Lateral gene transfer leaves lasting traces in Rhizaria

Jolien J.E. van Hooff[1§], Laura Eme[1]*

1. Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech Gif-sur-Yvette, France
§ Current address: Laboratory of Microbiology, Wageningen University, 6708WE Wageningen, The Netherlands

***Corresponding author: Laura Eme (laura.eme@universite-paris-saclay.fr)**

Abstract

Eukaryotic lineages acquire numerous prokaryotic genes via lateral gene transfer (LGT). However, LGT in eukaryotes holds many unknowns, especially its frequency, its long-term impact, and the importance of eukaryote-to-eukaryote LGT. LGT, and genome evolution in general, has not been rigorously studied in Rhizaria, which is a large and diverse eukaryotic clade whose members are mostly free-living, single-celled phagotrophs. We here explore LGT across Rhizaria since their origin until modern-day representatives, using a systematic, phylogenetic approach. On average, 30% of the genes present in current-day rhizarian genomes have originated through LGT at some point during the history of Rhizaria, which emerged about one billion years ago. We show that while LGTs are outnumbered by gene duplications, transferred genes themselves duplicate frequently, thereby amplifying their impact on the recipient lineage. Strikingly, eukaryote-derived LGTs were more prevalent than prokaryotic ones, and carry distinct signatures. Altogether, we here quantitatively and qualitatively reveal how LGT affected an entire eukaryotic phylum, thereby further demystifying LGT in eukaryotes.

Introduction

Over the last decade, eukaryotes have been shown to acquire novel genes via lateral gene transfer (LGT), in addition to better understood mechanisms like gene duplication[1–4]. A plethora of studies revealed that LGT contributed to capacities like hosting endosymbionts, gaining or reverting pathogenicity or thriving in low-oxygen environments[5–8]. Recently, LGT-derived genes were estimated to constitute 1% of eukaryotic gene inventories[1]. This estimate hints at a non-negligible role for LGT in eukaryotic evolution, and contrasts suggestions that LGT does not have an enduring, cumulative impact[9]. At the same time, its significance, compared to for example gene duplication, was suggested to be small[10–13].

These controversies highlight that many aspects of eukaryotic LGT remain enigmatic. Since most studies focused on recent LGTs, we cannot estimate the longer-term impact of LGT, specifically if older LGTs perpetuate to current-day genomes. Moreover, most studies assessed LGT into a small set of species, preventing us from comparing LGT frequencies among more distantly related organisms through a unified approach. In spite of long-held assumptions, we actually do not know how LGT frequency compares to other mechanisms for acquiring new genes, such as gene duplication[14]. Furthermore, eukaryote-to-eukaryote LGT has hardly been studied[1,2]. Finally, how transferred genes subsequently evolve remains an open question[4].

We sought to answer some of these questions by deeply and broadly interrogating LGT in eukaryotic microbes ('protists'), specifically Rhizaria. Rhizaria, an old, diverse and yet understudied clade, are interesting for eukaryotic LGT research. First, most LGT studies into eukaryotic microbes concerned pathogens, whereas Rhizaria are mostly free-living. Second, some publications suggested that particular rhizarians did incorporate transferred genes from prokaryotes[15,16] and/or eukaryotes[17], whereas in another none were found[18]. Finally, Rhizaria recently have turned out much more prevalent than anticipated, particularly in the oceans[19–21], rendering them important subjects of broad characterization, including their evolution.

In this work, we uncover how much LGT contributed to Rhizaria genomes by examining LGT from prokaryotic and eukaryotic donors and by comparing LGT to other evolutionary mechanisms. Using a phylogenetics-based approach, we show that on average 30% of rhizarian genes resulted from an LGT during rhizarian evolution. Furthermore, we characterize how genes evolve after LGT, suggesting that they duplicate frequently. Finally, we display hallmark LGTs to illustrate specific functional and evolutionary patterns. Overall, we present a comprehensive, quantitative and qualitative interrogation of LGT in the understudied Rhizaria, stressing the large influx of foreign genes in this predominantly microbial lineage.

Results

Rhizaria gained many genes through LGT from prokaryotes and eukaryotes

To investigate LGTs in Rhizaria, we used a phylogeny-based approach. We collected protein sequences of 29 Rhizaria (Supplementary Table 1). We searched for their orthologs in the SAR clade and for homologs in other eukaryotes, prokaryotes and viruses. The resulting gene families were subjected to phylogenetic inference (see Methods), yielding us 40,951 trees. In each tree, we identified the clusters of rhizarian sequences ('ancestral rhizarian nodes'), to determine their evolutionary origin. We discarded those displaying little signal or contaminants (Methods). We designated an ancestral rhizarian node to either have 1), a vertical origin, if it forms a monophyletic group with stramenopiles or alveolates, 2), an LGT origin, if it is nested in a non-SAR clade, or 3), to embody a *de novo* gene invention, if the tree only contains Rhizaria. We mapped LGTs, gene inventions and gene duplications onto the Rhizaria tree of life

(Methods). We detected 13,282 LGTs into Rhizaria since they diverged from stramenopiles and alveolates. These generated on average 30% of the examined proteins of Rhizaria (Figure 1, Supplementary Table 2), significantly more than the previously estimated 1%[1]. This estimate is consistent among species, whether they are represented by transcriptomic data, or by good-quality genomic data (*B. natans*: 24%, *P. brassicae*: 27%). If we conservatively ignore all ancestral rhizarian nodes that contain only one species, rigorously preventing falsely identifying contamination as LGT, we still identify 1,992 LGT events. In this strict analysis, LGT is found to be at the origin of ~20% of the examined proteins (Supplementary Figure 1). Notably, 9% of examined proteins result from an LGT into a deep branch of the rhizarian tree, such as in the ancestors of Reticulofilosa, Monadofilosa or Rotaliida, respectively (i.e., the labeled clades in Figure 1). This indicates that LGT has long-term and likely adaptive consequences in Rhizaria. Noticeably, because of numerous quality checks put in place to avoid false positives (see Methods), we here only examined 12% of the proteins. Such a limited data coverage is common for similar studies, since they typically include similar measures. Nonetheless, our LGT-tracing approach yields an estimate that is vastly higher than 1%, and we therefore posit that Rhizaria harbor many more LGTs than this number.

We sought to estimate the LGT rate by counting all LGTs along the evolutionary history of one of the species. As an example, we took *P. brassicaea*. We detected 726 LGTs in its lineage since Rhizaria diverged from Stramenopila and Alveolata. Using the mean minimal (1672,63 mya) and maximal (1985,53 mya) SAR divergence estimates[22], this comes down to 0.37 to 0.43 LGTs per million years, indeed lower than the previously refuted 1 LGT per million years[23]. Of note, this rate results from both LGT and subsequent gene losses thereof, which means it is underestimated based on what remains visible to us today. To assess the relative impact of LGT, we detected gene duplications (Figure 1), the mechanism many presume to have dominated eukaryotic evolution. We infer three times more duplication events than LGT events. In 36 out of 57 lineages, the number of duplications is higher. Noteworthy exceptions include the Rhizaria ancestor (181 LGTs, 7 duplications) and the Cercozoa ancestor (241 LGTs, 117 duplications). Nonetheless, we observe a positive correlation of LGT with duplication (Table 1, Spearman's correlation, $r=0.770$, $P<0.001$), as well as with invention (Spearman's correlation: $r=0.656$, $P<0.001$) and branch length (Spearman's correlation: $r=0.588$, $P<0.001$). This suggests that lineages that received many foreign genes also evolved considerably via other mechanisms, except gene loss. The loss rates might however have been overestimated due to proteome incompleteness (see Supplementary Table 1, BUSCO-scores).

To determine where the LGTs came from, we identified the clade in which the LGT nodes are nested. Globally, 48% of the identified LGTs were donated by eukaryotes, 39% by bacteria and fewer than 1% by archaea. The remainder was prokaryotic (2%) or undetermined (11%). Most lineages (46/56) received more eukaryotic than prokaryotic LGTs (Supplementary Figure 3, Supplementary Table 3). Likewise, LGTs that are shared among multiple Rhizaria were mostly donated by eukaryotes: 51%, compared to 26% (Bacteria), <1% (Archaea), 3% (undetermined prokaryotes). We confirm that *R. filosa* has many eukaryotic LGTs, though their overrepresentation is smaller than previously reported (58% versus 97%)[17]. In *P. chromatophora*, a unique primary phototropic species, we confirm a solid number of bacterial

LGTs (402/790, 51%), more than previously reported[16], and it also received many eukaryotic donations (299/790, 38%). We also traced viral genes in gene trees. LGT-originated ancestral rhizarian nodes more often have viral sequences in their gene trees than vertically inherited ones (18% versus 13%, P<0.001). LGT genes thus frequently have viral homologs, possibly suggesting that viruses mediated some LGT events.

The fates of laterally transferred genes

A transferred gene might take up a valuable role in the recipient lineage, but it needs to adapt to its new genomic and cellular environment. We sought to determine if and how LGTs might assimilate to the new host, for example through introns acquisition for LGT genes from prokaryotes, which initially lack spliceosomal introns[4]. In the species for which we have genomic data, most prokaryotic LGT-derived genes contained at least one intron (Figure 2A, *B. natans*: 81%, *P. brassicae*: 63%, *R. filosa*: 69%). However, in two species, they do lack introns significantly more often than vertically inherited genes (*B. natans*: P=0.006, *P. brassicae*: P=0.002). Due to these intron presences, we posit that most prokaryote-derived laterally transferred genes are well-integrated and expressed (for more details, see Supplementary Information, Supplementary Text: 'Introns in prokaryote-derived LGTs').

LGTs assimilation might also be indicated by the way they evolved after transfer. If they duplicate frequently, they may encode a beneficial function, which may profit from an increase in dosage[4]. If LGT genes are lost regularly, they are probably of limited value or maybe even harmful. Various LGT genes were reported to have duplicated in the recipient[7,8,24–26], but, to our knowledge, they have not been systematically compared to native genes. In most lineages (38/56), LGT genes duplicated more frequently than vertically inherited genes (Figure 2B). The differences were significant in 25 lineages, of which in 16, LGT genes duplicated more. This suggests that gene duplications may accrue the percentage of LGT-derived genes over time, provided that they are maintained. We performed a similar comparison for gene losses. In lineages in which we can observe gene loss (i.e., not in the leaves of the species tree), we see that most experienced more losses of vertically inherited genes (13/20 lineages, Figure 2C), but statistically significant differences are observed only in a fraction of them (5), whereas all lineages with more losses of LGT genes display a significant difference. Previous work in red algae also indicated that LGTs are more prone to gene loss[27]. Hence, LGT-derived genes do not accumulate severely because they are often lost. In addition to gene loss and duplication, we studied the evolutionary rates of LGTs and their tendency to gain or lose protein domains. LGT proteins seem to evolve significantly faster than vertically inherited genes in some lineages (Supplementary Information, Supplementary Text, 'LGT sequence divergence'). Our data suggest that they both gain and lose domains more often in the majority of lineages (Supplementary Information, Supplementary Text, 'Domain loss and gain'). Overall, we conclude that in most lineages, LGT genes have a more dynamic evolutionary history than vertically inherited genes, although no single pattern is shared across all lineages (except domain losses, but this may relate to a technical artifact, see Supplementary Information, Supplementary Text).

