RESEARCH

Open Access

Metatranscriptomes-based sequence similarity networks uncover genetic signatures within parasitic freshwater microbial eukaryotes

Arthur Monjot[1*], Jérémy Rousseau[2], Lucie Bittner[2,3] and Cécile Lepère[1*]

Abstract

Background Microbial eukaryotes play a crucial role in biochemical cycles and aquatic trophic food webs. Their taxonomic and functional diversity are increasingly well described due to recent advances in sequencing technologies. However, the vast amount of data produced by -omics approaches require data-driven methodologies to make predictions about these microorganisms' role within ecosystems. Using metatranscriptomics data, we employed a sequence similarity network-based approach to explore the metabolic specificities of microbial eukaryotes with different trophic modes in a freshwater ecosystem (Lake Pavin, France).

Results A total of 2,165,106 proteins were clustered in connected components enabling analysis of a great number of sequences without any references in public databases. This approach coupled with the use of an in-house trophic modes database improved the number of proteins considered by 42%. Our study confirmed the versatility of mixotrophic metabolisms with a large number of shared protein families among mixotrophic and phototrophic microorganisms as well as mixotrophic and heterotrophic microorganisms. Genetic similarities in proteins of saprotrophs and parasites also suggest that fungi-like organisms from Lake Pavin, such as Chytridiomycota and Oomycetes, exhibit a wide range of lifestyles, influenced by their degree of dependence on a host. This plasticity may occur at a fine taxonomic level (e.g., species level) and likely within a single organism in response to environmental parameters. While we observed a relative functional redundancy of primary metabolisms (e.g., amino acid and carbohydrate metabolism) nearly 130,000 protein families appeared to be trophic mode-specific. We found a particular specificity in obligate parasite-related Specific Protein Clusters, underscoring a high degree of specialization in these organisms.

Conclusions Although no universal marker for parasitism was identified, candidate genes can be proposed at a fine taxonomic scale. We notably provide several protein families that could serve as keys to understanding host-parasite interactions representing pathogenicity factors (e.g., involved in hijacking host resources, or associated with immune evasion mechanisms). All these protein families could offer valuable insights for developing antiparasitic treatments in health and economic contexts.

Keywords Sequence similarity network, Metatranscriptomic, Microbial eukaryotes, Freshwater ecosystems, Functional diversity, Parasites

*Correspondence:
Arthur Monjot
arthur.monjot.pro@gmail.com
Cécile Lepère
cecile.lepere@uca.fr
Full list of author information is available at the end of the article

Background

Aquatic ecosystems, such as oceans, lakes, and rivers, host a wide array of microbial eukaryotes that play essential roles in nutrient cycling, energy transfer, and ecosystem functioning. However, the functional diversity and ecological roles of these microbial eukaryotes remain largely unexplored. Traditional methods for studying microbial eukaryotes have limitations in capturing their functional potential and interactions within complex ecosystems. To overcome these limitations, researchers have turned to the use of omics approaches to predict the diversity and functional composition of microbial communities based on environmental data [1–4]. Metatranscriptomics is one of the best available approaches for acquiring extensive genetic and functional information from uncultured organisms isolated from the environment. While freshwater microbial eukaryotes have been rarely studied using RNA seq analysis [5–7], numerous metatranscriptomic studies have been conducted on marine microbial communities across spatial and temporal scales. These studies have led to significant advancements in our understanding of the physiology [8–12], nutritional modes [13, 14], and contributions to ocean biogeochemistry [15, 16] of these microbial eukaryotes.

As most current analyses depend on a species or function name, a considerable amount of newly generated sequences in environmental samples are overlooked. Microbial eukaryotes have indeed diverse and complex genomes, and this vast genes reservoir is for 40 to 60% without any match in functional databases [15, 17–19].

Current analytical approaches [20–24] generally do not include this uncharacterized fraction in downstream analyses, constraining their results to conserved pathways and housekeeping functions [21]. Sequence similarity network (SSN)-based approaches are powerful tools for analyzing the large amount of data produced by high-throughput sequencing and allow to study relationships between and within protein families (e.g., [25–33]). By clustering sequences while tuning appropriate similarity and overlaps, these networks allow large-scale omics comparisons in ecological studies, and notably enable the inclusion of functionally unannotated sequences in the global analysis (e.g., [17, 28]). Using transcriptomes or metagenomes, previous SSN-based studies unveiled protein families characteristic of specific organismal traits (e.g., toxicity and symbiosis capability of Dinoflagellata [34]) or environmental conditions [3].

Well studied cellular mechanisms such as photosynthesis, already benefit from markers which have been described from model organisms such as Diatoms, Cryptophytes, and Haptophytes (e.g., gene *psbO* [35]), and which can be used to target phototrophic organisms and their activities directly in the environment. Markers for other trophic modes can be more complicated to establish. However, phagotrophy is being increasingly studied; peptidases, proton pumps, and lysosome enzymes (cathepsin and rhodopsin) have been noted as candidate genes to target for this ecologically important feeding strategy (e.g., in heterotrophic Stramenopiles [36, 37]). Moreover, phagotrophy used by photosynthetic organisms (i.e., phago-mixotrophic organisms) has been the subject of many recent studies using experimental assays as well as in silico prediction models [14, 38, 39]. Although there is no clear marker for parasitism, parasites might be identified through the presence of various proteins involved in pathogenesis. The CRN domain or Crinkler proteins are, for instance, a class of effectors (proteins secreted into a host and modifying their behavior) known solely from parasites [40, 41]. However, the study of parasites is highly biased, suffering from the inherent difficulty in cultivating many species and the influence of an anthropocentric perspective, where the primary focus has been on studying parasites that infect humans or hosts of economic importance [42, 43].

In this study, the metabolic specificities of microbial eukaryotes representing a broad diversity of lifestyles (i.e., strict heterotrophic, photo-osmo-phago-mixotrophic, photo-osmo-mixotrophic, saprotrophic, and parasitic microorganisms) were explored in an understudied freshwater ecosystem using an original Sequence Similarity Network (SSN)-based approach. Metatranscriptomic dataset obtained from the meromictic Lake Pavin (France) was deeply investigated using this SSN-based approach, enabling the study of the abundant but understudied « microbial dark matter » (functionally and/or taxonomically uncharacterized sequences), along with depicting the potential shared genetic signatures of relatively little-known freshwater parasitic microorganisms.

Methods

Figure 1 presents the workflow used in this study.

Study site and dataset description

This study was conducted on a dataset acquired from the pristine meromictic Lake Pavin (Massif Central, France, 45° 29′ 45″ N, 2° 53′ 17″ E) [44]. This lake presents an exceptional opportunity to investigate diverse microbial communities occupying multiple ecological niches within a single ecosystem. It is indeed characterized by two permanently stratified water layers: an upper oxygenated layer (mixolimnion) extending from the surface to 60 m, and an anoxic lower layer (monimolimnion) extending from 60 to 92 m depth. Water sampling was carried out in both zones (i.e., at 9 m and 80 m) by day and night, at four contrasted periods in 2018 (April, June, September, and November) and for two size classes (0.8–10

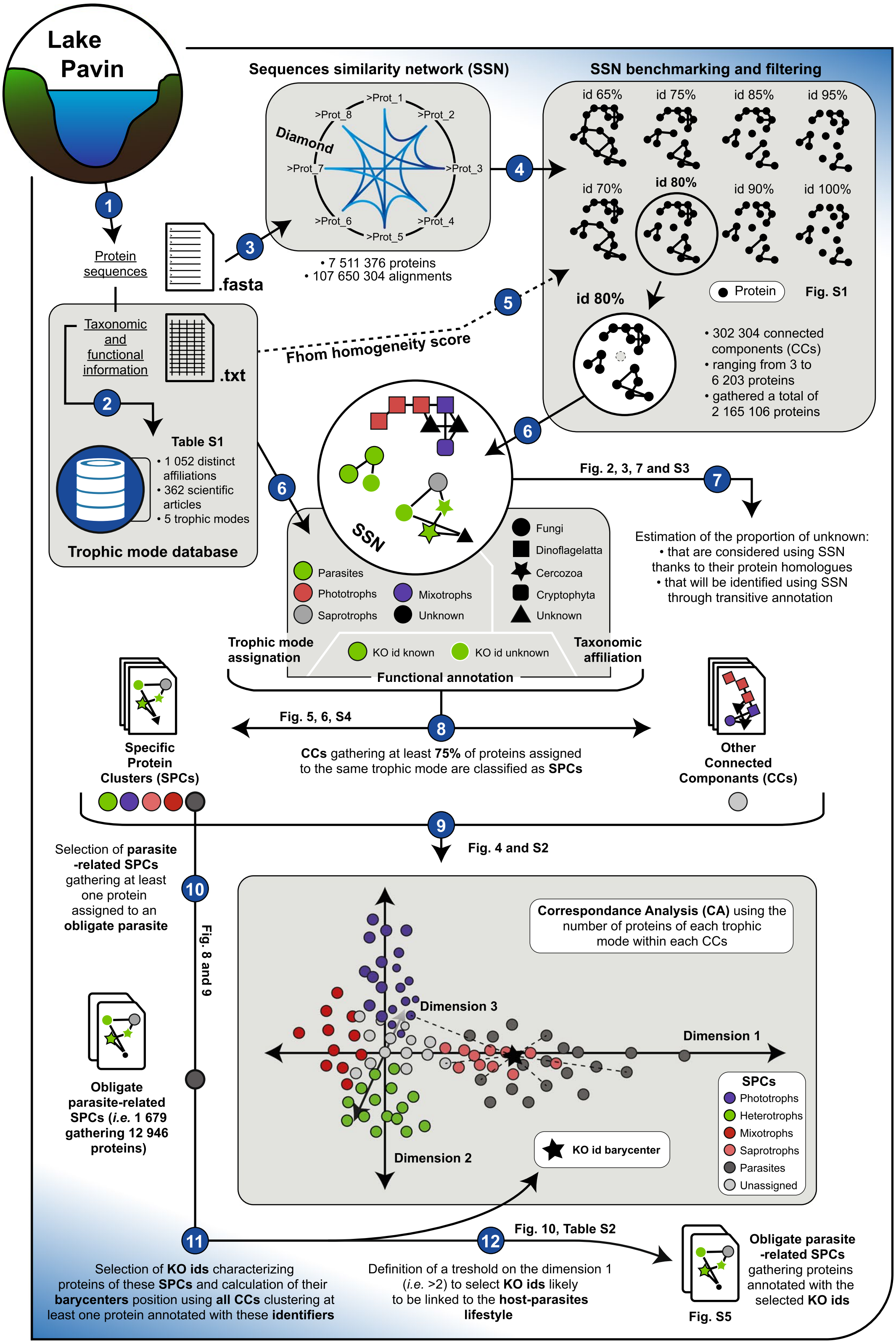
Lake Pavin
1
Protein sequences
.fasta
3
Sequences similarity network (SSN)
Diamond
>Prot_1
>Prot_2
>Prot_3
>Prot_4
>Prot_5
>Prot_6
>Prot_7
>Prot_8
• 7 511 376 proteins
• 107 650 304 alignments
4
SSN benchmarking and filtering
id 65%
id 75%
id 85%
id 95%
id 70%
id 80%
id 90%
id 100%
● Protein
Fig. S1
5
id 80%
• 302 304 connected components (CCs)
• ranging from 3 to 6 203 proteins
• gathered a total of 2 165 106 proteins
Taxonomic and functional information
.txt
Fhom homogeneity score
2
Table S1
• 1 052 distinct affiliations
• 362 scientific articles
• 5 trophic modes
Trophic mode database
6
6
SSN
Fig. 2, 3, 7 and S3
7
Estimation of the proportion of unknown:
• that are considered using SSN thanks to their protein homologues
• that will be identified using SSN through transitive annotation
Parasites
Phototrophs
Saprotrophs
Mixotrophs
Unknown
Fungi
Dinoflagelatta
Cercozoa
Cryptophyta
Unknown
Trophic mode assignation
KO id known
KO id unknown
Taxonomic affiliation
Functional annotation
Fig. 5, 6, S4
8
CCs gathering at least 75% of proteins assigned to the same trophic mode are classified as SPCs
Specific Protein Clusters (SPCs)
Other Connected Componants (CCs)
9
Fig. 4 and S2
Selection of parasite-related SPCs gathering at least one protein assigned to an obligate parasite
10
Fig. 8 and 9
Obligate parasite-related SPCs (i.e. 1 679 gathering 12 946 proteins)
Correspondance Analysis (CA) using the number of proteins of each trophic mode within each CCs
Dimension 3
Dimension 1
Dimension 2
KO id barycenter
SPCs
Phototrophs
Heterotrophs
Mixotrophs
Saprotrophs
Parasites
Unassigned
11
Selection of KO ids characterizing proteins of these SPCs and calculation of their barycenters position using all CCs clustering at least one protein annotated with these identifiers
12
Fig. 10, Table S2
Definition of a treshold on the dimension 1 (i.e. >2) to select KO ids likely to be linked to the host-parasites lifestyle
Obligate parasite -related SPCs gathering proteins annotated with the selected KO ids
Fig. S5

Fig. 1 Workflow of the sequence similarity network approach

and 10–50 µm), resulting in 32 samples as described in Monjot et al. [44]. The dataset analyzed in this study was obtained by Illumina NovaSeq 6000 (2 × 150 bp) sequencing (Illumina, San Diego, CA) and is publicly available under ENA accession number PRJEB61515.

