number of positions as well as its ratio to the number of columns in the alignment.

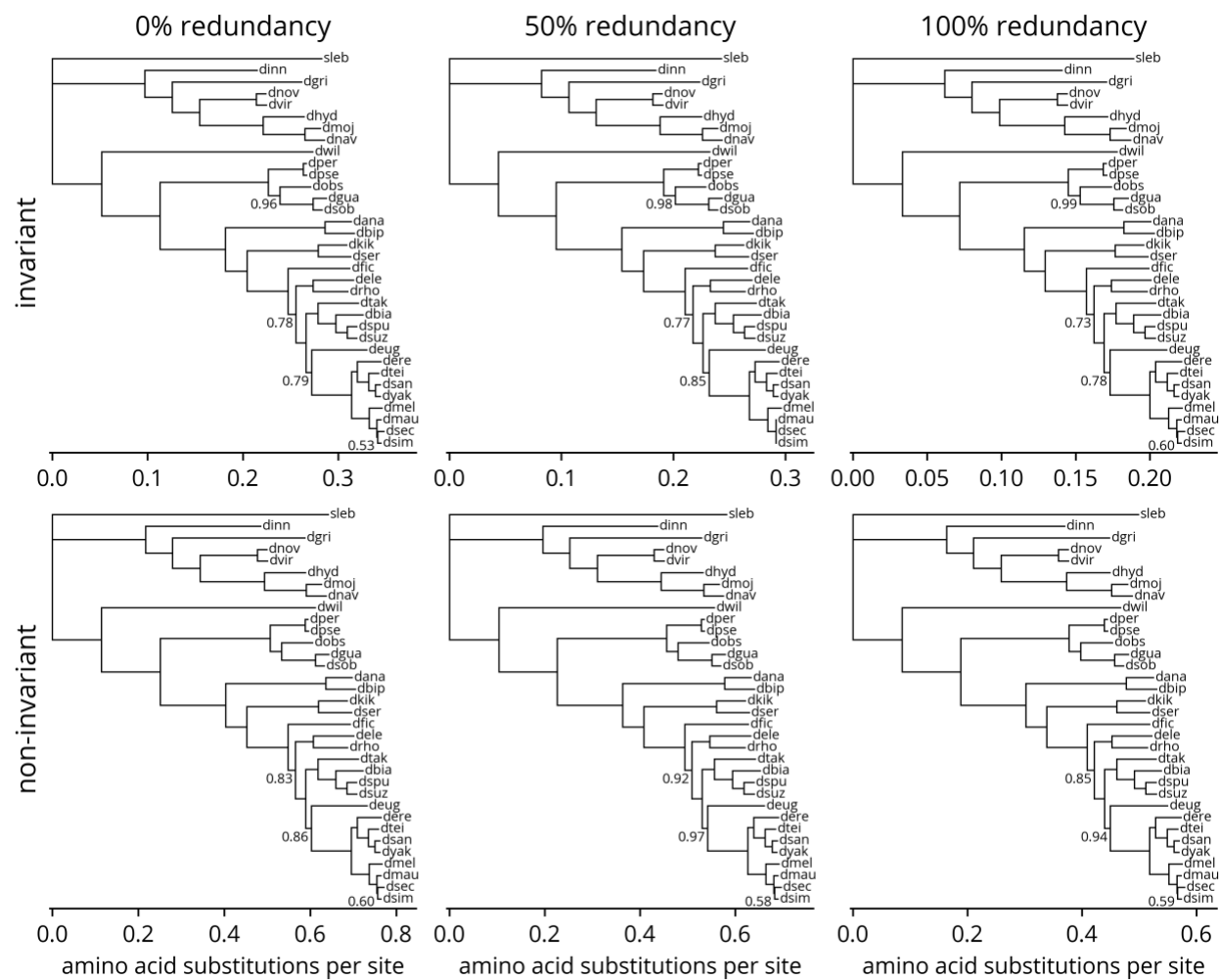


Figure S8: Phylogenetic trees fit to meta-alignments yielded by different sampling strategies under LG model.

