Guillem Marimon Fernandez

Searching for KEGG-Annotated Phagocytosis-Related Genes in Marine Microbial Eukaryotes

MASTER'S THESIS

Supervisors: Dr. Ramon Massana Molera (Institut de Ciències del Mar-CSIC) and Dr. Eduard Ocaña Pallarés (Institut de Biologia Evolutiva-CSIC UPF)

Tutor: Dr. Mireia Olivella García (UVIC-UCC)

Master's Degree in Health Data Science

Barcelona, 2025

Acknowledgements

I would like to thank Ramon Massana and Eduard Ocaña for their guidance and advice throughout this Master's Thesis. This project has allowed me to expand my knowledge in areas that were previously unfamiliar to me, and has helped me grow into a slightly better researcher. I would also like to extend my thanks to everyone at the ICM who supported me at different points during this project, from undergraduate students to doctoral and post-doctoral researchers. Their help was essential to the successful completion of this thesis.

Abstract

Phagocytosis is a key ecological and evolutionary process in marine microbial eukaryotes, yet its genetic basis remains poorly characterized, especially in uncultured protist lineages. In this project, we explored the potential of extending Kyoto Encyclopedia of Genes and Genomes (KEGG) based functional annotations to environmental protist genomes by focusing on genes characterized by KEGG to be present in the phagosome and lysosome pathways. We curated a list of 194 KEGG Orthologs (KOs) and evaluated their presence across two genomic datasets: high-quality reference genomes from EukProt's The Comparative Set (TCS), and a large collection of Single-cell Amplified Genomes (SAGs) from uncultured marine protists.

To detect these genes, we tested two genome annotation tools: eggNOG-mapper and HMMER's hmmsearch. While both tools were effective, HMMER proved to be more precise, especially in less complete genomes, making it the better method for high-confidence KO detection. We observed a strong correlation between genome completeness and KO recovery, and found that many phagocytosis-related genes annotated in animals are also present in marine protists.

Our results support the value of using SAGs and KEGG-based annotations to study functional traits in uncultured protists. With appropriate filtering and validation steps, we have shown how our approach can offer important insights into the evolutionary conservation and ecological diversity of phagocytosis-related genes in marine eukaryotes.

List of Figures

List of Tables

Index

1 Background and motivations

Phagocytosis is a fundamental process in marine ecosystems by which prey organisms are consumed and their biomass is incorporated into food webs. In fact, this process is especially important since protists (single-celled eukaryotes) use phagocytosis to feed on bacteria and other microorganisms. This makes them a central part of the microbial food web, linking microbial production to higher levels of the food chain and playing a major role in nutrient recycling [1]. However, studies searching for the genes behind this key ecological function in free-living phagotrophic protists are still limited, in part because of the lack of appropriate ecological models [2].

Hence, we were particularly interested in studying the genetic basis of phagocytosis in marine protists because, despite its ecological importance, there's still a lot we don't know about how this function is encoded in their genomes. In fact, while several studies in model organisms have helped identify genes involved in phagocytosis, marine protists are significantly more diverse and often poorly represented in reference databases [3].

When exploring databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) [4] to identify genes linked to phagocytosis, we often found ourselves questioning the accuracy of these annotations, especially when it comes to marine organisms. Some genes, such as peptidases or proton pumps, have been shown to be expressed during bacterivory in natural protist communities, suggesting they play important roles and are therefore more likely to be well annotated [5]. However, many other genes from uncultured or less-studied species lack clear functional descriptions, making it unclear whether their assigned functions are meaningful in this context [6].

Given this context, the main motivation behind this thesis was to test whether the genes typically associated with phagocytosis in databases are truly present in marine phagotrophs. For this purpose, we focused specifically on the KEGG database, which provides both functional gene annotations and curated pathway maps, making it a useful resource for linking genes to biological processes like phagocytosis.

2 Introduction

2.1 Phagocytosis

Phagocytosis is a fundamental cellular process by which eukaryotic cells engulf and digest particles such as bacteria, dead cells, organic debris, or even other eukaryotes. This mechanism is key for several important biological functions, including nutrient acquisition, immune defense, and cellular homeostasis [7].

In multicellular organisms, specialized immune cells like macrophages and neutrophils utilize phagocytosis to eliminate pathogens, damaged cells, or foreign particles, contributing to maintaining tissue health and immune surveillance. In unicellular eukaryotes, phagocytosis serves as the main source of nutrient intake, allowing these organisms to consume bacteria and other small particles to maintain all their metabolic and nutritional requirements [8].

2.1.1 Phagocytosis, a Fundamental Eukaryotic Process

From an evolutionary perspective, phagocytosis is hypothesized to have enabled the pivotal endosymbiotic events that led to the origin of mitochondria and chloroplasts. According to the endosymbiotic theory, a proto-eukaryotic cell likely internalized an alpha-proteobacterium, which evolved into the mitochondrion, followed later by the engulfment of a cyanobacterium-like organism, which led to the formation of a plastid (Martin et al., 2015). This makes phagocytosis not just a feeding mechanism, but also a driver of eukaryotic cellular complexity [9].

2.1.2 Phagocytosis in Multicellular Organisms

The phagocytosis process in multicellular organisms begins with the recognition of target particles by phagocytes. In animals, these cells possess surface receptors that are able to identify specific molecular patterns on pathogens or dying cells. When they are recognized, the phagocyte's plasma membrane extends around the particle, forming pseudopodia that eventually enclose the target within an internal vesicle called a phagosome [10].

After its internalization, the phagosome undergoes a maturation process in which it is fused with the lysosome, an organelle particularly rich in hydrolytic enzymes. When the phagosome is merged with the lyosome, they form a phagolysosome. Within the acidic environment of the phagolysosome, the engulfed material is degraded by lysosomal enzymes, effectively neutralizing potential threats and recycling cellular components [11].

The final stage of phagocytosis involves the resolution of the phagolysosome. Once degradation is complete, the resulting molecular fragments are either expelled from the cell through exocytosis or recycled for various cellular functions. This process plays a crucial role in immune signaling, as some of the degraded pathogen components are presented on the phagocyte's surface in a mechanism known as antigen presentation. By doing so, the phagocyte not only eliminates harmful entities, but also activates other immune cells, effectively connecting the innate and adaptive branches of the immune system [12].

2.1.3 Phagocytosis in Protists

As mentioned earlier, phagocytosis in protists primarily serves as a feeding mechanism rather than an immune response. Consequently, the recognition of targets in protists is generally non-specific and is often based on physical and chemical cues from the environment. For instance, protists may detect and move toward higher concentrations of certain chemicals released by potential prey, a behavior known as chemotaxis. In these cases, recognition occurs upon direct contact, when the protist extends its membrane around the particle to form a phagosome, initiating the subsequent steps previously described for multicellular organisms [13].

2.2 KEGG Database

To develop our project we focused on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which has become a central resource for pathway-based gene annotation across a wide range of species. Created in 1995 by Professor Minoru Kanehisa at Kyoto University, the KEGG database is a bioinformatics resource developed to facilitate the understanding of high-level functional biological systems. Initially, KEGG was established under the Japanese Human Genome Program to provide a reference knowledge base for linking genomes to life and the

environment. Nowadays, it integrates genomic, chemical, and systemic functional information, offering a quality platform for the interpretation of genome sequences and other high-throughput data.

All this data is comprised into multiple interconnected databases, including KEGG PATHWAY, KEGG GENES, KEGG COMPOUND, and KEGG ORTHOLOGY, among others. The one that we mainly explore for this project was the KEGG PATHWAY database. It essentially contains manually curated pathway maps representing molecular interaction and reaction networks, such as metabolic pathways, regulatory pathways, and disease pathways.

Thanks to their well-structured representation of biological knowledge, KEGG pathway maps have become a powerful tool for researchers to visualize and interpret the complex biological processes that occur within cells. In particular, by mapping experimental results to these curated pathways, scientists can identify active biological functions, uncover disease mechanisms, and explore potential therapeutic targets. This integrative framework is especially valuable in fields such as systems biology, functional genomics, and drug development, where understanding interactions at a system level is essential.

2.3 Datasets

To try to answer this question, we decided to use two different genomic datasets: one from uncultured species - our Single-cell Amplified Genomes (SAGs) - and one from cultured species, EukProt's The Comparative Set (TCS).

2.3.1 SAGs

In recent years, the use of Single-cell Amplified Genomes (SAGs) has emerged as a key method for studying marine protists that cannot be cultured [14]. This approach builds on earlier advances in single-cell genomics, which were originally developed to investigate uncultivated microbes in other environments. Specifically, the study by Marcy et al. in 2007 [15] was the first to propose single-cell sequencing as a powerful tool for analyzing complex microbial ecosystems without the need for cultivation. In that work, the authors designed and fabricated a microfluidic chip with the ability to perform isolation and amplification of microbial cells. In order to test their proposal they worked with microbes found in the human subgingival crevice, which were later sequenced using RNAseq. Their entire pipeline is different from ours, but the reasoning behind the use of single-cell genomics is essentially the same: the impossibility to culture.

If we talk about protists, SAGs are genomes obtained from individual microbial cells. They are mainly generated using a combination of microfluidics, fluorescence-activated cell sorting (FACS) cytometry, and whole genome amplification (WGA) techniques, followed by a Next Generation Sequencing (NGS) technique to obtain reads.

The first paper describing the obtention of SAGs from a marine sample was published back in 2007 by Stepanauskas, R. et al. [16]. They explained how the use of multiple displacement amplification (MDA) followed by PCR screening resulted in the construction of a pilot library of 11 SAGs from the Gulf of Maine.

Since then, the field has evolved considerably and nowadays, thanks to Next Generation Sequencing (NGS) techniques and other new technologies, both sequencing process and analysis pipelines are far more efficient and reliable. It is because of this reason that SAGs have become increasingly important for the study of microbial diversity and function in complex ecosystems.

In particular, our SAGs come from the Blanes Bay Microbiome Observatory (BBMO). This observatory project has been the main source of data for marine-related publications since 1950, being one of the most important Spanish projects for the gathering of marine data.

The location of the observatory is not a coincidence. This strategic location at the beginning of Costa Brava is a very good representative of an oligotrophic (relatively nutrient-poor) coastal ecosystem that has been relatively unaffected by human and terrestrial influences. This was also the exact place where the pioneer of ecological and oceanographic research Ramon Margalef chose to carry out his first studies on marine microorganisms, published more than 75 years ago. Microbial data from the site goes back to 1992, and they have a special project where microbial data has been sampled monthly for more than 20 years, helping in the study of the coastal Mediterranean Sea [17] [18].

2.3.2 EukProt's The Comparative Set (TCS)

In addition to our SAGs, we also decided to include a set of high-quality reference genomes in order to provide a broader phylogenetic context for our analyses. For this purpose, we used EukProt v3 [19], a comprehensive database comprising genome-scale predicted proteomes from a wide array of eukaryotic lineages. This dataset mainly consists of data from cultured organisms, ensuring the reliability and quality.

3 Objectives

This project was designed with the aim to achieve three main objectives:

1. To investigate the current annotation of the phagosome and lysosome pathways in the KEGG database. This included compiling all KEGG Orthology (KO) identifiers associated with these pathways and evaluating their taxonomic coverage.
2. To study in depth the phagosome and lysosome pathway diagrams presented in KEGG. Having and understanding on the different steps of phagocytosis should allow us to classify our KOs into different categories based on their function. Our aim is to later evaluate which categories are more or less represented across our genomes.
3. To assess the presence of these annotated genes in both reference and environmental datasets. Specifically, we aimed to detect a curated list of 194 phagosome and lysosome related KOs in two genomic resources: (1) a high-quality set of 165 reference genomes from EukProt's The Comparative Set (TCS), and (2) our own collection of Single-cell Amplified Genomes (SAGs) obtained from marine microbial eukaryotes.

4 State-of-the-art

4.1 The Uncultured Majority

The subject of marine microbes has been around for a long time, but it is nowadays that, thanks to Next Generation Sequencing (NGS) techniques, the field has experienced a significant increase in popularity and lots of possibilities have opened. Regardless of this fact, the difficulties of

culturing microorganisms for their later sequencing and analysis is still one of the big limitations in the microbial ecology field.

This special subject has been addressed as "The uncultured microbial majority" [20], and it refers to the fact that the majority of microbial species present in the natural environment have not yet been successfully cultured and studied in a laboratory setting. This is especially true for microbial species that inhabit complex environments such as soil, water, and the human body.

Several facts explain why these cultures can be so hard to develop, one of them being the complexity of marine ecosystems. Marine environments are incredibly diverse and complex, and minor differences in both physical and chemical parameters can easily influence microbial growth and survival. This makes it really challenging for researchers to recreate these conditions for the culturing of each specific type of microbe. And, when successful, often these cultures exhibit prolonged growth rates and low densities, which makes them difficult to be sequenced.

To solve this problem, the use of Single-cell Amplified Genomes (SAGs) [21] arose as one of the main solutions. This technique was key because it allowed researchers to study the genomes of uncultured microbes by sequencing their DNA directly from environmental samples, solving the problem of culture bias and providing a very good insight into their diversity and functional capabilities. As a result, we are gradually gaining a better understanding of the uncultured microbial majority and their importance in natural ecosystems.

4.2 Phagocytosis in Marine Microbial Eukaryotes

In marine environments, phagocytosis is a central mechanism in the transfer of biomass and energy through microbial food webs. Unicellular eukaryotes, especially heterotrophic protists like ciliates, dinoflagellates, and choanoflagellates consume bacteria and small plankton, mediating the flux of organic matter and nutrients across trophic levels [1]. This microbial loop allows the recycling of dissolved organic carbon into the food web, maintaining ecosystem productivity.

Recent research has emphasized the underappreciated contribution of phagocytosis to carbon acquisition in marine nanoplankton. For instance, studies modeling the trophic strategies of marine plankton suggest that heterotrophy via phagocytosis accounts for a substantial fraction of their carbon intake, sometimes exceeding photosynthetic contributions, especially under nutrient limitation [22] [23].

Thus, phagocytosis not only supports survival strategies at the single-cell level but also plays a critical role in larger-scale ecological processes like carbon cycling, nutrient remineralization, and food web structuring in ocean ecosystems.

4.3 Challenges in Studying Phagocytosis-Related Genes

Despite its ecological and evolutionary importance, the molecular basis of phagocytosis still remains poorly characterized outside of well-characterized model organisms like *Dictyostelium discoideum*, *Entamoeba histolytica*, and mammalian models like *Mus musculus*. In particular, marine microbial eukaryotes, and especially uncultured lineages, are significantly underrepresented in functional genomic databases, making it difficult to track phagocytosis-related gene families across environmental diversity [24].

As early mentioned, in recent years the rise of single-cell genomics and environmental sequencing has enabled the discovery of thousands of novel eukaryotic lineages, but its functional annotation remains a bottleneck. In such datasets, the identification of phagocytosis-related genes is

complicated by the multifactorial nature of the process, which involves cytoskeletal remodeling, vesicle trafficking, membrane signaling, and digestion pathways [25]. Additionally, homologous gene families can diverge significantly across taxa, which complicates orthology inference.

Consequently, developing targeted approaches for identifying conserved phagocytosis genes, such as leveraging curated KEGG Orthologs (KOs) or HMM profiles, is essential for bridging this gap and understanding how these vital processes operate in natural protist assemblages [26] [27].

5 Methods

All High-Performance computing analyses were run between the Marine Bioinformatics Service (MARBITS) of the Institut de Ciències del Mar (ICM-CSIC) in Barcelona, and in Centro de Supercomputación de Galicia (CESGA).

5.1 Single-cell Amplified Genomes (SAGs) Obtention

We worked with Single-cell Amplified Genomes (SAGs) collected in two different years: 2018 and 2020. In both cases, the process began with the collection of 50 liters of seawater from Blanes Bay, located in the northwest Mediterranean. A subsample of this water was then subjected to Flow Cytometry Sorting [28] at the Parc de Recerca Biomèdica de Barcelona (PRBB). During this process, individual cells were sorted into 384-well plates based on three categories defined by cell size and autofluorescence: Pico-Heterotrophic (PicoHF), Nano-Heterotrophic (NanoHF), and Pico-Phototrophic (PicoPF), each representing distinct microbial functional groups.

In 2018, the amplification was performed in Alacant, Spain, using Multiple Displacement Amplification (MDA). The sequencing step was done with Illumina in Barcelona, at Centre Recerca Genomica (CRG) in Parc de Recerca Biomedica de Barcelona (PRBB). This resulted in 81 SAGs.

In 2020, several plates were sorted but just two plates were selected for further processing. They were sent to two different facilities to perform distinct amplification techniques. One plate was sent to Leuven, Belgium, where Whole Genome Amplification (WGA) was performed, followed by Illumina sequencing. This resulted in 381 SAGs. The second plate was sent to Alacant, where amplification was done using MDA. Illumina sequencing was later performed also at CRG in PRBB. This resulted in 384 SAGs.