Unigene catalog, protein prediction, and annotations

The metatranscriptome-derived unigene catalog was obtained as described in Carradec et al. [15] and Monjot et al. [44]. Briefly, paired-end reads from each metatranscriptomic sample were assembled using velvet (v1.2.07) with a kmer size of 89 as described in Carradec et al. [15]. Isoform detection was performed using Oases (v0.2.08). Contigs smaller than 150 bp were removed from further analysis. Contig redundancy was removed using CD-HIT-EST (v4.6.1), with the following parameters: *-id 95 -aS 90* (95% nucleic identity over 90% of the length of the smallest sequence) as described in Carradec et al. [15]. For each cluster of contigs, the longest sequence was kept as reference for the unigenes catalog.

Proteins were predicted from all unigenes with *Transdecoder.LongOrfs* followed by *TransDecoder.Predict* (v5.5.0) using the default parameters. Then, unigenes without predicted protein were used for a second run with a minimum protein length of 70 *(-m)*. Finally, the predicted proteins were tested against the AntiFam database (v7.0) [45] with *hmmsearch* using the *–cut_ga* parameter [46].

The KEGG Orthology (KO) identifiers were assigned by KoFamScan (v1.3.0) with the KO's HMM profiles (2022–01-03 release) [47]. For proteins without significant hit, the best hit with an e-value < 1e − 5 was retained as described in Hu et al. [13].

Taxonomic affiliation was performed on proteins with the MMseqs2 suite (v407b315) [48], against the MetaEuk database [49] using mmseqs taxonomy and the parameters *–tax-lineage 1 –lca-mode 2 –max-seqs 100 -e 0.00001 -s 6 –max-accept 100.* The dataset was cleaned of contaminants by excluding proteins affiliated to Bacteria, Archaea, and Viruses.

An in-house trophic mode association table was developed using 362 scientific articles. This table links 1052 distinct taxonomic affiliations to five main trophic modes: phototrophic (considered as photo-osmo-mixotrophic, capable of mixotrophy by photosynthesis coupled to osmotrophy), mixotrophic (photo-osmo-phago-mixotrophic, capable of mixotrophy by photosynthesis coupled to osmotrophy and phagotrophy), heterotrophic (dependent on organic matter from other organisms as a source of nutrients), saprotrophic (dependent on dead or decomposing organic matter as a source of nutrients), and parasitic (facultative and obligate, living in association with and at the expense of one or more hosts, partially or throughout their lives) (Table S1). Proteins affiliated to organisms referenced in literature as multicellular (i.e., Metazoan, some Basidiomycota and Ascomycota, Rhodophyta, Magnoliopsida, Pinopsida, and Polypodiopsida) were removed (Table S1).

SSN building

A sequence similarity network (SSN) where vertices correspond to sequences and edges represent the similarity and coverage between pairs of sequences was built. Diamond (v2.1.7) [50] was used in *blastp* mode to compute the percentage of similarity between every pair of proteins detected in the metatranscriptomic dataset, with the options *-e 1e-5 –sensitive.* Diamond output was filtered using 80% identity and 80% coverage threshold. This coverage threshold is commonly used in SSN-based studies [34] and identity threshold is usually determined on the basis of maximum functional homogeneity between linked proteins [3]. The filtered output was used to build SSN with the igraph-python library (v0.10.4) [51]. The resulted SSN gathered singleton (i.e., vertices without any homology with other sequences) and CCs (connected components: subgraphs composed of at least two vertices disconnected from the rest of the network). Only CCs characterized by at least three vertices were kept for the rest of the analysis. The resulting SSN was finally composed of 302,304 CCs, including 2,165,106 proteins (5,317,694 proteins clustered in CCs characterized by less than 3 proteins were excluded from the analysis).

CC analysis

Statistical and CC analysis were performed using R (v4.3.3) [52]. Each CC was characterized by its protein sequences, which were taxonomically affiliated, trophically assigned, and functionally annotated against KEGG database. Using these annotations, functional homogeneity of each CC was assessed by computing a homogeneity score (Fhom) as described in Faure et al. [3]. The means of Fhom scores as well as CC number and their annotations were compared for eight similarity thresholds (i.e., 65%, 70%, 75%, 80%, 85%, 90%, 95%, and 100%) (Fig. S1). The intermediary 80% identity threshold was selected to maximize the functional homogeneity between linked proteins relative to the total number of CCs while minimizing the number of functionally unannotated components.

CCs were linked to a trophic mode when a majority of their proteins was associated to the same trophic mode. Considering potential taxonomic affiliation errors leading to mis-association of proteins, the proportion of 75% among the total assigned proteins of the component was chosen. These specific components were called SPCs (Specific Protein Clusters) hereafter.

Correspondence analysis (CA) was realized to describe the relationship between CCs and trophic modes with the FactoMineR package (v2.6) [53] and using the number of proteins associated to each trophic mode for each CC. Wilcoxon signed-rank paired test was realized to measure the significance of the difference between trophic modes using the number of proteins within each CC.

CCs which clustered both unknown (trophically, functionally, and/or taxonomically uncharacterized) and known proteins characterized by the same functional annotation, taxonomic affiliation, and trophic mode assignment were used to estimate the number of unknown proteins that could be by extension labeled with the same information (trophic, functional, or taxonomic).

SPCs related to parasitism

This study focused on the SPCs related to parasitism. Among them, SPCs which contained at least one protein associated to an obligate parasite were kept in order to select proteins potentially involved in host-parasite relationships. Barycenter positions on the CA were computed for each KO id (KEGG Orthology identifier)-related to metabolic pathway ($\text{KEGG}_{\text{PATHWAY}}$ = Metabolism) detected among obligate parasite-related SPCs. Briefly, the barycenter positions (x_b, y_b, and z_b) were calculated for each KO id using the positions (x_i, y_i, and z_i) of n CCs (clustering at least one protein annotated with this identifier, among the totality of CCs of the data set) following the formulae: $x_b = \frac{\sum_{i=1}^{n} x_i}{n}; y_b = \frac{\sum_{i=1}^{n} y_i}{n}; z_b = \frac{\sum_{i=1}^{n} z_i}{n}$. We used the first three dimensions which collectively explain 65.7% of the variance (Fig. S2). These dimensions provide the best differentiation between parasites and the other trophic modes. These barycenters therefore represent the gravity centers of each KO id within a tri-dimensional space (i.e., CA). The obligate parasite-related SPCs were graphically represented using ggraph (v2.1.0) [54], tydigraph (v1.2.3) [55], and the igraph R package (v1.5.0) [51]. Upset plot were generated using the ggupset R package (v0.3.0) (https://github.com/const-ae/ggupset).

Results

Sequence similarity network statistics

Almost 10 million transcripts were retrieved from the Lake Pavin metatranscriptomic dataset. From these transcripts, 7,511,376 proteins have been predicted and compared to each other (i.e., 107,650,304 diamond alignments) to build a sequence similarity network (SSN). After filtration steps, 302,304 connected components (CCs) were obtained, ranging from 3 to 6203 proteins (Fig. S3) and gathered a total of 2,165,106 proteins. Among them, 32.5% were taxonomically affiliated and 18.5% were assigned to a trophic mode (703,930 and 401,570 proteins, respectively; Fig. 2). Trophically unassigned proteins were predominantly affiliated to Alveolata, Stramenopiles, Chlorophyta, and Opisthokonta, accounting, for example, for more than half of the Alveolata-affiliated proteins. Among trophically assigned proteins, mixotrophs dominated (n = 156,485) and were mainly affiliated to Alveolata, Euglenozoa, Stramenopiles, Haptista, and Cryptophyta. Phototrophs affiliated to Chlorophyta, Stramenopiles, and Dinoflagellata comprised 136,946 assigned proteins. Heterotrophs (n = 77,955) represented about half the number of proteins assigned to mixotrophs. These heterotrophs included a diverse range of organisms from various groups: Stramenopiles, Alveolata, Opisthokonta, Euglenozoa, Cryptophyta, and Amoebozoa. The proportion of proteins associated with parasites and saprotrophs was low (24,232 and 5952, respectively) and the majority was affiliated to Opisthokonta. Proteins assigned to parasites were also found among Stramenopiles, Amoebozoa, Alveolata, Heterolobosea, Metamonada, and Rhizaria. Facultative parasites mainly affiliated to Opisthokonta, Stramenopiles, and Amoebozoa were represented by 20,152 proteins, while those with an obligatory lifestyle accounted for only 4080 proteins and were found in majority among Alveolata, Rhizaria, Euglenozoa, Opisthokonta, and Stramenopiles (Fig. 2).

CC distribution among trophic modes

More than half of CCs grouped proteins without any trophic assignment (i.e., 162,001 CCs). The remaining CCs clustered proteins (i.e., a total of 1,400,022 sequences) assigned to one (129,151 CCs) and up to the five trophic modes (162 CCs). Eighteen thousand nine hundred sixty CCs were composed solely of trophically assigned proteins, while 121,343 CCs contained both assigned and unassigned proteins (Fig. 3). Among the specific CCs (i.e., all trophically assigned proteins within the component were assigned to the same trophic mode): (i) 52,682 (including 7007 CCs for which all the proteins were assigned) were linked to mixotrophs, (ii) 42,561 (6258) to phototrophs, (iii) 25,389 (3458) to heterotrophs, (iv) 6920 (1203) to parasites, and (v) 1599 (215) to saprotrophs.

CCs linked to multiple trophic modes suggest shared common features. Mixotrophs shared 4171 CCs with phototrophs and 2498 with heterotrophs (exclusively). Parasites shared a significant number of CCs with saprotrophs (i.e., 1103) and heterotrophs (490) (exclusively; Fig. 3).