LGTs have been hypothesized to be enriched in gene-poor, heterochromatic genome regions, prohibiting them from jeopardizing the integrity of the recipient's genes[4]. We studied the genomic environment of LGTs in *B. natans* and *P. brassicae* by measuring the genes' flanking intergenic region (FIR, i.e., the distance to its neighbors). In *B. natans*, we did not observe any significant differences in the FIRs for LGTs compared to vertically inherited genes (Figure 2D). In *P. brassicae*, we noticed that LGTs have larger FIRs than vertical genes (median FIR LGTs: 628, median FIR vertically inherited genes: 505, P=0.012). In addition to the complete sets of LGTs, we compared the FIRs of LGTs acquired at different evolutionary timepoints to those of native genes, allowing us to search for signals of LGTs moving to more gene-dense regions over time[4]. In *P. brassicae,* the recent, species-specific LGTs are located in the most gene-poor regions (median FIR LGTs:777, P<0.001, Supplementary Figure 4E). Hence, LGTs might be initially enriched in gene-poor regions and move to gene-richer, possibly transcriptionally active regions thereafter. However, based on these two species we cannot conclude that this is a common trajectory.

Transferred proteins harbor donor-specific signatures

To better understand the potential functional contribution of LGTs to the recipient cell, and mechanisms and constraints of the transfer process itself, we broadly characterized LGT genes (Supplementary Table 5). We predicted features for sequences and for their ancestors, represented by internal nodes in the reconciled gene trees (see Methods). Various studies reported that LGT genes tend to be relatively short[4,28]. Such a length bias could be caused by the higher chance that a short gene stays intact or does not disrupt essential genes in the recipient genome. Like previous studies[15,29], we observe that LGT genes (median length: 226 amino acids) are shorter than native ones (Figure 3A, median length: 248 amino acids, P<0.001). However, eukaryote-derived LGT genes are actually longer (median length: 258, P<0.001). The bias toward shorter sequences in previous studies probably resulted from their focus on prokaryotic LGTs, which indeed have short lengths in our results (median length: 192, P<0.001). Hence, engulfing and integrating larger DNA segments might not necessarily form a major hurdle for LGT.

To get a grasp of what roles LGTs play, we predicted their cellular localizations using DeepLoc[30] and their functions using EggNOG-mapper[31]. Similar to protein length biases, these functional signatures are different for LGTs donated by prokaryotes and eukaryotes. Prokaryote-derived genes are predicted to be strongly overrepresented in the extracellular environment (2.6-fold higher average localization probability than vertically inherited genes, P<0.001), whereas eukaryotic-derived genes localize relatively often to the nucleus (1.4-fold higher average localization probability, P<0.001, Figure 3B). In fact, their localization propensities display an inverted pattern. It does not surprise that prokaryotic LGTs do not localize to eukaryotic organelles, and that their localizations in general strongly deviate from native genes, stronger than those of eukaryotic LGTs. This stronger deviation also explains why the pattern across LGTs (both eukaryotic and prokaryotic) mostly resembles that of the prokaryotic LGTs (Supplementary Figure 5C). We observe a similar trend for COG functional categories (Figure 3C, Supplementary Figure 5D), although interestingly here, both eukaryotic and prokaryotic

LGTs are overrepresented in ‘Defense mechanisms’ (prokaryotic LGTs: 2.4-fold enrichment, P<0.001, eukaryotic LGTs: 3.1-fold enrichment, P<0.001). Possibly, this hints at a specific preference for the Rhizaria to adopt foreign genes that relate to (genomic) parasite avoidance. However, most COG categories have overrepresentations of LGT genes over vertically inherited genes, rather than the other way around (Figure 3C, Supplementary Figure 5D, most have a score > 1), which might be due to EggNOG-mapper failing to annotate vertically inherited rhizarian genes, possibly because Rhizaria are poorly represented in the EggNOG database. Altogether, these results indicate that previous characterizations of LGT into eukaryotes, in which extracellular localizations and metabolic functions stood out, stem from a focus on prokaryote-to-eukaryote transfer. Indeed, prokaryotic diversity mainly manifests itself through metabolism, while eukaryotic diversity is often morphological, and this difference seems to be reflected in the genes they donated to Rhizaria.

Illustrative LGTs

To exemplify some of the patterns of LGT in Rhizaria, we selected various LGTs from our inventory. We selected them for carrying function-related features that were specifically enriched in LGTs in a particular lineage, compared to vertically inherited genes (see Methods), and confirmed the LGT origin of the sequences concerned by inferring a maximum likelihood phylogeny with a more sophisticated evolutionary model (LG+C60). We observed that the foraminiferan *Elphidium margaritaceum* displayed, surprisingly, an overrepresentation of the COG category ‘Defense mechanisms’ in its LGTs, and that these frequently seem donated by Opisthokonta, such as animals (Figure 4A). The protoplast feeder *L. vorax* has many LGTs predicted to localize to the extracellular environment, and these typically come from either plants (Figure 4B) or bacteria (Figure 4C). Indeed, some descendant sequences of these LGTs harbor signal peptides for secretion, though the ones that lack it may comprise partial sequences. After transfer, the gene may have been duplicated (Figure 4B). In Foraminifera, we observed that LGTs distinctively map to the COG category ‘Nucleotide metabolism and transport’, putatively often of prokaryotic provenance. Interestingly, our selected example also might have been transferred to other eukaryotes, i.e., alveolates and Discoba (Kinetoplastida), possibly also involving a transfer among these two (Figure 4D). In Rhizaria, the gene seems to have been duplicated, particularly in *R. filosa*, which is represented by genomic data. Interestingly, some rhizarian sequences show homology to giant viral genes (GVOGs), potentially indicating that they may have entered the foraminiferan ancestor via a viral agent.

Discussion

In this study, we demonstrate that LGT contributes significantly to the gene contents of eukaryotes, as exemplified by the unicellular-dominated Rhizaria, where they gave rise to on average 30% of the proteins that we could scrutinize phylogenetically. This severely exceeds the previously suggested 1%[1]. In fact, this percentage approximates estimates of the LGT contents of prokaryotic genomes[32]. We ascribe this unexpectedly high percentage to our phylogenetically deep investigations and to our inclusion of eukaryote-derived LGTs.

Importantly, our findings argue for considering LGT as an evolutionary mechanism shaping eukaryotic genomes. While gene duplications occur more frequently, LGTs are non-negligible, as indicated by the over 13,000 identified LGT events across the Rhizaria tree of life. Disregarding LGT leads to overestimating species-specific genes, or to reconstructing large ancestral genomes combined with many gene loss events. Indeed, various eukaryotes have been assigned many lineage-specific genes. This phenomenon may partially be explained by LGT from prokaryotes, in addition to extensive sequence divergence and a lack of genome sampling. Moreover, many regard the reconstructed genomes of the last eukaryotic common ancestor (LECA) and other ancestors as inconceivably large, also given the many subsequent gene losses they require. Both would be alleviated by allowing for eukaryote-to-eukaryote LGT while reconstructing such ancestors, which actually seems more common than prokaryote-to-eukaryote LGT.

We cannot exclude that some of the LGTs reported here are the result of contaminant sequences, in spite of measures to filter these out. Yet, our most conservative phylogenetics-based estimate came down to 20% of genes as derived from LGT. This estimate contains minimal contamination, since it excludes species-specific genes. In fact, it likely represents a severe underestimation, because many species in our dataset are distantly related to one another, as indicated by the long terminal branches in the species tree, which indeed correlate with the number of species-specific LGTs (Supplementary Figure 9). We hence need a more extensive sampling in Rhizaria[14,33] to fully appreciate the impact of LGT in this clade. Whereas Rhizaria datasets may be prone to contamination due to their feeding behavior and (endo)symbiotic interactions, these lifestyles also facilitate the acquisition of foreign genes. Such interaction-mediated LGT may strengthen the interaction itself[34].

These results raise questions about eukaryote-to-eukaryote LGT, which is both abundant and different from prokaryote-to-eukaryote LGT. A key question is whether eukaryote-derived genes operate in typical eukaryotic cellular processes, because integration into such a process might be complex. After all, typically many different players are involved, so the transferred gene would need to establish many interactions. Nonetheless, LGT genes from eukaryotes might have some advantages for functional integration. If they have introns, they might suffer less from silencing, such as via the HUSH complex[35]. Generally speaking, closer related species might exchange genes via LGT more frequently, as reported in grasses[36,37]. For exactly this reason, our LGT estimates are still limited, since we did not detect LGTs from stramenopiles, alveolates or from other rhizarians. Detecting LGTs among closer related lineages is inherently challenging, because gene trees often do not allow for reliable discrimination of LGT from vertical inheritance.

Our study yielded no clear cues about potential mechanisms of eukaryotic LGTs, an important gap in our current knowledge[2]. Viruses have been proposed to serve as transfer agents, also in Rhizaria[38,39]. We observed that many LGT-derived proteins have homologs in viruses, indicated by viral sequences in LGT-associated gene trees, which may suggest transfer via viruses. A recent large-scale investigation uncovered that viruses and eukaryotes exchange genes

frequently[40]. Also, giant viruses integrate into eukaryotic genomes, providing a large 'foreign' coding capacity[41]. However, the two genomes that we analyzed did not clearly have LGTs co-localizing with viruses. Possibly, we can discern potential mechanisms once we have a better and more diverse understanding of the lifestyles of Rhizaria, to assess if certain features are associated with LGT frequency.

Our survey of LGT into Rhizaria provides insights into the genome evolutionary dynamics of an enigmatic clade of eukaryotes, and a resource for potential functional characterization of specific LGT-derived genes. The ongoing genomic characterization of diverse eukaryotes combined with our qualitative automated approach for LGT detection opens up the possibility to systematically assess the patterns reported here across a range of other eukaryotic clades. In turn, this will allow for a complete picture of the impact of LGT across the tree of eukaryotes.

Figures

Figure 1. Stramenopila-Alveolata-Rhizaria (SAR) species tree with the LGT, duplication and invention events across the rhizarian tree of life (left panel). The number of LGT events is given in absolute numbers on each branch, whereas the relative frequencies of LGT, duplication and invention events are depicted in the bar chart on top of the branch. TThe contributions of vertical inheritance, LGT and invention to each species' gene inventories is visualized by the stacked bar charts (right panel). The bar charts present the percentages among the studied proteins. The percentage of studied proteins relative to all proteins is indicated by the blue gradient dot at the very right. The events were derived from a phylogenetics-based interrogation that traced the ultimate origin of a given gene family in the rhizarian tree of life, which could either be 'vertical' (vertically inherited from the common ancestor of the SAR), 'LGT' (via a lateral gene transfer event from a non-SAR lineage) or 'invention' (a *de novo* invented gene in the Rhizaria, apparently absent from any other lineages). The numbers of gene duplications were extracted from gene tree - species tree reconciliation with GeneRax[42] (Methods).

Figure 2. Evolutionary fates and genomic contexts of LGTs. A. Intron presences of vertically inherited and LGT-derived genes, presented as the percentage of examined genes that has at least one intron. Intron presence was only assessed for three species with appropriate genomic data, namely *B. natans*, *P. brassicae* and *R. filosa*. P-values were obtained with the chi-square test of independence (Methods). B,C. Gene duplications and losses (mean of normalized values, see Methods) for vertically inherited genes and LGTs genes across lineages in the rhizarian tree of life. Note that 'Ascetosporea' did not receive any LGTs (Figure 1), hence this branch has no data here. Losses could only be inferred for non-terminal branches. P-values obtained with the Mann-Whitney U-test (Methods). D. Distributions of the flanking intergenic regions (FIRs) for LGTs and for vertically inherited genes in *B. natans* and *P. brassicae*. P-values obtained with the Mann-Whitney U-test (Methods).

Figure 3. Features of LGT proteins. A-C. Input comprises predicted annotations of nodes in single gene trees, reflecting 'ancestral rhizarian nodes' upon gene entry in the Rhizaria clade (Methods, 'Characterizing LGT evolution and function using gene evolutionary histories'). A.

