The Ribosomal Operon Database (ROD): A full-length rDNA operon database extracted from genome assemblies

Anders K. Krabberød*, Embla Stokke, Ella Thoen, Inger Skrede, Håvard Kauserud

Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, Norway

*Corresponding author

Abstract

Current rDNA reference sequence databases are tailored towards shorter DNA markers, such as parts of the 16/18S marker or the ITS region. However, due to advances in long-read DNA sequencing technologies, longer stretches of the rDNA operon are increasingly used in environmental sequencing studies to increase the phylogenetic resolution. There is, therefore, a growing need for longer rDNA reference sequences. Here, we present the Ribosomal Operon Database (ROD), which includes eukaryotic full-length rDNA operons fished from publicly available genome assemblies. Full-length operons were detected in 34.1% of the 34,701 examined eukaryotic genome assemblies from NCBI. In most cases (53.1%), more than one operon variant was detected, which can be due to intragenomic operon copy variability, allelic variation in non-haploid genomes, or technical errors from the sequencing and assembly process. The highest copy number found was 5,947 in *Zea mays*. In total, 453,697 unique operons were detected, with 69,480 operon clusters remaining after intragenomic clustering at 99% sequence identity. The operon length varied extensively across eukaryotes, ranging from 4,136 to 16,463 bp, which will lead to considerable PCR bias during PCR amplification of the entire operon. Clustering the full-length operons revealed that the different parts (i.e., 18S, 28S, the hypervariable region V4 of 18S, and ITS) provide divergent taxonomic resolution, with 18S and the V4 region being the most conserved. The Ribosomal Operon Database (ROD) will be updated regularly to provide an increasing number of full-length rDNA operons to the scientific community.

Keywords: DNA metabarcoding; amplicon sequencing; environmental sequencing; reference sequences; ribosomal DNA

Introduction

DNA metabarcoding, high-throughput sequencing of short DNA regions amplified from environmental DNA followed by taxonomic annotation using known reference sequences, is a powerful approach to explore microbial communities (Taberlet et al., 2018). Different parts of the ribosomal rDNA operon are the predominant markers used in DNA metabarcoding studies, such as the V4 region commonly used for microeukaryotes (Hadziavdic et al., 2014; Krabberød et al., 2022; Logares et al., 2012) or the ITS regions commonly used for fungi (Schoch et al., 2012). In the last decade and a half, relatively short DNA markers within the 200-400 bp range have typically been used, mainly due to inherent length limitations of relevant sequencing technologies. For taxonomic annotation of these markers, designated sequence databases such as PR2 (Guillou et al., 2013), RDP (Maidak et al., 2001), SILVA (Quast et al., 2012), and UNITE (Nilsson et al., 2019) have been constructed, including millions of relatively short rDNA reference sequences.

In contrast to short-read DNA metabarcoding, long-read sequencing technologies make it possible to sequence longer stretches of the rDNA region (Jamy et al., 2020), even the entire rDNA operon (Gaonkar & Campbell, 2024; Hooper et al., 2023; Wurzbacher et al., 2019), with high precision (Callahan et al., 2019). This calls for new reference sequences covering the whole rDNA operon. One way to construct full-length operon reference sequences is to extract this information from genome assemblies; with increasing speed and precision, genome assemblies are becoming publicly available, driven by large-scale initiatives such as the Earth Biogenome Project (Lewin et al., 2022), ERGA

(Formenti et al., 2022), Mycocosm (Grigoriev et al., 2014) and Darwin Tree of Life (Darwin Tree of Life Project Consortium, 2022). However, finding and extracting complete, full-length operons from genome assemblies can be difficult since technical errors may occur during sequencing and bioinformatics processing, rendering the ribosomal operon incomplete. Moreover, since the rDNA region is a multi-copy marker, intragenomic variants can exist (Lindner et al., 2013; Simon & Weiss, 2008) due to incomplete homogenization of the repeats (Elder & Turner, 1995; Ganley & Kobayashi, 2007).