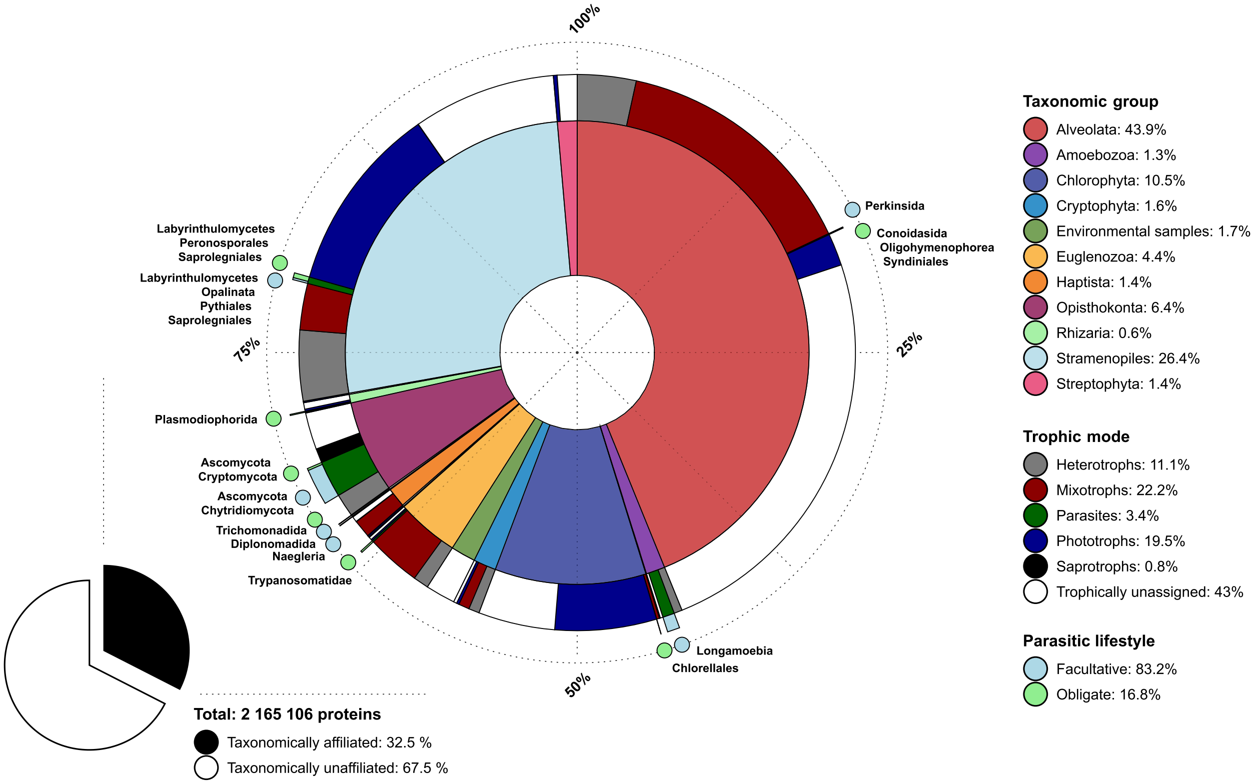

Fig. 2 Taxonomic diversity and trophic assignment of proteins integrated in the sequence similarity network. From inner to outer circles, colors represent taxonomic affiliation, trophic assignment, and parasitic lifestyle. The proportion of each trophic mode is documented for every eukaryotic lineage. Information specific to parasites is also provided (i.e., their lifestyle, which is categorized based on the degree of host dependence: obligate or facultative). The most abundant parasite groups integrated in the SSN are specified for each parasitic lifestyle within each taxonomic group. The term "Environmental samples" is a classification used by the NCBI taxonomy database and refers to sequences that have been obtained directly from the environment, without specific identification of the source organism

Using the number of proteins associated with each trophic mode within each CC, correspondence analysis (CA) showed significant differences (Wilcoxon signed-rank p-value < 0.001) between almost all of trophic modes (with the three first dimensions explaining 65.7% of the variance) (Fig. 4, S2). The first dimension dissociated parasites and saprotrophs from the other trophic modes, while second and third dimensions split phototrophs, heterotrophs, and mixotrophs.

Taxonomic specificities and genetic basis of microbial eukaryotic trophic modes

Mixotrophs, phototrophs, and heterotrophs were represented by numerous Specific Protein Clusters (SPCs). SPCs are Connected Components for which at least 75% of assigned proteins were assigned to the same trophic mode. The study found 53,962 SPCs for mixotrophs (containing 403,174 proteins), 43,291 SPCs for phototrophs (334,474 proteins), and 26,063 SPCs for heterotrophs (181,251 proteins). Nearly half of these protein sequences were identified by their taxonomic classification, as shown in Fig. 5. Although some taxa appeared in more than one of these three trophic modes (e.g., Dinophyceae, Chromulinales), taxonomic diversity greatly differed at low taxonomic levels (< Class) (Fig. 5). Parasites and saprotrophs were characterized by a lower number of SPCs: 7236 and 1647, respectively, grouping only 49,148 and 9034 proteins with an average taxonomic affiliation rate of 46.9%. Proteins clustered in SPCs of both trophic modes mainly belonged to Chytridiomycota and Ascomycota (Opisthokonta: Fungi). Parasites were also characterized by numerous proteins affiliated to Longamoebia (Amoebozoa: Discosea), Saprolegniales, Peronosporales, and Labyrinthulomycetes (Stramenopiles: Oomycetes and Bigyra), Trypanosomatidae (Euglenozoa: Kinetoplastida), and Plasmodiophorida (Rhizaria: Endomyxa).

Functional annotation against the KEGG database allowed the identification of more than half of the proteins clustered in SPCs for phototrophs, heterotrophs, saprotrophs, and parasites (Fig. S4) (i.e., 53.9% of the total proteins were annotated with a KO id (KEGG Orthology identifier)). Mixotrophs were the least annotated mode,

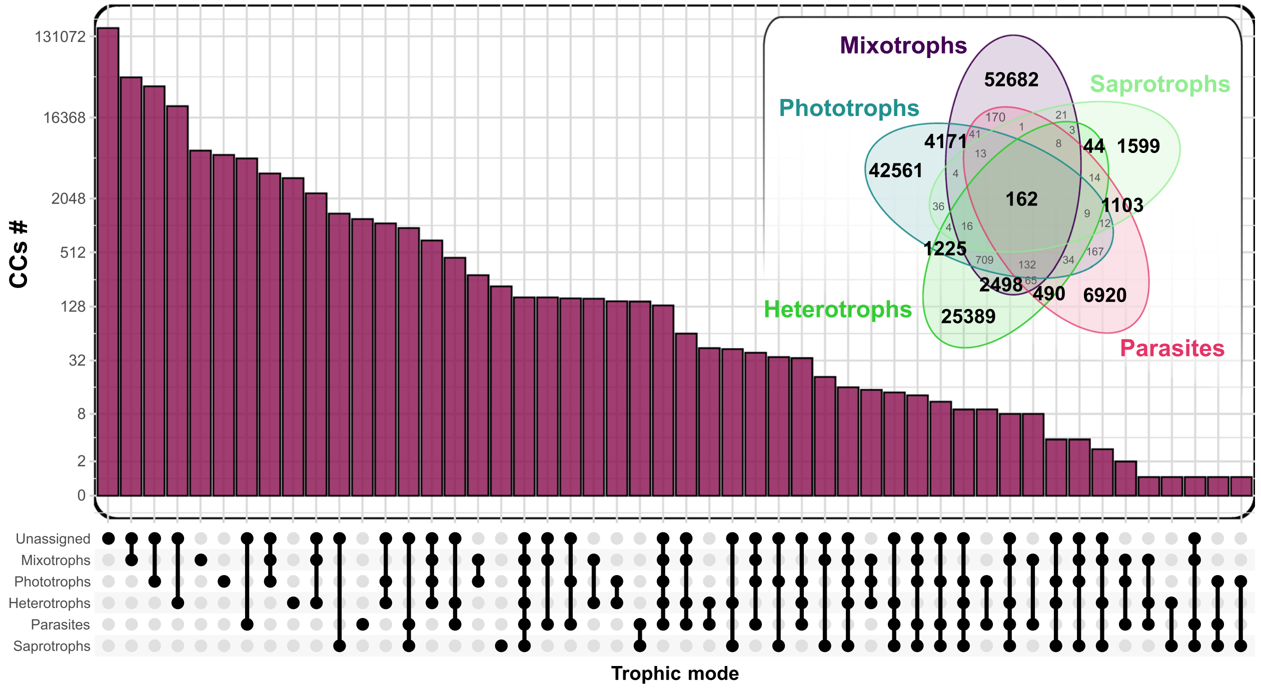

Fig. 3 Upset plot and Venn diagram displaying shared and unique connected components across trophic modes. The occurrence of each combination of proteins trophic mode assignment within CCs are displayed on the bar plot. The *Y* axis is represented on a pseudo-logarithmic scale (base 2)

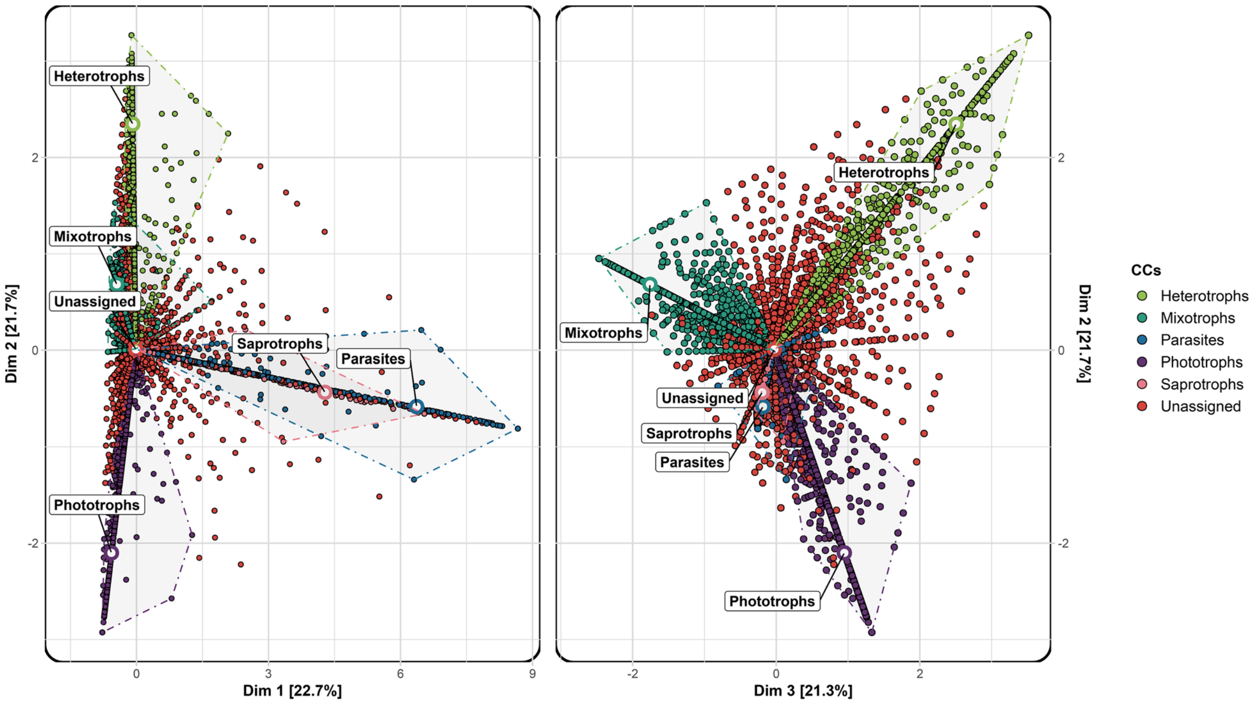

Fig. 4 Correspondence analysis testing the relationship between CCs and trophic modes. This analysis is processed using the number of proteins associated to each trophic mode for each CC. CCs with a 75% specificity to a trophic mode, i.e., SPCs, are labelled with different colors while non-specific components (specified as unassigned) are filled in red