Distributions of protein lengths of vertically inherited genes, all LGTs, LGT genes from prokaryotes and LGTs from eukaryotes. P-values were derived from a Mann-Whitney U test on the distributions of the probabilities for LGT-derived versus vertically inherited proteins (Methods). B. Localization propensities of LGT proteins over vertically inherited proteins for their average probability to localize to each organelle, as calculated by DeepLoc[30]. If relative localization propensity > 1, this indicates that the LGTs have a stronger predicted propensity to localize to that cellular compartment than vertically inherited genes. The LGTs were split into their putative donor affiliations (prokaryotic or eukaryotic). P-values were derived from a Mann-Whitney U test on the distributions of the probabilities for LGT-derived versus vertically inherited proteins (Methods). C. Fold-enrichment of LGT proteins over vertically inherited genes regarding their labeling of a COG functional category based on EggNOG-mapper[31]. The input did not comprise actual proteins, but the predicted annotations of nodes in single gene trees (Methods, 'Characterizing LGT evolution and function using gene evolutionary histories'). The LGTs were split into their putative donor affiliations (prokaryotic or eukaryotic). P-values result from a chi-square test of independence of the proportion of LGT proteins with a certain annotation and of vertically inherited proteins with it.

Figure 4. LGT phylogenies. A-D. Maximum-likelihood phylogenies inferred for selected LGTs with IQ-TREE[43] (Methods). The sequences of the selected LGTs are highlighted in light blue. The 'description' and 'cog category' comprise the EggNOG-mapper annotations of the sequences of the LGTs, if any[31]. The presence of a signal peptide was predicted with TargetP[44] and the homology to a giant viral gene was based on profile HMMs from GVOG[45].

Tables

Table 1. Correlations between the numbers of different evolutionary events (LGT, duplication, invention and loss) as well as the branch length in the species tree, measured across the branches in the rhizarian tree of life.

Event type 1	**Event type 2**	**Spearman correlation coefficient**	**Spearman correlation *P*-value**
LGT	duplication	0.77	*P*<0.001
LGT	invention	0.66	*P*<0.001
LGT	branch length	0.59	*P*<0.001
LGT	loss	-0.07	*P*=0.86
duplication	invention	0.60	*P*<0.001
duplication	branch length	0.46	*P*=0.002
duplication	loss	-0.06	*P*=0.90
invention	branch length	0.46	*P*=0.002
invention	loss	0.14	*P*=0.54
branch length	loss	0.09	*P*=0.78

Methods

Assembling a Rhizaria and sister clades dataset

We established a dataset of predicted protein sequences of 29 Rhizaria lineages, coming from both genomic (four) and transcriptomic data (25). These included publicly available data and data shared by other researchers via personal communication. For all details on the data sources of individual lineages, see Supplementary Table 1. In addition to Rhizaria, we likewise collected data from Stramenopila and Alveolata (together named Halvaria), because of their relatively poor representation in generic databases such as GenBank. Such a deep sampling of Rhizaria's sister clades is supposed to yield us the highest likelihood of detecting vertical descent for Rhizaria genes, in case there is any. The Rhizaria predicted proteomes' completenesses range from 1.6% (*Gromia sphaerica*, a transcriptome) to 89.0% (*P. brassicae*, a genome), with a median of 57%, as estimated by BUSCO (version 4.0.5, lineage dataset eukaryota_odb10)[46]. While four of our species' datasets were derived from genomic data, we did not possess genome sequences themselves from one species (*Globobulimina sp.*), and therefore we excluded it from subsequent analyses for which such genome sequences and annotated genome information were required, such as genome validation of LGTs, gene density and intron examinations (see below), and treated it as a 'transcriptomic' dataset. Furthermore, for one of the three species with genomic data, *R. filosa*, the genome was very fragmented (e.g., the median number of genes per scaffold was one), and therefore we treated it as a 'transcriptomic' dataset as well in most analyses, except for the examination of introns. Finally, we point out that we had originally included another species, *Euglypha rotunda*, in our Rhizaria dataset. However, because we identified a suspiciously high number of LGTs in this species, we eliminated it later. Consequently, we also had to eliminate some sequences from the *L. vorax* proteome (see Supplementary Information, Supplementary Text, 'Eliminating *Euglypha rotunda* and some *Leptophrys vorax* sequences').

Inferring a Rhizaria species phylogeny

To infer a species phylogeny of the species in our dataset, we employed PhyloFisher, a new and sophisticated phylogenetic pipeline for eukaryotic species tree inference[47]. In brief, it consists of software that allows one to search, nominate and select orthologs of 240 marker genes in one's own species, which subsequently can be used to create a concatenated alignment that serves as input for phylogenetic inference software. We ran the first steps of PhyloFisher using its provided scripts: config.py, fisher.py, informant.py, working_dataset_constructor.py, sgt_constructor.py and forest.py, all with default settings. Because this dataset (PhyloFisher v.1.0 dataset) contains various Rhizaria, we were able to use the 'phylogenetically-informed' route in the 'fisher.py' step, supplying the rhizarian species in that dataset as 'Blast Seed'. We used ParaSorter, a graphical user interface tool which is part of

PhyloFisher, to (re)designate the Rhizaria sequences from our dataset as ortholog, paralog or contamination, and if necessary also from the PhyloFisher dataset. Subsequently, we used 'apply_to_db.py' to update the PhyloFisher dataset with our own taxa and select_orthologs.py to eliminate two marker genes for which, using the single gene trees, selecting ortholog was very complicated (H2A and PYGB). Finally, prep_final_dataset.py and matrix_constructor.py were used to build the concatenated alignment. The coverages of the 29 focal Rhizaria lineages in this alignment can be found in Supplementary Table 1 ('Supermatrix coverage'). The eukaryotic species tree including these 29 Rhizaria was inferred using IQ-TREE v.2.0.3[43] under the LG+G4+C60+F model and with ultrafast bootstrapping with 1000 replicates[48]. We rooted the tree on the branch uniting Obazoa+Amoebozoa+CRuMs and Metamonada+Discoba. Within Rhizaria, all branches are well-supported (>80%), except for the one containing Radiolaria (*Lithomelissa setosa* and *Sticholonche zanclea*, 21%). Moreover, *Mikrocytos mackini* clustered within the Metamonada, but this represents a long-branch artifact[49]. Therefore, in the Rhizaria phylogeny that we used to map LGT events (for example displayed in Figure 1), we pruned the *Mikrocytos mackini* branch and grafted it as sister to *G. sphaerica*, according to the position obtained before[49]. The entire original phylogeny is depicted in Supplementary Figure 2.

Establishing and expanding gene families

To maximize the chances of finding orthologs of the protein sequences in Rhizaria in other SAR-lineages, we combined our collection of 418 SAR taxa (still including *E. rotunda*, see Supplementary Information, Supplementary Text, 'Eliminating *Euglypha rotunda* and some *Leptophrys vorax* sequences') and used OrthoFinder (version 2.3.8) to establish orthogroups of their protein sequences[50]. From all of these orthogroups, we selected the ones containing the Rhizaria of our dataset (n=341035). To find potential orthologs not present in Stramenopila or Alveolata, or to find xenologs, that is, homologs related to the Rhizaria sequences via LGT, we assembled an 'outgroup' dataset. For non-SAR eukaryotes and viruses, we collected such sequences from NR. To limit the computational burden of the homology searches, downsampled these sequences to 90% identity using CD-hit (version 4.8.1)[51]. For Bacteria and Archaea, we used data from GTDB release 89[52], and similarly reduced these sequences with CD-hit at a cut-off of 70% sequence identity. Subsequently, we searched for homologs of the rhizarian sequences in the orthogroups using DIAMOND v2.0.9 in the 'ultra-sensitive' mode[53]. We did this separately for seven different search databases, since the outgroup sequences were split into Archaea, Bacteria, Metazoa, Viridiplantae, Fungi, all other eukaryotes, and viruses. In each of these searches, the maximum number of target sequences was set to 2000. Note that we did not allow *all* rhizarian sequences to serve as a query: only the ones that were shorter than twice the median of its orthogroups were used to search with, because searching with abnormally long sequences might gather hits that are not homologous to other sequences in the orthogroup, for example because the very long sequence has an additional protein domain, relative to the other orthogroup members. Finally, we added the outgroup hits across all seven outgroup datasets to the orthogroups of the rhizarian sequence queries.

Large-scale inference of single gene trees

To allow for inferring single gene trees with which we could detect LGTs, we selected the expanded orthogroups with at least 15 and no more than 6000 members. We attempted to shrink orthogroups with more than 6000 members by downsampling their non-SAR sequences with DIAMOND blastp, using the orthogroup's rhizarian sequences as query and setting the –max-target-seq parameter to 500. The large orthogroups that, after downsampling, had no more than 6000 members were included after all. We aligned the sequences in each orthogroup using mafft v7.407[54] in the 'auto' mode and trimmed the alignments with BMGE v1.12[55] using the following parameters: -m BLOSUM30 -b 3 -g 0.7 -h 0.5. We required the resulting multiple sequence alignments to contain at least 50 positions. Also, we demanded individual sequences in the alignment to have no more than 80% gaps in the alignment; otherwise they would be removed from it. If, as a result, the alignment maintained fewer than 15 sequences, it would be excluded from phylogenetic inference. Finally, we inferred single gene trees for orthogroups using IQ-TREE v2.0.3[43] using the LG+F+R5 model and applying the SH-like approximate likelihood ratio test with 1000 replicates (-aLRT 1000) to obtain branch support values[56]. All 40951 unprocessed single gene trees can be found in Supplementary Dataset 1.

Detecting LGT with single gene trees

The majority of patterns we attribute to LGTs in this manuscript were deduced from LGTs that we detected using the 40951 single gene trees. We screened these trees for LGTs using ETE3[57]. In short, for each clade of rhizarian sequences in these trees ('ancestral rhizarian nodes'), we determined its evolutionary origin, which may either be 'vertical' (reflecting vertical inheritance), 'lateral' (reflecting an LGT event) or 'invention' (possibly indicating a novel, rhizaria-specific gene). We determined the origin by considering the taxonomic affiliation of the parent nodes of the ancestral rhizarian node, which itself was determined while ignoring the rhizarian sequences themselves. In brief, if either the first or the second parent (see more detailed explanation below for parent identification) was constituted by SAR sequences, we deduced a vertical origin. Alternatively, if they consisted of prokaryotes, or of eukaryotes from a specific, non-SAR group, we interpreted this as a lateral origin. Finally, if the tree only contains rhizarian sequences, we designate this ancestral rhizarian node, which equals the whole tree, as an invention.

A more detailed description of the analysis is provided here and in Supplementary Figure 7. The analysis of the single gene trees starts with annotating each leaf in the tree. For each leaf, its species identity was found, as well as its taxonomy. The latter was derived from NCBI Taxonomy[58], for eukaryotes and viruses, (downloaded on October 4th, 2021) or from GTDB[52], for prokaryotes (release 89). We also assigned an initial group label to each leaf: 'prokaryotes', 'eukaryotes' (i.e., non-SAR eukaryotes), 'halvarians' (Stramenopila, Alveolata), 'rhizaria', 'viruses' or 'unknown' (Supplementary Figure 7B). If the tree lacked rhizarian sequences (which might be the case if they were removed from the alignment after trimming, see 'Inferring single gene trees'), we did not further analyze it. Since we assume that in most cases viruses are not the ultimate donors of LGTs, but mere transfer agents, we remove the viral sequences from the

tree, while recording the proportion of viral sequences this orthogroup has. We also remove the 'unknown' sequences from the tree.If the remaining tree consists of 80% or more of rhizarian sequences, we call an invention, and we consider the non-rhizarian sequences results of contamination or of LGTs in which Rhizaria served as donors.