5.2 SAGs pipeline

Both Leuven and Alacant SAGs from 2020 were processed using the the pipeline shown in Figure 1. SAGs from 2018 were processed by David López Escardó (CSIC-Istitut de Biologia Evolutiva (IBE)) following a similar pipeline. However, the final outputs from both pipelines are essentially the same.

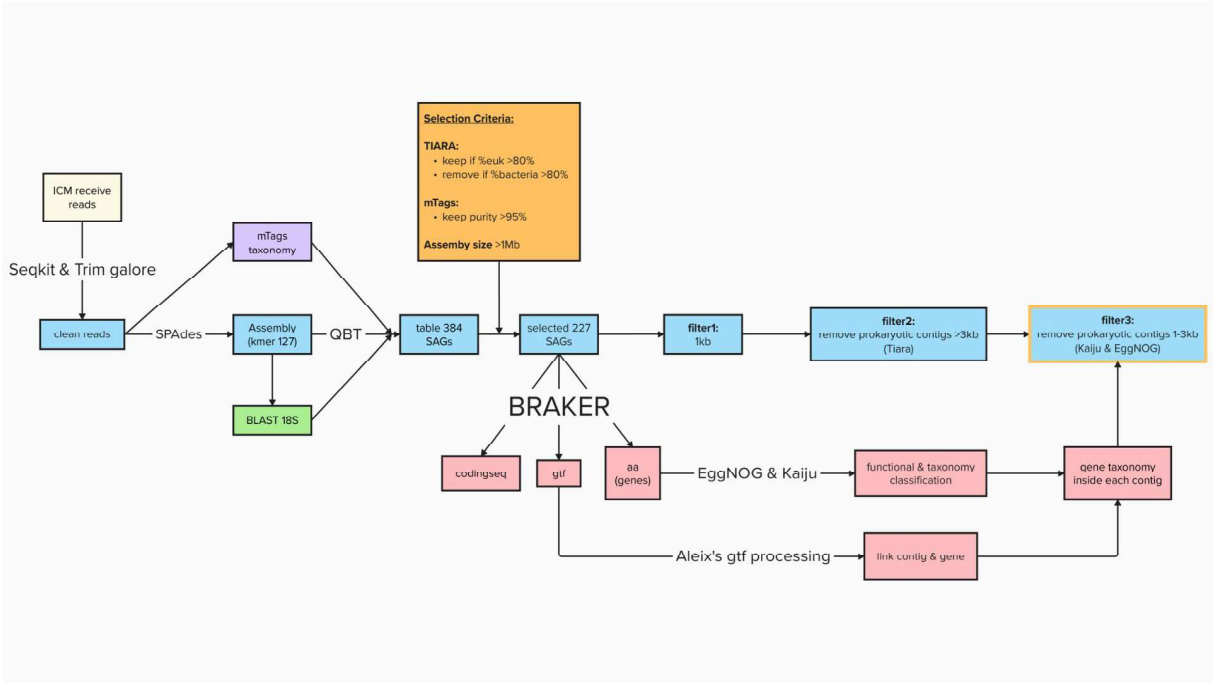

Figure 1: Overview of the workflow from read preprocessing to gene-level taxonomy classification. Steps include quality filtering, assembly, filtering, annotation (Kaiju, EggNOG), and gene-linking using BRAKER outputs.

5.2.1 Preprocessing

The first thing that needed to be done was preprocessing the data that came from the sequencer. We performed a first approach to the data using the seqkit tool [29]. This simple but very versatile program allowed us to briefly see what we had inside our reads. Overall seqkit is a very powerful tool that has lots of different applications, all regarding the quick visualization and summary of information from genomic data sets. After our seqkit analysis, the results were checked to be correct and we proceeded to trim the adapter from our sequences. These adapters are placed to facilitate the amplification step by Illumina. To perform such a task we used Trim Galore (Krueger F. 2012).

Trim Galore is a very simple program. It just needs to know (or not) the sequence of these adapters and it will remove them from our sequences. In our case we did not specify any adapter sequence, so the program takes the first 13 bp of Illumina standard adapters ('AGATCGGAAGAGC') by default, which is suitable for both ends of paired-end libraries. After the trimming process, we did a second seqkit in order to confirm that the trimming was successful.

The last step of the preprocessing section was to concatenate both lane1 and lane2 reads. The sequencing center in Belgium conducted two trials at sequencing these SAGs, and they were called lane1 and lane2. Given this situation we needed to concatenate all line1 and line2 reads in a single file.

At this point two possible paths opened in our pipeline: to perform a de-novo assembly of each SAG, or to follow the mTags pipeline. Both are independent from one another, but their results can be related, so we followed them in parallel and joined their outputs at the end of each pipeline.

5.2.2 Assembly pipeline

SPAdes (St. Petersburg genome assembler) [30] was the program selected to perform our de-novo assembly. SPAdes has become one of the most popular assembler tools over recent years due to its ability to handle a diverse range of datasets with different variable coverage and qualities, and each dataset has its own pipeline to ensure the best assembled output possible.

In our case, since we had single-cell reads from Illumina, we only needed to input both forward and reverse reads from our concatenation step and tell the program that our data was from single-cell reads. After two days we finally obtained our results, which consisted of several files and reports regarding the performance of our assembly. Still, the most important files for us were the fasta formatted files. There were two files: a contigs fasta file, and a scaffold fasta file.

Contigs represent the initial building blocks of a genome. They are consensus DNA sequences generated from the overlap of several smaller reads. The union of multiple contigs connected by gaps forms scaffolds. And as mentioned earlier, both contigs and scaffolds are stored in a fasta file each.

The analysis of the assembly was done using three different programs. They are often run at the same time because the combination of their results generates a very good overall summary of the performance of our assembly.

Quast (QUAlity of Assessment Tools) [31], which is also developed by St. Petersburg's Academic University (like SPAdes) is a quality assessment tool for the evaluation and comparison of multiple genome assemblies. It can evaluate assemblies both with or without a reference genome, and it produces multiple reports, summary tables, and plots, which end up being very useful for an easy visualization of results.

BUSCO (Benchmarking Universal Single-Copy Orthologs) [32] is a tool that uses gene prediction as a measure to assess the quality of genome assemblies. It compares our assembly against a database of (255 in our case) highly conserved single-copy orthologs (BUSCOs) that are expected to be present in all eukaryotic genomes. Since these BUSCO genes are unique (single-copy) orthologs across a wide range of species, they are considered to be very good indicators of genome completeness and quality. The completeness score is calculated taking into account BUSCOs whose sequence was either found complete or fragmented. And just by looking at this score we can get a good hint at the performance of our assembly.

Tiara [33] is a quite complex deep-learning-based approach for the identification of eukaryotic sequences in the metagenomic data. It is also the program used in the cleaning step of our pipeline (see later). Tiara works in two steps: in the first step, it classifies sequences into six classes: three prokaryotes being Archaea, Bacteria, and Prokarya (sequences undistinguishable between Bacteria and Archaea), two classes of eukaryotes being Eukarya (nuclear genomes and organelles), mitochondrial and plastidial genomes and unknowns for sequences that could not be properly classified. The second step consists of a deeper classification of the organellar sequences into classes representing plastidial genomes, mitochondrial genomes, and unknowns for the possible unclassified sequences.

This classification is done by using a deep neural network, which is a type of machine learning algorithm specially designed to model complex relationships in data. This neural network is trained with a large dataset of 8220 genomic sequences from the three domains of life Eukarya, Bacteria, and Archaea.

After the three programs finish and give their results, we summarize and reformat these results as three separate reports. Finally, we merge the most valuable fields of these reports into a

single table containing the final information about our SAGs.

This table will be later expanded with the results from the OTU table obtained at the end of the mTags pipeline.

5.2.3 Taxonomic classification using mTags pipeline

The mTags pipeline consists of two steps: first the extraction of mTags using BLAST [34], and then the classification of those mTags in an OTU table format.

mTags (metagenomic Tags) are short 18S rDNA metagenomic Illumina reads from a given sample (150 bp long in our case). These mTags are extracted using BLAST of the reads against a database. In our case, we used the "eukaryotes V4" database [35], a custom database of eukaryotic V4 18S rDNA sequences.

The database was first built using 97% clustered V4 sequences from previous environmental high throughput sequencing (HTS) studies in European coastal systems: Blanes Bay Microbial Observatory [36], BioMarKs project [37], and the water column of the global ocean sampled during the Malaspina cruise [38]. Then, the database was complemented with trimmed 97% clustered V4 sequences from SILVA SSU 128 [39] that were not found in environmental HTS data sets. All sequences were manually curated and classified at 3 taxonomic levels: OTU 97 level (Operational Taxonomic Units of sequences clustered at 97% similarity), taxonomic group (in general a formal Class), and supergroup.

With BLAST, each mTag is characterized as 3-word "references" depending on how similar they are to the V4 regions. The database has these 3 fields sorted in a way that if any OTU could not be characterized from the mTAG, the OTU field would have "NA", and at least we can have the taxonomic group. In the case that the taxonomic group characterization fails, the taxonomic group field will also have "NA" and the supergroup field will hopefully become the main field. With this method, we can easily identify which are the most useful mTags.

Once the database is created and we have our mTags referenced by these 3 fields, an OTU table is built. This OTU table has one column for each SAG and one row for each reference found in any of the SAGs. When a reference is present in one of the SAGs, the cell in the table where they match will have the number of times that particular reference appears in that SAG. Note that each SAG can have more than one different reference and each reference can be present multiple times in more than one SAG. This results in a big and messy table, which needs to be processed if we want to perform any proper analysis.

The processing of the table was done using R, and it was a challenging step because we wanted to summarize the OTU table in a very specific and more user-friendly format. To do so, rows and columns needed to be rearranged and some calculations needed to be done. It was very important for us to be able to easily visualize numbers and percentages of agreement among the multiple characterizations from a single SAG.

Questions such as "Which is the main OTU, if there is any?" or "Is this number of matching OTUs enough to consider this OTU the main OTU, if maybe there is more agreement regarding the taxonomic group?" needed to be answered just by looking at the new processed table, which made us think about plenty of different formats for the table.

Once we had this table containing all the important information about each SAG, we had a great view of each of the 384 SAGs. With this nice perspective, we could easily distinguish and select good from bad SAGs.

5.2.4 SAG Selection

The main selection criteria was what we called the "purity percentage", which is for each SAG, the percentage of mTags belonging to the most common (main) characterized eukaryotic group. We selected SAGs having $> 96\%$ purity with more than one assigned mTAG. Other properties such as BUSCO completeness and bacteria percentage from Tiara were also taken into account.

5.2.5 Cleaning pipeline - 3 Filters

Having this subset of the best 227 SAGs, we design a cleaning pipeline consisting of 3 filters.

Filter 1

The first filter is the simple. It consist of just filtering out contigs smaller than 1000 base pairs (bp). This was done using Seqkit's [29] filtering option.

Filter 2

The second filter is entirely based on Tiara [33] results. It consists of filtering out all cotings that Tiara consider to be prokaryotic contigs. Notice that Tiara only classifies contigs that are larger than 3000 bp.

Filter 3

The last filter focuses only on those contigs in the range of 1000 bp and 3000 bp. We used Kaiju and eggNOG to classify them and later filter out those for which we have strong evidence that they are prokaryotic contigs. Both programs need genes in order to classify them, so we used BRAKER2 (2.1.6) to predict genes for our SAGs.

BRAKER2 [40] is a fully automated pipeline designed for the annotation of protein-coding genes in novel eukaryotic genomes. It integrates two primary gene prediction tools: GeneMark and AUGUSTUS. Specifically, BRAKER2 uses GeneMark-EP+ for initial gene prediction and training, and AUGUSTUS for final gene prediction, incorporating extrinsic evidence such as RNA-Seq data and protein homology information.

The pipeline operates by first using GeneMark-EP+ to predict genes based on protein homology data, generating initial gene models. These models are then used to train AUGUSTUS, which performs the final gene prediction step, integrating both initial predictions and any available extrinsic evidence. This approach allows BRAKER2 to achieve high gene prediction accuracy even in the absence of closely related reference annotations or RNA-Seq data.

Once we have genes predicted for each of our SAGs, we pass them through eggNOG and Kaiju to annotate them.

eggNOG-mapper-v2 [41] is a robust and efficient program designed for large-scale functional annotation of novel sequences, including those from genomes, transcriptomes, and metagenomes. It uses precomputed orthologous groups and phylogenies from its own database to assign functional information to query sequences based on orthology relationships, which are generally more reliable than simple homology-based methods.

We employed eggNOG-mapper to annotate the predicted protein sequences generated by BRAKER2. As a result, we obtained a big table containing valuable information such as KEGG Orthology (KO) identifiers, Clusters of Orthologous Groups (COG) classifications, Gene Ontology (GO) terms, and NCBI taxonomic identifiers to determine the likely origin of each gene. By cross-

referencing these identifiers with a manually curated internal database, we classified each gene as being of eukaryotic or bacterial origin.

To complement the functional annotations provided by eggNOG-mapper and to obtain an independent assessment of gene origins, we used Kaiju [42], a fast and sensitive tool for taxonomic classification of high-throughput sequencing data. While Kaiju is commonly used for classifying sequencing reads, it is equally applicable to assembled contigs, as it assigns taxonomic labels based on translated protein sequences, regardless of sequence length.

Kaiju operates by translating input nucleotide sequences into amino acid sequences across all six reading frames and then searching for matches against a reference protein database using a memory-efficient implementation of the Burrows-Wheeler Transform (BWT) algorithm. In our case, we used the NCBI BLAST non-redundant protein database (nr) as the reference, which provides broad taxonomic coverage and is particularly effective for detecting homologs across diverse organisms.

This protein-level classification approach allows for an enhanced sensitivity, particularly for identifying sequences from organisms that may be underrepresented in nucleotide-based reference databases. By assigning taxonomic identifiers to each contig based on the best matches in the nr database, we were able to determine the likely origin of each sequence. This approach allowed us to obtain a second independent opinion to complement the assignments provided by eggNOG-mapper, increasing our confidence in the overall classification and enhancing the robustness of our of our taxonomic filtering strategy.

The final step involved classifying each of our contigs based on the independent outputs of the two annotation tools. Once all contigs were annotated and their taxonomic origin inferred, we established a strict filtering criterion for identifying and removing likely contaminant sequences. Specifically, we decided to remove only those contigs that were consistently classified as prokaryotic by both programs. In contrast, we kept any contig for which the two tools disagreed or where only one of them indicated a prokaryotic origin while the other provided no assignment. This conservative approach ensured that we minimized the risk of erroneously discarding eukaryotic sequences due to tool-specific limitations or possible database gaps.

5.2.6 Coassembly

With our set of SAGs fully cleaned using the three filtering strategies described above, we wanted to study the possibility of performing coassemblies. This process involves merging the raw sequencing reads from multiple SAGs prior to assembly, with the goal of improving contiguity and recovering more complete genomic regions.

BLAST

First, we performed a BLAST [34] search of all 227 SAGs against a reference database constructed from the SAGs themselves. In addition, the 74 Leuven SAGs were also blasted against the same database.

A second database was made combining the 227 Alacant and 81 2018 SAGs. Then, the same 81 SAGs were blasted against this database.

By combining all BLAST results with 18S rRNA gene information obtained from the mTags pipeline, we were able to group SAGs into clusters representing the same or closely related taxa.

ANI

To fully ensure the similarity between the SAGs grouped within each cluster, we calculated their Average Nucleotide Identity (ANI) [42]. ANI provides a robust measure of nucleotide-level similarity between genomes and is widely used to assess whether sequences belong to the same species or closely related lineages. In our context, it served as an additional check to validate that SAGs grouped for coassembly were indeed highly similar at the genomic level.

Our ANI calculation pipeline started by first preparing a BLAST-compatible nucleotide databases for each SAG. For this, we filtered scaffolds to retain only those longer than 1,000 base pairs using `seqkit` [29], ensuring high-confidence alignments. These scaffolds were then splited into 1,000 bp fragments with 10 bp overlaps using `pyfasta split` [43], which helped standardize sequence lengths and improve comparability across SAGs. The resulting fragments were used to build nucleotide BLAST databases for each SAG using `makeblastdb` [34], enabling efficient and consistent pairwise comparisons.

After generating the BLAST databases for each SAG, we proceeded with the ANI calculations by performing all-vs-all pairwise comparisons. For each SAG within a cluster, its 1K-mers were used as queries in a BLAST search against all other SAGs in the same group. We used `blastn` [34] with stringent filters to retain only high-confidence hits, including a minimum query coverage of 70%, a maximum of one high-scoring pair per alignment, and an e-value threshold of 1e-5. The output of each BLAST comparison was then parsed to compute the total length of aligned sequences and the identity-weighted percentage of matching nucleotides. These values were used to calculate the ANI between each pair of SAGs, and the results were saved in per-sample summary files.