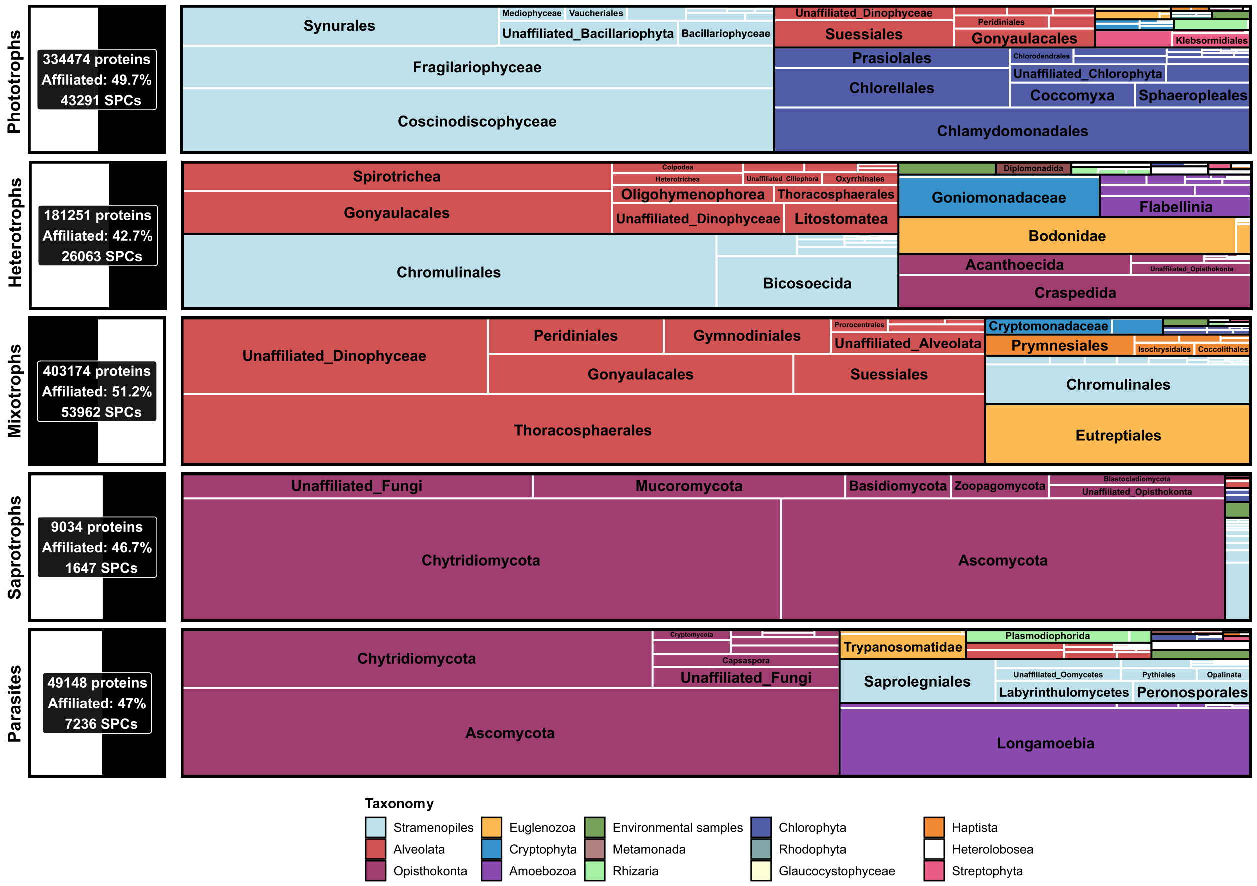

Fig. 5 Taxonomic affiliation of proteins gathered in Specific Proteins Clusters (SPCs) of each trophic mode. Number of SPCs, total number of proteins within SPCs, and affiliation rate are displayed on the horizontal histograms (on the left) for each tropic mode. The taxonomic affiliation of proteins is reported on the right and the size of boxes shows the proportion of proteins taxonomically affiliated to each eukaryotic lineage

with only 46% of functionally annotated proteins. Metabolisms-related proteins accounted for almost a third of annotations, ranging from 11.4% (mixotrophs) to 21.5% (saprotrophs).

Contrary to the high variability of taxonomic affiliations between the SPCs of each trophic mode, their functional annotations varied less considering the B level (i.e., intermediary level in classification of functional pathway in KEGG database) of KEGG metabolism annotations (Fig. S4). Amino acids, carbohydrates, lipids, and energy metabolism were the most common across all trophic modes.

However, more variations between trophic modes were recorded at the C level (Fig. 6). Although heterotrophs and parasites showed an important proportion of proteins related to lipid metabolisms, such as those involved in fatty acid degradation, some specificities were detected: (i) heterotrophs were characterized by numerous proteins involved in sphingolipid and glycerophospholipid metabolism; (ii) parasites seem to allocate a significant portion of their metabolic capacity to the biosynthesis of unsaturated fatty acids and glycerophospholipid metabolism. Saprotrophs SPCs also clustered proteins involved in glycerophospholipid as well as sulfur metabolism and prodigiosin biosynthesis. Finally, a great proportion of proteins linked to photosynthetic organisms (phototrophs and mixotrophs) are involved in photosynthesis (e.g., antenna proteins) and carotenoid biosynthesis (Fig. 6).

Using SSN to consider the unknown

Evaluating homology between proteins obtained from metatranscriptomes enables the inclusion of unknown sequences (i.e., no functional annotation and/or no taxonomic affiliation) in the analysis. As a result, 57.8% of total sequences were taken into account in the analysis (1,252,112 distributed in 124,455 CCs), while only 15.7% (340,482 proteins) were actually functionally annotated and taxonomically affiliated (Fig. 7). Similarly, 1,400,022 sequences (64.6% of the dataset) were linked to trophic mode information while 18.5% were initially labelled (Fig. 7).

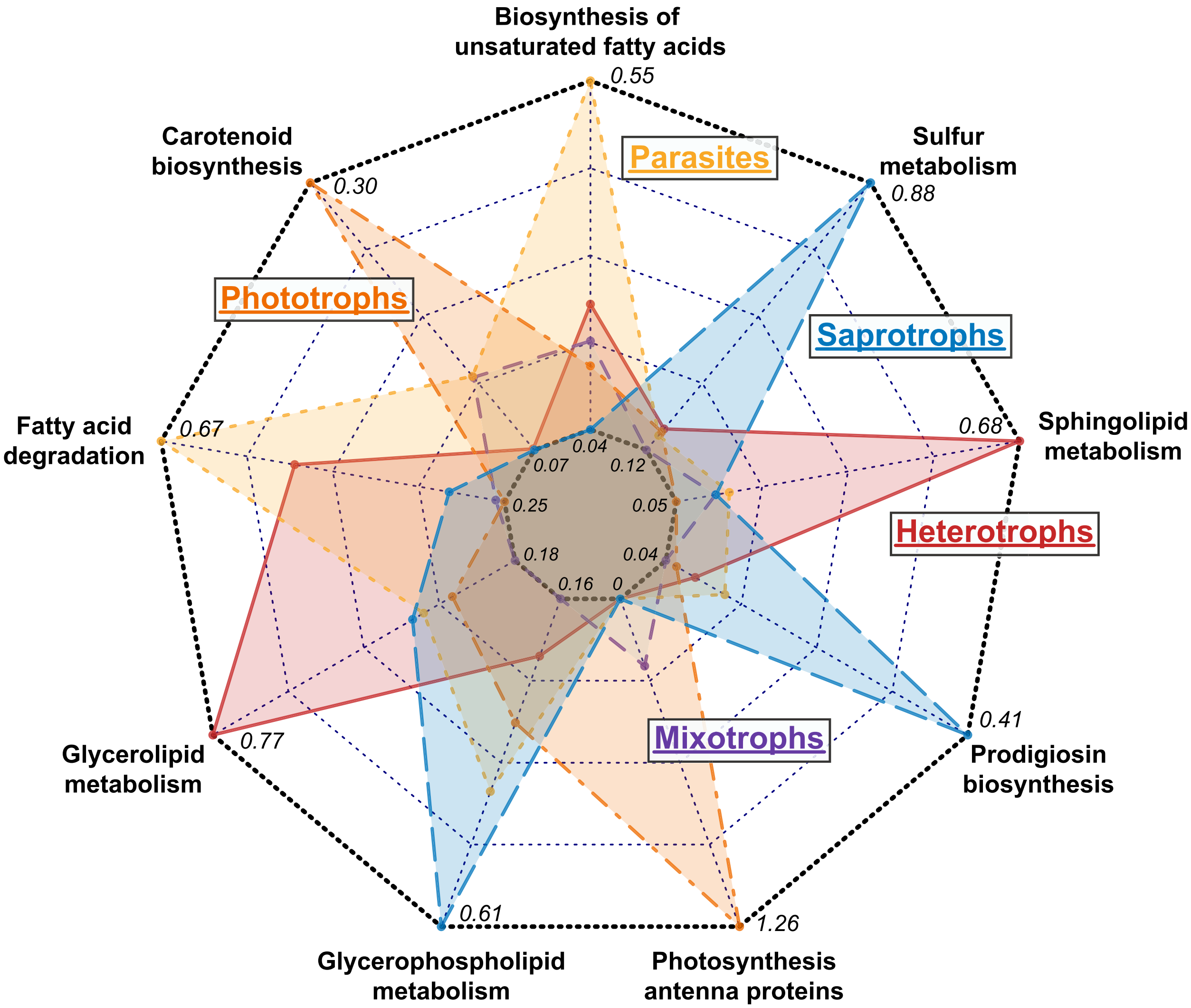

Fig. 6 Proportion of proteins functionally annotated with the most fluctuating KEGG metabolic pathways for each trophic mode. Values at the center are the minimum proportion value for this KEGG category while those at the exterior represent maximum proportion value

Furthermore, assuming that unknown proteins, which grouped with known proteins characterized by the same taxonomic, functional, and trophic information were closely related, we could estimate the number of unknown proteins that can be identified through the use of SSN (i.e., by transitive transfer of annotations). This re-evaluation allowed for an increase in the taxonomic affiliation rate of the proteins by an average of 6% at all taxonomic levels (i.e., from 25.09 to 31.96% (Class/Order), from 29.95 to 36.32% (Phylum) and from 32.51 to 38.74% (Division)). Additionally, the functional annotation improved from 53.91 to 57.38% and the trophic assignment showed an 8% improvement (from 18.54 to 26.18%).

SPCs related to parasitism

Parasite SPCs were reduced to 1679 when selecting those that grouped at least one protein affiliated with an obligate parasite. They clustered 12,946 proteins, of which 35.6% were affiliated (i.e., 4604 proteins). Among them, 62% were assigned to obligate parasites (2850), 9.2% to facultative parasites (including Amoebozoa, Opisthokonta, and Stramenopiles) (425), 1.6% to saprotrophs (e.g., Opisthokonta and Stramenopiles) (72), and

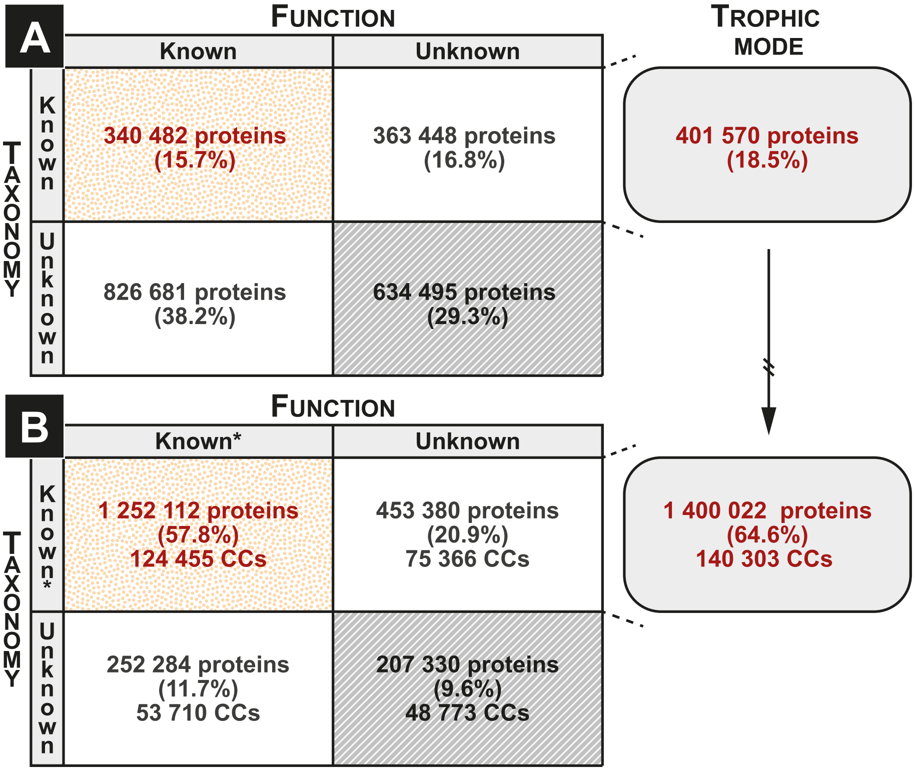

Fig. 7 Taxonomic affiliation, functional annotation, and trophic assignment statistics of proteins using traditional metatranscriptomic approach (**A**) and with the implementation of SSN-based approach (**B**). Statistics of taxonomic affiliation, functional annotation, and trophic mode assignation of proteins are compared with (**B**) or without (**A**) the implementation of SSN-based approach. *: proteins are defined as "Known" when they clustered within the same CCs of proteins that are taxonomically affiliated, functionally annotated, and/or trophically assigned. They are not annotated using transitive transfer of annotation but only considered in the analysis

26.8% remained trophically unassigned (1229) (Fig. 8). This selection also reduced distinct functional annotations from 3268 to 1143 KO ids, which were likely to be involved in host-parasite relationships. Among the selected SPCs, 496 had no protein functional annotation (e.g., SPC n°18 (Fig. 9)), 837 were annotated to a unique KO id (e.g., SPC n°33938), and 346 had multiple annotations (e.g., SPC n°32689).