Alternatively, we sought to determine whether the rhizarian sequences in this tree were of vertical or lateral origin. We did this for each ancestral rhizarian node in the tree, which we identified as follows. First, we root the tree on a random, non-rhizarian leaf that also has a non-rhizarian sister. We subsequently find all rhizarian nodes in the tree (Supplementary Figure 7C). Considering that some instances of polyphyletic rhizaria might be artifacts, we implemented two measures to merge rhizarian nodes that are separated in the tree, but which likely belong together. First, we merged rhizarian nodes if, upon unison, they contained only a single sequence that was non-rhizarian. This single sequence is then ignored in the subsequent steps (Supplementary Figure 7D). Second, we merged the rhizarian nodes if we considered their separation to be poorly supported, which is the case if there are fewer than two well-supported branches between them. Note that our cut-off for a well-supported branch is 0.8 (SH-aLRT support values). We merged such rhizarian nodes by pruning one and grafting it into the base of the other, keeping the non-rhizarian sister clades intact (Supplementary Figure 7E). Whereas the first measure alleviates the impact of potential LGT or contamination from Rhizaria to other clades, the second corrects probable gene tree anomalies. After these operations, we counted the total number of resulting ancestral rhizarian nodes (Supplementary Figure 7F). If there are more than five of such nodes in the processed tree, we regarded the tree to be too unresolved to draw conclusions about the origins of these nodes, and therefore we did not analyze it. Alternatively, we set out to deduce the origin of each of its ancestral rhizarian nodes.

For each ancestral rhizarian node in the tree, we first assessed whether it might consist of contaminant sequences. Specifically, if it has sequences from a single species only, we checked if these had a suspiciously high identity to any stramenopile or alveolate sequence (>90%), or to any other eukaryotic or prokaryotic sequence (>80%), as determined by a DIAMOND BLASTP search with parameters ultra-sensitive -k 1 --max-hsps 1 --query-cover 70 (Supplementary Figures 7G, 10). To have a very large search database, and thus to have a high sensitivity for potential contamination, we here employed our own SAR dataset as input for the stramenopile and alveolate homology searches, and NR (download date: Apr 13, 2020, SAR and viruses excluded in the homology searches) for other eukaryotic and prokaryotic homology searches. If the ancestral rhizarian node does not pass this contamination check, i.e., if it contained at least one sequence with such a suspiciously identical hit, we discarded it. Alternatively, we unrooted the tree first, and rerooted it to get an optimal insight into the putative origin of the ancestral rhizarian node. For this purpose, we first annotated the internal and external nodes in the tree (Supplementary Figure 7H), using NCBI Taxonomy and GTDB's taxonomy. For a node to be annotated, we require its branch to be well-supported (>0.8 SH-like support). In principle, we defined the clade of a node as the last common ancestor of the species of its leaves. However, to avoid unlikely high-level annotations, such as 'Bacteria' or 'Eukayota', we allowed for a maximum of 20% of leafs from another domain (prokaryotes and eukaryotes), phylum (prokaryotes) or supergroup (eukaryotes), if this would yield a lower-level,

and thus more specific, annotation. For example, we would annotate a node that contains 20 Bilateria leafs (eukaryotes, supergroup Opisthokonta) and three Euglenozoa leaves (eukaryotes, supergroup Discoba), as 'Bilateria'. Allowing for such 'interspersing' sequences reflects the possibility that the node's composition was affected by LGT, contamination or other artifacts, and therefore improves the representation of the actual lineage. After annotating, we inspected the identities of both sisters of the ancestral rhizarian node in an unrooted configuration. If both of them are Stramenopila, Alveolata or SAR, we rooted the tree on the ancestral rhizarian node, in line with the position of Rhizaria in the eukaryotic tree of life, relative to Stramenopila and Alveolata. Alternatively, we selected the sister that has the fewest SAR sequences. Within that sister, we searched for the leaf with the longest branch and we rooted the tree on that leaf. Subsequently, we identified the parents in the tree that would inform us on the origin of the ancestral rhizarian node (Supplementary Figure 7I). From the ancestral node, we moved upward in the tree and stored the first and the second node that have a support value above 0.8. In a few cases, only a single parent was found.

For each parent, we determined its group identity ('SAR', 'eukaryotes', 'prokaryotes' or 'mix', if it contains prokaryotic and eukaryotic sequences) and its specific clade (for example 'Metazoa'), not taking the rhizarian sequences into account. We used this information to detect the origin of the ancestral rhizarian node (Supplementary Figure 7J). If either the first parent or the second parent has 'SAR' as a group identity, we inferred the ancestral rhizarian node to have a vertical origin. Also, if there is no SAR, but the eukaryotic parents contain a wide variety of eukaryotes, we inferred a vertical origin. In this case, the gene might have been lost from stramenopiles and alveolates. Alternatively, if the parents encompass prokaryotes, or eukaryotes from a specific group (not 'Eukaryota', but for example 'Metazoa'), or a mix of prokaryotes and eukaryotes, we inferred it to have a lateral origin, i.e. to be acquired by Rhizaria through LGT. The precise determination of the ancestral rhizarian node origin can be found in our decision schemes (Supplementary Table 7). In case of a lateral origin, we selected the putative donor with the identity of the first or second parent, depending on their group compositions. To prevent incorrect origins of highly divergent sequences, we discarded the origin if an ancestral rhizarian node has a very long branch to its first sister (Supplementary Figure 7K). Specifically, we eliminated it if the median tip-to-tip distance between the ancestral rhizarian node's leaves and the leaves of the first sister is larger than four.

After the analysis of all single gene trees, we exploited the available genome data of *B. natans* and *P. brassicae* to validate the lateral origins that we identified in these species, i.e., to ensure that they were not derived from contamination (Supplementary Figure 7L). For each ancestral rhizarian node with a lateral origin that is specific to *B. natans* or *P. brassicae*, we required that its corresponding genes are located on a scaffold that contained at least one gene for which we inferred a vertical origin. Such a vertical origin in the respective species might either stem from a vertical origin of the entire ancestral rhizarian node that the gene is part of, or from an ancient lateral origin, due to which the species itself inherited the gene vertically.

In total, our analysis of the single gene trees allowed us to determine the origins of 12% of the proteins in our 29 rhizarian lineages. In Supplementary Figure 6, we show the consecutive steps

in our workflow that eliminated sequences, due to which we could not infer the origins of the other 88%.

Comprehensive reconstruction of gene evolutionary histories

Following the detection of the origins of ancestral rhizarian nodes using gene phylogenies, we carried out an overall reconstruction of the evolution of these rhizarian genes in the Rhizaria tree of life, including gene duplications and losses. We gathered the rhizarian protein sequences of the ancestral rhizarian nodes / clusters. If an ancestral rhizarian node or cluster contained at least three sequences, we aligned them using MAFFT 'linsi' and trimmed the alignment with BMGE (settings: -m BLOSUM45 -b 3 -g 0.7 -h 0.5). We used GeneRax v2.0.1[42] to infer rhizaria-only gene trees and to simultaneously reconcile them with the Rhizaria species tree, using the LG+G substitution model and the UndatedDL reconciliation model. For the ancestral rhizarian nodes with fewer than three sequences, we either employed Notung v2.9.1.5[59] (two sequences) or artificially created a 'branch' of this sequence itself. As input for Notung, we used the detached subtree corresponding to the ancestral rhizarian node. We annotated all resulting reconciled single gene trees with the information collected during the LGT detection, most importantly the type of origin (vertical, lateral or invention) and, in case of a lateral origin, the donor clade.

We used the annotated, reconciled single gene trees for the reconstruction, allowing us to project the frequencies of all kinds of events, including LGTs, duplications and losses, onto the Rhizaria species tree. First, we revised the reconciled rhizarian single gene trees by adding branches that were currently missing, but that our LGT detection analysis indirectly inferred. That is, if we, from the LGT detection, designated the rhizarian origin of the single gene tree to have been vertical, but the root of the reconciled single gene tree was annotated as belonging to a younger branch than Rhizaria (for example 'Foraminifera'), we added sister branches to tree at the top level, creating these missing internal nodes (in this case: 'Retaria' and 'Rhizaria'). This way, we pushed back the ancestral rhizarian node to the rhizarian common ancestor. The sister branches correspond to 'lost' branches, and were also annotated as such, along with the hypothesized clade in which they were lost ('Radiolaria' and 'Cercozoa', in the example). Similarly, internal nodes were added *within* the single gene tree, if certain internal nodes in the species tree did not have a representative in it. Also here, such an absence corresponded to a gene loss in a rhizarian lineage, which was added as a lost branch. Note that the second type of 'missing' internal nodes as well as lost branches are present in NHX-formatted trees generated by Notung reconciliation, but the ones generated by GeneRax did not yet have this feature. All branches that were added were given a length of zero, since we have no meaningful way of estimating such branch lengths. Subsequently, for each internal and external (only non-lost ones) node in each revised reconciled gene tree, we collected the following information: origin (vertical, lateral, invention or duplication, depending on the result of the LGT detection, or on the parent node being a duplication node), the donor clade (in case of a lateral origin), whether or not the node itself is a duplication node, the median distance from this node to the leaves of the tree, the total number of gene duplications in its descendants and the total number of losses. We also defined the 'residence' of the gene, reflecting the node's presence along the species

tree: it entails the summed length of the branches along the species tree for which we have the strongest evidence of the presence of the gene in them, namely in the form of actual sequences. For example, the residence of a gene which is only represented by sequences in two distantly related species such *B. natans* and *P. brassicae*, would in their common ancestor be the sum of the distances from that internal node in the species tree to *B. natans* and to *P. brassicae*, respectively, which was collected by pruning the species tree for these two species. This residence metric allowed us to correct for missing data when studying duplication and loss frequencies (see 'Characterizing LGT using gene evolutionary histories' below) and to come up with a normalized frequency of such duplication and loss events. We also collected the median root to tip distance from this pruned species tree, which allows us to correct the value of the median distance of the node.

Finally, we combine the information of all reconciled single gene trees to count the LGTs, duplications and losses in all branches of the species tree. In addition, we use them to estimate which genes were present in the ancestors of the extant Rhizaria in our dataset.

Characterizing LGT evolution and function using gene evolutionary histories

We used the reconciled single gene trees to examine if genes from LGTs are different from vertically inherited genes with respect to their evolution and their functional and structural features. Generally speaking, we here use the information of the genes that we inferred to have been present in a particular branch in the rhizarian species tree, as a result of the reconstruction described above. In such an analysis, we compared all genes in that branch with an ultimately 'vertical' origin in the Rhizaria (reflecting the result of the detection) to the ones that were gained through LGT in that specific branch. Consequently, we for example discarded genes that were present due to vertical inheritance from its parent, but that were ultimately acquired by rhizarians through LGT, albeit in an older ancestor. We reasoned that such genes could distort the signal, because they are difficult to classify as either vertical or lateral. Possibly, they have already adapted to the rhizarian host and behave more 'vertical-like'.