By visualizing the resulting ANI values and stating a threshold of 97%, we were able to validate the coherence of each taxonomic cluster. SAGs that did not meet this threshold were considered outliers and were excluded from the group.

Coassembly

The main coassembly was done by merging together all the reads that belonged to each taxa. Then, the exact assembly and cleaning pipeline previously explained were followed to obtain the final coassemblies [44].

After completing the coassembly process, we needed to reorganize all of our SAGs because while plenty of them were now assembled together, there were also SAGs that remained as individuals. This new organization constituted what we call our "Final Genomes".

5.3 EukProt's The Comparative Set (TCS)

We downloaded the 196 proteomes designated as part of "The Comparative Set" (TCS) [19], a curated subset intended to support cross-lineage comparisons. However, prior to preparing our custom database, we decided to exclude 31 of these 196 proteomes due to recent findings of contamination from other eukaryotic or prokaryotic genomic sources. This issue was communicated to us by Eduard Ocaña-Pallarès, a postdoctoral researcher in Dr. Daniel J. Richter's group, which coordinates the development of the EukProt database. These 31 proteomes are scheduled to be removed in the forthcoming EukProt v4 release due to the identified contamination.

Consequently, we excluded these 31 proteomes from our study and retained 165 high-confidence species to use in our analyses.

5.4 Trophic Modes

Trophic modes refer to the strategies organisms employ to acquire energy and nutrients, fundamentally shaping their ecological roles. These modes are pivotal in determining how organisms interact within ecosystems, influencing energy flow and nutrient cycling. In our case, having these annotations was critical to contextualize and validate the presence of our phagocytosis-related genes.

Sofia Bejar, a fellow student from our group, conducted an exhaustive literature review to manually annotate the trophic modes of the majority of our species. Her work allowed us to classify our species into six different trophic modes:

1. **Phagotroph**: Organisms that ingest particulate food matter, such as bacteria or other cells, through phagocytosis. This mode is common among many protists and certain animal cells.
2. **Osmotroph**: Organisms that absorb dissolved organic compounds across their membranes. This mode is typical of fungi and some protists, allowing them to assimilate nutrients directly from their environment.
3. **Photosynthetic**: Organisms capable of converting light energy into chemical energy via photosynthesis. This group includes plants, algae, and certain bacteria, which serve as primary producers in ecosystems.
4. **Parasite**: Organisms that live on or within a host organism, deriving nutrients at the host's expense. These parasitic relationships can significantly impact host populations and community dynamics.
5. **Mixed**: Organisms that combine multiple trophic strategies such as mixotrophs, that can both do the photosynthesis and ingest food particles. This flexibility allows them to easily adapt to varying environmental conditions.
6. **Unknown**: Species for which the trophic mode could not be determined due to insufficient or inconclusive information.

5.5 KEGG's Current Annotation Study

We conducted an in-depth analysis of the species currently annotated in KEGG as having genes related to phagocytosis. Through this exploration we found a clear taxonomic bias toward animal species. That led us to wonder if these genes that are mostly annotated in animals could be also present in marine microbial eukaryotes. This question became central to our project, as it reflects on how far we can trust existing annotations when studying organisms that might be less studied.

To access all available organisms in KEGG, we used the `KEGGREST` [45] package in R to retrieve the complete list of annotated species. Additionally, for each KO involved in the phagosome and lysosome pathways, we manually obtained its taxonomic distribution. This distribution refers to the number of organisms in which a given KO is annotated. We extracted this information by navigating to the taxonomy section of each KO entry and downloading the "Summary view" of Group 3, which provides a hierarchical classification across three taxonomic levels.

5.6 KEGG's Phagosome and Lysosome pathway diagrams deep dive

Since this project is focused on the phagosome and lysosome functions, KEGG became an key resource for us. That is the reason why the study of their pathway diagrams was very imporant for us.

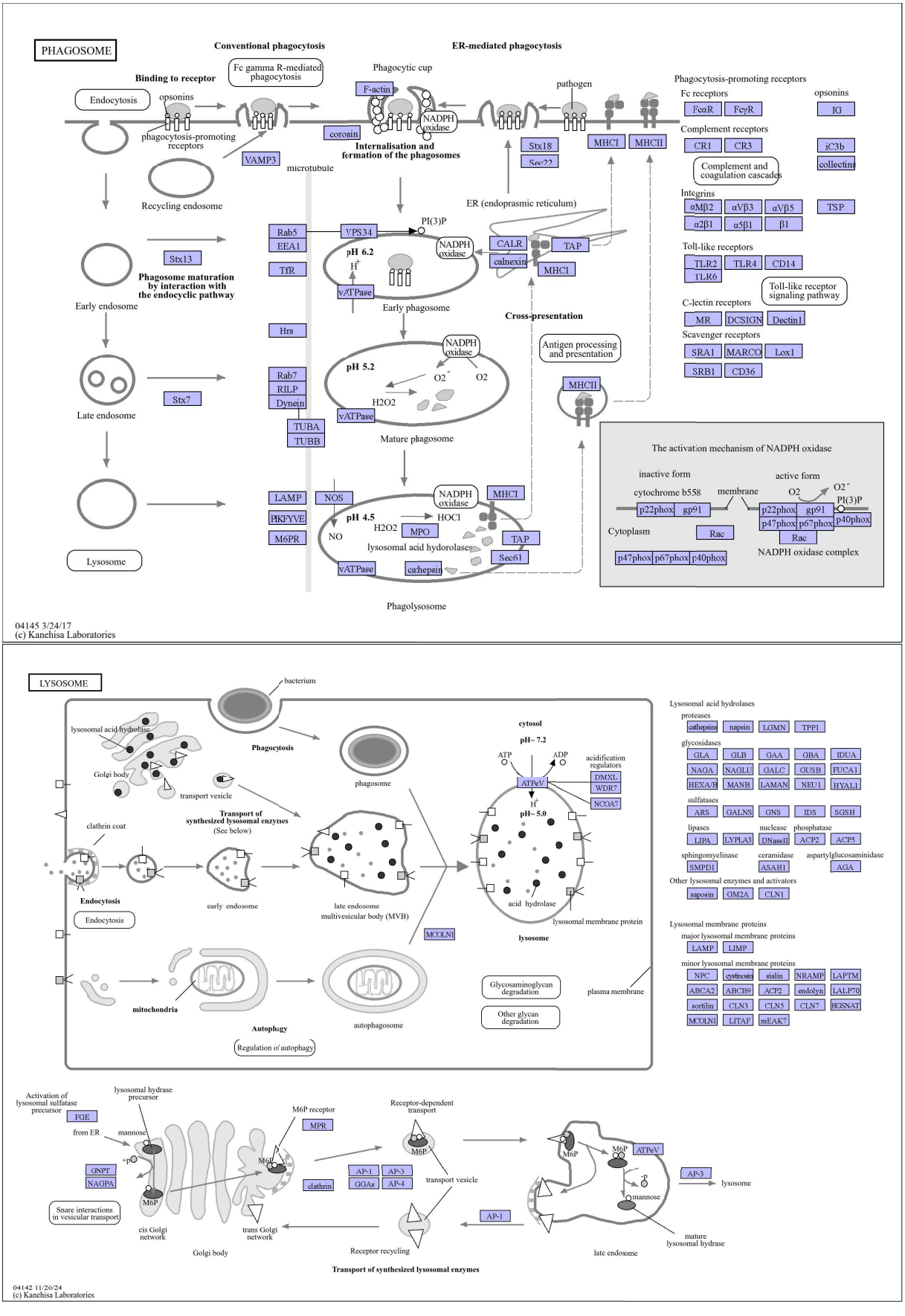

Figure 2: KEGG reference diagrams for the phagosome (top) and lysosome (bottom) pathways.

5.6.1 Phagosome

The KEGG Phagosome pathway diagram (hsa04145) is organized to illustrate the sequential phases of phagocytosis, visually separating the stages of receptor engagement, phagosome formation, and phagosome maturation/digestion.

1. **Phagocytosis-promoting Receptors (Top)**

   On the cell surface (top of the diagram), multiple phagocytosis-promoting receptors bind to particles or pathogens (often via opsonins like antibodies or complement fragments). This corresponds to the reception phase, where specific receptors on the phagocyte surface recognize ligands on the particle surface, initiating phagocytosis. On the right side, the diagram highlights groups of surface receptors such as *Fc receptors*, *complement receptors*, *integrins*, *Toll-like receptors*, and *scavenger receptors*, along with *opsonin* molecules like *immunoglobulins*, *collectins*, and *thrombospondin*. These receptor–ligand interactions are the ones that trigger endocytosis and the formation of the phagocytic cup, an actin-rich invagination of the membrane that eventually closes off to internalize the particle into a phagosome. [46] [47]

2. **Conventional Phagocytosis (Middle-Left)**

   Toward the middle-left of the diagram, the focus shifts to internalization and phagosome formation. This section, labeled as **"Conventional phagocytosis"**, highlights how the phagocytic cup closes into a membrane-bound vesicle called the nascent phagosome, a process driven by F-actin polymerization. Actin regulators like *coronin* and various signaling molecules gather at the cup, and the *NADPH oxidase* complex may start assembling on the phagosomal membrane, as shown by the label "*NADPH oxidase*" at the cup. Once the particle is engulfed, a sealed phagosome forms and begins to mature by fusing with or receiving material from endocytic compartments. The diagram shows this with an arrow linking the nascent phagosome to early endosomes via *Syntaxin13*, followed by fusion with late endosomes via *Syntaxin7*, and eventually with a lysosome. [48]

   This maturation process is illustrated through separate circular compartments labeled "Early phagosome" (pH around 6.2), "Mature phagosome" (pH around 5.2), and "Phagolysosome" (pH around 4.5), showing the gradual acidification caused by proton pumps like *v-ATPase*. Each stage includes specific molecular players. Early phagosomes recruit *Rab5*, *EEA1*, and *Vps34*, a PI3-kinase that produces *PI3P*. Late phagosomes recruit *Rab7*, *RILP*, and *dynein* for transport along microtubules, and also acquire *LAMP* proteins on their membrane. Both the drop in pH and the arrival of digestive enzymes are key hallmarks of the maturation process. [49] [11]

   The final phagolysosome, formed after lysosomal fusion, is enriched with acid hydrolases and with reactive forms of oxygen and nitrogen. These molecules, known as *reactive oxygen species (ROS)* and reactive nitrogen species *(RNS)*, contribute to creating a harsh antimicrobial environment that breaks down the engulfed microbe.

3. **ER-mediated Phagocytosis (Middle-Right)**

   The middle-right side shows an alternate phagocytosis pathway called **"ER-mediated phagocytosis"**, which is especially relevant in antigen-presenting cells. In this branch, the endoplasmic reticulum (ER) membrane contributes to forming the phagosome. During this process, ER-resident proteins such as *Syntaxin 18* and *Sec22*, which are SNAREs, mediate fusion between the ER and the membrane of the nascent phagosome. As a result, the phagosome can acquire ER proteins like *calnexin* (*CANX*) and *calreticulin* (*CALR*). [50] [51]

This pathway is linked to antigen cross-presentation, hence we can observe how the diagram shows peptides from the phagosome being transferred to the cytosol through the *Sec61* translocon, and then imported into the ER lumen by the *TAP* transporter, where they are loaded onto *MHC class I* molecules with the help of associated chaperones. These peptides are eventually presented at the cell surface via *MHC I*, as indicated by dashed arrows labeled "Cross-presentation"

Meanwhile, the conventional pathway delivers peptides to *MHC class II* loading compartments. In this case, phagolysosomes correspond to *MHC II* antigen processing, and the diagram shows *MHC II* presenting degraded antigen peptides at the cell surface.

5.6.2 Lysosome

The KEGG Lysosome pathway diagram (hsa04142) appears as a simpler diagram compared to the phagosome one. It is organized into four distinct sections, each illustrating key processes that lead to the formation of the lysosome:

1. **Synthesis and Transport of Lysosomal Enzymes (Bottom)**

   At the bottom of the diagram, the pathway begins with the synthesis of lysosomal hydrolases in the rough endoplasmic reticulum (*ER*). These enzymes undergo post-translational modifications in the Golgi apparatus, where *mannose-6-phosphate (M6P)* residues are added. *M6P* serves as a targeting signal, allowing the enzymes to bind to *M6P* receptors in the trans-Golgi network. The enzyme–receptor complexes are then packaged into *clathrin*-coated vesicles and transported to late endosomes. Upon reaching the acidic environment of the late endosome, the enzymes dissociate from the receptors and are delivered to the lysosome, where they become active in degrading various substrates. [52]

2. **Substrate Delivery to Lysosomes (Center)**

   Moving to the center of the diagram, the focus shifts to the different pathways through which substrates are delivered to lysosomes. There are three primary routes:

   - Phagocytosis: as previously explained, specialized cells such as macrophages engulf large particles like pathogens or cellular debris, forming phagosomes that subsequently fuse with lysosomes [53].
   - Endocytosis: this process involves the internalization of extracellular materials through the formation of vesicles at the plasma membrane [54].
   - Autophagy: damaged organelles and portions of the cytoplasm are sequestered into double-membrane vesicles called autophagosomes, which then fuse with lysosomes to form autolysosomes for degradation [55].

3. **Lysosomal Acid Hydrolases (Top right)**

   The top right portion of the diagram lists lysosomal acid hydrolases, which are enzymes responsible for breaking down various macromolecules within the lysosome. This enzymatic activity is represented in the center of the diagram, where we can imagine that after substrates are delivered to the lysosome via endocytosis, phagocytosis, or autophagy, they will be degraded in the lysosome by these hydrolases. The resulting breakdown products are then transported back into the cytosol to be reused, completing the recycling process [56].

4. **Lysosomal Membrane Proteins (Bottom right)**

   The bottom right section of the diagram lists all the key lysosomal membrane proteins and transporters. *Lysosome-associated membrane proteins (LAMPs)* contribute to lysosomal stability and fusion events. These are accompanied by transporters embedded in the lysosomal membrane, which facilitate the export of degradation products to the cytosol [57].

5.7 KO selection

We compiled a list of KEGG Orthology (KO) identifiers associated with the phagosome and lysosome pathways by manually retrieving all annotated KOs from each pathway's entry in the KEGG database. This process involved reviewing the pathway diagrams and extracting every KO mentioned. In total, we identified 85 unique KOs for the phagosome pathway and 99 unique KOs for the lysosome pathway, with 10 KOs shared between both. This resulted in a final list of 194 distinct KOs that we targeted in our analyses.

5.8 KO Classification

Thanks to the previously explained KEGG's phagosome and lysosome pathways deep dive, we were able to classify the 194 KOs using a two-level scheme. The first level was a general classification that grouped KOs into three broad functional categories: *Reception*, *Vacuole Formation*, and *Digestion*.

We then refined this scheme into a more detailed classification that captured additional functional information. The final set of categories included:

1. **Reception**: Proteins involved in recognizing and binding external particles or pathogens.
2. **Phagosome maturation**: Proteins that help phagosomes fuse with endosomes and lysosomes, turning them into acidic compartments that can digest what they engulfed.
3. **Lysosome membrane proteins**: Integral membrane proteins that maintain lysosomal structure and function.
4. **Digestion**: Enzymes that degrade internalized material once the phagosome fuses with the lysosome.
5. **Digestive enzyme transport**: Proteins responsible for trafficking hydrolases to lysosomes.
6. **Proton pump**: Proteins that acidify the lysosomal lumen.
7. **Oxygen radicals**: Enzymes that produce *reactive oxygen species (ROS)* to kill engulfed microbes.

5.9 Hidden Markov Models for KO finding

5.9.1 Hidden Markov Models

Hidden Markov Models (HMMs) [58] are statistical models used to represent systems that transition between a series of hidden states, each producing observable outputs. In bioinformatics,

HMMs are particularly valuable for modeling biological sequences such as DNA, RNA, or proteins, where the underlying biological processes (like gene structures or protein domains) are not directly observable but can be inferred from the sequence data [59].

An HMM comprises a finite set of states, each associated with a probability distribution over possible output symbols (emissions). Transitions between states are governed by a set of probabilities, defining the likelihood of moving from one state to another. The model is termed "hidden" because the sequence of states through which the model passes is not directly observable; only the emitted symbols are visible.

In our particular context regarding gene identification, HMMs are employed to model the statistical properties of different regions within genes, such as exons, introns, and intergenic spaces. By training an HMM on known gene structures, the model can learn the typical patterns and characteristics of these regions. Once trained, the HMM can be used to predict gene locations in unannotated sequences by identifying regions that statistically resemble the learned patterns.