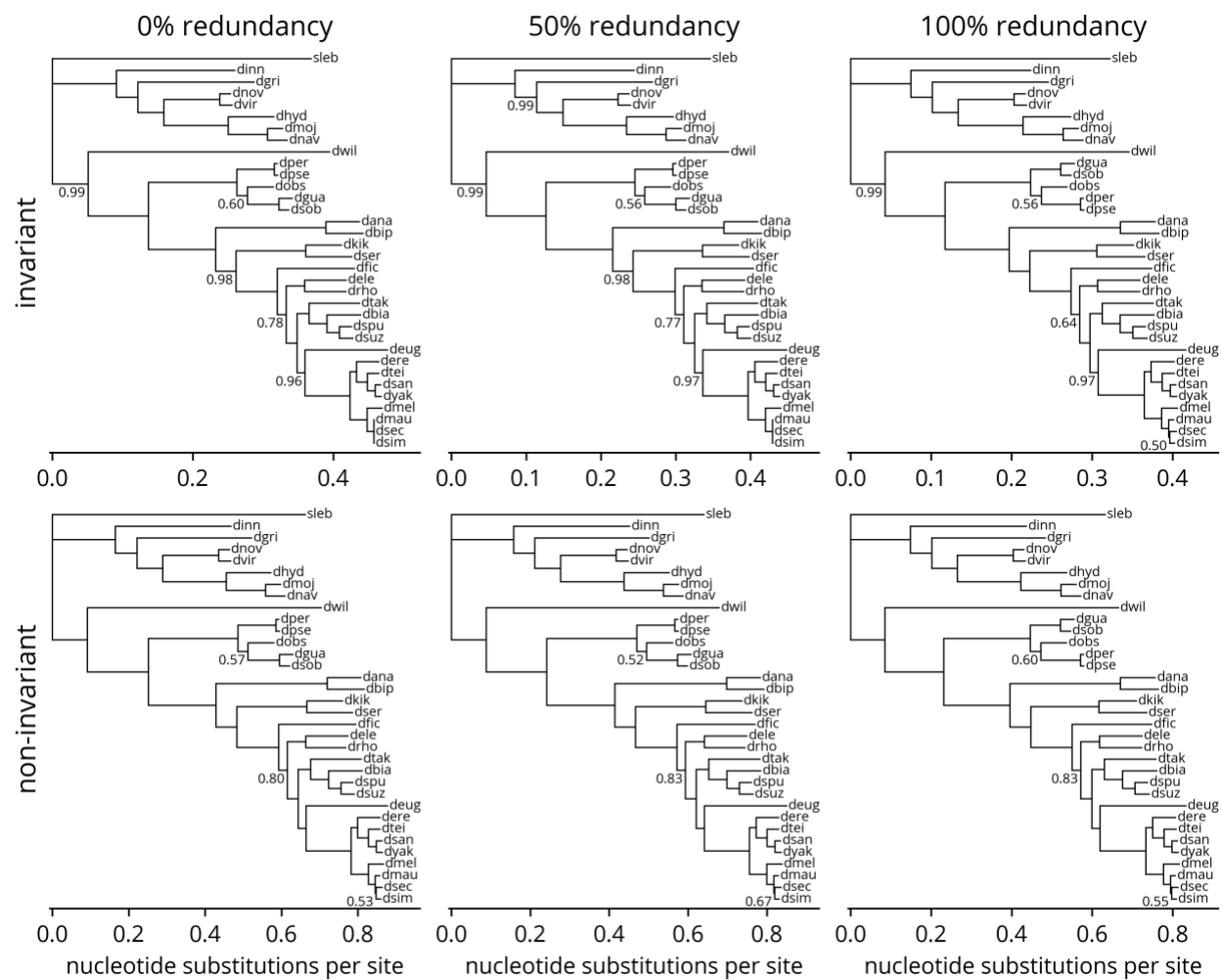


Figure S9: Phylogenetic trees fit by different sampling strategies under GTR model.

Table S1: **Genome annotations.**

Species	Species ID	Taxon ID	Version	Source
<i>Drosophila ananassae</i>	dana	7217	102	NCBI
<i>Drosophila biarmipes</i>	dbia	125945	102	NCBI
<i>Drosophila bipectinata</i>	dbip	42026	102	NCBI
<i>Drosophila elegans</i>	dele	30023	102	NCBI
<i>Drosophila erecta</i>	dere	7220	101	NCBI
<i>Drosophila eugracilis</i>	deug	29029	102	NCBI
<i>Drosophila ficusphila</i>	dfic	30025	102	NCBI
<i>Drosophila grimshawi</i>	dgri	7222	103	NCBI
<i>Drosophila guanche</i>	dgua	7266	100	NCBI
<i>Drosophila hydei</i>	dhyd	7224	101	NCBI
<i>Drosophila innubila</i>	dinn	198719	100	NCBI
<i>Drosophila kikkawai</i>	dkik	30033	102	NCBI
<i>Drosophila mauritiana</i>	dmau	7226	100	NCBI
<i>Drosophila melanogaster</i>	dmel	7227	FB2022_02	FlyBase
<i>Drosophila mojavensis</i>	dmoj	7230	102	NCBI
<i>Drosophila navojoa</i>	dnav	7232	101	NCBI
<i>Drosophila novamexicana</i>	dnov	47314	100	NCBI
<i>Drosophila obscura</i>	dobs	7282	101	NCBI
<i>Drosophila persimilis</i>	dper	7234	101	NCBI
<i>Drosophila pseudoobscura</i>	dpse	7237	104	NCBI
<i>Drosophila rhopaloa</i>	drho	1041015	102	NCBI
<i>Drosophila santomea</i>	dsan	129105	101	NCBI
<i>Drosophila sechellia</i>	dsec	7238	101	NCBI
<i>Drosophila serrata</i>	dser	7274	100	NCBI
<i>Drosophila simulans</i>	dsim	7240	103	NCBI
<i>Drosophila subobscura</i>	dsob	7241	100	NCBI
<i>Drosophila subpulchrella</i>	dspu	1486046	100	NCBI
<i>Drosophila suzukii</i>	dsuz	28584	102	NCBI
<i>Drosophila takahashii</i>	dtak	29030	102	NCBI
<i>Drosophila teissieri</i>	dtei	7243	100	NCBI
<i>Drosophila virilis</i>	dvir	7244	103	NCBI
<i>Drosophila willistoni</i>	dwil	7260	102	NCBI
<i>Drosophila yakuba</i>	dyak	7245	102	NCBI
<i>Scaptodrosophila lebanonensis</i>	sleb	7225	100	NCBI

Table S2: **Phylogenetic diversity criteria.**

Species IDs	Minimum number
dinn, dgri, dhyd	2
dnov, dvir	1
dmoj, dnav	1
dper, dpse	1
dgua, dsob	1
dana, dbip	1
dkik, dser	1
dele, drho	1
dtak, dbia	1
dspu, dsuz	1
dere, dtei	1
dsan, dyak	1
dmel	1
dmau, dsec, dsim	1