We tested if LGTs display different evolutionary dynamics, i.e., if they differ from vertically inherited genes with regard to their duplication and loss tendencies. For each branch in the species tree, we examined its vertically inherited and laterally acquired genes and counted the numbers of gene duplications and gene losses in their descendants. We normalized these counts using the above-described 'residence', which allows us to A) correct for potentially missing data (e.g., as a results of incomplete predicted protein datasets), and B) account for the fact that genes absent from many branches inherently have fewer opportunities for undergoing these evolutionary events, but that this is not necessarily reflective of their tendency to undergo such events per se. For the normalized losses, we furthermore only take as a numerator the number of losses that we were able to detect in the initial reconciled single gene tree, not the ones that are due to the pushing back of vertically inherited nodes to Rhizaria (see 'Comprehensive reconstruction of gene evolutionary histories'). We assessed whether the two types of gene origin had different normalized counts of duplications and losses by performing a

two-sided Mann-Whitney U test, provided that they both had a distribution, i.e., more than one unique normalized count. For each type of origin, we calculated the mean normalized count and the differences between these means. In a similar manner, we analyzed the branch lengths of LGTs, using the median branch length from the root to the tips, and normalized these with a root-to-tip median branch length from the pruned 'residence' species tree (see 'Comprehensive reconstruction of gene evolutionary histories').

Similarly, we investigated the dynamics of protein domain gain and losses in vertically inherited and laterally derived genes. For this, we used the Pfam annotations of the nodes in the reconciled single gene trees, the gathering of which is described below. Since we were not only interested in domain gains and losses within the Rhizaria, but also in their evolution before they entered the rhizarian lineage, we also inferred which Pfam domains likely had been present in the first parent of the node uniting the rhizaria in the original single gene trees (see 'Detecting LGTs with single gene trees); this parent also contains sequences from non-rhizarian species. To find the Pfam compositions of the parent, we performed HMMER's hmmscan onto its descendant non-rhizarian leaves using Pfam-A 3.1b2[60] (this is also the version used by InterProScan v5.48-83.0, which was used to annotate the rhizarian sequences, see below). We subsequently annotated these parents the same way we annotated the internal nodes in the reconciled (rhizaria-only) single gene trees, which is described below. Finally, for each node in the rhizarian gene tree we collected the domain gains and losses that it experienced before (that is, relative to its parent) and after (in all of its descendants). Similar to the gene duplications and losses, we normalized the counts of the domain gains and losses in its descendants using the gene's residence. We also studied the differences between the domain dynamics of the two origins categories in a similar manner as for the gene duplications and losses.

To study the structural and functional properties of LGTs, we first predicted such features for all rhizarian protein sequences in our dataset. These included intrinsic protein disorder with MobiDB-lite v3.10.0[61], coiled-coil regions and many protein family annotations from InterProScan v5.48-83.0[62], protein subcellular localization predictions with SignalP v5.0b[63], TargetP v2.0[44] and DeepLoc v1.0[30], transmembrane domains with Phobius v1.01[64] and TMHMM v2.0c[65], viral and NCLDV affinities using VOGDB and GVOG datasets from ViralRecall v2.0[45], carbohydrate enzyme activities with dbCAN2[66] and COG functional categories with eggNOG-mapper v2.1.5[31]. For each ancestral rhizarian node with an inferred origin (see 'Detecting LGT with single gene trees') we sought to annotate the internal nodes of its reconciled (rhizaria-only) single gene tree. We collected all features of its leaves. Dependent on the datatype of a given feature, we used a specific approach to project it onto the internal nodes. If it was an 'object', such as the Pfam annotation, we applied Dollo parsimony, with the additional requirement that it needed to be present in at least 10% of the internal node's leaves. If it was a numeric value, such as the protein length, we collected those of all the internal node's leaves and used their median to annotate the internal node. For each feature annotation, we compared its presence (boolean data) or values (numeric data) in LGTs to vertically inherited genes, either separately for the individual branches in the species tree, or combined, and tested their differences for statistical significance with either the chi-square test for a contingency table

(boolean data) or with the Mann-Whitney U test (numeric data). For the first type of data, we reported the numbers of genes (branch-specific analysis) or ancestral rhizarian nodes (combined analysis) that were predicted to possess this feature. For the second type of data, we reported the median value. Note that for the internal node annotations, and for the subsequent analyses, we used two datasets of features: one with largely qualitative exponents of the features and one with largely quantitative exponents of these features (Supplementary Table 5). For Figure 3B, the quantitative data containing the DeepLoc localization probabilities were used (that is, the median values), whereas for Figure 3C, we used the qualitative data of COG categories.

Characterizing LGT intron content and genomic context using individual sequences

Various of the analyses presented here, particularly those pertaining to genomic contexts of genes, were executed at the level of the individual protein sequence, and not at the level of gene representants (i.e., nodes) in reconciled single gene trees. For these analyses, we checked for every rhizarian protein in our dataset if its ultimate origin in Rhizaria (vertical, lateral or invention), meaning that we searched for the 'ancestral rhizarian node' that it was part of. In addition, in case of a lateral or invention origin, we transferred the information about the phylogenetic time point of acquisition, i.e., the branch in the rhizarian species tree that acquired or invented it.

For the examination of intron acquisition, we extracted the following gene information from the gff files of the genome assemblies of *B. natans*, *P. brassicae* and *R. filosa*: the number of introns, the total length of the introns, the proportion of the gene that is intronic and the average length of the introns. We retrieved the median numbers of these four estimates as well as the proportion of genes with at least one intron, all for both vertically inherited and laterally acquired proteins, and for each species individually. In addition, we splitted the proteins that were laterally acquired by their time point of acquisition, allowing us to study patterns in 'young' and 'old' LGTs separately, and tested if they intron properties were different from vertically inherited genes with the Mann-Whitney U or chi-square test. In addition to the overall comparison, we selected the LGTs that were donated by eukaryotes as well as those donated by prokaryotes and compared them to vertically inherited genes separately.

Likewise, we studied the genomic context of LGTs, but herein we were not able to analyze the genome of *R. filosa*, because it is too fragmented to obtain any meaningful results. Therefore, we studied gene densities, viral abundance, NCLDV abundance and transposable element (TE) abundance for *B. natans* and *P. brassicae*. We operationalized these properties by calculating the distance to the nearest gene, virus-annotated gene (based on VOG HMM profiles), giant virus-annotated gene (based on GVOG HMM profiles) and TE, respectively, always screening both strands. TE annotations of the genomes were obtained with RepeatModeler v2.0.1 with the LTR structural discovery pipeline[67] and RepeatMasker v4.1.1[68] with the Dfam TE Tools Container v1.2[69]. We calculated the medians of all these distances for vertically inherited and for laterally acquired proteins, where the latter were also separated based on their time point of

acquisition. We tested the differences in the distributions of these distances using the Mann-Whitney U test.

Selection and gene tree inference of illustrative LGTs

For each COG category (based on EggNOG-mapper) and subcellular localization (based on DeepLoc), we determined which lineages of the Rhizaria tree of life display a particular overrepresentation in its LGTs, relative to its vertically inherited genes. Note that here, we used the predicted annotations of the 'ancestral rhizarian nodes', i.e., the node uniting the rhizarian sequences representing the LGT sequences (see 'Characterizing LGT evolution and function using gene evolutionary histories'). Subsequently, we checked if these often have particular donors, and if so, which. We sampled the LGTs in the lineages that have the functional overrepresentation and the frequent donor, and inspected the corresponding gene phylogenies. We confirmed the LGTs by again inferring a phylogeny of their orthogroup with IQ-TREE, now using a more sophisticated evolutionary model (LG+C60) and with 1000 ultrafast bootstraps[48]. For this, we downsampled very large orthogroups by keeping only 1000 non-rhizarian sequences that are closest to the rhizarian sequences (i.e., having the shortest branches to) in the original gene tree. We detect LGTs in the resulting trees as described in 'Detecting LGTs with single gene trees', though we increased the branch length cut-off to eight, since all branches are substantially longer under this more complex evolutionary model. After validation and manual inspection of the phylogenies, we selected four of them for Figure 4, which reflect diverse functions and evolutionary patterns of LGTs of our inventory.

Statistics and data visualization

We used the Python toolkit SciPy[70] for all statistical analyses performed in this study. If applicable, such as in the case of the Mann-Whitney U test, we used the 'two-tailed' version of that test. Across all statistical tests, we performed a P-value correction for multiple testing errors, namely the Benjamini-Hochberg procedure, which decreases the false discovery rate (FDR). For this purpose, we used the python package statsmodels[71]. The thereafter applied significance threshold was $P<0.05$. All figures were generated with R's ggplot2 package and with the ggtree package[72], except Figure 4, which was made with iTOL[73].

Data availability

Supplementary Datasets 1-3 can be found at
https://figshare.com/projects/Lateral_gene_transfers_LGTs_in_Rhizaria/158240

Acknowledgements

This work was supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant to L.E. for Macro-EpiK, call ID ERC-2018-STG, 803151). Bioinformatic analyses were run on Amoeba, the IFB Core

Cluster and the ABiMS Cluster. We thank the following people for sharing sequence data of various SAR species: Christian Woehle and Alexandra-Sophie Roy (Kiel University), Rebecca Gast (Woods Hole Oceanographic Institution), Nick Irwin and Varsha Mathur (University of Oxford), Fabien Burki (Uppsala University), Denis Tikhonenkov (Papanin Institute for Biology of Inland Waters), Jürgen Strassert (Leibniz Institute of Freshwater Ecology and Inland Fisheries), Tonje Marita Bjerkan Heggeset (SINTEF Materials and Chemistry), Kristina Terpis and Chris Lane (The University of Rhode Island), and Matthew Brown (Mississippi State University). We thank the PhyloFisher team (Mississippi State University) for providing support in inferring the eukaryotic species phylogeny, Edward Susko (Dalhousie University) for advice on implementing multiple testing correction, and Michael Seidl (Utrecht University) for advice on genomic density estimates.

Author contributions

Conceptualization: J.J.E.v.H. and L.E.; investigation: J.J.E.v.H.; result analysis and interpretation: J.J.E.v.H. and L.E. supervision: L.E.; writing: J.J.E.v.H. and L.E. Funding acquisition: L.E.

Competing interests

The authors declare no competing interests.

Supplementary Material

All supplementary figures, tables, text and datasets can be found as Supplementary Material alongside this preprint. Figure and table captions, dataset descriptions and supplemental text can be found in the file ‘Supplementary Information’.

References

1. Van Etten J, Bhattacharya D. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends Genet*. 2020;36(12):915-925.

2. Sibbald SJ, Eme L, Archibald JM, Roger AJ. Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends Parasitol*. Published online August 19, 2020.

3. Gabaldón T. Patterns and impacts of nonvertical evolution in eukaryotes: a paradigm shift. *Ann N Y Acad Sci*. Published online 2020.

4. Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol*. 2018;16(2):67-79.

5. Husnik F, Nikoh N, Koga R, et al. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*. 2013;153(7):1567-1578.

6. Shi-Kunne X, van Kooten M, Depotter JRL, Thomma BPHJ, Seidl MF. The Genome of the Fungal Pathogen Verticillium dahliae Reveals Extensive Bacterial to Fungal Gene Transfer. *Genome Biol Evol*. 2019;11(3):855-868.

7. Eme L, Gentekaki E, Curtis B, Archibald JM, Roger AJ. Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut. *Curr Biol*. 2017;27(6):807-820.

8. Xu F, Jerlström-Hultqvist J, Kolisko M, et al. On the reversibility of parasitism: adaptation to a free-living lifestyle via gene acquisitions in the diplomonad Trepomonas sp. PC1. *BMC Biol*. 2016;14:62.

9. Ku C, Nelson-Sathi S, Roettger M, et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*. 2015;524(7566):427-432.

10. Tria FDK, Brueckner J, Skejo J, et al. Gene Duplications Trace Mitochondria to the Onset of Eukaryote Complexity. *Genome Biol Evol*. 2021;13(5).

11. Tria FDK, Martin WF. Gene duplications are at least 50 times less frequent than gene transfers in prokaryotic genomes. *Genome Biol Evol*. Published online October 1, 2021.

12. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genet*. 2011;7(1):e1001284.