Profile HMMs (or HMM profiles) [60] are a specialized form of HMMs used to model multiple sequence alignments of protein or nucleotide sequences. They capture position-specific information about conserved regions, insertions, and deletions within a sequence family.

The construction of a profile HMM involves aligning a set of related sequences to identify conserved regions. From this alignment, the model is built by assigning match, insertion, and deletion states corresponding to the alignment columns. Each state is associated with emission probabilities reflecting the frequency of each residue at that position, and transition probabilities indicating the likelihood of moving between states.

5.9.2 HHMER's hmmsearch

To identify the 194 KEGG Orthology (KO) identifiers related to the phagosome and lysosome pathways within our genomes, we used HMMER's `hmmsearch` [61]. In order to detect KOs in our genomes, this tool requires a Hidden Markov Model (HMM) profile for each KO of interest. We obtained these HMM profiles from the KOfam database, which is part of the KofamKOALA resource developed by the Kyoto University Bioinformatics Center [62]. The profiles are publicly accessible and can be easily downloaded from the GenomeNet FTP server in the form of a compressed folder containing HMM profiles for all KEGG Orthologs, summing up to more than 26,000 entries.

KEGG's HMM profiles are constructed through a systematic process designed to accurately assign KEGG Orthologs (KOs) to protein sequences. Initially, for each KO, KEGG collects a set of protein sequences annotated with that specific KO identifier. To ensure diversity and reduce redundancy, these sequences undergo clustering using CD-HIT with a 100% identity threshold, effectively removing identical sequences. The resulting non-redundant sequences are then aligned using MAFFT, a multiple sequence alignment tool, to identify conserved regions across the sequences. From these alignments, profile HMMs are generated using HMMER's hmmbuild program.

To enhance the accuracy of KO assignments, KOfamKOALA employs adaptive score thresholds for each HMM. This involves dividing the non-redundant sequences of a KO into three groups. Two groups are used to build the HMM, while the third serves as a positive dataset. Sequences from other KOs constitute the negative dataset. Using HMMER's `hmmsearch`, similarity scores (bit scores) between the HMM and sequences in both datasets are calculated. A threshold score is determined to maximize the F-measure, balancing precision and recall. This process is repeated three times, each time rotating the positive dataset, and the average of the three

threshold scores is set as the adaptive threshold for that KO.

The processing of `hmmsearch` output was carried out using R. The core of the approach is a "dynamic" filter that compares the score of each match reported by `hmmsearch` to the KO-specific threshold provided by KofamKOALA. Matches with scores higher than the corresponding threshold were kept, while those that did not pass the filter were discarded.

5.10 eggNOG-mapper for KO finding

As previously discussed, eggNOG-mapper [41] provides comprehensive functional annotations, including taxonomic classifications, that is why in our earlier analyses we primarily used it for taxonomic assignment. However, eggNOG-mapper also assigns KO identifiers to genes, offering an additional layer of functional insight. Thus, in order to complement our HMMER-based KO identifications, we used eggNOG-mapper-v2 to annotate our genomes. This approach allowed us to cross-validate the presence of the 194 phagosome and lysosome-related KOs identified by HMMER.

eggNOG-mapper assigns KOs by first identifying orthologous groups (OGs) based on homology to precomputed entries in the eggNOG database. It basically aligns the query sequences using fast tools such as DIAMOND, MMseqs2, or HMMER, and then transfers KO annotations from the best-matching orthologs. This orthology-based method improves annotation accuracy by using evolutionary context rather than simple sequence similarity.

By integrating eggNOG-mapper into our workflow, we obtained an additional independent set of KO annotations. Comparing these with our results from HMMER allowed us to assess the consistency and reliability of KO assignments across these two different methodologies.

Unlike HMMER, where we specifically searched for a predefined set of 194 HMM profiles, eggNOG-mapper takes a much broader approach by assigning all possible KOs to each gene based on orthology. Given this, we post-processed the eggNOG-mapper output to retain only the KO annotations that matched our curated list of 194 KOs, and discarding the rest.

5.11 eggNOG vs hmmsearch program comparison

Compared to eggNOG-mapper, we can say that the HMMER-based approach is generally more strict. While eggNOG-mapper transfers KO annotations based on best ortholog matches, it prioritizes coverage and speed, which can result in broader but occasionally less specific assignments. In contrast, HMMER's `hmmsearch` uses curated Hidden Markov Models with predefined threshold scores for each KO, ensuring that only high-confidence matches are kept. This harsh filtering helps minimize false positives but may also miss true but divergent homologs.

To study this, we selected 10 model organisms from EukProt's The Comparative Set (TCS) [19]:

- ***Dictyostelium discoideum***: A social amoeba widely used as a model for phagocytosis and cell signaling.
- ***Danio rerio (zebrafish)***: A vertebrate model organism popular in developmental and genetic studies.
- ***Homo sapiens***: The human genome serves as a reference for understanding phagocytosis in immune cells.

- ***Caenorhabditis elegans***: A nematode model used extensively in developmental biology and neurobiology.
- ***Drosophila melanogaster***: A fruit fly model organism used for genetic, developmental, and immune research.
- ***Saccharomyces cerevisiae***: Budding yeast used as a model for eukaryotic cell biology and molecular genetics.
- ***Schizosaccharomyces pombe***: Fission yeast, complementary to S. cerevisiae, with different cellular mechanisms.
- ***Chlamydomonas reinhardtii***: A unicellular green alga model for studying photosynthesis and flagellar motility.
- ***Arabidopsis thaliana***: A flowering plant model organism widely used in plant genetics and physiology.
- ***Tetrahymena thermophila***: A ciliate protist used in studies of nuclear dimorphism and membrane trafficking.

For each species, we retrieved the phagosome and lysosome-related KOs that are annotated in the KEGG database, treating this as the reference set of KOs each model organisms should have. We then compared the KO predictions made by HMMER and eggNOG-mapper for each genome, evaluating how well each method recovered the known KOs. In addition, we assessed which of the 194 target KOs from our curated list were detected by each method, allowing us to identify both shared and method-specific annotations, as well as potential false positives or false negatives. Finally, we also combined the results from both methods to evaluate whether their union could improve results.

For this comparison, we computed the following metrics:

- **TP (True Positive)**: The KO was present in both the tool's output and the KEGG pathway for that species (i.e., the tool correctly predicted it).
- **FN (False Negative)**: The KO was present in the KEGG pathway but not detected by the tool (i.e., the tool missed it).
- **TN (True Negative)**: The KO was found in neither the tool's output nor the KEGG pathway (i.e., the tool correctly did not predict it).
- **FP (False Positive)**: The KO was present in the tool's output but not in the KEGG pathway for that species (i.e., the tool incorrectly predicted it).

From these values, we then calculated the following evaluation metrics:

- **Accuracy**: Proportion of correct predictions among all cases: $(TP + TN)/(TP + TN + FP + FN)$.
- **Precision**: Proportion of predicted positives that are actually correct: $TP/(TP + FP)$.
- **Recall**: Proportion of actual positives that were correctly identified: $TP/(TP + FN)$.
- **F1-score**: Harmonic mean of precision and recall, balancing both: 2·(Precision·Recall)/(Precision+Recall).

5.12 Tubulin quality control

As a way to validate the effectiveness of our KO search strategies and assess the completeness of our genomes, we wanted to detect the presence of tubulin genes. Tubulins (classified by us as part of the *Vacuole Formation group*), particularly the α- and β-subunits, play a central role in the eukaryotic cytoskeleton and are found across virtually all eukaryotic organisms. Their presence is essential for various cellular processes, including maintaining cell shape, enabling intracellular transport, and facilitating cell division. Due to their consistent expression and critical role in cellular function, tubulins serve as reliable housekeeping genes and are commonly used as a control in gene expression studies. Therefore, identifying tubulins in our genomes served as a form of internal quality control, confirming both the success of our KO identification methods and the overall quality and completeness of our genomic data. [63]

5.13 Heatmap

To visualize the distribution of our KOs across genomes, we used a heatmap as a way to summarize our data into a large binary presence/absence matrix with a clear and interpretable format.

We generated this heatmap using the `pheatmap` package in R [64]. This tool allowed for flexible heatmap creation, including hierarchical clustering of rows and columns, annotation bars for metadata, and color customization. `pheatmap` automatically computes distance matrices and performs clustering (by default using Euclidean distance and complete linkage), although these options can be customized. The values in the matrix are binary (1 for presence, 0 for absence), and we applied hierarchical clustering to both dimensions to identify co-occurring genes and similar genomic profiles. We also added annotation bars to indicate functional categories (at two levels of detail) and taxonomy (with trophic modes), which helped contextualize the observed patterns.

6 Results and Discussion

6.1 SAG Selection

Using our criteria based on "purity", BUSCO and Tiara results, we ended up selecting 227 SAGs out of the 384 Alacant SAGs, 74 out of the 381 Lauven SAGs, and 75 out of the 81 SAGs from 2018. Table 1 summarizes the number of SAGs sequenced and those that were kept after selection.

Table 1: Summary of SAGs sequenced and retained after quality filtering.

Individual SAGs	**Year 2018**	**2020–Alacant**	**2020–Leuven**
SAGs sequenced	81	384	381
SAGs selected	75	227	31

6.2 Coassemblies

Figure 3 shows the change in assembly size (A), measured for contigs larger than 1000bp, and BUSCO completeness (B) between individual SAGs (red) and their corresponding coassemblies

(blue).

In total, 46 coassemblies were generated: 8 combining SAGs from 2018, 33 combining SAGs from 2020, and 5 that combine SAGs from both years. While improvements are consistently observed in both metrics, the increase in genome completeness is particularly striking. In terms of assembly size, a modest increase is actually a good sign, as it suggests that redundancy is being reduced and that the coassembly is primarily adding new, non-redundant genomic information. Together, these results highlight not only the importance of producing coassemblies to improve genome quality, but also the reliability of our approach in doing so.

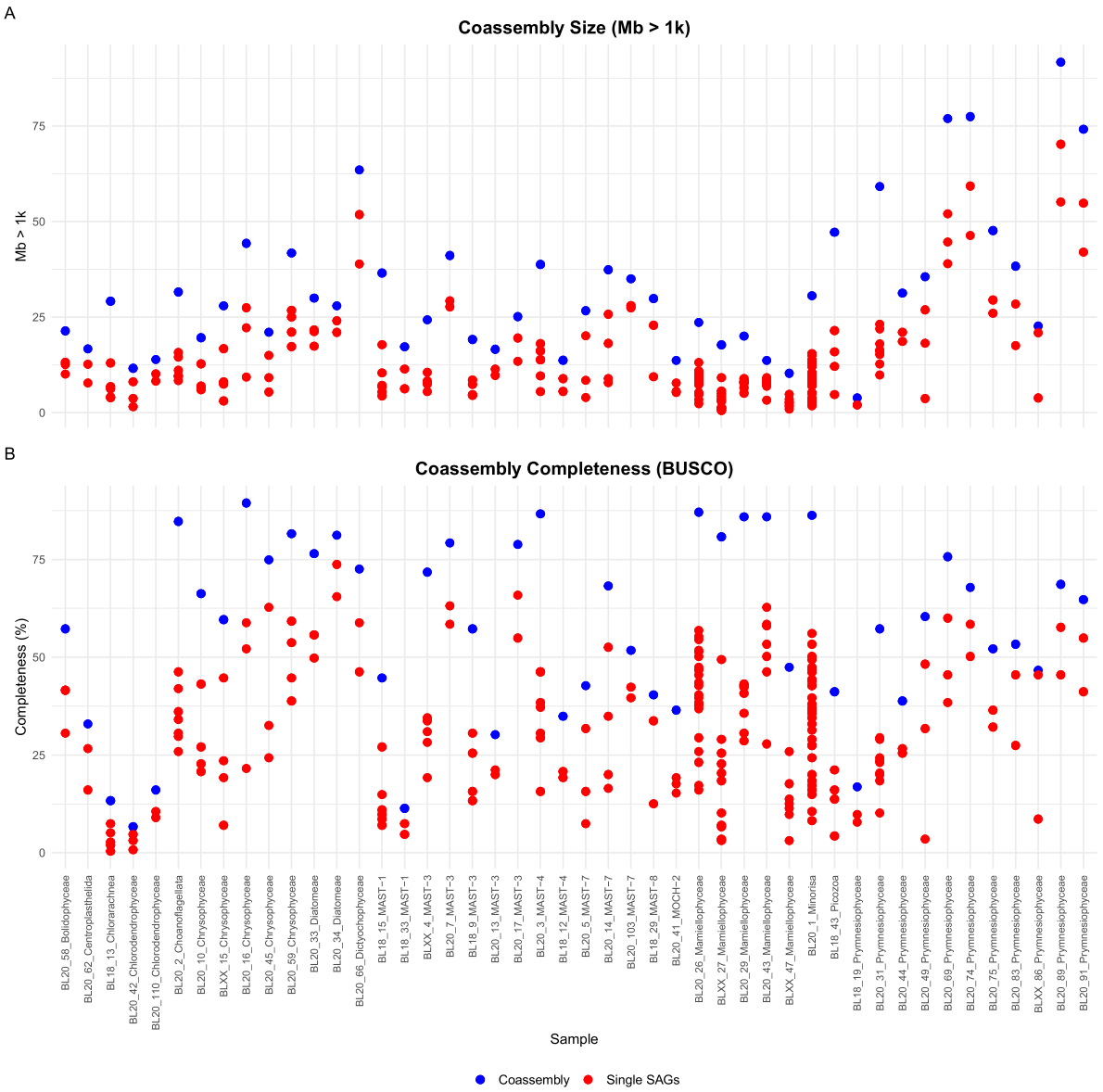

Figure 3: Coassembly quality metrics across several taxonomic groups. **(A)** Total assembly size (Mb) for each SAG and coassembly, after filtering out contigs shorter than 1000 bp. **(B)** BUSCO-based completeness percentage for the same assemblies. Blue dots represent coassemblies; red dots represent the individual SAGs that compose each coassembly.

6.3 Final Genomes from SAGs

Table 2 shows a brief summary of how our final dataset is distributed:

Table 2: Final distribution of SAGs after coassembly. "Both" refers to SAGs whose coassembly includes SAGs from both 2018 and 2020.

Year	Coassembled	Individual	Total SAGs
2018	8	29	37
2020	33	80	113
Both	5		5
Total			**155**

6.4 FG & TCS taxonomy and BUSCO completess overview

Table 3 provides a detailed overview of the number of genomes and average BUSCO completeness scores for each taxonomic group, separated by origin: from our 155 Final Genomes (FG) or the 165 genomes in EukProt's Comparative Set (TCS). In total, the dataset includes 320 genomes distributed across 26 taxonomic groups. For each group, we report in this table the number of genomes and the mean BUSCO score for both FG and TCS, along with the total number of genomes in that group. Groups represented by only one or two genomes were grouped under "Others." The "DeepBranching" category refers to taxa that fall into early-diverging or poorly resolved branches of the eukaryotic tree, making it difficult to confidently assign them to any well-established clade.

Figure 4 provides a detailed view of the contribution of each dataset to the different taxonomic groups shown in Table 3. Our Final Genomes from SAGs primarily contribute to Prymnesiophyceae, Ochrophyta, MAST, and green algae. Most of the remaining groups are mainly represented by genomes from the TCS dataset. MALV and Chlorarachniophyta are exclusively represented by SAGs, which makes sense given that these are predominantly marine lineages and tend to be underrepresented in cultured genome collections such as TCS.

Table 3: Summary of sample counts and BUSCO completeness per taxonomic group, across TCS and FG genomes.