Using longer stretches of the rDNA operon as a DNA metabarcoding marker can also pose extra challenges during PCR, partly due to its length, making it harder for the polymerase enzyme to read and copy the entire region. Moreover, various PCR biases, such as operon length differences among the target taxa, can skew the results. Although not thoroughly studied across the Eukaryotic Tree of Life, a considerable length variation occurs between lineages (Hall et al., 2022). One reason for this variation is introns, which have been shown to be prevalent in both 18S and 28S (Bhattacharya et al., 2000; Cannone et al., 2002; Hedberg & Johansen, 2013; Karpov et al., 2017).

Different parts of the rDNA operon mutate and evolve at different rates (Clark, 1987; Hall et al., 2022; Hillis & Dixon, 1991). Compared to the relatively conserved 18S and 28S regions, the ITS region evolves rapidly and can provide highly resolved taxonomic information down to the species level (Kauserud, 2023; Schoch et al., 2012). Different parts of the 18S and 28S genes also evolve at varying rates based on their position in the final rRNA molecule; in 18S, there are nine regions known to be hypervariable (named V1 to V9) (Hadziavdic et al., 2014; Nelles et al., 1984; Rajilić-Stojanović et al., 2009), while in 28S, there are twelve named D1-D12 (Chilton et al., 2003; Hassouna et al., 1984). These regions have different lengths and variability.

For instance, the V4 region in 18S shows significant sequence variability (Hadziavdic et al., 2014) and has been extensively used as a marker in DNA metabarcoding projects (Hadziavdic et al., 2014). D1 and D2 of the 28S have similar properties and have been used to identify species (Sonnenberg et al., 2007). Knowing the level of variability and which taxonomic resolution the markers provide is crucial in DNA metabarcoding analyses. However, we have limited knowledge about the relative resolution of the different parts of the rDNA operon and how this may vary between taxonomic groups.

In this study, we introduce the full-length Ribosomal Operon Database (ROD) constructed from publicly available genome assemblies of eukaryotes. In addition to providing a high number of full-length operon reference sequences for the scientific community, we report on different features of the rDNA operons, including multiple operon variants in many genomes and length differences across eukaryotes. We also compare the level of molecular variation and how this affects the number of sequence clusters across different parts of the rDNA operon (18S, 28S, V4, ITS1, and ITS2), as well as the entire operon.

Materials and methods

Fishing reference operons

Initially, Barrnap (Seemann, n.d.) was used to identify the ribosomal RNA genes 18S, 5.8, and 28S from a selection of approximately 1000 genomes from Phycocosm (Grigoriev et al., 2021) and Mycocosm (Grigoriev et al., 2014). Following the initial identification, the output from Barrnap was analyzed to manually extract genomic regions where 18S, 5.8S, and 28S were arranged as a continuous operon on the same contig or scaffold, and a new Hidden Markov Model (HMM) was built to represent the complete operon. This model was constructed with hmmer3 (*HMMER 3.4 (Aug 2023)*, n.d.) from an MAFFT (Katoh & Standley, 2013) alignment of 940 full-length ribosomal operons, representing a wide array of eukaryotic lineages and covering the major branches of the eukaryotic tree of life.

All eukaryotic genomes available in NCBI as of December 2023, in total 34,701, were downloaded and screened with the new full-length operon model using *hmmscan*. Regions matching the ribosomal operon were found in 17,107 of the genomes. To maintain the focus on complete, full-length operons, Barrnap (Seemann, n.d.) was used to search for operons with incomplete 18S or 28S; in the cases where either of these genes were marked as "partial," the operons were excluded from further analysis. After the filtration step, we were left with 453,697 operon sequences that we could say with reasonable confidence were full length.

To simplify the database, we utilized clustering to group operons present in multiple copies within genomes. Specifically, we used VSEARCH (Rognes et al., 2016) to cluster these operons within each genome assembly at a 99% similarity threshold.