Although some SPCs grouped proteins with different taxonomic affiliations (at the class level such as the SPCs n°32689 and n°18 grouping respectively Oomycetes and Fungi, or at the phylum level such as the SPC n°33938 which clustered Cercozoa, Discosea, and Fungi) the majority clustered few proteins (≈7.3 proteins/SPC) with similar taxonomic affiliations (e.g., SPCs n°24890 and n°29238 (Fig. 9)).

To identify proteins likely to be involved in host-parasite interactions, barycenter of each KO id characterizing proteins of obligate parasite-related SPCs were computed on the CA (Fig. 10). This strategy enabled to overcome protein filtration bias by considering functionally unannotated proteins as well as proteins clustered in SPCs of other trophic modes. By selecting all KO ids whose barycenter coordinates on axis 1 were among the most important (>2), 40 identifiers likely to be involved in host-parasite interactions were obtained. SPCs containing these identifiers (i.e., $n=74$, Fig. S5) clustered very few proteins (≈6.7 proteins/SPC) characterized by similar taxonomic affiliations, suggesting that there is probably no overall protein family shared by all parasitic microbial eukaryotes.

Among these KO ids (complete list in Table S2), two were linked to uncharacterized proteins in KEGG (i.e., K09795/K09983), 12 were not clearly related to parasitism in literature (e.g., K13514, K10971, K14263), and 26 were linked to host-parasite interaction (12 already described and 14 probably linked but not yet described).

Among those already described, we found proteins affiliated to Trypanosomatidae and involved in the well-known mechanism of immunogenicity and extracellular survival of these microorganisms (i.e., K20656/K12167 [56, 57]). Others affiliated with Fungi were related to resistance against toxic compounds (e.g., K06141, K09043

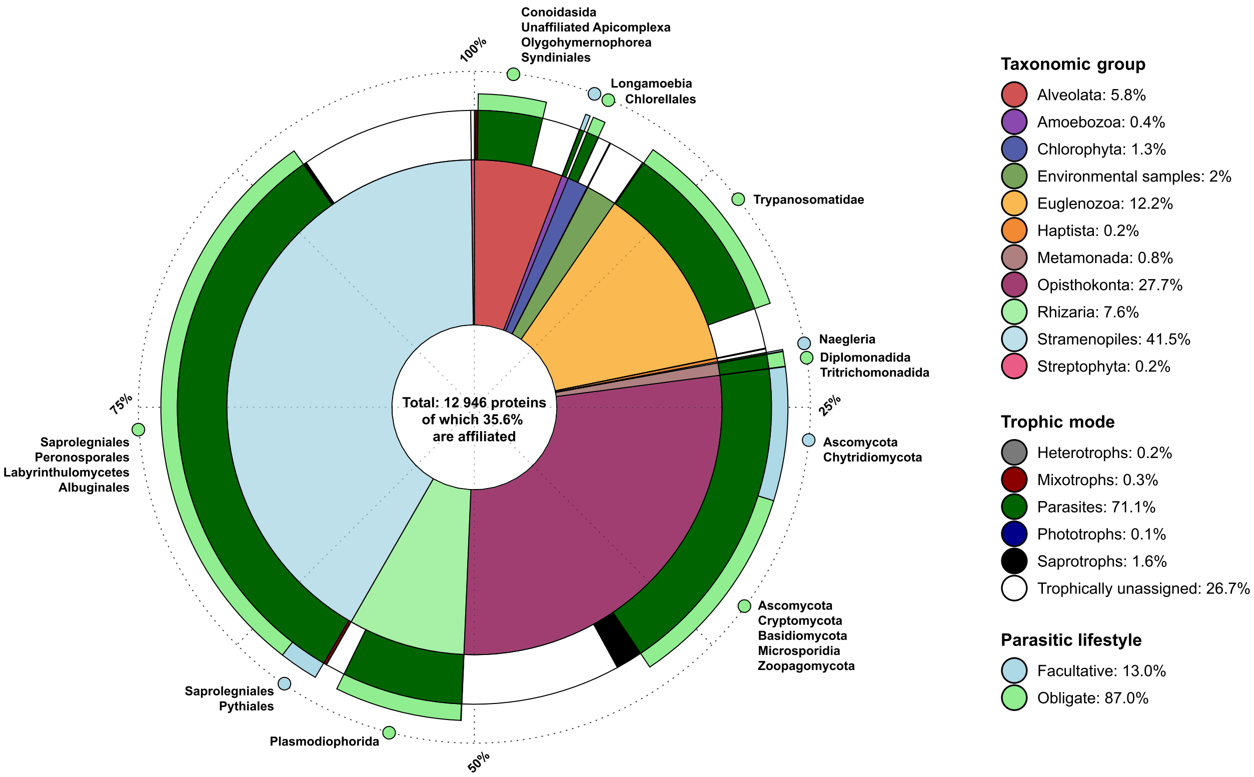

Fig. 8 Taxonomic diversity and trophic assignment of proteins gathered in SPCs characterized by at least one protein assigned to obligate parasite. From inner to outer circles, colors represent taxonomic affiliation, trophic assignment, and parasitic lifestyle. The most abundant parasite groups are specified for each parasitic lifestyle within each taxonomic group

[58, 59]) or described as stimulating parasite metabolism (e.g., K09240, K18278) (Table S2).

Finally, the other KO ids, although not yet described in a host-parasite interaction, were characterized by functions that may contribute to establishing this type of interaction or by features correlated with parasitism. We found, among others, some (i) related to anti-apoptotic agents and affiliated to Plasmodiophorida and Ascomycota (K20637/K17968 [60, 61]), (ii) involved in DNA repair process (K19465, K10791 [62]) with multiple affiliations (Conoidasida, Ascomycota and Oomycetes), (iii) corresponding to genes whose distribution within fungi-like organism diversity (e.g., Oomycetes, Fungi, Labyrinthulomycetes) seems to depend on their trophic capacity (saprotrophic or parasitic) (i.e., K07414, K21200, and K15629 which belong to cytochrome P450 family [63]) (Table S2).

Discussion

Identify the unknown

The proportion of unknown (i.e., taxonomically and/or functionally uncharacterized) in environmental datasets is high when examining functional diversity of eukaryotes and is rapidly increasing due to the deluge of sequences characteristics of the post-genomic era [64, 65]. Recent extensive sequencing initiatives, such as the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP, https://www.ncbi.nlm.nih.gov/bioproject/248394, [1]) and the global ocean atlas of eukaryotic genes retrieved from TARA Ocean expeditions [15], report that an average of almost 50% of sequences remain without related sequences [34, 49]. Estimates of the unknown are even higher when considering the few large-scale metatranscriptomic studies conducted in freshwater ecosystems [5, 7], indicating that 85% of all predicted proteins lack both known functions and assigned taxonomies [6].

By taking into account both known and unknown proteins, our SSN-based workflow enables improvement in the number of sequences analyzed by 42.1%, while traditional approaches would have used only 15.7% of the dataset by excluding proteins without functional annotation and taxonomic affiliation (Fig. 7). Furthermore, it allows consideration of significantly more sequences that were neither functionally annotated nor taxonomically affiliated (decreasing from 29.3 to 9.6%) in accordance with previous SSN-based study which link 15% more proteins to taxonomy or function [3].

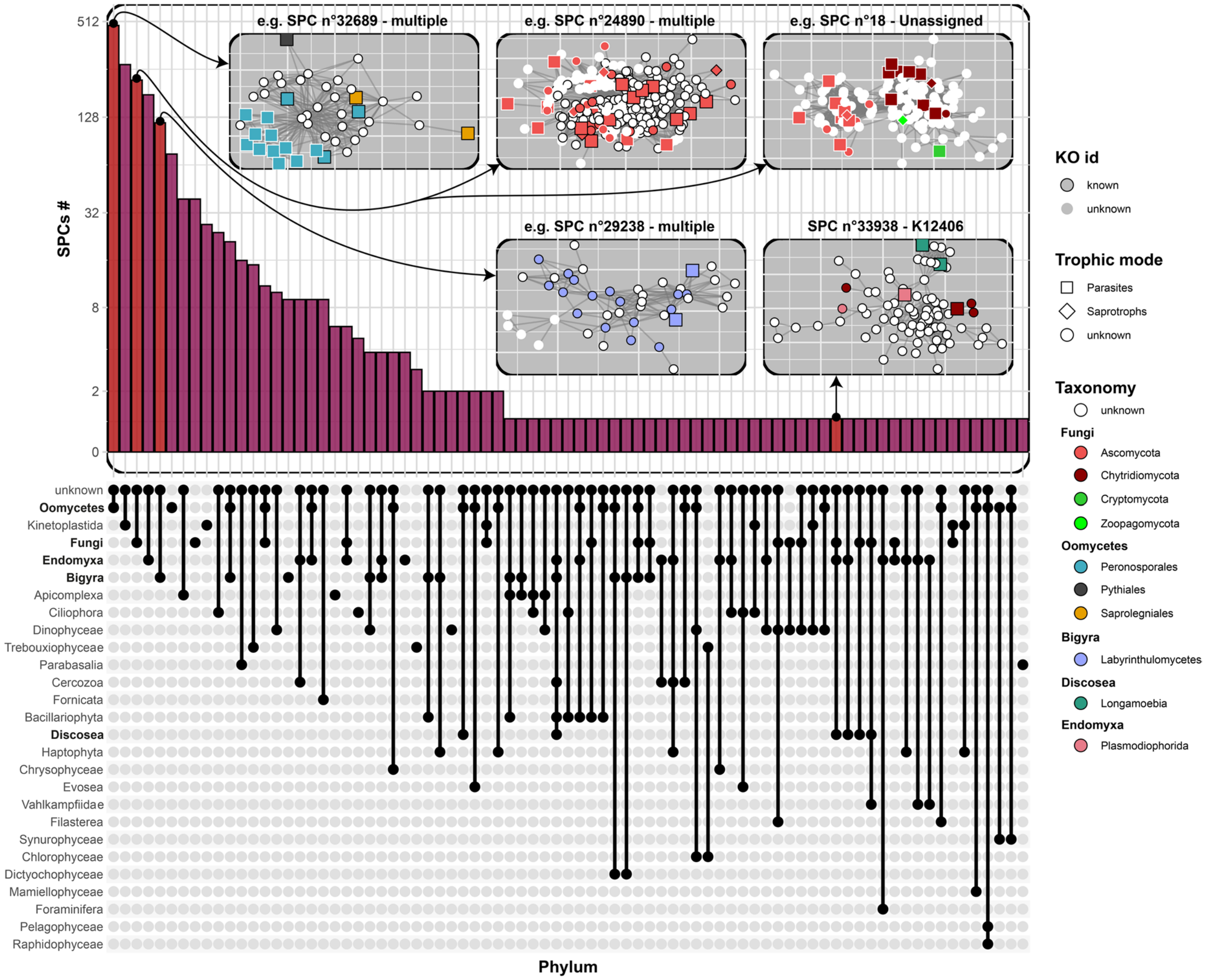

Fig. 9 Upset plot displaying shared and unique obligate parasite-related SPCs across taxonomic phyla. The occurrence of each combination of protein taxonomic affiliation within obligate parasite- related SPCs are displayed on the bar plot. Only parasite-related SPCs which clustered at least one protein assigned to an obligate parasite are considered. The *Y* axis is represented on a pseudo-logarithmic scale (base 2). Examples of SPCs with various taxonomic combination features are represented on the top of the figure by sub-graphs gathering proteins (colored points) which are characterized by taxonomic affiliation, trophic assignment, and functional annotation (KO id). The mention "multiple" means that proteins of the SPC are annotated with multiple KO ids. The identification number of SPCs is referenced on the top of each sub-graph. Black outlined points represent a protein annotated with the KO id referenced at the top of the sub-graph (e.g., SPC n°33938). Bold phyla are those represented within example SPCs

Assuming that functional homogeneity is strictly maintained within connected components (i.e., each CC would group proteins sharing the same functional characteristics), we estimate that an average of 6% of unknown proteins function could be identified by extrapolating information from known proteins. However, the transitive transfer of annotations remains debatable [66, 67] and could be completed by further SSN-based prediction tools to substantially improve the consistency and accuracy of annotations. These tools could include methods on connectedness to refine annotations [68] or the construction of a probabilistic graphical model to assess the relevance of annotations [69].