Taxonomic Group	Samples TCS	BUSCO average TCS	Samples FG	BUSCO average FG	Total Samples
Prymnesiophyceae	5	68.4	47	37.1	52
Ochrophyta	5	81.8	27	48.2	32
MAST	3	75.0	25	37.9	28
Green algae	9	82.5	15	37.5	24
DeepBranching	15	76.7	1	7.1	16
Discoba	15	75.7	1	42.0	16
Fungi	14	94.6	1	32.9	15
Metazoa	14	98.0	1	3.1	15
Choanoflagellata	5	93.2	6	31.3	11
Bigyra	7	76.5	2	23.9	9
MALV	0	0.0	9	14.0	9
Others	7	79.5	2	26.7	9
Amoebozoa	8	81.8	0	0.0	8
Cercozoa	5	81.6	3	19.9	8
Metamonada	8	51.5	0	0.0	8
Rhodophyta	8	66.9	0	0.0	8
Streptophyta	8	96.9	0	0.0	8
Unicellular holozoa	7	90.0	1	32.9	8
Apicomplexa	6	58.4	0	0.0	6
Heterotrophic archaeplastida	2	89.2	4	19.5	6
Centroplasthelida	3	87.6	2	23.9	5
Ciliophora	4	77.4	1	5.1	5
Peronosporomycetes	5	93.8	0	0.0	5
Chlorarachnea	0	0.0	3	40.8	3
Dinoflagellata	2	77.1	1	5.9	3
Unknown	0	0.0	3	4.8	3

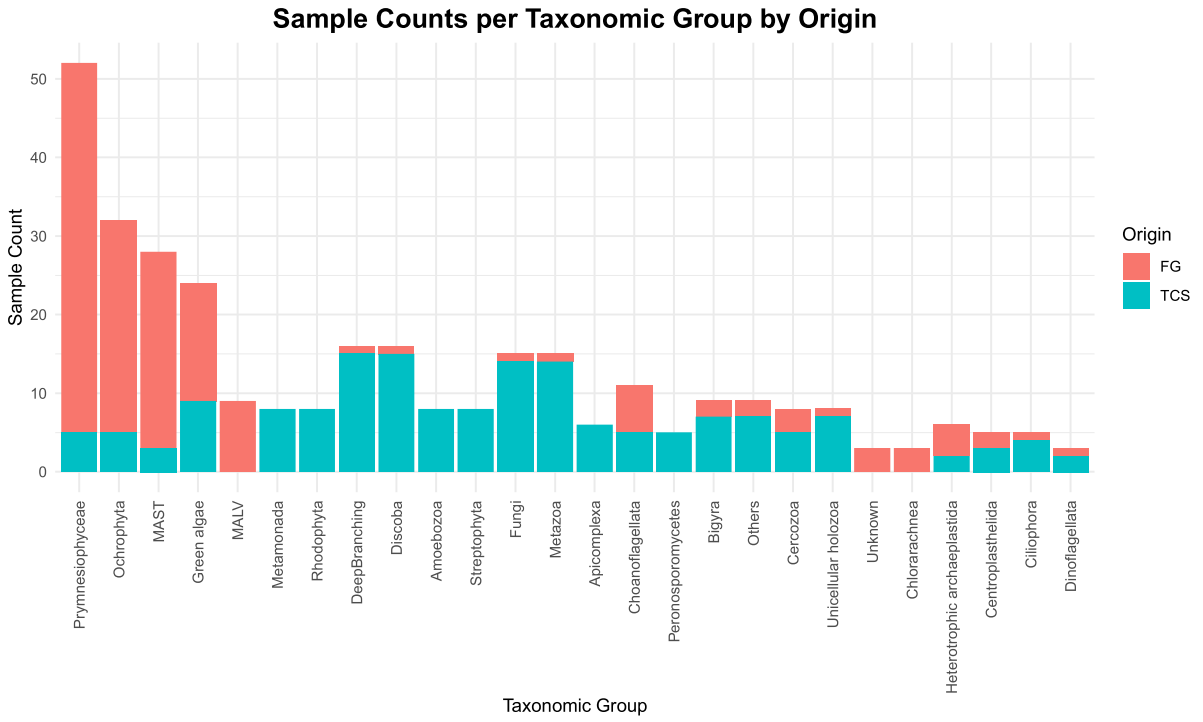

Figure 4: Distribution of samples across taxonomic groups, colored by their origin (TCS or FG).

Figure 5 shows the BUSCO completeness of each individual genome, corresponding to the taxonomic groups summarized in Table 3. As expected, TCS genomes tend to be far more complete overall. If we focus on those taxonomic groups that are primarily represented by our Final Genomes (e.g., Prymnesiophyceae, Ochrophyta, MAST)), BUSCO completeness is more variable. Still, several genomes reach high completeness levels above 75%.

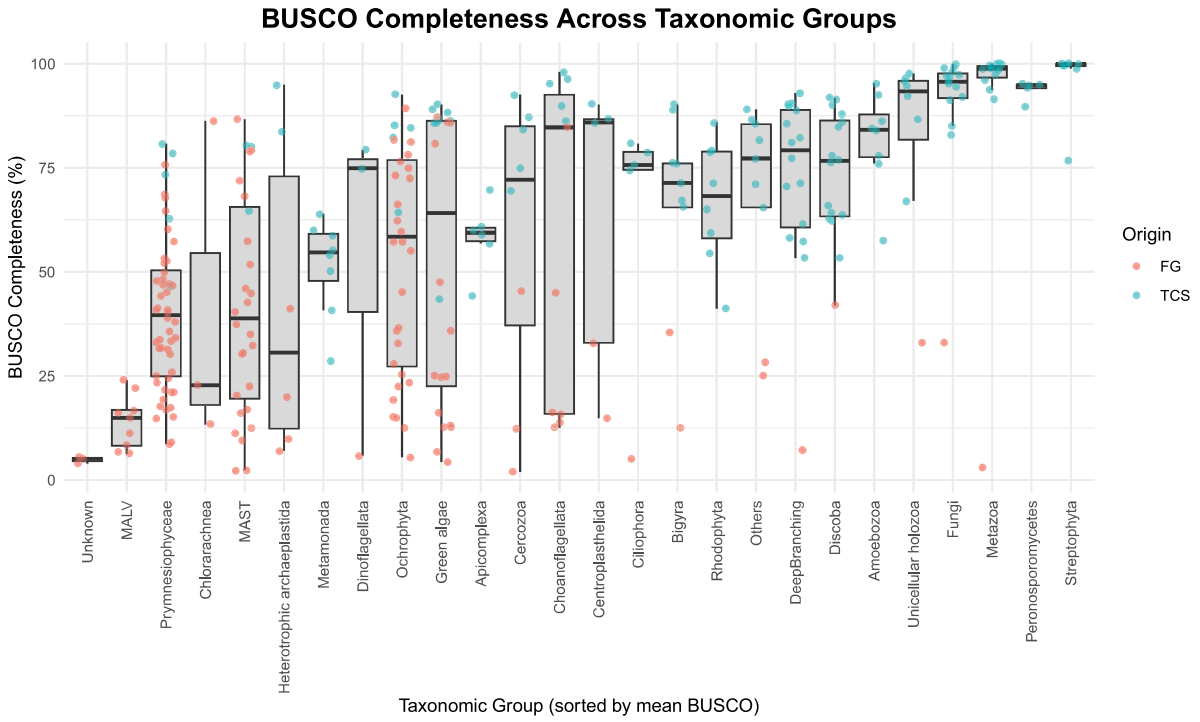

Figure 5: BUSCO completeness across taxonomic groups, grouped by sample origin and sorted by mean BUSCO completness.

6.5 Trophic Modes

Table 4 shows the distribution of trophic modes across our species, for both TCS and FG combined.

Table 4: Number of species annotated under each trophic mode, across TCS and FG.

Trophic Mode	Count
Mixed	7
Osmotroph	18
Parasite	43
Phagotroph	115
Photosynthetic	133
Unknown	4

6.6 KEGG Current Annotation Study

Table 5 summarizes the number of organisms annotated in KEGG across major taxonomic groups.

Table 5: Number of organisms annotated in KEGG by high-level taxonomic group.

Phylum	**Number of Organisms**
Animals	743
Archaea	449
Bacteria	9252
Fungi	198
Plants	170
Protists	70

Since our study focuses on eukaryotic microbial species, we excluded Bacteria and Archaea from our analysis. Additionally, within the Plants group, we separated green algae from other plant taxa. This distinction is important because green algae, particularly marine species like *Ulva* and *Chlamydomonas* play a crucial role in marine ecosystems. They are significant contributors to primary production, forming the base of the marine food web, and are involved in carbon cycling and nutrient dynamics.

Table 6 shows the final count of organisms we worked with, after discarding prokaryotes and splitting green algae from the rest of the plants.

Table 6: Final count of eukaryotic organisms considered in our study, after applying filtering criteria and taxonomic adjustments.

Group	**Number of Organisms**
Animals	743
Fungi	198
Green algae	11
Plants	159
Protists	70

6.6.1 Annotation of the 194 KOs across taxonomic categories

Figure 6 shows an overview of how the 194 KOs are annotated across our five selected KEGG taxonomic categories. We can easily see how all of our KOs are annotated in all animals. In fungi, green algae, and plants, approximately 45% of the KOs are annotated. For protists, which is the group we are more interested about, the annotation rate is slightly higher, with around 65% of our KOs being annotated.

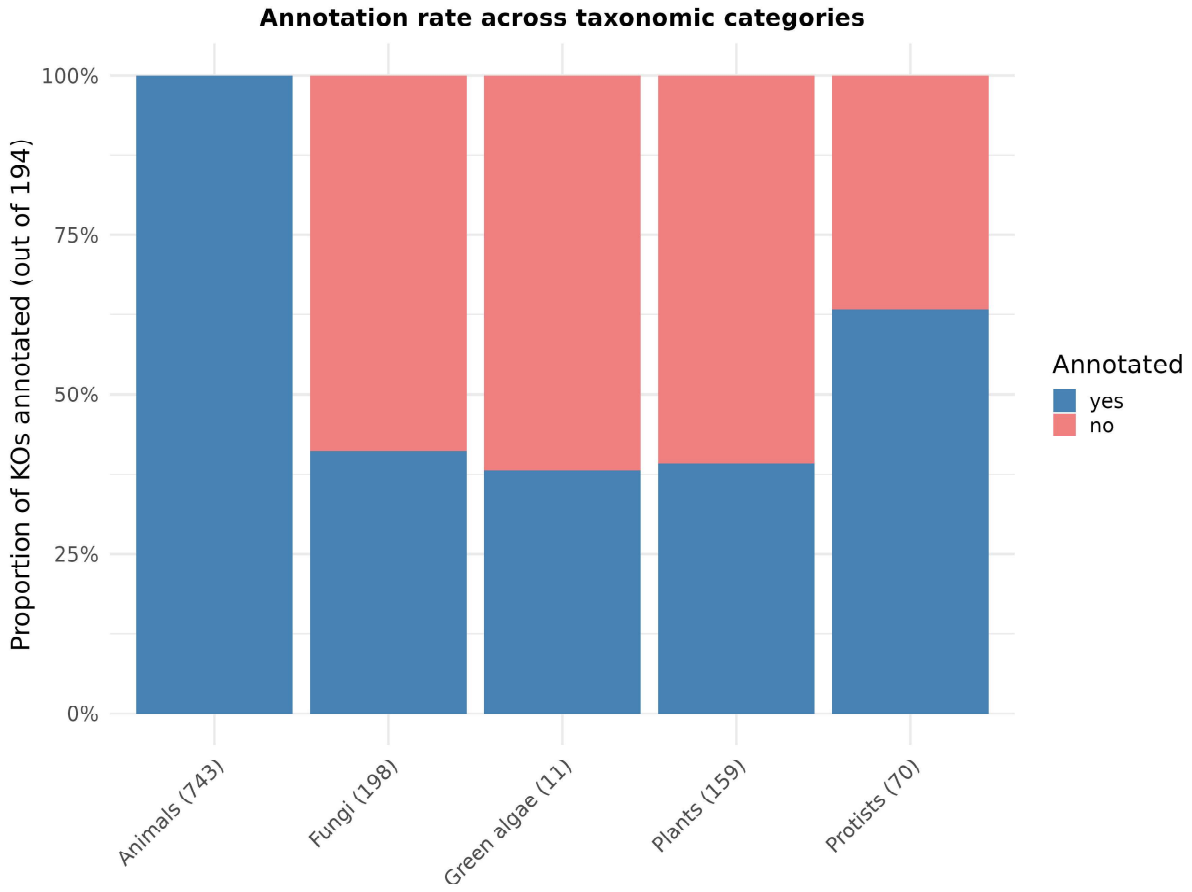

Figure 6: Proportion of annotated KOs across different eukaryotic groups. Each bar represents the 194 selected KOs and they are divided into annotated and unannotated fractions per group.

6.6.2 Individual KO annotation distribution per taxonomic group

Figure 7 shows how KO annotations are distributed across each taxonomic group. In animals, most KOs are annotated across many of KEGG's animal genomes, typically between 400 and 700 of them. In green algae, plants, and fungi, annotations are more sparse; however, when a KO is annotated, it tends to be present in nearly all species within that group. In contrast, protists show a different pattern: most KOs are annotated in only about 20 or less of the 70 protist genomes included in KEGG.

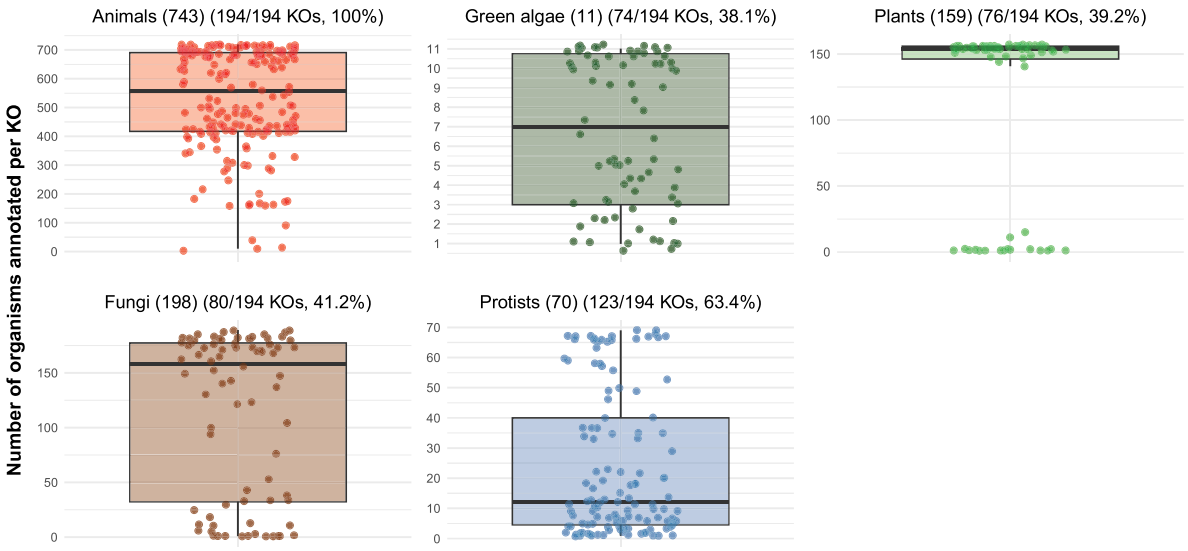

Figure 7: Boxplots with jitter showing the distribution of KO annotations across five eukaryotic groups. Each dot represents a KEGG Ortholog (KO), and its position on the y-axis indicates the number of organisms within that group in which the KO is annotated.

6.7 KO Classification

Table 7 presents the two-level classification that we assigned to our set of 194 KOs.

Table 7: Breakdown of main classifications and subcategories with their counts.

Main Classification	**Subcategory**	**Count**
Digestion (100)	Digestive enzymes	53
	Digestive enzyme transport	22
	Proton pump	17
	Oxygen radicals	8
Reception (36)	Reception	36
Vacuole Formation (58)	Phagosome Maturation	34
	Lysosome membrane proteins	24
Total		**194**

Figure 8 shows the same KEGG pathway diagrams from Figure 2 but now colored according to our classification.