Eliminating contaminants

Several measures were taken to eliminate contaminants and incorrectly annotated genomes. This was done by comparing the centroid representatives from the clustered operons to the best hits in NCBI GenBank using blastn (Altschul et al., 1990). For each operon, 50 hits were retrieved from the NCBI nucleotide database and sorted by bitscore, e-value, and identity. The top hit for each query was then selected.

Taxonomic information was standardized using the R package *taxonomizr* (Sherrill-Mix, 2023) for both the retrieved genomes and the blast hit. The taxonomy identifier (TaxId) from the genomes in NCBI and the best hits for each centroid were used in the function *getTaxonomy(),* and the option for desired taxonomic levels was set to retrieve *superkingdom*, *kingdom*, *phylum*, *subphylum*, *class*, *order*, *family*, *genus*, and *species*. A comparative analysis of taxonomic assignments revealed possible discrepancies between the genome annotations and blast results, particularly in the higher taxonomic levels from *superkingdom* down to *order*. Operons with conflicting taxonomic assignments at these levels were excluded from the database to maintain its integrity. Additionally, operons with incomplete taxonomic information were quality-controlled by manually examining the top 50 blast hits, and a taxonomic assignment was selected to determine whether it was the majority among those matches. A total of 8614 centroids representing 34,015 ribosomal operons from 2,500 genomes were removed as contaminants. For 1,075 genomes, we uncovered one or more contaminant operons in addition to the operon corresponding to the expected taxonomy of the sequenced genome.

Streamlining taxonomy

Finally, the taxonomic assignments for the remaining operons were standardized to 9 different levels, following the setup in PR2 (Guillou et al., 2013). The levels used in ROD are *domain, supergroup division, subdivision, class, order, family, genus,* and *species*. The higher levels *domain*, *supergroup*, *division*, and *subdivision* correspond roughly to *superkingdom, kingdom, phylum,* and *subphylum* in NCBI but are adjusted to recent suggestions for the organization of the major branches of the Tree of Life (Adl et al., 2019; Burki et al., 2019). In addition to the taxonomic levels, ROD contains the GenBank *assembly_id* and the *taxid* from NCBI for each genome. After checking for length, clustering, and removing contaminants, the database comprised 69,480 operon variants, representing 453,697 copies of the ribosomal operon from 11,935 genomes.

Genetic distance

The operon variants were aligned separately for each genome assembly with more than three variants using MAFFT (Katoh & Standley, 2013). After aligning, the genetic distance was calculated for each alignment using the *dist.dna()* function (Paradis & Schliep, 2019) with the TN93 model, which assumes distinct rates for both kinds of transition (A ↔ G versus C ↔ T) and transversions (Tamura & Nei, 1993), and the base rates were estimated from the data. The distance calculation was weighted by the geometric mean of the size of the variants, i.e.

$$D_{weigthed} = \frac{D_{original}}{\sqrt{size_A \times size_B}}$$

where $size_A$ and $size_B$ represent the copy number of two given operon variants and $D_{original}$ the unweighted genetic distance between the two. The average genetic distance was then calculated for each subdivision on the Tree of Life (Table 1).

Clustering analyses

To compare the level of molecular variation and its impact on the number of retrieved clusters, we extracted different parts (18S, 28S, V4, ITS1, and ITS2) from the entire operons. To retrieve these regions, we first built an alignment of the operon variants using MAFFT. Next, we marked the boundaries of each region based on a set of primers and the identity score for the alignment. The primers used to identify the V4 region were 3NDf (Cavalier-Smith et al., 2009), and TAReukREV3 (Stoeck et al., 2010); the end of 18S (which is also the start of ITS1) was confirmed by EukB (Medlin et al., 1988), while the primer fITS7 was used to split ITS1 and ITS2 in 5.8S (Ihrmark et al., 2012). Finally, the beginning of 28S/end of ITS2 was confirmed with the primer ITS4 (White et al., 1990) (Fig. 1)

In addition to regions 18S, 28S, ITS1, ITS2, and subregion V4, a superregion ITS1-5.8S-ITS2 was extracted and clustered. Full intragenomic operon centroids and subsectioned datasets were processed separately. Following a precursory clean-up to remove sequences with more than 20 consecutive N-bases, sequences were dereplicated and clustered using VSEARCH (Rognes et al. 2016). Clustering was performed from 95.0% to 100% with 0.5 increments. Cluster centroids were used as representatives to collect taxonomic annotation for each cluster and compared at 97.0% similarity, as this is a frequently applied threshold in DNA metabarcoding.