13. Colnaghi M, Lane N, Pomiankowski A. Genome expansion in early eukaryotes drove the transition from lateral gene transfer to meiotic sex. *Elife*. 2020;9.

14. Sibbald SJ, Archibald JM. More protist genomes needed. *Nature Ecology &Amp; Evolution*. 2017;1:0145.

15. Fan X, Qiu H, Han W, et al. Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. *Science Advances*. 2020;6(18):eaba0111.

16. Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. *Proc Natl Acad Sci U S A*. 2016;113(43):12214-12219.

17. Glöckner G, Hülsmann N, Schleicher M, et al. The genome of the foraminiferan Reticulomyxa filosa. *Curr Biol*. 2014;24(1):11-18.

18. Stjelja S, Fogelqvist J, Tellgren-Roth C, Dixelius C. The architecture of the Plasmodiophora brassicae nuclear and mitochondrial genomes. *Sci Rep*. 2019;9(1):15753.

19. del Campo J, Not F, Forn I, Sieracki ME, Massana R. Taming the smallest predators of the oceans. *ISME J*. 2013;7(2):351-358.

20. Guidi L, Chaffron S, Bittner L, et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*. 2016;532(7600):465-470.

21. Biard T, Stemmann L, Picheral M, et al. In situ imaging reveals the biomass of giant protists in the global ocean. *Nature*. 2016;532(7600):504-507.

22. Strassert JFH, Irisarri I, Williams TA, Burki F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat Commun*. 2021;12(1):1-13.

23. Martin William F. Too Much Eukaryote LGT. *Bioessays*. 2017;39(12):1700115.

24. Vancaester E, Depuydt T, Osuna-Cruz CM, Vandepoele K. Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms. *Mol Biol Evol*. 2020;37(11):3243-3257.

25. Shin NR, Doucet D, Pauchet Y. Duplication of horizontally acquired GH5_2 enzymes played a central role in the evolution of longhorned beetles. *Mol Biol Evol*. Published online June 28, 2022.

26. Siddique S, Radakovic ZS, Hiltl C, et al. The genome and lifestage-specific transcriptomes of a plant-parasitic nematode and its host reveal susceptibility genes involved in trans-kingdom synthesis of vitamin B5. *Nat Commun*. 2022;13(1):6190.

27. Rossoni AW, Price DC, Seger M, et al. The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *Elife*. 2019;8:e45017.

28. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. Published online November 12, 2021.

29. Ciach MA, Pawłowska J, Muszewska A. Horizontal gene transfer in 44 early diverging fungi favors short, metabolic, extracellular proteins from associated bacteria. *bioRxiv*. Published online December 6, 2021:2021.12.02.471044.

30. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*. 2017;33(21):3387-3395.

31. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol*. Published online October 1, 2021.

32. Corel E, Méheust R, Watson AK, McInerney JO, Lopez P, Bapteste E. Bipartite Network Analysis of Gene Sharings in the Microbial World. *Mol Biol Evol*. 2018;35(4):899-913.

33. Burki F, Keeling PJ. Rhizaria. *Curr Biol*. 2014;24(3):R103-R107.

34. Gilbert C, Maumus F. Multiple horizontal acquisitions of plant genes in the whitefly Bemisia tabaci. *bioRxiv*. Published online January 12, 2022:2022.01.12.476015.

35. Seczynska M, Bloor S, Cuesta SM, Lehner PJ. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature*. Published online November 18, 2021.

36. Hibdige SGS, Raimondeau P, Christin PA, Dunning LT. Widespread lateral gene transfer among grasses. *New Phytol*. Published online April 22, 2021.

37. Dunning LT, Olofsson JK, Parisod C, et al. Lateral transfers of large DNA fragments spread functional genes among grasses. *Proc Natl Acad Sci U S A*. 2019;116(10):4416-4425.

38. Matsuo M, Katahata A, Tachikawa M, et al. Large DNA virus promoted the endosymbiotic evolution to make a photosynthetic eukaryote. *bioRxiv*. Published online October 18, 2019:809541.

39. Lhee D, Lee J, Ettahi K, et al. Amoeba Genome Reveals Dominant Host Contribution to Plastid Endosymbiosis. *Mol Biol Evol*. 2021;38(2):344-357.

40. Irwin NAT, Pittis AA, Richards TA, Keeling PJ. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol*. Published online December 31, 2021.

41. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature*. 2020;588(7836):141-145.

42. Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Mol Biol Evol*. 2020;37(9):2763-2774.

43. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*. 2020;37(5):1530-1534.

44. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, et al. Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*. 2019;2(5):e201900429.

45. Aylward FO, Moniruzzaman M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in ’Omic Data. *Viruses*. 2021;13(2).

46. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*. 2021;38(10):4647-4654.

47. Tice AK, Žihala D, Pánek T, et al. PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLoS Biol*. 2021;19(8):e3001365.

48. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*. 2018;35(2):518-522.

49. Burki F, Corradi N, Sierra R, et al. Phylogenomics of the Intracellular Parasite Mikrocytos mackini Reveals Evidence for a Mitosome in Rhizaria. *Curr Biol*. 2013;23(16):1541-1547.

50. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):1-14.

51. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152.

52. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*. 2020;38(9):1079-1086.

53. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021;18(4):366-368.

54. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780.

55. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol*. 2010;10:210.

56. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307-321.

57. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol*. 2016;33(6):1635-1638.

58. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* . 2020;2020.

59. Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*. 2012;28(18):i409-i415.

60. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412-D419.

61. Necci M, Piovesan D, Clementel D, Dosztányi Z, Tosatto SCE. MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. *Bioinformatics*. Published online December 16, 2020.

62. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236-1240.

63. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019;37(4):420-423.

64. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*. 2004;338(5):1027-1036.

65. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *J Mol Biol*. 2001;305(3):567-580.

66. Zhang H, Yohe T, Huang L, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018;46(W1):W95-W101.

67. Flynn JM, Hubley R, Goubert C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;117(17):9451-9457.

68. RepeatMasker Home Page.

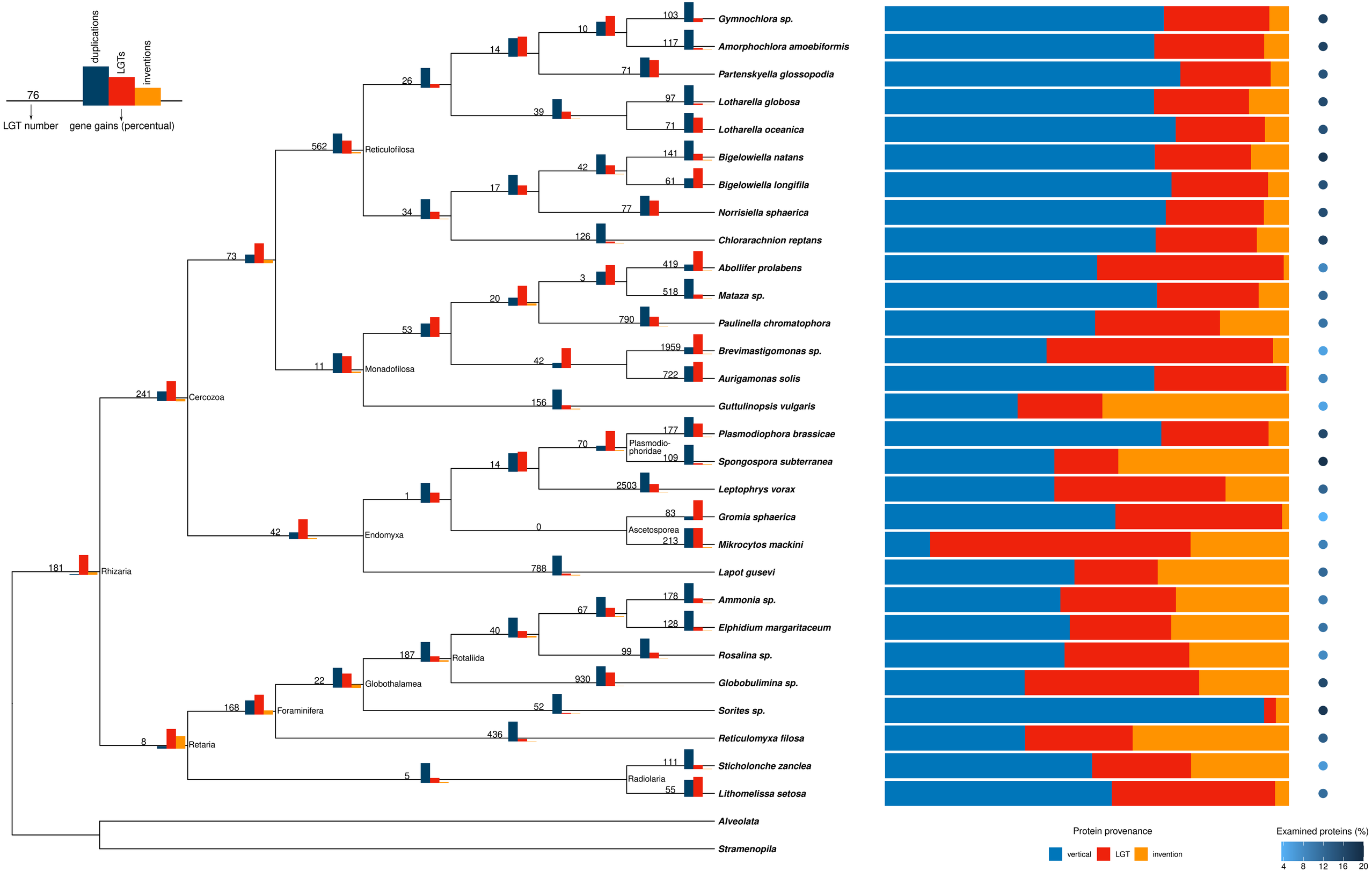
69. *TETools: Dfam Transposable Element Tools Docker Container*. Github