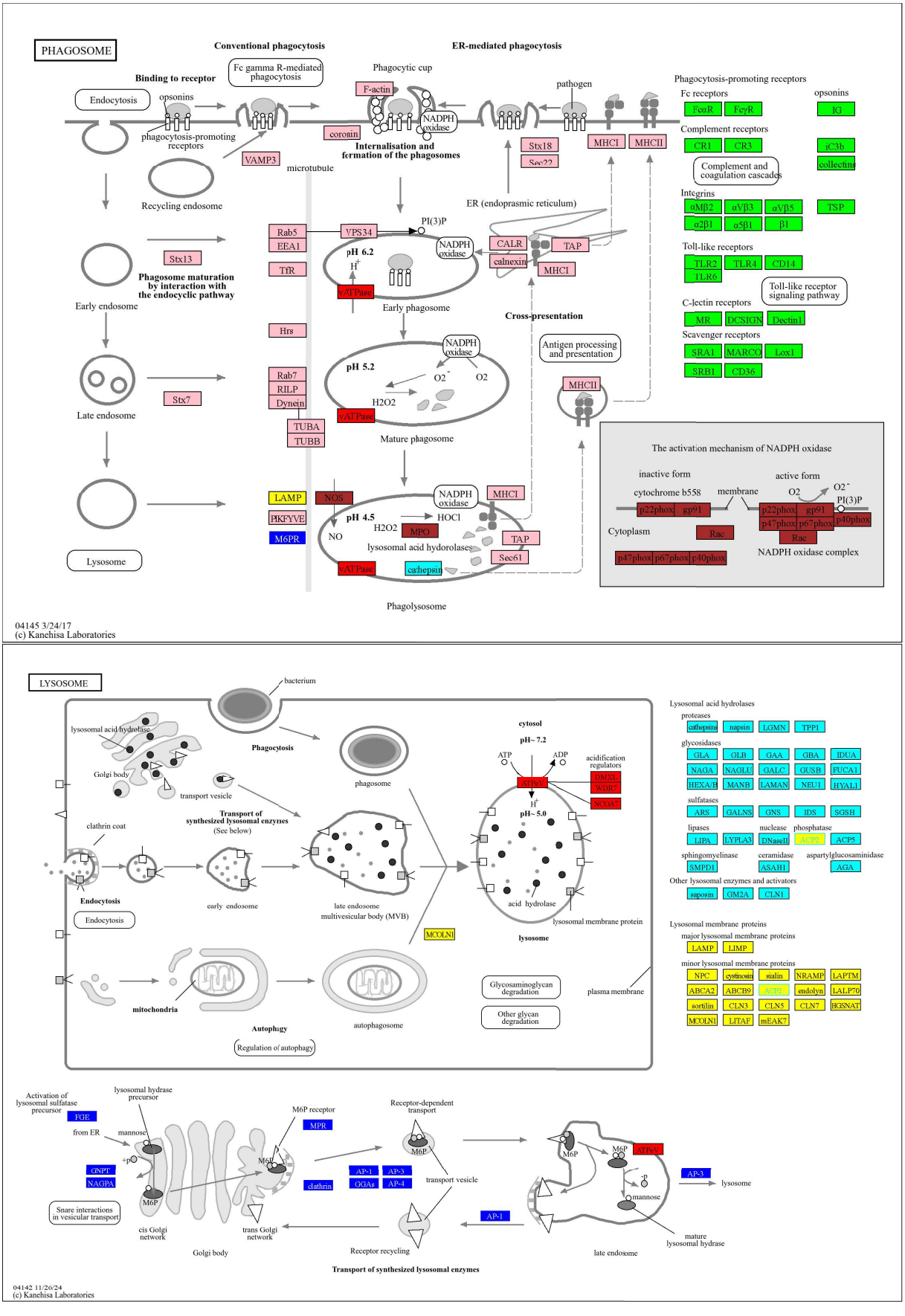

Figure 8: KEGG pathway maps highlighting functional categories of genes. (Top) Lysosome pathway (ko04142), (Bottom) Phagosome pathway (ko04145). Gene functions are color-coded as follows: *Reception* (green), *Phagosome Maturation* (pink), *Lysosome membrane proteins* (yellow), *Proton pump* (red), *Digestive enzyme transport* (blue), *Oxygen radicals* (brown), and *Digestive enzymes* (cyan).

6.8 eggNOG vs hmmsearch program comparison

6.8.1 eggNOG

Table 8: KO detection performance metrics per species using eggNOG-mapper.

Species	Code	KO Species Total	Found	TP	FP	FN	TN	Accuracy	Precision	Recall	F1 Score
Dictyostelium discoideum	ddi	75	77	72	5	3	114	0.9841	0.9351	0.9600	0.9474
Danio rerio	dre	169	179	165	14	4	11	0.9778	0.9218	0.9763	0.9483
Homo sapiens	hsa	192	143	143	0	49	2	0.7474	1	0.7448	0.8537
Caenorhabditis elegans	cel	88	112	84	28	4	78	0.9759	0.7500	0.9545	0.8400
Drosophila melanogaster	dme	94	118	90	28	4	72	0.9759	0.7627	0.9574	0.8491
Saccharomyces cerevisiae	sce	45	45	42	3	3	146	0.9843	0.9333	0.9333	0.9333
Schizosaccharomyces pombe	spo	42	44	40	4	2	148	0.9895	0.9091	0.9524	0.9302
Chlamydomonas reinhardtii	cre	41	57	38	19	3	134	0.9829	0.6667	0.9268	0.7755
Arabidopsis thaliana	ath	60	66	56	10	4	124	0.9783	0.8485	0.9333	0.8889
Tetrahymena thermophila	tet	66	70	63	7	3	121	0.9840	0.9000	0.9545	0.9265
Mean								**0.9580**	**0.8627**	**0.9294**	**0.8893**

Table 8 has the results that we obtained when using only eggNOG to find our 194 KOs in 10 eukaryotic model organisms. We can say that eggNOG's overall performance was high, with a mean accuracy of 95.8%, precision of 86.2%, recall of 92.9%, and an F1-score of 88.9%.

If we go more in depth, we can see how Accuracy was consistently high ($> 97\%$) for most species, except for **Homo sapiens** (74.7%), which showed a large number of false negatives (FN = 49) and very few true negatives (TN = 2). This may suggest that eggNOG missed a significant number of KOs in *Homo sapiens* likely due to incomplete mapping or database mismatches.

Precision was perfect (1) for *Homo sapiens*, indicating no false positives at all. However, recall was substantially lower (74.4%), showing that while the tool was conservative (avoiding overprediction), it failed to detect many real KOs.

For species like *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Tetrahymena thermophila*, performance was excellent across all metrics, with balanced high precision and recall, indicating reliable annotation by eggNOG. For other species like *Chlamydomonas reinhardtii*, we had a moderate precision (66.6%) but still high recall (92.6%), suggesting that many real KOs were detected, but at the cost of more false positives.

These results are consistent with our expectations for eggNOG. It is generally reliable for functional KO prediction across diverse eukaryotic species, but performance can vary by organism. In particular, the high recall that we observe across species suggests that eggNOG is effective at recovering most true KOs, but the variability in precision reflects differences in KO representation or annotation completeness within KEGG.

6.8.2 HMMER's hmmsearch

Table 9: KO detection performance metrics per species using HMMER's hmmsearch.

Species	Code	KO Species Total	Found	TP	FP	FN	TN	Accuracy	Precision	Recall	F1 Score
Dictyostelium discoideum	ddi	75	69	63	6	12	113	0.9362	0.9130	0.8400	0.8750
Danio rerio	dre	169	170	168	2	1	23	0.9948	0.9882	0.9941	0.9912
Homo sapiens	hsa	192	190	190	0	2	2	0.9897	1	0.9896	0.9948
Caenorhabditis elegans	cel	88	82	76	6	12	100	0.9362	0.9268	0.8636	0.8941
Drosophila melanogaster	dme	94	94	91	3	3	97	0.9843	0.9681	0.9681	0.9681
Saccharomyces cerevisiae	sce	45	41	40	1	5	148	0.9741	0.9756	0.8889	0.9302
Schizosaccharomyces pombe	spo	42	40	39	1	3	151	0.9845	0.9750	0.9286	0.9512
Chlamydomonas reinhardtii	cre	41	37	34	3	7	150	0.9634	0.9189	0.8293	0.8718
Arabidopsis thaliana	ath	60	63	58	5	2	129	0.9894	0.9206	0.9667	0.9431
Tetrahymena thermophila	tet	66	52	50	2	16	126	0.9167	0.9615	0.7576	0.8475
Mean								**0.9669**	**0.9548**	**0.9026**	**0.9267**

Table 9 shows the performance metrics obtained using HMMER's `hmmsearch` to detect our set of 194 KOs across the same 10 eukaryotic model species used with eggNOG. These results can now be directly compared with those obtained using eggNOG.

Overall, HMMER performed slightly better than eggNOG, with a mean accuracy of 96.7% (vs. 95.8% in eggNOG), precision of 95.5% (vs. 86.3%), and F1-score of 92.7% (vs. 88.9%). Although eggNOG had slightly higher recall (92.9%) than HMMER (90.3%), HMMER showed a more balanced performance, especially in terms of precision.

Focusing on *Homo sapiens*, HMMER corrected a key weakness observed in the eggNOG results. While eggNOG showed low recall (74.5%) for *Homo sapiens*, HMMER achieved nearly perfect scores (Precision = 1, Recall = 98.9%, F1 = 99.4%), highlighting its superior sensitivity for this species. Something similar happens in *Danio rerio*, where HMMER achieved slightly better precision (98.8% vs. 92.2%) while maintaining very high recall in both methods ($> 97\%$).

For other species like *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*, both tools performed well, but HMMER again offered a more balanced tradeoff between precision and recall , avoiding the tendency that eggNOG had to overpredict in some cases (e.g., FP = 28 in *Drosophila melanogaster* using eggNOG vs. FP = 3 with HMMER).

On the other hand, in species like *Tetrahymena thermophila* and *Chlamydomonas reinhardtii*, both tools had a quite reduced recall. However, HMMER still achieved higher precision in both cases (96.2% and 91.9%, respectively), suggesting a better filtering of false positives.

These results are also consistent with our expectations for HMMER, which we initially favored for KO detection based on prior experience and its interpretability. However, to avoid bias, we systematically compared its performance with eggNOG. In summary, we found that HMMER offers greater precision and a more balanced performance across species compared to eggNOG, which tended to favor recall at the expense of precision in some cases. HMMER appears more conservative but accurate, especially valuable in species where avoiding false positives is important. The improved performance in *Homo sapiens* and *Drosophila further* illustrates its strength in high-quality genome contexts.

6.8.3 eggNOG + hmmsearch

Table 10: Performance comparison using both programs (eggNOG and hmmsearch).

Species	Code	KO Species Total	Only eggnog	Only hmmer	Both	Total	TP	FP	FN	TN	Accuracy	Precision	Recall	F1 Score
Dictyostelium discoideum	ddi	75	15	7	62	84	74	10	1	109	0.9433	0.8810	0.9867	0.9308
Danio rerio	dre	169	14	5	165	184	169	15	0	10	0.9227	0.9185	1	0.9575
Homo sapiens	hsa	192	2	49	141	192	192	0	0	2	1	1	1	1
Caenorhabditis elegans	cel	88	35	5	77	117	87	30	1	76	0.8402	0.7436	0.9886	0.8488
Drosophila melanogaster	dme	94	28	4	90	122	94	28	0	72	0.8557	0.7705	1	0.8704
Saccharomyces cerevisiae	sce	45	6	2	39	47	43	4	2	145	0.9691	0.9149	0.9556	0.9348
Schizosaccharomyces pombe	spo	42	5	1	39	45	41	4	1	148	0.9742	0.9111	0.9762	0.9425
Chlamydomonas reinhardtii	cre	41	20	0	37	57	38	19	3	134	0.8866	0.6667	0.9268	0.7755
Arabidopsis thaliana	ath	60	11	8	55	74	60	14	0	120	0.9278	0.8108	1	0.8955
Tetrahymena thermophila	tet	66	21	3	49	73	64	9	2	119	0.9433	0.8767	0.9697	0.9209
Mean											**0.9263**	**0.8494**	**0.9804**	**0.9077**

In Table 10 we present performance metrics obtained after combining the outputs from eggNOG and HMMER. In this approach, a KO is considered detected if it was found by either tool. Additionaly, the table also includes additional columns showing the number of KOs uniquely found by eggNOG, uniquely found by HMMER, and those identified by both.

As a general analysis, we see how the combined approach shows a very high recall of 98.0%, which is significantly higher than using eggNOG (92.9%) or HMMER (90.3%) alone. This indicates that merging results from both tools is effective at recovering almost all true KOs.

Precision decreased to 84.9% compared to HMMER's 95.5% and eggNOG's 86.3%, as expected since merging predictions should lead to an increase in the number of false positives.

Accuracy (92.6%) and F1-score (90.8%) remain strong but both are slightly lower than HMMER's performance. This demonstrates again that while the combined strategy helps us recover more KOs, it also introduces more false positives, which reduces precision.

Seems like the benefit of combining tools is especially clear in species like *Homo sapiens*, where all 192 KOs were correctly identified with no false positives or false negatives, resulting in perfect scores across all metrics. This resolves the high false negative rate observed for *Homo sapiens* when using eggNOG alone.

In contrast, species such as *Caenorhabditis elegans* and *Drosophila melanogaster* show an increase in false positives, resulting in lower precision (∼74–77%) even though recall reached 100% for *Drosophila melanogaster*. These cases highlight the cost of merging predictions, showing that overprediction becomes more common in species with complex or less well-curated annotations.

Chlamydomonas reinhardtii continues to show reduced precision (66.7%), as seen previously, and *Tetrahymena thermophila* reaches a strong F1-score (0.92) with improved recall (96.9%) compared to individual tools.

We can end this whole section concluding that combining KO predictions from eggNOG and HMMER results in a highly sensitive annotation strategy that is capable of detecting nearly all true KOs across species. And while this approach sacrifices some precision due to a rise in false positives, it is particularly effective in ensuring comprehensive KO recovery, making it a strong option when the we aim to have the maximum coverage possible. However, for contexts like our project where high precision is critical, we believe that using only HMMER is the best option.

Figure 9 contains two plots that visually summarize the KO detection results described above, but this time applied to both of our full datasets rather than just the 10 model species. As expected, the TCS genomes show a higher number of KO hits overall, since they are higher-quality genomes from well-annotated species, as previously discussed. Additionaly, in TCS most KOs are detected by both eggNOG and HMMER (red), indicating strong agreement between

tools.

Our Final Genomes from SAGs also show a good number of KO hits, although not as many as TCS. Most of these hits are detected by eggNOG alone (green), indicating that eggNOG is the main contributor in this dataset. We also observe a decent number of KOs found by both tools (red), while HMMER-only hits (blue) remain low. This actually happens in both datasets, indicating that most of HMMER's value may come from confirming predictions also made by eggNOG.

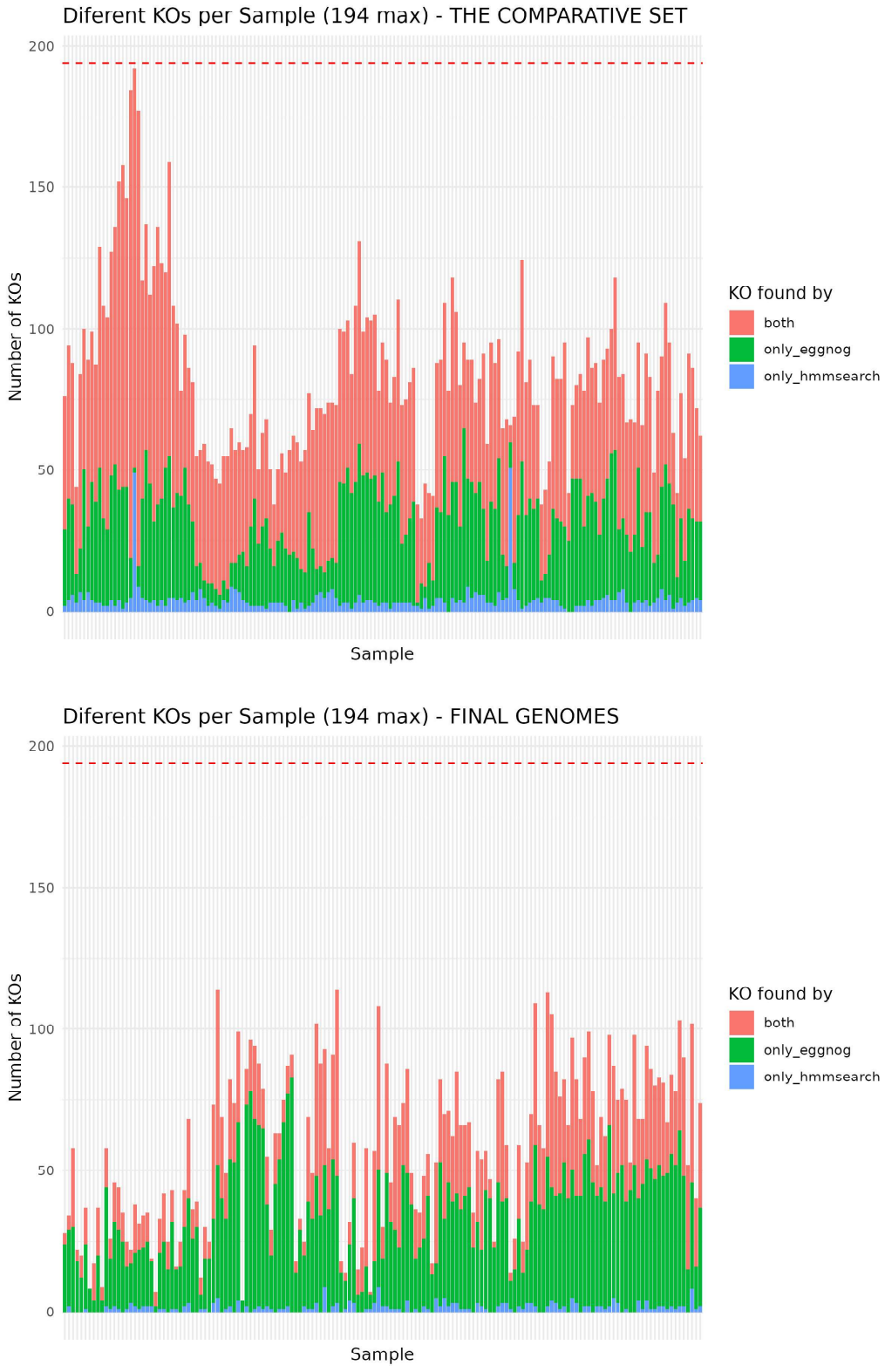

Figure 9: KO recovery per sample using eggNOG and HMMER. **Top:** 165 genomes from The Comparative Set (TCS). **Bottom:** 155 Final Genomes (FG) from our SAG dataset. KOs are grouped by detection method: only by eggNOG (green), only by HMMER (blue), or by both (red).