Choosing a representative sequence

As a final resource, we chose one representative sequence from each genome assembly with the highest copy number of all the operon variants from a genome if more than one variant existed. In a few genome assemblies (68), two or more variants with the highest copy number were equally abundant, and they were not included in the reduced reference dataset in ROD.

Data visualization

Visualization of the number of genomes, operon variants, copy numbers, and genetic distance on the eukaryotic Tree of Life was done with the package *metacoder* (Foster et al., 2017) in R v4.3.3 (R Core Team, 2024). The ridgeplots were made in R with *ggplot2* (Wickham, 2016) and *ggridges* (Wilke, 2024), and the colors were based on the package *wesanderson* (Ram & Wickham, 2023).

Results

While a signal for the ribosomal operon was found in 17,107 out of the 34,701 genome assemblies, full-length rDNA operon(s) were detected in 11,935 genome assemblies (34.4%). The operons were first clustered intragenomically at 99.0% sequence similarity to remove tentative technical errors from sequencing and assembly, and the number of clusters and the size of each cluster were registered. The number of clusters represents the number of operon variants per genome, and the size represents the number of copies found for a given variant. The set of operon variants (69,480), as well as a reference set including only the most abundant operon variant per genome (11,867), are available in the ROD database (https://github.com/krabberod/ROD). The release used in this study is ROD v1.0 (GoldenRod), which is also deposited on Zenodo (https://zenodo.org/records/10951604).

Figure 2 shows the distribution of the detected operons across major taxonomic groups. Fungi, which are intensively sequenced partly due to their small genomes, represent 55.5% of the genomes in ROD, followed by Metazoa (22.7%) and Viridiplantae, 17.0% (Fig. 2, Table 1). A single copy of the operon was found in 5,601 (46.9%) of all the assemblies, while 53.1% had multiple copies. Among the genomes with multiple copies, most had between 2-10 copies (3,385), 2,241 had between 11 and 100 copies, 708 had more than 100 copies, and 92 genomes had more than 1,000 copies (Fig. 2). Most copies were observed in genome assemblies of Viridiplantae (Fig. 2, Table 1). The highest copy number, 5,947, was found in the genome of *Zea mays* (genome assembly id GCA_024505845 (Zhao et al., 2022)). Also, many Metazoa possessed several copies (Table 1), with the insect *Clusia tigrina* having the highest number with 4,691 copies (genome assembly id GCA_920105625 (Crowley et al., 2023)).

The number of operon variants per genome varied considerably between major taxonomic groups (Fig. 2). While 8,356 out of the 11,935 genome assemblies with a full-length operon had a single operon variant, 2,733 had between 2-10 variants, 730 between 11 and 100 variants, and 117 had more than 100 variants. As with the copy numbers, Viridiplantae included the most variant numbers (Fig. 2). Two plant species were the only genomes with more than 1000 operon variants: *Thlaspi arvense* with 1,017 variants (genome assembly id GCA_018983045 (Geng et al., 2021)) and *Silybum marianum* with 2,097 variants (genome assembly id GCA_001541825.1 (Shim et al., 2023))(Table 1 and Supplementary table 1). In contrast, fungal genome assemblies generally had fewer variants; the highest number was 81, found in the ascomycete *Leptosphaeria maculans* (genome assembly id GCA_022343315 (Tian et al., 2024)).).