Functional diversity of microbial eukaryotic trophic modes

Only 162 CCs encompass all trophic modes (≈0.05% of all CCs; Fig. 3) and may be recognized as constituting the core proteome of all eukaryotes. These results align with previous studies; for example, [34] stated

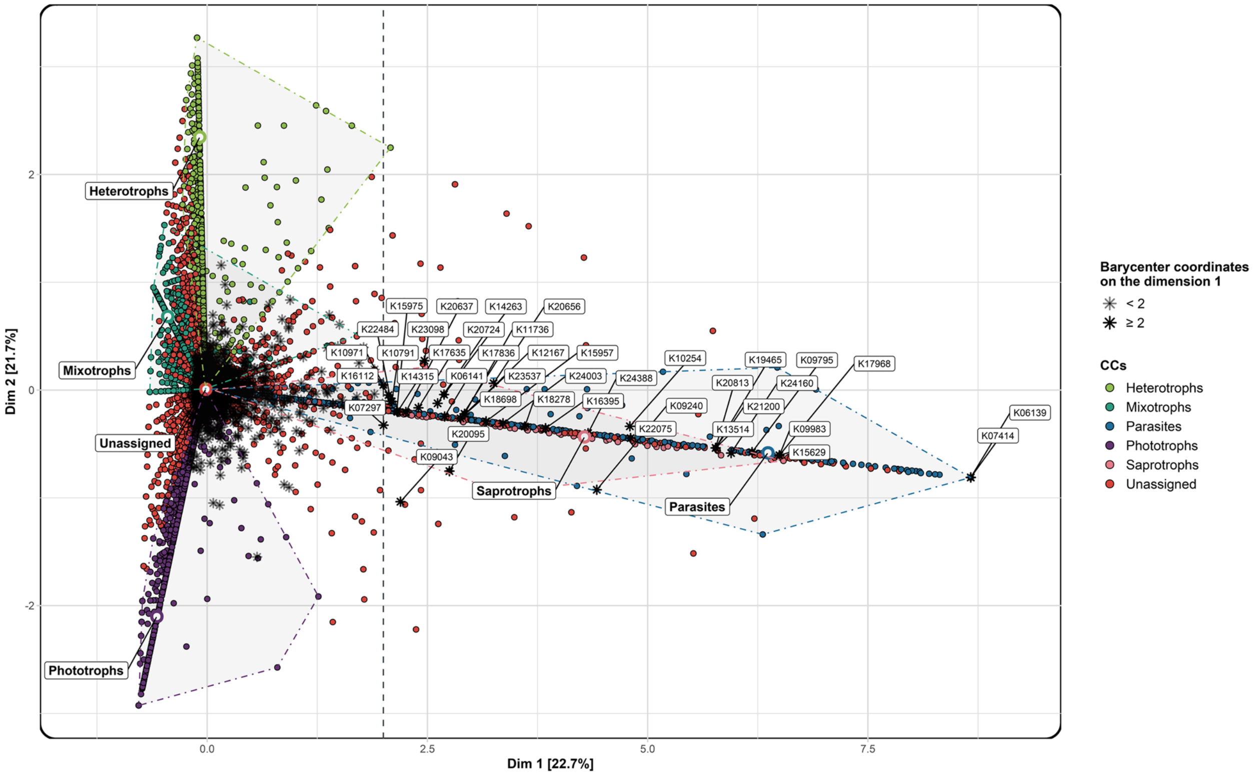

Fig. 10 Correspondence analysis testing the relationship between CCs and trophic modes with the addition of barycenter of KO ids correlated with parasitism. Barycenters representing the gravity centers of each KO id within a tri-dimensional space are represented by stars. Only barycenters of KO ids characterizing SPCs which contained at least one protein associated to an obligate parasite are kept. Those with coordinates on the dimension 1 (dimension discriminating parasite-related SPCs and the others) exceeding 2 highlight the KO ids likely to be involved in host-parasite interactions and are represented by darker stars. SPCs related to each trophic mode are labelled with different colors while unspecific components (specified as unassigned) are filled in red

that the core proteome of Dinoflagellata was composed of 252 CCs (≈0.07% of all CCs), while another study at the eukaryotic domain scale reported 255 single-copy orthologs in at least 90% of 70 species (i.e., BUSCO set for eukaryotes [70, 71]). Although representing a small proportion of the network (3.69%), CCs connecting multiple trophic modes suggest metabolic similarity, as observed among phototrophs, mixotrophs, and heterotrophs. Mixotrophs exhibit many CCs solely shared with phototrophs (i.e., 4171) and heterotrophs (2498) (Fig. 3), in accordance with their metabolic versatility, being able to combine ability of both specialists (e.g., photosynthesis, osmotrophy, and phagotrophy [72, 73]). This finding aligns with a previous study of Lambert et al. [14] which explored shared gene families between mixotrophic, heterotrophic, and phototrophic marine protists. In contrast, heterotrophs and phototrophs have fewer common CCs (i.e., 1225 CCs exclusively grouping these two trophic modes). Parasites and saprotrophs also exhibit proteins that are clustered in common CCs. This clustering may be consistent with their potentially shared metabolic pathways, as well as their closely related taxonomy (e.g., Fungi). Indeed, the lifestyles of Fungi affiliated with Chytridiomycota and Ascomycota in freshwater environments span a large spectrum, ranging from obligate parasite to saprotrophs (Table S1). They also exhibit high variability from one species to another within the same genus [74–76].

The high quantity of Specific Protein Clusters (SPCs) (43.7% of the CCs) suggests that microorganisms of single trophic modes possess numerous specific genes involved in distinct metabolic pathways and/or are affiliated with highly divergent taxonomy. Within these SPCs, an average of 47.5% of proteins are taxonomically affiliated, and 55.6% are functionally annotated (Fig. 5 and S4). While taxonomy appeared highly divergent depending on the trophic assignment (e.g., Opisthokonta and Amoebozoa are absent from photosynthetic groups, while Haptista, exclusively comprise photosynthetic members), functional annotations varied much less, at least at a high level (i.e., level B: $\mathrm{KEGG}_{\mathrm{PATHWAY}}$ = Metabolism). Categories taking part in primary metabolisms such as amino acids, carbohydrate, or lipid metabolic pathways remain stable across all trophic modes, indicating a relative functional redundancy which guarantees the resilience and stability of ecosystem functioning.

These relatively low differences should be considered with the understanding that we neither examined relative gene expression nor considered all genes, but analyzed the richness of those mapped in metabolic pathway categories in the KEGG database. Nevertheless, some functional variations turned out to be significant at a finer level (i.e., C: $\text{KEGG}_{\text{PATHWAY}}=\text{Metabolism}$) and appear to reflect the functional specificities of trophic modes (Fig. 6). For instance, in line with their trophic abilities, phototrophs and mixotrophs are characterized by numerous genes involved in photosynthesis as well as in carotenoid biosynthesis as described in Lambert et al. [14]. In contrast, strict heterotrophs exhibit a substantial portion of their genes dedicated to fatty acid processes, including sphingolipid and glycerolipid metabolism, along with fatty acid degradation, which are likely associated with their phagocytosis activity [77] and the synthesis of biological membranes [78]. Numerous genes involved in sulfur metabolism were found within saprotrophs, suggesting their implication in the sulfur cycle. The sulfur cycle is an important biochemical cycle in Lake Pavin; however, the majority of the research involves bacterial species [79, 80]. Yet, assimilation of conventional sources of sulfur by eukaryotes, such as Fungi, has already been demonstrated (e.g., in association with plant or involved in human pathologies) [81, 82] but their real contribution to biogeochemical cycles in freshwater ecosystems remains to be explored [83]. Finally, parasites have extensive genes involved in fatty acid metabolism, which aligns with the expectation that lipid metabolism is highly developed within parasites of highly divergent taxonomy [84–87]. This development is necessary to establish host-parasite interactions and involves a high diversity of genes.

Toward a functional marker of parasitism

In this study, we consider that SPCs enable the investigation of functional protein composition and specificities within trophic modes of microbial eukaryotes. It appears relevant to incorporate them into the process of identifying potential genetic markers associated to parasitism. Parasitic microorganisms are abundant and diverse in aquatic ecosystems; however, they remain enigmatic (in term of role and life cycle) and understudied. This is primarily due to the challenges associated with culturing parasites in association with their hosts, their relatively small size, and the difficulty of isolating rare parasites with very few free-living forms [88]. The use of environmental DNA/RNA offers an alternative for assessing parasite diversity and abundance as well as potential functions [88]. In this context, we focused on the 1679 SPCs clustering at least one protein assigned to an obligate parasite, aiming to identify genes likely playing a role in host-parasite interactions and propose potential marker candidates. Among the 4604 taxonomically annotated proteins gathered in the selected SPCs, a majority are assigned to parasites (71.2%) with a limited number characterized by alternative assignments (i.e., 1.6% corresponding to saprotrophs) (Fig. 8). This suggests that parasite-assigned proteins may be relatively specific, and thus that derived functions are not shared by other trophic modes. These findings support the proposal that proteins involved in host-parasite interactions resulted from a long evolutionary process, constantly adapting to host pressures to ensure survival and reproductive success. Facultative parasite-assigned proteins represent a small proportion (≈9.2%) and are affiliated mainly with Opisthokonta and Stramenopiles, as well as, to a lesser extent, Amoebozoa. Amoebozoa only include facultative representatives, which explains the low number of proteins in SPCs (Figs. 2 and 8). Opisthokonta and Stramenopiles group together both obligate and facultative parasites, indicating a relative genetic similarity between proteins of each trophic mode. Indeed, there is often a fine line between obligate, facultative, and saprotrophic organisms, and dissimilarity could be slight, reporting multiple degrees depending on the nature of the interaction (e.g., Oomycetes and Fungi may be characterized as facultative and obligate parasites, biotrophic, necrotrophic, hemi-biotrophic or saprophytic) [74–76, 89, 90].

In contrast, the other divisions (i.e., mainly Alveolata, Euglenozoa, Metamonada, and Rhizaria (Fig. 8)) are only represented in SPCs by proteins assigned to obligate parasites. This suggests that selected parasite-related proteins are involved in host-parasite interactions and have no homolog in free-living organisms of similar divisions. This is consistent with the fact that these divisions gather distinct phylogenetic levels of organisms known to be exclusively obligate parasite, such as Apicomplexa and Parabasalia phyla or Syndiniales, Plasmodiophorida, and Trypanosomatida orders.