6.9 Tubulin quality control

Figure 10 shows the results of our tubulin-based quality control analysis, performed using only hmmsearch. To keep the analysis clean and consistent, we used only hmmsearch, which we consider the most robust and reliable tool based on the performance observed in our comparisons.

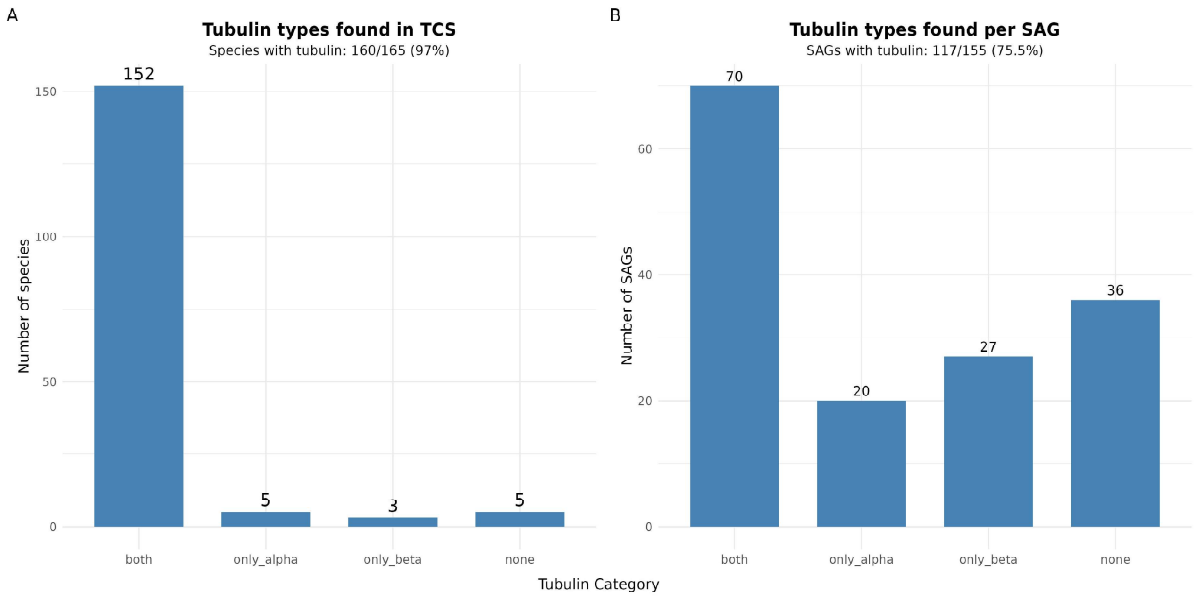

Figure 10: Tubulin classification across species and SAGs. **(A)** Tubulin types found in TCS. **(B)** Tubulin types found in our SAGs.

In Plot A, we observe that 97% of TCS genomes (160 out of 165) contain tubulin genes, with the vast majority (152 species) having both alpha and beta tubulin. Only a small number of species contain only one type or none at all. This confirms once again the completeness and high quality of the TCS genomes.

Plot B shows more variable results across our 155 Final Genomes (SAGs): while 117 SAGs (75.5%) contain at least one tubulin gene, only 70 have both types. A good fraction of SAGs contain only alpha (20) or only beta (27), and 36 SAGs show no detectable tubulin. These results are consistent with the expected lower completeness of SAGs, and support the use of tubulin presence as a straightforward quality indicator.

6.10 Number of founds per BUSCO completeness

Figure 11 shows the relationship between genome completeness (BUSCO score) and the number of KOs detected across all genomes using only hmmsearch. Each point represents a genome, colored by taxonomic group and shaped by origin, depending if the genome comes from our SAG Final Genomes (green circles) or from The Comparative Set (red triangles).

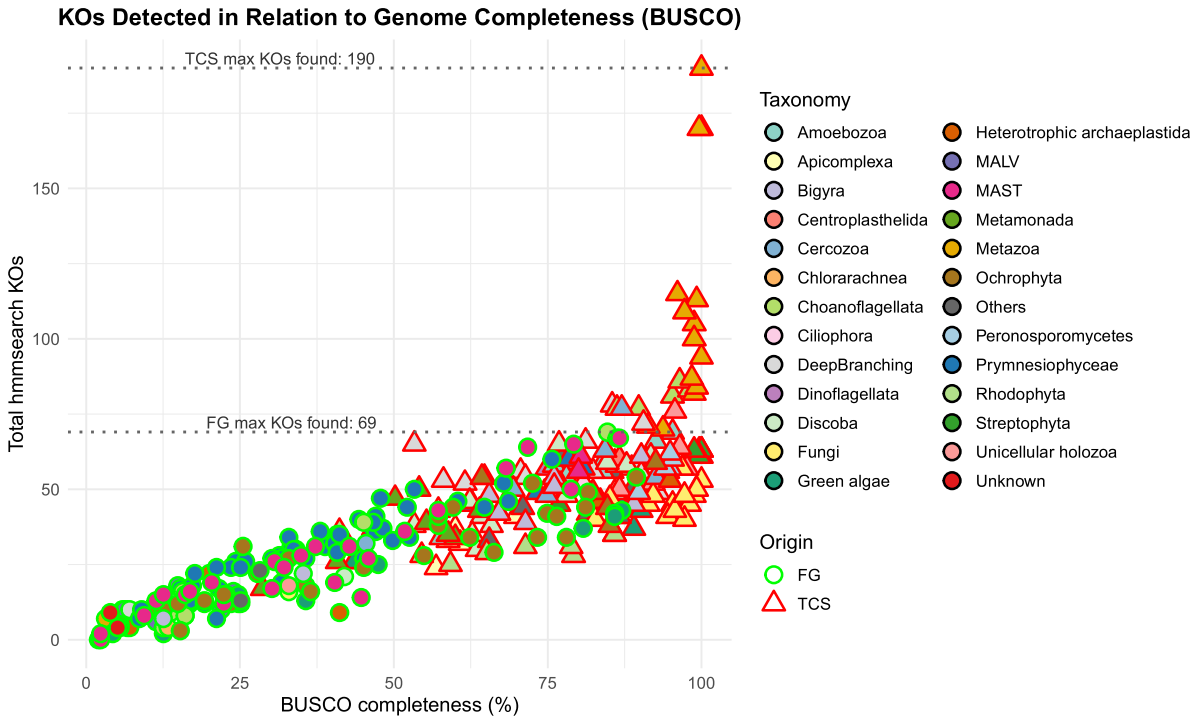

Figure 11: KOs detected in relation to genome completeness (BUSCO). BUSCO completeness vs. KO count with taxonomy and origin. Dotted lines indicate the maximum KO count per origin.

As expected, we observe a clear positive trend: genomes with higher BUSCO completeness tend to have more KOs detected. TCS genomes, which are generally of higher quality, cluster in the upper right region of the plot, with many reaching near-complete BUSCO scores. In contrast, our Final Genomes show greater variability, with many clustering at lower completeness values and recovering fewer KOs, while a small fraction reaches relatively high KO counts.

Since we are only using HMMER's `hmmsearch`, the number of detected KOs is generally lower compared to when eggNOG annotations are also included. However, we chose to rely solely on hmmsearch, as we believe it is the tool that provides more reliable results.

6.11 Heatmap

In Figure 12, we show a heatmap representing the distribution of our KOs across a subset of our genomes. To focus on high-confidence detections, this subset only included genes found with hmmsearch in genomes with more than 50% BUSCO completeness. After applying this filter, our dataset included 36 genomes from FG and 159 from TCS, meaning that most of the data shown in the heatmap comes from TCS, but FG genomes are also represented.

The heatmap clusteres both rows (genomes) and columns (KOs), which helps highlight presence patterns. At the top, two annotation bars display our functional classifications. The first one (categories_3) groups KOs into the 3 main big categories that we previously discussed (Reception, Digestion, and Vacuole Formation). The second annotation bar (categories_7) shows the break down of the 3 big categories into 7 smaller and more specific ones.

If we losely observe cthis classification bars, we can see how reception KOs are only present in a small group of Metazoans (*Homo sapiens* among them), which makes sense because *Homo*

sapiens have reception for immunologic purposes, while the rest of our groups should not have them because they dont do phagocytosis by immuno, they do it just to eat so ehy eat whatever they find, they dont need to recognize them.

If we closely observe the top classification bars, we can see how reception KOs are only present in a small group of Metazoans (animals), with *Homo sapiens* among them. This makes biological sense becasue in animals, phagocytosis often plays an immunological role, where cells must recognize and respond to specific external targets (e.g., pathogens). In contrast, the rest of the eukaryotes in our dataset just engage in phagocytosis mainly for feeding purposes. These organisms don't need such specific recognition mechanisms, they just engulf whatever particles or prey they encounter.

Proton pump genes and digestive enzymes appear to be the most conserved KOs across all taxonomic groups, highlighting their essential role in the phagocytic process. This can be explained as proton pumps are key in the phagolysosome step, one of the most critical step for the effective digestion of the engulfed material, which is done by digestive enzymes.

If we look at the taxonomy bar on the left side, we can generally see that species from the same taxonomic group tend to cluster together. This suggests that closely related organisms often share similar patterns of KO presence and absence, reflecting shared evolutionary history or similar functional requirements. If we focus on taxonomy alone, we see that the clustering is not perfectly clean, as several groups show a bit of variability. However, since we also have trophic mode information of every group, we can actually see how species with the same trophic strategy often cluster together. This suggests that the is still biologically meaningful, just not strictly taxonomic. If we focus on each color, it is clear that the main "core" of KO presence comes from phagotrophic species. We also observe a distinct cluster composed mostly of photosynthetic organisms, showing that different functional gene patterns are associated with different trophic modes.

To conclude, we have seen how this heatmap gives us a clear overview of how key functions are distributed across our dataset, revealing both a conserved "core" set of genes shared by many groups, and other set of genes whose presence varies more between species. In the future, we believe that these patterns could help us better understand the functional evolution of phagocytosis in our set protist.

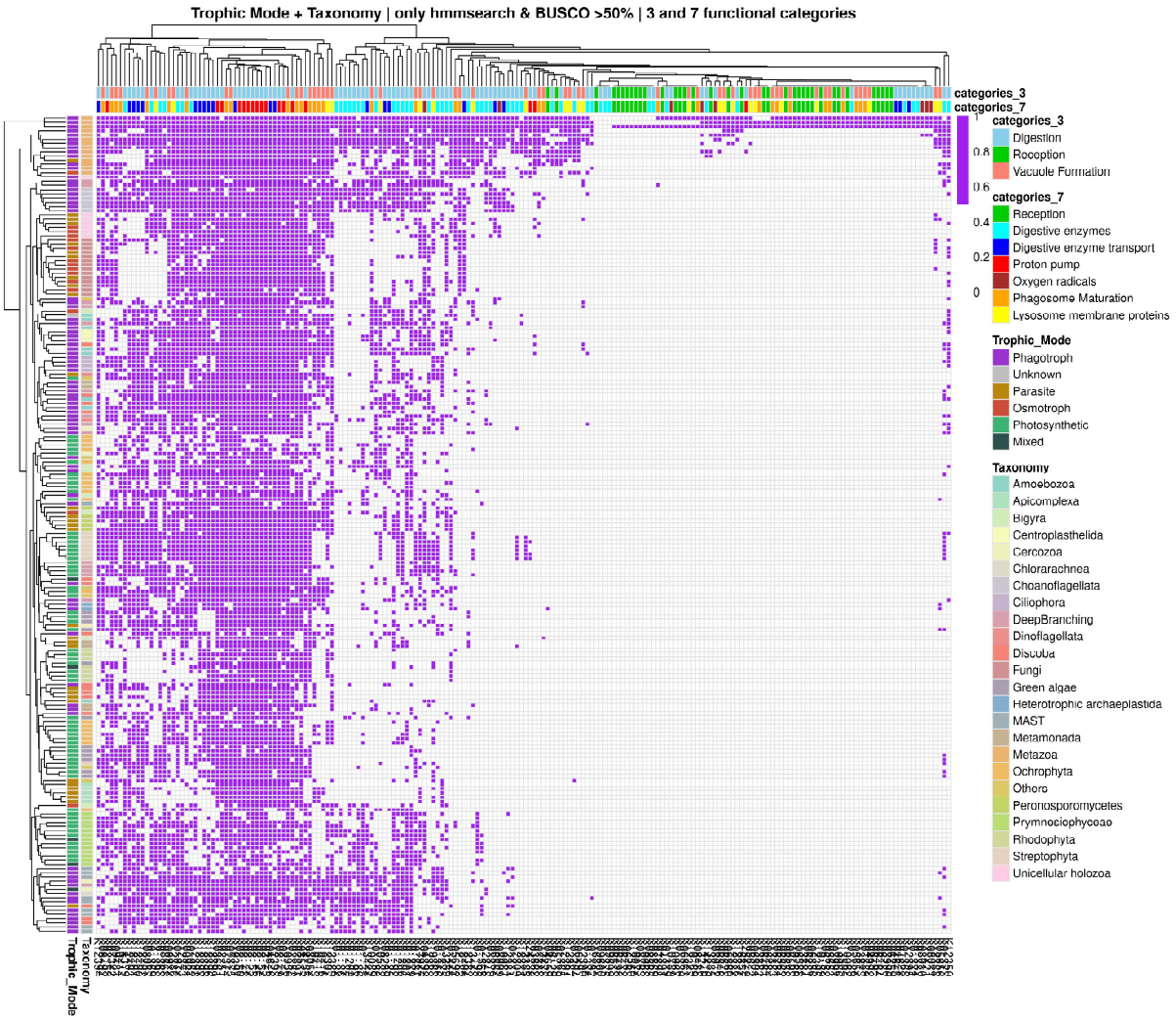

Figure 12: Functional KO presence across species based on hmmsearch-only results filtered by BUSCO completeness $> 50\%$. Columns are annotated with two functional classification layers: a broader 3-category level and a more detailed 7-category level. Row colors represent taxonmic groups and their trophic mode.

7 Conclusions

This project aimed to assess the presence and distribution of phagocytosis-related genes in marine microbial eukaryotes. Using a curated set of 194 KEGG Orthologs (KOs) extracted from KEGG pathway maps, we systematically explored their occurrence across both well-characterized reference genomes (TCS) and newly assembled Single-cell Amplified Genomes (SAGs).

Our results lead to several key conclusions. First, we observed a clear taxonomic bias in KEGG's current annotation, with animal species dominating the phagocytosis-related KO landscape. While this is not surprising given the historic focus of molecular biology on model organisms, it raises concern and makes us wonder if KEGG annotations can be really extrapolated to less-studied lineages like protists. In fact, we found that while animals had nearly complete coverage of all 194 KOs, the annotation rate dropped significantly in other groups, especially protists, which averaged only 65% coverage.

Despite this, our analyses demonstrated that many of these supposedly "animal-specific" genes are, in fact, detectable in marine microbial eukaryotes. Using our SAGs pipeline followed by the formation of coassemblies, we recovered a substantial portion of phagocytosis-related KOs from our SAGs, especially when genome completeness (assessed by BUSCO) was high.

The comparison between eggNOG and hmmsearch as KO identification tools revealed the strengths and weaknesses of each program. eggNOG offered broader coverage, often detecting more KOs, but it came with a higher false-positive rate. HMMER's hmmsearch, on the other hand, provided higher precision and more correct matches. Notably, combining the results of both methods increased overall recall to 98%, but also introduced more false positives, highlighting the classic precision-recall tradeoff. We conclude that for contexts like ours where KO presence must be confirmed with high confidence, HMMER's `hmmsearch` should remain as the preferred tool.

One particularly informative result came from the tubulin quality control analysis. Tubulin genes, essential in eukaryotic cells, were detected in 97% of TCS genomes and in 75.5% of Final Genomes coming from our SAGs. This finding reinforces the interpretation that SAG assemblies, while powerful tools for exploring uncultured diversity, are often not that complete and require careful validation. However, having found tubulin in three-quarters of our SAGs is still a strong result and reaffirms that our pipelines to build and refine SAGs are reliable.

As we expected, our findings showed a strong positive correlation between genome completeness and KO recovery. This again remarks the importance of high-quality assemblies for functional inference and highlights the utility of SAG coassemblies in approaching reference-level genome quality for uncultured taxa.