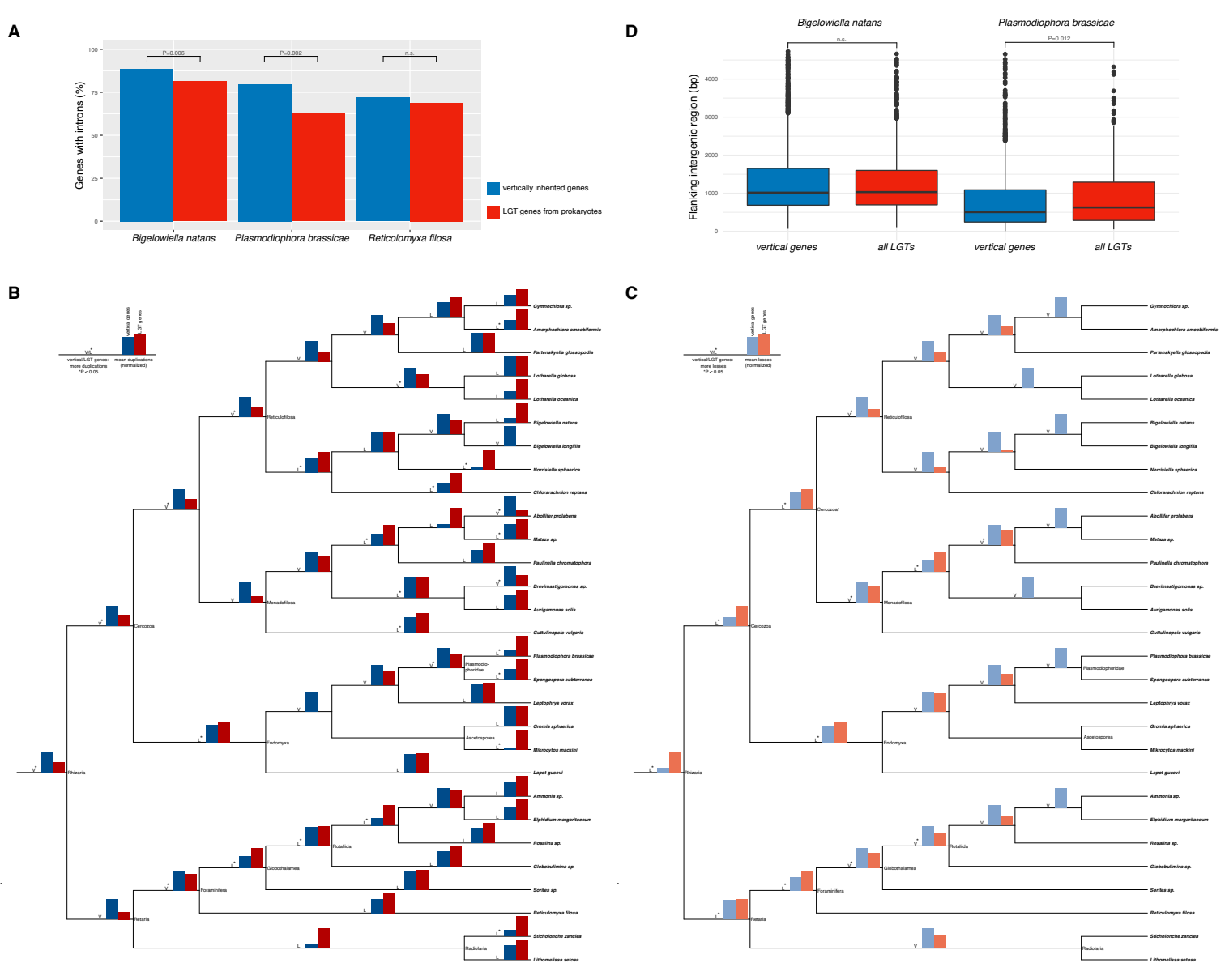
70. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272.

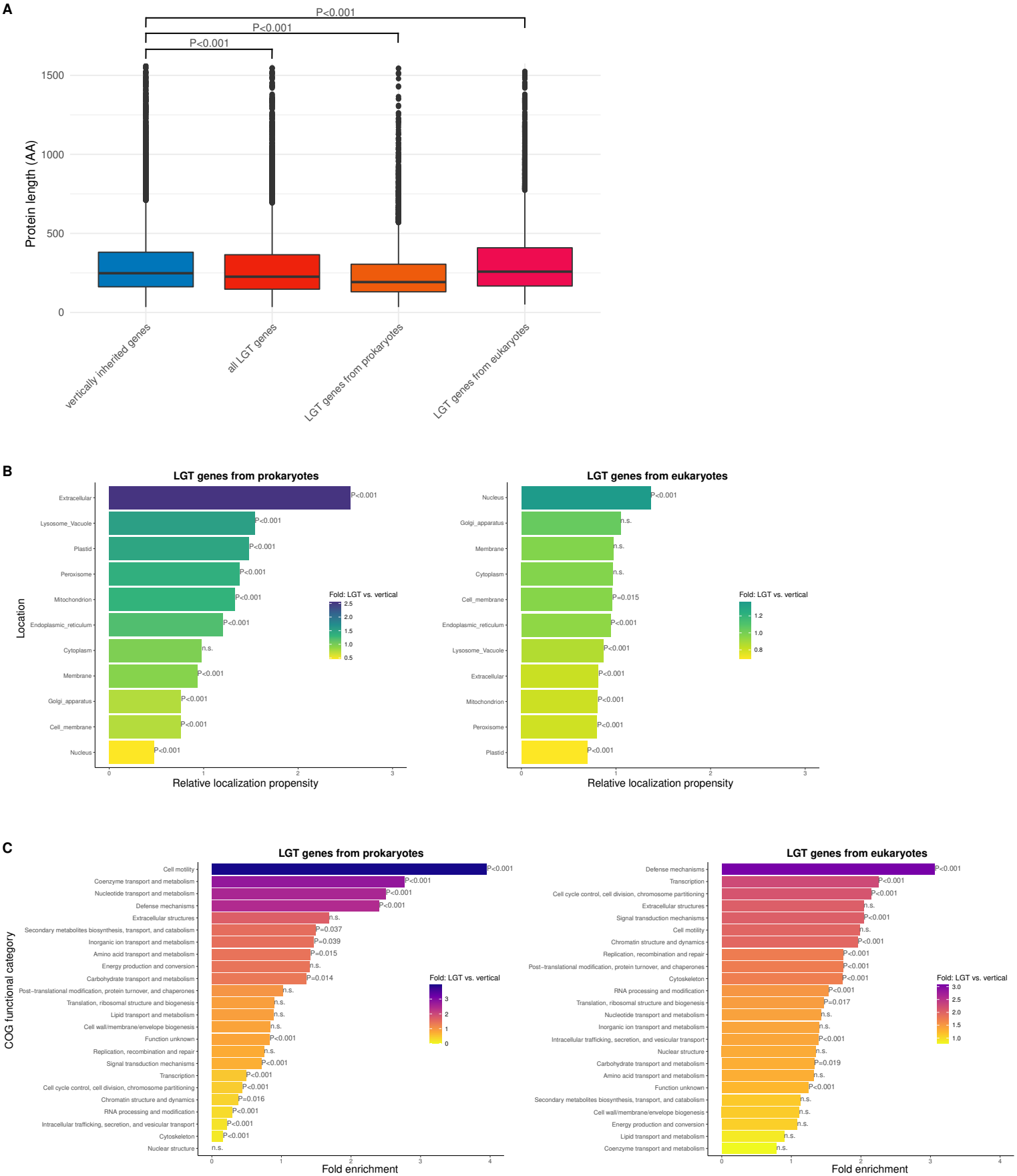
71. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference*. SciPy; 2010.

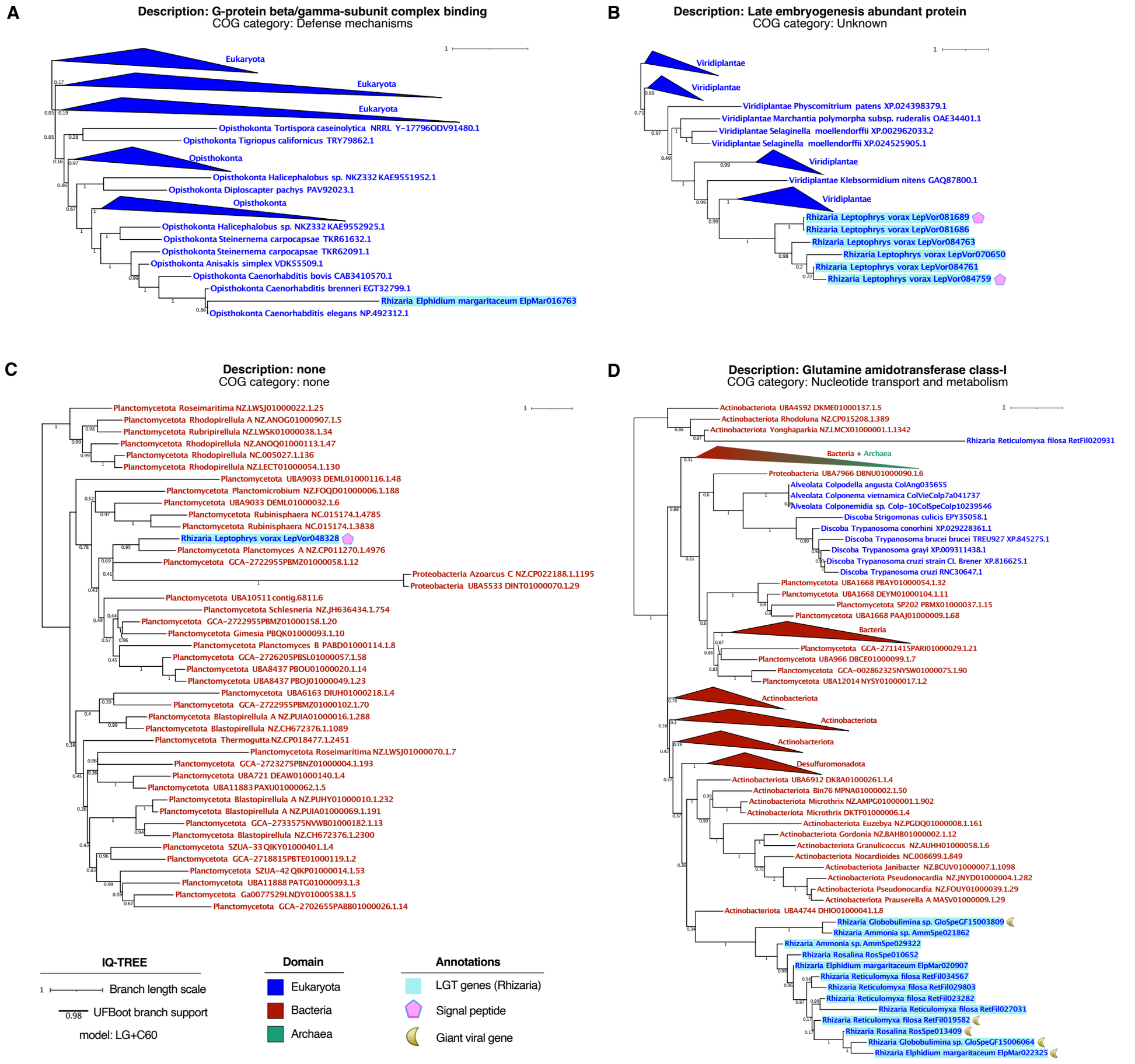
72. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. 2017;8(1):28-36.

73. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49(W1):W293-W296.


duplications
LGTs
inventions
76
LGT number
gene gains (percentual)
Gymnochlora sp.
Amorphochlora amoebiformis
Partenskyella glossopodia
Lotharella globosa
Lotharella oceanica
Bigelowiella natans
Bigelowiella longifila
Norrisiella sphaerica
Chlorarachnion reptans
Abollifer prolabens
Mataza sp.
Paulinella chromatophora
Brevimastigomonas sp.
Aurigamonas solis
Guttulinopsis vulgaris
Plasmodiophora brassicae
Spongospora subterranea
Leptophrys vorax
Gromia sphaerica
Mikrocytos mackini
Lapot gusevi
Ammonia sp.
Elphidium margaritaceum
Rosalina sp.
Globobulimina sp.
Sorites sp.
Reticulomyxa filosa
Sticholonche zanclea
Lithomelissa setosa
Alveolata
Stramenopila
Reticulofilosa
Monadofilosa
Cercozoa
Rhizaria
Endomyxa
Plasmodio-
phoridae
Ascetosporea
Rotaliida
Globothalamea
Foraminifera
Retaria
Radiolaria
Protein provenance
vertical
LGT
invention
Examined proteins (%)
4
8
12
16
20