Not one SPC grouping together all phyla with parasite relatives has been identified, thus preventing the identification of a universal marker for parasitism (Fig. 9). While it might be argued that a less stringent identity threshold could enhance the identification of potential parasitism marker genes, there is uncertainty regarding the possibility of other trophic modes also being included in parasite-related clusters, undermining the effectiveness of the strategy. Nevertheless, by examining the link between KO id barycenters and parasite-related SPCs (Fig. 10), we identified candidate genes that could be referenced as parasitism marker for specific phyla (Fig. 10, S5, Table S2). For example, we reported genes linked with the well-known mechanism of immunogenicity and extracellular survival of Trypanosoma coding for plasmanylethanolamine desaturase (involved in the biosynthesis of glycosylphosphatidylinositol, known as variant surface glycoproteins (VSG) of Kinetoplastida [56, 91, 92]) as well as encoding E3 ubiquitin-protein ligase (involved in SUMOylation process, which positively regulate VSG expression [57]). We furthermore reported several genes linked to the pathogenicity of fungi-like organisms, such as MFS multidrug transporters or glyoxalase protein associated with toxic compound detoxification (in Ascomycota and Labyrinthulomycetes, respectively) [58, 93]. Other genes correspond to virulence factors in Ascomycota (zinc-cluster proteins [94], sodium/hydrogen antiporter [95], thiamine precursor [96]; or proline-specific permease in Oomycetes [97, 98]) (Table S2).

Other genes, not yet described in a host-parasite interaction but coding for proteins whose functions may coincide with parasitic lifestyle, such as anti-apoptotic agents, have been highlighted for Ascomycota, Plasmodiophorida, and Oomycetes. Genes involved in DNA repair processes are found within Labyrinthulomycetes, Conoidasida, Ascomycota, and Oomycetes. Genes encoding cytochrome P450 family proteins, which are described as distributed within fungi-like organisms depending on their trophic capacity [63], are found in Labyrinthulomycetes and Oomycetes. All the genes reported within parasite species that can cause diseases in humans or affect species of economic interest could represent interesting lines for consideration in a health or economic context for the development of antiparasitic treatments [92, 93, 99, 100].

Conclusion

The use of the SSN-based strategy to study trophic modes of microbial eukaryotes proves to be highly relevant. It enables easy examination of genetic signatures across distant phyla that may share the same ecological function in the environment. It also allows consideration of the "unknown" of the dataset and therefore improves the number of sequences analyzed.

This study uncovered a significant number of shared protein families among mixotrophic and phototrophic microorganisms as well as mixotrophic and heterotrophic microorganisms, highlighting the metabolic versatility of mixotrophs. Similarly, we observed shared protein families between saprotrophs and parasites. These findings suggest that many microbial eukaryotes traditionally classified as facultative parasites, such as Chytridiomycota, may adopt a saprophytic lifestyle under certain environmental conditions. This lifestyle variability is not limited to individual species but extends to entire phyla known to contain parasitic members (e.g., Fungi, Oomycetes). Such adaptability likely plays a crucial role in the ecological success and resilience of these microorganisms in Lake Pavin and across diverse environmental conditions.

The high degree of specialization in parasitic organisms is particularly evident in the specificity of obligate parasite-related Specific Protein Clusters (SPCs) and the significant proportion of parasitic protein involved in lipid metabolism. This specialization is the result of long-term evolutionary processes punctuated by rapid adaptations driven by the "arms race" phenomenon between hosts and parasites. Although no universal marker for parasitism was identified, candidate genes emerged at a fine taxonomic scale. This finding suggests that there is no evolutionary convergence of proteins induced solely by the parasitic lifestyle, at least not at high sequence similarity level (80% protein identity).

Overall, our study sheds new insights into the understanding of eukaryotic ecological role within aquatic ecosystems and provides several candidate protein families that could serve as keys to understanding host-parasites interactions regardless of the availability of these proteins in public databases.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-024-02027-0.

Supplementary Material 1: Figure S1. Statistics of sequence similarity clustering using different similarity threshold for each type of information (i.e., functional annotation or trophic assignment). From the top to the bottom: $\bar{X}_{Nannot}$/CC=Mean of distinct functional and trophic annotations number by CC; $\bar{X}_{Fhom}$=Mean of functional homogeneity scores; $\bar{X}_{Fhom}$/N_{CC}=Mean of functional homogeneity scores divided by the total number of CCs; NA%=proportion of CCs clustering proteins without any functional and trophic information. At the bottom: N=total number of CCs; $\bar{X}$=mean number of proteins by CC.

Supplementary Material 2: Figure S2. Histogram reporting the percentage of explained variances for each dimension of the correspondence analysis provided in figures 4 and 8.

Supplementary Material 3: Figure S3. Histogram (top) and boxplot (bottom) reporting the number of CCs in relation with their protein's abundance. Each bar (top) depicts the number of CCs grouping a specific range of proteins, segmented into intervals of 50. For instance, the first bar indicates that 299,357 CCs grouped between 3 and 50 proteins while the last one shows that only one CC grouped 6203 proteins. Boxplot (bottom) displays the maximum and the mean protein's abundance as well as the global distribution of CCs along their protein's abundance. The Y axis is represented on a pseudo-logarithmic scale (base 2).

Supplementary Material 4: Figure S4. Functional annotation of proteins gathered in Specific Proteins Clusters of each trophic mode. The number of SPCs, number of proteins within SPCs, annotation rate against KEGG database as well as proportion of proteins annotated with KEGG metabolism category ($KEGG_{PATHWAY}$=Metabolism) are displayed on the horizontal histograms for each trophic mode. The precise functional annotation of proteins is reported on the right and the size of boxes shows the proportion of proteins functionally annotated to each KEGG metabolism.

Supplementary Material 5: Figure S5. SPCs representation gathering proteins annotated with KO ids which centroids are highly correlated to parasitism. The mention "multiple" means that proteins of the SPC are annotated with more than 3 KO ids. The identification number of SPCs is referenced on the top of each sub-graph. Black outlined points represent a protein annotated with the KO id referenced at the top of the sub-graph.

Supplementary Material 6: Table S1. Association table linking taxonomy to trophic modes. Taxonomic ranks are specified with the following code: d_(Division), k_(Kingdom), p_(Phylum), c_(Class), o_(Order), f_(Family) and g_(Genus). Trophic modes are identified using the following codes: PHOTO: photo-osmo-mixotrophs, MIXO: photo-osmo-phago-mixotrophs, HET: heterotrophs, SAP: saprotrophs and PARA: parasites. Each trophic mode assignment is described using 6 additional columns. Mixotrophs are characterized according to the Mixoplankton Database (MDB [39]; see column "Type 1" and "Description of mixotrophy", respectively). For parasites, we provide information on their degree of parasitism (obligate versus facultative) in the "Type 1" column. The "Type 2" column offers additional classification details. The nature of their interactions with hosts is documented in the "Description of endophytic interaction" and "Description of biotic interaction and pathology" columns. The "Host/symbiont" column identifies the organisms they parasitize or form symbiotic relationships with. References supporting each information are provided in the columns"References_x". The "Retained in the analysis" column indicates whether proteins affiliated with these taxa were included in the Sequence Similarity Network (SSN) generation process.

Supplementary Material 7: Table S2. Function and potential link with parasitic lifestyle of KO ids which coordinates of centroids are highly correlated to parasitism.

Acknowledgements

The authors are grateful to Genoscope (Evry, France) for sequencing data as well as the Mésocentre Clermont Auvergne University and AuBi platform for providing help, computing and storage resources. The work of AM was supported by a PhD fellowship funded by the 'Ministère de l'Enseignement Supérieur de la Recherche et de l'Innovation' and by France Génomique (ANR-10-INBS-09-08).

LB acknowledges the Institut Universitaire de France for her 5-year nomination as Junior Member (2020–2025).

Authors' contributions

A.M. conceptualized and processed all analyzes, built trophic mode database, contributed in the availability of scripts and was the major contributor in writing the manuscript. L.B. conceptualized the SSN methodology and contributed in writing the manuscript. J.R. conceptualized the SSN methodology and contributed in the availability of scripts. C.L. supervised the project and was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Funding

France Génomique (ANR-10-INBS-09-08), Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, Institut Universitaire de France.

Data availability

All data and material are available in this publication and additional information online (supplementary material and supplementary figures). Sequencing data are archived at ENA under accession number PRJEB61515.

All scripts and procedure necessary to reproduce our analysis are publicly available at https://github.com/amonjot/SSN_Monjot_2024.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]CNRS, Laboratoire Microorganismes: Génome Et Environnement, Université Clermont Auvergne, Clermont-Ferrand 63000, France. [2]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université Des Antilles, Paris, France. [3]Institut Universitaire de France, Paris, France.

Received: 13 June 2024 Accepted: 31 December 2024
Published online: 06 February 2025

References

1. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. Roberts RG, editor. PLoS Biol. 2014;12:e1001889.
2. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348:1261605–1261605.
3. Faure E, Ayata S-D, Bittner L. Towards omics-based predictions of planktonic functional composition from environmental data. Nat Commun. 2021;12:4361.
4. Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics. 2022;2:100123.
5. Grossmann L, Beisser D, Bock C, Chatzinotas A, Jensen M, Preisfeld A, et al. Trade-off between taxon diversity and functional diversity in European lake ecosystems. Mol Ecol. 2016;25:5876–88.
6. Trench-Fiol S, Fink P. Metatranscriptomics From a Small Aquatic System: Microeukaryotic Community Functions Through the Diurnal Cycle. Front Microbiol. 2020;11:1006.
7. Li L, Delgado-Viscogliosi P, Gerphagnon M, Viscogliosi E, Christaki U, Sime-Ngando T, et al. Taxonomic and functional dynamics during chytrid epidemics in an aquatic ecosystem. Mol Ecol. 2022;31:5618–34.
8. Alexander H, Rouco M, Haley ST, Wilson ST, Karl DM, Dyhrman ST. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. Proc Natl Acad Sci USA. 2015;112:E5972–9.
9. Pearson GA, Lago-Leston A, Cánovas F, Cox CJ, Verret F, Lasternas S, et al. Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula. ISME J. 2015;9:2275–89.
10. Lampe RH, Mann EL, Cohen NR, Till CP, Thamatrakoln K, Brzezinski MA, et al. Different iron storage strategies among bloom-forming diatoms. Proc Natl Acad Sci USA. 2018;115:E12275–84.
11. Caputi L, Carradec Q, Eveillard D, Kirilovsky A, Pelletier E, Pierella Karlusich JJ, et al. Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. Glob Biogeochem Cycles. 2019;33:391–419.
12. Kolody BC, McCrow JP, Allen LZ, Aylward FO, Fontanez KM, Moustafa A, et al. Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. ISME J. 2019;13:2817–33.
13. Hu SK, Liu Z, Alexander H, Campbell V, Connell PE, Dyhrman ST, et al. Shifting metabolic priorities among key protistan taxa within and below the euphotic zone. Environ Microbiol. 2018;20:2865–79.
14. Lambert BS, Groussman RD, Schatz MJ, Coesel SN, Durham BP, Alverson AJ, et al. The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. Proc Natl Acad Sci USA. 2021;119:e2100916119.
15. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018;9:373.
16. Cohen NR, McIlvin MR, Moran DM, Held NA, Saunders JK, Hawco NJ, et al. Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. Nat Microbiol. 2021;6:173–86.
17. Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome Biol Evol. 2018;10:707–15.

18. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021;39:105–14.
19. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. eLife. 2022;11:e67667.
20. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34:2115–22.
21. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35:833–44.
22. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15:962–8.
23. Chen IMA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. 2019;47:D666-77.
24. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 2019;48:D570–8.
25. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. Jordan IK, editor. PLoS One. 2009;4:e4345.
26. Bittner L, Halary S, Payri C, Cruaud C, De Reviers B, Lopez P, et al. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biol Direct. 2010;5:47.
27. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci USA. 2013;110:E1594–603.
28. Cheng S, Karkar S, Bapteste E, Yee N, Falkowski P, Bhattacharya D. Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front Ecol Evol. 2014;2:1–13.
29. Forster D, Bittner L, Karkar S, Dunthorn M, Romac S, Audic S, et al. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol. 2015;13:16.
30. Corel E, Lopez P, Méheust R, Bapteste E. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol. 2016;24:224–37.
31. Méheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci USA. 2016;113:3579–84.
32. Copp JN, Akiva E, Babbitt PC, Tokuriki N. Revealing unexplored sequence-function space using sequence similarity networks. Biochemistry. 2018;57:4651–62.
33. Rizos I, Debeljak P, Finet T, Klein D, Ayata S-D, Not F, et al. Beyond the limits of the unassigned protist microbiome: inferring large-scale spatio-temporal patterns of Syndiniales marine parasites. ISME Commun. 2023;3:16.
34. Meng A, Corre E, Probert I, Gutierrez-Rodriguez A, Siano R, Annamale A, et al. Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Mol Ecol. 2018;27:2365–80.
35. Pierella Karlusich JJ, Pelletier E, Zinger L, Lombard F, Zingone A, Colin S, et al. A robust approach to estimate relative phytoplankton cell abundances from metagenomes. Mol Ecol Resour. 2023;23:16–40.
36. Labarre A, Obiol A, Wilken S, Forn I, Massana R. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. Limnol Oceanogr. 2020;65:149–60.
37. Massana R, Labarre A, López-Escardó D, Obiol A, Bucchini F, Hackl T, et al. Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. ISME J. 2021;15:154–67.
38. Bock NA, Charvet S, Burns J, Gyaltshen Y, Rozenberg A, Duhamel S, et al. Experimental identification and in silico prediction of bacterivory in green algae. ISME J. 2021;15:1987–2000.
39. Mitra A, Caron DA, Faure E, Flynn KJ, Leles SG, Hansen PJ, et al. The Mixoplankton Database (MDB): Diversity of photo-phago-trophic plankton in form, function, and distribution across the global ocean. J Eukaryotic Microbiol. 2023;70:e12972.
40. James TY, Pelin A, Bonen L, Ahrendt S, Sain D, Corradi N, et al. Shared signatures of parasitism and phylogenomics unite cryptomycota and microsporidia. Curr Biol. 2013;23:1548–53.
41. Quiroz Velasquez PF, Abiff SK, Fins KC, Conway QB, Salazar NC, Delgado AP, et al. Transcriptome analysis of the entomopathogenic oomycete Lagenidium giganteum reveals putative virulence factors. Appl Environ Microbiol. 2014;80:6427–36.
42. del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. The others: our biased perspective of eukaryotic genomes. Trends Ecol Evol. 2014;29:252–9.
43. Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, et al. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Community Journal. 2022;2:e56.
44. Monjot A, Bronner G, Courtine D, Cruaud C, Da Silva C, Aury J, et al. Functional diversity of microbial eukaryotes in a meromictic lake: coupling between metatranscriptomic and a trait-based approach. Environ Microbiol. 2023;25:3406–22.
45. Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A. AntiFam: a tool to help identify spurious ORFs in protein annotation. Database. 2012;2012:1–5.
46. Eddy SR. Accelerated profile HMM searches. Pearson WR, editor. PLoS Comput Biol. 2011;7:e1002195.
47. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
48. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
49. Levy Karin E, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8:48.
50. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
51. Csardi G, Nepusz T. The igraph software package for complex network research. InterJ Complex Syst. 2006;1695:1–9.
52. R Development Core Team. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.
53. Lê S, Josse J, Husson F. FactoMineR An R Package for Multivariate Analysis. J Stat Soft. 2008;25:1–18.
54. Lin Pedersen T. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. 2022. Available from: https://ggraph.data-imaginist.com, https://github.com/thomasp85/ggraph.
55. Lin Pedersen T. tidygraph: A Tidy API for Graph Manipulation. 2023. Available from: https://tidygraph.data-imaginist.com, https://github.com/thomasp85/tidygraph.
56. Villas Bôas MHS, Lara LS, Wait R, Barreto BE. Identification of plasmenylethanolamine as a major component of the phospholipids of strain DM 28c of Trypanosoma cruzi. Mol Biochem Parasitol. 1999;98:175–86.
57. López-Farfán D, Bart J-M, Rojas-Barros DI, Navarro M. SUMOylation by the E3 Ligase TbSIZ1/PIAS1 Positively Regulates VSG Expression in Trypanosoma brucei. Hill KL, editor. PLoS Pathog. 2014;10:e1004545.
58. Costa C, Dias PJ, SÃ¡-Correia I, Teixeira MC. MFS multidrug transporters in pathogenic fungi: do they have real clinical impact? Front Physiol. 2014;5:1–8.
59. So YS, Maeng S, Yang DH, Kim H, Lee KT, Yu SR, et al. Regulatory Mechanism of the Atypical AP-1-Like Transcription Factor Yap1 in Cryptococcus neoformans. Mitchell AP, editor. mSphere. 2019;4:e00785-19.
60. Potting C, Tatsuta T, König T, Haag M, Wai T, Aaltonen MJ, et al. TRIAP1/PRELI complexes prevent apoptosis by mediating intramitochondrial transport of phosphatidic acid. Cell Metab. 2013;18:287–95.
61. He H, Huang J, Wu S, Jiang S, Liang L, Liu Y, et al. The roles of GTPase-activating proteins in regulated cell death and tumor immunity. J Hematol Oncol. 2021;14:171.
62. Smolarz B, Wilczyński J, Nowakowska D. DNA repair mechanisms and Toxoplasma gondii infection. Arch Microbiol. 2014;196:1–8.
63. Sello MM, Jafta N, Nelson DR, Chen W, Yu J-H, Parvez M, et al. Diversity and evolution of cytochrome P450 monooxygenases in Oomycetes. Sci Rep. 2015;5:11572.

64. Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res. 2017;45:11495–514.
65. Wyman SK, Avila-Herrera A, Nayfach S, Pollard KS. A most wanted list of conserved microbial protein families with no known domains. PLoS One. 2018;13:e0205749.
66. Promponas VJ, Iliopoulos I, Ouzounis CA. Annotation inconsistencies beyond sequence similarity-based function prediction – phylogeny and genome structure. Stand in Genomic Sci. 2015;10:108.
67. de Crécy-Lagard V, de Amorin Hegedus R, Arighi C, Babor J, Bateman A, Blaby I, et al. A roadmap for the functional annotation of protein families: a community perspective. Database (Oxford). 2022;2022:baac062.
68. Hornung BVH, Terrapon N. An objective criterion to evaluate sequence-similarity networks helps in dividing the protein family sequence space. Kolodny R, editor. PLoS Comput Biol. 2023;19:e1010881.
69. Yunes JM, Babbitt PC. Effusion: prediction of protein function from sequence similarity networks. Hancock J, editor. Bioinformatics. 2019;35:442–51.
70. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
71. 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. Kelley J, editor. Mol Biol Evol. 2021;38:4647–54.
72. Edwards KF. Mixotrophy in nanoflagellates across environmental gradients in the ocean. Proc Natl Acad Sci USA. 2019;116:6211–20.
73. Ward BA. Mixotroph ecology: More than the sum of its parts. Proc Natl Acad Sci USA. 2019;116:5846–8.
74. Shearer C. The freshwater Ascomycetes. Nova Hedwigia. 1993;56:1–33.
75. Frenken T, Alacid E, Berger SA, Bourne EC, Gerphagnon M, Grossart H-P, et al. Integrating chytrid fungal parasites into plankton ecology: research gaps and needs: Research needs in plankton chytridiomycosis. Environ Microbiol. 2017;19:3802–22.
76. Van den Wyngaert S, Ganzert L, Seto K, Rojas-Jimenez K, Agha R, Berger SA, et al. Seasonality of parasitic and saprotrophic zoosporic fungi: linking sequence data to ecological traits. ISME J. 2022;16:2242–54.
77. Lima S, Milstien S, Spiegel S. Sphingosine and sphingosine kinase 1 involvement in endocytic membrane trafficking. J Biol Chem. 2017;292:3074–88.
78. de Carvalho CCCR, Caramujo MJ. The various roles of fatty acids. Molecules. 2018;23:2583.
79. Biderre-Petit C, Boucher D, Kuever J, Alberic P, Jézéquel D, Chebance B, et al. Identification of Sulfur-Cycle Prokaryotes in a Low-Sulfate Lake (Lake Pavin) Using aprA and 16S rRNA Gene Markers. Microb Ecol. 2011;61:313–27.
80. Berg JS, Jézéquel D, Duverger A, Lamy D, Laberty-Robert C, Miot J. Microbial diversity involved in iron and cryptic sulfur cycling in the ferruginous, low-sulfate waters of Lake Pavin. PLoS One. 2019;14:e0212787.
81. Linder T. Assimilation of alternative sulfur sources in fungi. World J Microbiol Biotechnol. 2018;34:51.
82. Amich J. Sulfur metabolism as a promising source of new antifungal targets. J Fungi (Basel). 2022;8:295.
83. Krauss G-J, Solé M, Krauss G, Schlosser D, Wesenberg D, Bärlocher F. Fungi in freshwaters: ecology, physiology and biochemical potential. FEMS Microbiol Rev. 2011;35:620–51.
84. Ramakrishnan S, Serricchio M, Striepen B, Bütikofer P. Lipid synthesis in protozoan parasites: a comparison between kinetoplastids and apicomplexans. Prog Lipid Res. 2013;52:488–512.
85. Bi K, He Z, Gao Z, Zhao Y, Fu Y, Cheng J, et al. Integrated omics study of lipid droplets from Plasmodiophora brassicae. Sci Rep. 2016;6:36965.
86. Laundon D, Chrismas N, Bird K, Thomas S, Mock T, Cunliffe M. A cellular and molecular atlas reveals the basis of chytrid development. eLife. 2022;11:e73933.
87. Shunmugam S, Arnold C-S, Dass S, Katris NJ, Botté CY. The flexibility of Apicomplexa parasites in lipid metabolism. PLoS Pathog. 2022;18:e1010313.
88. Bass D, Stentiford GD, Littlewood DTJ, Hartikainen H. Diverse applications of environmental DNA methods in parasitology. Trends Parasitol. 2015;31:499–513.
89. Zhao Z, Liu H, Wang C, Xu J-R. Correction: Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 2013;15:6.
90. Fiore-Donno AM, Bonkowski M. Different community compositions between obligate and facultative oomycete plant parasites in a landscape-scale metabarcoding survey. Biol Fertil Soils. 2021;57:245–56.
91. Menon AK, Eppinger M, Mayor S, Schwarz RT. Phosphatidylethanolamine is the donor of the terminal phosphoethanolamine group in trypanosome glycosylphosphatidylinositols. EMBO J. 1993;12:1907–14.
92. Hong Y, Kinoshita T. Trypanosome Glycosylphosphatidylinositol Biosynthesis. Korean J Parasitol. 2009;47:197.
93. Chauhan SC, Madhubala R. Glyoxalase I gene deletion mutants of Leishmania donovani exhibit reduced methylglyoxal detoxification. PLoS One. 2009;4:e6805.
94. John E, Singh KB, Oliver RP, Tan K-C. Transcription factor lineages in plant-pathogenic fungi, connecting diversity with fungal virulence. Fungal Genet Biol. 2022;161:103712.
95. Jimenez V, Mesones S. Down the membrane hole: Ion channels in protozoan parasites. Kafsack BFC, editor. PLoS Pathog. 2022;18:e1011004.
96. Jin D, Sun B, Zhao W, Ma J, Zhou Q, Han X, et al. Thiamine-biosynthesis genes Bbpyr and Bbthi are required for conidial production and cell wall integrity of the entomopathogenic fungus Beauveria bassiana. J Invertebr Pathol. 2021;184:107639.
97. Berg JA, Hermans FWK, Beenders F, Abedinpour H, Vriezen WH, Visser RGF, et al. The amino acid permease (*AAP*) genes *CsAAP2A* and *SlAAP5A / B* are required for oomycete susceptibility in cucumber and tomato. Mol Plant Pathol. 2021;22:658–72.
98. Garbe E, Miramón P, Gerwien F, Ueberschaar N, Hansske-Braun L, Brandt P, et al. GNP2 Encodes a High-Specificity Proline Permease in Candida albicans. Hogan DA, editor. mBio. 2022;13:e03142-21.
99. Stijlemans B, Baral TN, Guilliams M, Brys L, Korf J, Drennan M, et al. A glycosylphosphatidylinositol-based treatment alleviates trypanosomiasis-associated immunopathology. J Immunol. 2007;179:4003–14.
100. Warrilow AGS, Hull CM, Rolley NJ, Parker JE, Nes WD, Smith SN, et al. Clotrimazole as a potent agent for treating the oomycete fish pathogen Saprolegnia parasitica through inhibition of sterol 14α-demethylase (CYP51). Appl Environ Microbiol. 2014;80:6154–66.

Publisher's Note