Finally, producing a heatmap helped us bring together all of our results into an interpretable visualization. By classifying KO presence/absence patterns based on both taxonomy and function, we were able to observe how species from the same taxonomic group and trophic mode tend to share similar sets of genes. These shared patterns highlight that many KOs, especially those related to core phagocytic functions, are quite conserved within several taxonomic groups.

In conclusion, our project demonstrated both the feasibility and the value of extending KEGG-based gene annotations to environmental protist genomes, specially if an appropriate care is taken in method selection and data validation. It also highlighted the potential of SAGs, when properly filtered, cleaned and assembled, to shed light on the genomic machinery behind ecologically important traits like phagocytosis in organisms that were previously impossible to study.

8 Future work

Our project provided a broad overview of the presence or absence of phagocytosis-related KOs across protist genomes. However, our analysis was mainly limited to binary presence/absence data, which, while useful as a starting point, leaves us with several important questions that we aim to explore in the near future.

One promising direction would be to explore the 71 KOs (out of the 194) that are not annotated for any protist in KEGG. We could make a more specific investigation to determine whether these missing KOs are also absent in our SAGs. And if they are indeed missing in our data, we believe it would be valuable to identify which protist genomes from KEGG contain them in order to assess whether those species differ significantly from the ones we have in our SAGs. This deeper study could help clarify whether their absence reflects genuine biological differences, issues with genome completeness, or limitations in the annotation process.

We think it would also be valuable to go beyond the binary question of whether a KO is present or not, and instead study how present it is by looking into copy number, functional importance, or even enrichment of specific KOs within certain taxonomic groups.

We also noted that most KOs appear to map to a single gene in the genome, suggesting a one-to-one relationship, but in a few rare cases, it seems that multiple KOs may be assigned to the same gene. We believe it would be interesting to explore if these cases are due to annotation artifacts or reflect real biological features such as multifunctional proteins or rare gene fusion events.

Another area worth exploring is the systematic absence of expected core genes in certain samples. For example, the lack of tubulin in 25% of our SAGs.

Finally, we think that future studies could benefit from a more targeted focus on specific taxonomic groups. For example, groups like MAST showed interesting patterns in KO presence and absence, thus we think that understanding which KOs are consistently found or missing in such lineages could provide valuable insights into their unique biology. In a similar way, examining sample-level differences may reveal why some SAGs contain significantly more KOs than others, and whether these differences relate to species identity, genome quality, or even ecological adaptations.

9 Limitations

One of the main limitations of our study, as discussed in the early sections, was the nature of our data, which was entirely composed by unculturable protist species. Since these organisms cannot be grown in laboratory conditions, we must rely on Single Amplified Genomes (SAGs) to study them. And while SAGs have allowed us to explore genomes that would otherwise be inaccessible, they are far from perfect and they come with several challenges. Specifically, it is very difficult to obtain complete genomes from SAGs due to biases in amplification and uneven coverage. As a result, this made it harder for us to obtain strong conclusions about the presence or absence patterns of specific KOs across our species. If we had access to more complete genomes from a broader range of protists, we could study which genes are consistently present or absent across lineages with more confidence.

10 Ethical and Social Impact

This project addresses an important ethical concern in science: the overrepresentation of certain taxa in biological databases, which can bias downstream analyses and limit the global scientific understanding of life. It also promotes open science by relying on public genomic resources (e.g., EukProt, KEGG, KOfam, eggNOG) and transparent pipelines, ensuring that the methods and results can be reproduced and built upon.

From a social perspective, our project highlights the importance of microbial biodiversity in sustaining marine ecosystems, which are critical for human well-being. A better understanding of microbial functions such as phagocytosis and nutrient cycling has implications for food webs, climate regulation, and ocean health. This issues are increasingly relevant in the face of environmental crises and society's dependence on marine resources.

11 Environmental Sustainability

Our project is connected to sustainability through its focus on marine microbial eukaryotes and their roles in oceanic nutrient cycling and carbon sequestration. With our work on examining genes related to phagocytosis, we indirectly explored how these protists contribute to the microbial loop and the biological carbon pump, key components of global biogeochemical cycles.

Additionally, the use of Single-cell Amplified Genomes (SAGs) and computational annotation avoids invasive sampling or the need for continuous cultivation in laboratory conditions, minimizing the ecological footprint of the research. This non-disruptive approach aligns with the principles of sustainable science, particularly in marine environments where preserving ecosystem integrity is key.

We also contribute to long-term environmental monitoring efforts, particularly through the use of SAGs derived from long-term observatories like Blanes Bay Microbial Observatory. By improving functional genomic insight into these data, the research helps ensure that monitoring programs are not only taxonomically rich but also functionally informative, which is critical for guiding future conservation efforts.

References

[1] Evelyn B Sherr and Barry F Sherr. Significance of predation by protists in aquatic microbial food webs. Technical report, 2002.

[2] Ramon Massana, Aurelie Labarre, David López-Escardó, Aleix Obiol, François Bucchini, Thomas Hackl, Matthias G. Fischer, Klaas Vandepoele, Denis V. Tikhonenkov, Filip Husnik, and Patrick J. Keeling. Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. *ISME Journal*, 15:154–167, 1 2021.

[3] Jackie L. Collier and Joshua S. Rest. Swimming, gliding, and rolling toward the mainstream: Cell biology of marine protists, 2019.

[4] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. Technical report, 2000.

[5] Aleix Obiol, David López-Escardó, Eric D. Salomaki, Monika M. Wiśniewska, Irene Forn, Elisabet Sà, Dolors Vaqué, Martin Kolísko, and Ramon Massana. Gene expression dynamics of natural assemblages of heterotrophic flagellates during bacterivory. *Microbiome*, 11, 12 2023.

[6] Aurelie Labarre, David López-Escardó, Francisco Latorre, Guy Leonard, François Bucchini, Aleix Obiol, Corinne Cruaud, Michael E. Sieracki, Olivier Jaillon, Patrick Wincker, Klaas Vandepoele, Ramiro Logares, and Ramon Massana. Comparative genomics reveals new functional insights in uncultured mast species. *ISME Journal*, 15:1767–1781, 6 2021.

[7] Roberto J. Botelho and Sergio Grinstein. Phagocytosis, 7 2011.

[8] Jens Boenigk and Hartmut Arndt. Bacterivory by heterotrophic flagellates: community structure and feeding strategies. Technical report, 2002.

[9] William F. Martin, Sriram Garg, and Verena Zimorski. Endosymbiotic theories for eukaryote origin, 9 2015.

[10] Alan Aderem and David M Underhill. Mechanisms of phagocytosis in macrophages. Technical report, 1999.

[11] Otilia V Vieira, Roberto J Botelho, and Sergio Grinstein. Phagosome maturation : aging gracefully. Technical report, 2002.

[12] Ariel Savina and Sebastian Amigorena. Phagocytosis and antigen presentation in dendritic cells, 10 2007.

[13] Miao Pan, Matthew P. Neilson, Alexander M. Grunfeld, Phillip Cruz, Xi Wen, Robert H. Insall, and Tian Jin. A g-protein-coupled chemoattractant receptor recognizes lipopolysaccharide for bacterial phagocytosis. *PLoS Biology*, 16, 5 2018.

[14] Victor de Jager and Roland J. Siezen. Single-cell genomics: Unravelling the genomes of unculturable microorganisms. *Microbial Biotechnology*, 4:431–437, 7 2011.

[15] Yann Marcy, Cleber Ouverney, Elisabeth M Bik, Tina Lö Sekann § ¶, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, Darren Platt, Philip Hugenholtz, David A Relman, and Stephen R Quake. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated tm7 microbes from the human mouth. Technical report, 2007.

[16] Ramunas Stepanauskas, Michael E Sieracki, and David M Karl. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Technical report, 2007.

[17] Pierre E Galand, Carmen Gutiérrez-Provecho, Ramon Massana, Josep M Gasol, and Emilio O Casamayor. Inter-annual recurrence of archaeal assemblages in the coastal nw mediterranean sea (blanes bay microbial observatory). 55:2117–2125, 2010.

[18] Ina Maria Deutschmann, Anders K. Krabberød, Francisco Latorre, Erwan Delage, Cèlia Marrasé, Vanessa Balagué, Josep M. Gasol, Ramon Massana, Damien Eveillard, Samuel Chaffron, and Ramiro Logares. Disentangling temporal associations in marine microbial networks. *Microbiome*, 11, 12 2023.

[19] Daniel J Richter, Cédric Berney, Jürgen F H Strassert, Yu-Ping Poh, Emily K Herman, Sergio A Muñoz-Gómez, Jeremy G Wideman, Fabien Burki, and Colomban De Vargas. M e r s e n n e peer community journal section: Genomics eukprot: A database of genome-scale predicted proteins across the diversity of eukaryotes. 2022.

[20] Michael S. Rappé and Stephen J. Giovannoni. The uncultured microbial majority, 2003.

[21] Anne-Kristin Kaster and Morgan S Sobol. Microbial single-cell omics: the crux of the matter.

[22] Kyle F. Edwards, Qian Li, Kelsey A. McBeain, Christopher R. Schvarcz, and Grieg F. Steward. Trophic strategies explain the ocean niches of small eukaryotic phytoplankton. *Proceedings of the Royal Society B: Biological Sciences*, 290, 1 2023.

[23] Ben A. Ward and Michael J. Follows. Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. *Proceedings of the National Academy of Sciences of the United States of America*, 113:2958–2963, 3 2016.

[24] Fabien Burki, Andrew J. Roger, Matthew W. Brown, and Alastair G.B. Simpson. The new tree of eukaryotes, 1 2020.

[25] Daniel J. Richter and Nicole King. The genomic and cellular foundations of animal origins, 2013.

[26] Aleix Obiol Plana. Return of the marine heterotrophic flagellates: diversity, distribution and gene expression patterns. Technical report.

[27] Quentin Carradec, Eric Pelletier, Corinne Da Silva, Adriana Alberti, Yoann Seeleuthner, Romain Blanc-Mathieu, Gipsi Lima-Mendez, Fabio Rocha, Leila Tirichine, Karine Labadie, Amos Kirilovsky, Alexis Bertrand, Stefan Engelen, Mohammed Amin Madoui, Raphaël Méheust, Julie Poulain, Sarah Romac, Daniel J. Richter, Genki Yoshikawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Sarah Searson, Silvia G. Acinas, Emmanuel Boss, Michael Follows, Gabriel Gorsky, Nigel Grimsley, Lee Karp-Boss, Uros Krzic, Stephane Pesant, Emmanuel G. Reynaud, Christian Sardet, Mike Sieracki, Sabrina Speich, Lars Stemmann, Didier Velayoudon, Jean Weissenbach, Olivier Jaillon, Jean Marc Aury, Eric Karsenti, Matthew B. Sullivan, Shinichi Sunagawa, Peer Bork, Fabrice Not, Pascal Hingamp, Jeroen Raes, Lionel Guidi, Hiroyuki Ogata, Colomban De Vargas, Daniele Iudicone, Chris Bowler, and Patrick Wincker. A global ocean atlas of eukaryotic genes. *Nature Communications*, 9, 12 2018.

[28] Urania Christaki, Claude Courties, Ramon Massana, Philippe Catala, Philippe Lebaron, Josep M. Gasol, and Mikhail V. Zubkov. Optimized routine flow cytometric enumeration of heterotrophic flagellates using sybr green i. *Limnology and Oceanography: Methods*, 9:329–339, 2011.

[29] Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. Seqkit: A cross-platform and ultrafast toolkit for fasta/q file manipulation. *PLoS ONE*, 11, 10 2016.

[30] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19:455–477, 5 2012.

[31] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: Quality assessment tool for genome assemblies. *Bioinformatics*, 29:1072–1075, 4 2013.

[32] Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. Busco: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31:3210–3212, 10 2015.

[33] Michał Karlicki, Stanisław Antonowicz, and Anna Karnkowska. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, 38:344–350, 1 2022.

[34] Scott McGinnis and Thomas L. Madden. Blast: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32, 7 2004.

[35] Aleix Obiol, Caterina R. Giner, Pablo Sánchez, Carlos M. Duarte, Silvia G. Acinas, and Ramon Massana. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, 20:718–731, 5 2020.

[36] Caterina R. Giner, Vanessa Balagué, Anders K. Krabberød, Isabel Ferrera, Albert Reñé, Esther Garcés, Josep M. Gasol, Ramiro Logares, and Ramon Massana. Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology*, 28:923–935, 3 2019.

[37] Ramon Massana, Angélique Gobet, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Aurélie Chambouvet, Richard Christen, Jean Michel Claverie, Johan Decelle, John R. Dolan, Micah Dunthorn, Bente Edvardsen, Irene Forn, Dominik Forster, Laure Guillou, Olivier Jaillon, Wiebe H.C.F. Kooistra, Ramiro Logares, Frédéric Mahé, Fabrice

Not, Hiroyuki Ogata, Jan Pawlowski, Massimo C. Pernice, Ian Probert, Sarah Romac, Thomas Richards, Sébastien Santini, Kamran Shalchian-Tabrizi, Raffaele Siano, Nathalie Simon, Thorsten Stoeck, Daniel Vaulot, Adriana Zingone, and Colomban de Vargas. Marine protist diversity in european coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, 17:4035–4049, 10 2015.

[38] Caterina R. Giner, Massimo C. Pernice, Vanessa Balagué, Carlos M. Duarte, Josep M. Gasol, Ramiro Logares, and Ramon Massana. Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. *ISME Journal*, 14:437–449, 2 2020.

[39] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41, 1 2013.

[40] Tomáš Brůna, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. Braker2: Automatic eukaryotic genome annotation with genemark-ep+ and augustus supported by a protein database. *NAR Genomics and Bioinformatics*, 3:1–11, 3 2021.

[41] Carlos P. Cantalapiedra, Ana Hernandez-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, 38:5825–5829, 2021.

[42] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, 7, 4 2016.

[43] Konstantinos T Konstantinidis and James M Tiedje. Genomic insights that advance the species definition for prokaryotes. Technical report, 2005.

[44] Masato Kogawa, Masahito Hosokawa, Yohei Nishikawa, Kazuki Mori, and Haruko Takeyama. Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Scientific Reports*, 8, 12 2018.

[45] Dan Tenenbaum, Jeremy Volkening, and Bioconductor Package Maintainer. *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*, 2025. R package version 1.49.0.

[46] Carlos Rosales and Eileen Uribe-Querol. Phagocytosis: A fundamental process in immunity, 2017.

[47] Justin J. Lim, Sergio Grinstein, and Ziv Roth. Diversity and versatility of phagocytosis: Roles in innate immunity, tissue remodeling, and homeostasis. *Frontiers in Cellular and Infection Microbiology*, 7, 5 2017.

[48] Valentin Jaumouillé and Sergio Grinstein. Molecular mechanisms of phagosome formation. *Microbiology Spectrum*, 4, 5 2016.

[49] Johannes Westman and Sergio Grinstein. Determinants of phagosomal ph during host-pathogen interactions, 1 2021.

[50] Pierre Guermonprez, Loredana Saveanu, Monique Kleijmeer, Jean Davoust, Peter van Endert, and Sebastian Amigorena. Er–phagosome fusion defines an mhc class i cross-presentation compartment in dendritic cells. *Nature*, 425(6956):397–402, 2003.

[51] Ignacio Cebrian, Geraldine Visentin, Nicolas Blanchard, Mabel Jouve, Alexandre Bobard, Catarina Moita, Jost Enninga, Luis F. Moita, Sebastian Amigorena, and Ariel Savina. Sec22b regulates phagosomal maturation and antigen crosspresentation by dendritic cells. *Cell*, 147:1355–1368, 12 2011.

[52] Maria Francisca Coutinho, Maria João Prata, and Sandra Alves. Mannose-6-phosphate pathway: A review on its role in lysosomal function and dysfunction, 4 2012.

[53] Carmine Settembre, Alessandro Fraldi, Diego L. Medina, and Andrea Ballabio. Signals from the lysosome: A control centre for cellular clearance and energy metabolism, 5 2013.

[54] Jatta Huotari and Ari Helenius. Endosome maturation, 8 2011.

[55] Noboru Mizushima and Masaaki Komatsu. Autophagy: Renovation of cells and tissues, 11 2011.

[56] Marion M. Bradford. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72:248–254, 1976.

[57] Eeva Liisa Eskelinen. Roles of lamp-1 and lamp-2 in lysosome biogenesis and autophagy, 10 2006.

[58] Lawrence R. Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[59] Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. Technical report, 2009.

[60] Sean R Eddy. Profile hidden markov models. Technical report.

[61] Sean R. Eddy. Accelerated profile hmm searches. *PLoS Computational Biology*, 7, 2011.

[62] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36:2251–2252, 4 2020.

[63] Pavla Binarová and Jack Tuszynski. Tubulin: Structure, functions and roles in disease, 10 2019.

[64] Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2025. R package version 1.0.13.