A
Genes with introns (%)
100
75
50
25
0
P=0.006
P=0.002
n.s.
Bigelowiella natans
Plasmodiophora brassicae
Reticulomyxa filosa
vertically inherited genes
LGT genes from prokaryotes
D
Bigelowiella natans
Plasmodiophora brassicae
n.s.
P=0.012
Flanking intergenic region (bp)
4000
3000
2000
1000
0
vertical genes
all LGTs
vertical genes
all LGTs
B
vertical genes
LGT genes
vertical/LGT genes: more duplications *P < 0.05
mean duplications (normalized)
Gymnochlora sp.
Amorphochlora amoebiformis
Partenskyella glossopodia
Lotharella globosa
Lotharella oceanica
Bigelowiella natans
Bigelowiella longifila
Norrisiella sphaerica
Chlorarachnion reptans
Abollifer prolabens
Metaxa sp.
Paulinella chromatophora
Brevimastigomonas sp.
Aurigamonas solis
Guttulinopsis vulgaris
Plasmodiophora brassicae
Spongospora subterranea
Leptophrya vorax
Gromia sphaerica
Mikrocytos mackini
Lapot gusevi
Ammonia sp.
Elphidium margaritaceum
Rosalina sp.
Globobulimina sp.
Sorites sp.
Reticulomyxa filosa
Sticholonche zanclea
Lithomelissa setosa
Reticulofilosa
Monadofilosa
Cercozoa
Plasmodiophoridae
Ascetosporea
Endomyxa
Rhizaria
Rotalida
Globothalamea
Foraminifera
Retaria
Radiolaria
C
vertical genes
LGT genes
vertical/LGT genes: more losses *P < 0.05
mean losses (normalized)
Cercozoa1


A
P<0.001
P<0.001
P<0.001
Protein length (AA)
1500
1000
500
0
vertically inherited genes
all LGT genes
LGT genes from prokaryotes
LGT genes from eukaryotes
B
LGT genes from prokaryotes
Extracellular
Lysosome_Vacuole
Plastid
Peroxisome
Mitochondrion
Endoplasmic_reticulum
Cytoplasm
Membrane
Golgi_apparatus
Cell_membrane
Nucleus
Location
Relative localization propensity
Fold: LGT vs. vertical
LGT genes from eukaryotes
Nucleus
Golgi_apparatus
Membrane
Cytoplasm
Cell_membrane
Endoplasmic_reticulum
Lysosome_Vacuole
Extracellular
Mitochondrion
Peroxisome
Plastid
Relative localization propensity
Fold: LGT vs. vertical
C
LGT genes from prokaryotes
Cell motility
Coenzyme transport and metabolism
Nucleotide transport and metabolism
Defense mechanisms
Extracellular structures
Secondary metabolites biosynthesis, transport, and catabolism
Inorganic ion transport and metabolism
Amino acid transport and metabolism
Energy production and conversion
Carbohydrate transport and metabolism
Post-translational modification, protein turnover, and chaperones
Translation, ribosomal structure and biogenesis
Lipid transport and metabolism
Cell wall/membrane/envelope biogenesis
Function unknown
Replication, recombination and repair
Signal transduction mechanisms
Transcription
Cell cycle control, cell division, chromosome partitioning
Chromatin structure and dynamics
RNA processing and modification
Intracellular trafficking, secretion, and vesicular transport
Cytoskeleton
Nuclear structure
COG functional category
Fold enrichment
Fold: LGT vs. vertical
LGT genes from eukaryotes
Defense mechanisms
Transcription
Cell cycle control, cell division, chromosome partitioning
Extracellular structures
Signal transduction mechanisms
Cell motility
Chromatin structure and dynamics
Replication, recombination and repair
Post-translational modification, protein turnover, and chaperones
Cytoskeleton
RNA processing and modification
Translation, ribosomal structure and biogenesis
Nucleotide transport and metabolism
Inorganic ion transport and metabolism
Intracellular trafficking, secretion, and vesicular transport
Nuclear structure
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Function unknown
Secondary metabolites biosynthesis, transport, and catabolism
Cell wall/membrane/envelope biogenesis
Energy production and conversion
Lipid transport and metabolism
Coenzyme transport and metabolism
Fold enrichment
Fold: LGT vs. vertical


A
Description: G-protein beta/gamma-subunit complex binding
COG category: Defense mechanisms
Eukaryota
Eukaryota
Eukaryota
Opisthokonta Tortispora caseinolytica NRRL Y-17796ODV91480.1
Opisthokonta Tigriopus californicus TRY79862.1
Opisthokonta
Opisthokonta Halicephalobus sp. NKZ332 KAE9551952.1
Opisthokonta Diploscapter pachys PAV92023.1
Opisthokonta
Opisthokonta Halicephalobus sp. NKZ332 KAE9552925.1
Opisthokonta Steinernema carpocapsae TKR61632.1
Opisthokonta Steinernema carpocapsae TKR62091.1
Opisthokonta Anisakis simplex VDK55509.1
Opisthokonta Caenorhabditis bovis CAB3410570.1
Opisthokonta Caenorhabditis brenneri EGT32799.1
Rhizaria Elphidium margaritaceum ElpMar016763
Opisthokonta Caenorhabditis elegans NP.492312.1
B
Description: Late embryogenesis abundant protein
COG category: Unknown
Viridiplantae
Viridiplantae
Viridiplantae Physcomitrium patens XP.024398379.1
Viridiplantae Marchantia polymorpha subsp. ruderalis OAE34401.1
Viridiplantae Selaginella moellendorffii XP.002962033.2
Viridiplantae Selaginella moellendorffii XP.024525905.1
Viridiplantae
Viridiplantae Klebsormidium nitens GAQ87800.1
Viridiplantae
Rhizaria Leptophrys vorax LepVor081689
Rhizaria Leptophrys vorax LepVor081686
Rhizaria Leptophrys vorax LepVor084763
Rhizaria Leptophrys vorax LepVor070650
Rhizaria Leptophrys vorax LepVor084761
Rhizaria Leptophrys vorax LepVor084759
C
Description: none
COG category: none
Planctomycetota Roseimaritima NZ.LWSJ01000022.1.25
Planctomycetota Rhodopirellula A NZ.ANOG01000907.1.5
Planctomycetota Rubripirellula NZ.LWSK01000038.1.34
Planctomycetota Rhodopirellula NZ.ANOQ01000113.1.47
Planctomycetota Rhodopirellula NC.005027.1.136
Planctomycetota Rhodopirellula NZ.LECT01000054.1.130
Planctomycetota UBA9033 DEML01000116.1.48
Planctomycetota Planctomicrobium NZ.FOQD01000006.1.188
Planctomycetota UBA9033 DEML01000032.1.6
Planctomycetota Rubinisphaera NC.015174.1.4785
Planctomycetota Rubinisphaera NC.015174.1.3838
Rhizaria Leptophrys vorax LepVor048328
Planctomycetota Planctomyces A NZ.CP011270.1.4976
Planctomycetota GCA-2722955PBMZ01000058.1.12
Proteobacteria Azoarcus C NZ.CP022188.1.1195
Proteobacteria UBA5533 DINT01000070.1.29
Planctomycetota UBA10511 contig.6811.6
Planctomycetota Schlesneria NZ.JH636434.1.754
Planctomycetota GCA-2722955PBMZ01000158.1.20
Planctomycetota Gimesia PBQK01000093.1.10
Planctomycetota Planctomyces B PABD01000114.1.8
Planctomycetota GCA-2726205PBSL01000057.1.58
Planctomycetota UBA8437 PBOU01000020.1.14
Planctomycetota UBA8437 PBOJ01000049.1.23
Planctomycetota UBA6163 DIUH01000218.1.4
Planctomycetota GCA-2722955PBMZ01000102.1.70
Planctomycetota Blastopirellula A NZ.PUIA01000016.1.288
Planctomycetota Blastopirellula NZ.CH672376.1.1089
Planctomycetota Thermogutta NZ.CP018477.1.2451
Planctomycetota Roseimaritima NZ.LWSJ01000070.1.7
Planctomycetota GCA-2723275PBNZ01000004.1.193
Planctomycetota UBA721 DEAW01000140.1.4
Planctomycetota UBA11883 PAXU01000062.1.5
Planctomycetota Blastopirellula A NZ.PUHY01000010.1.232
Planctomycetota Blastopirellula A NZ.PUIA01000069.1.191
Planctomycetota GCA-2733575NVWB01000182.1.13
Planctomycetota Blastopirellula NZ.CH672376.1.2300
Planctomycetota SZUA-33 QIKY01000401.1.4
Planctomycetota GCA-2718815PBTE01000119.1.2
Planctomycetota SZUA-42 QIKP01000014.1.53
Planctomycetota UBA11888 PATG01000093.1.3
Planctomycetota Ga0077529LNDY01000538.1.5
Planctomycetota GCA-2702655PABB01000026.1.14
D
Description: Glutamine amidotransferase class-I
COG category: Nucleotide transport and metabolism
Actinobacteriota UBA4592 DKME01000137.1.5
Actinobacteriota Rhodoluna NZ.CP015208.1.389
Actinobacteriota Yonghaparkia NZ.LMCX01000001.1.1342
Rhizaria Reticulomyxa filosa RetFil020931
Bacteria + Archaea
Proteobacteria UBA7966 DBNU01000090.1.6
Alveolata Colpodella angusta ColAng035655
Alveolata Colponema vietnamica ColVieColp7a041737
Alveolata Colponemidia sp. Colp-10ColSpeColp10239546
Discoba Strigomonas culicis EPY35058.1
Discoba Trypanosoma conorhini XP.029228361.1
Discoba Trypanosoma brucei brucei TREU927 XP.845275.1
Discoba Trypanosoma grayi XP.009311438.1
Discoba Trypanosoma cruzi strain CL Brener XP.816625.1
Discoba Trypanosoma cruzi RNC30647.1
Planctomycetota UBA1668 PBAY01000054.1.32
Planctomycetota UBA1668 DEYM01000104.1.11
Planctomycetota SP202 PBMX01000037.1.15
Planctomycetota UBA1668 PAAJ01000009.1.68
Bacteria
Planctomycetota GCA-2711415PARI01000029.1.21
Planctomycetota UBA966 DBCE01000099.1.7
Planctomycetota GCA-002862325NYSW01000075.1.90
Planctomycetota UBA12014 NYSY01000017.1.2
Actinobacteriota
Actinobacteriota
Actinobacteriota
Desulfuromonadota
Actinobacteriota UBA6912 DKBA01000261.1.4
Actinobacteriota Bin76 MPNA01000002.1.50
Actinobacteriota Microthrix NZ.AMPG01000001.1.902
Actinobacteriota Microthrix DKTF01000006.1.4
Actinobacteriota Euzebya NZ.PGDQ01000008.1.161
Actinobacteriota Gordonia NZ.BAHB01000002.1.12
Actinobacteriota Granulicoccus NZ.AUHH01000058.1.6
Actinobacteriota Nocardioides NC.008699.1.849
Actinobacteriota Janibacter NZ.BCUV01000007.1.1098
Actinobacteriota Pseudonocardia NZ.JNYD01000004.1.282
Actinobacteriota Pseudonocardia NZ.FOUY01000039.1.29
Actinobacteriota Prauserella A MASV01000009.1.29
Actinobacteriota UBA4744 DHIO01000041.1.8
Rhizaria Globobulimina sp. GloSpeGF15003809
Rhizaria Ammonia sp. AmmSpe021862
Rhizaria Ammonia sp. AmmSpe029322
Rhizaria Rosalina RosSpe010652
Rhizaria Elphidium margaritaceum ElpMar020907
Rhizaria Reticulomyxa filosa RetFil034567
Rhizaria Reticulomyxa filosa RetFil029803
Rhizaria Reticulomyxa filosa RetFil023282
Rhizaria Reticulomyxa filosa RetFil027031
Rhizaria Reticulomyxa filosa RetFil019582
Rhizaria Rosalina RosSpe013409
Rhizaria Globobulimina sp. GloSpeGF15006064
Rhizaria Elphidium margaritaceum ElpMar022325
IQ-TREE
Branch length scale
UFBoot branch support
model: LG+C60
Domain
Eukaryota
Bacteria
Archaea
Annotations
LGT genes (Rhizaria)
Signal peptide
Giant viral gene