To chart the amount of intragenomic variation, we calculated the genetic distance between all operon copies for each assembly separately (Fig. 2). Among the most abundant taxonomic groups sampled, Viridiplantae possessed the highest average genetic distance between copies (0.033). In comparison, the different sub-groups in Metazoa and Fungi had, on average, between 0.001-0.027 (Metazoa) and 0-0.12 (Fungi) sequence distance between copies (Table 1).

The highest genetic divergence for a single assembly was observed in the assembly of the fungi *Neurospora* sp. CHS-2018c. Here, we found 30 operon variants, all single copies, with an average genetic distance of 0.215 (supplementary table 1).

The length of the operon varied extensively across the organismal groups (Fig.3). The shortest operon found was 4,136 bp in the plant *Capsicum annuum* (assembly id GCA_030864225, (Estrada et al., 2024)), and the longest was 16,463 bp in the large bee-fly *Bombylius major* (assembly id GCA_932526495 (Lawniczak et al., 2023)). However, 86.0% of all operons had a length between 5,000 and 6,500 bp, and the median length was 5,779 bp (Fig. 3). Most of the operons in Fungi and Viridiplantae ranged between 5,500-6,100 bp in length (Fig. 3). The length variation was more extensive in Metazoa, where some groups, like Acanthocephala with an average operon length of 10,776 bp, possessed long operons.

To investigate the effect of sequence clustering thresholds on the retrieved number of clusters using full-length operons as compared to the 18S and 28S, as well as the shorter V4, ITS, ITS1, and ITS2 regions, these parts were clustered separately across various sequence identity thresholds (Fig. 4a). Across all clustering thresholds, the 18S region, as well as the V4 region, provided the lowest numbers of clusters. In contrast, the more variable ITS region and the full-length operon provided the most (Fig. 4a). At a 97% clustering threshold, a clearly lower number of unique species and genera were retrieved through 18S and V4, as compared to the other regions (Fig. 4b). At this clustering level, ITS provided the highest taxonomic resolution followed by full-length operon data.

Discussion

Recently, long-read sequencing approaches have been adopted in microbial ecology and DNA metabarcoding to obtain more detailed phylogenetic information about the microbial communities under investigation (Jamy et al., 2020, 2022; Tedersoo et al., 2020). More solid phylogenetic information is especially beneficial for poorly studied organisms that can be hard to place into a broader phylogenetic context. However, for long-read DNA metabarcoding analyses, we need reference sequences that cover longer stretches of the rDNA region, preferably the entire operon. Motivated by this, we provide full-length rDNA operon reference sequences assembled into the Ribosomal Operon Database (ROD). In addition to functioning as reference sequences for taxonomic annotation, the sequences in the ROD database may also fulfill other needs, such as reference sequences during chimera-checking, primer design, or phylogenetic binning. ROD currently includes 69,480 operon variants and their copy numbers extracted from 11,935 genomes. In addition, a reference set with only one operon (i.e., the most frequent) per genome is made available, which might be more useful in some investigations.

The reference sequence data have undergone multiple error correction steps, where contaminant and tentatively spurious sequences have been parsed out. Nevertheless, we cannot entirely dismiss the possibility of spurious operons in the database. For example, we observed that many of the genome assemblies (29.3% of the full-length operons) were contaminated by other organisms, and partial operons. Due to improved sequencing and bioinformatics methods, we envision that future genome assemblies will be of even higher quality.

We argue that full-length rDNA operon reference sequences should not be established based on concatenated sequences from different sub-parts, e.g., 18S and 28S, derived from multiple specimens, as this could lead to erroneous concatenation of different operon variants. Moreover, high-quality reference sequences should preferably be obtained from reference material, including specimens or cultures. Reference sequences obtained from environmental sequence data that have been assigned a taxonomy based on other reference sequences can also be useful but should be treated with caution since this represents indirect evidence. In such cases, errors can quickly spread and multiply in the reference data.

Multiple operon variants were detected in about 30.0 % of the genomes analyzed, with some exhibiting high quantities of variants. Multiple variants might appear for different reasons, including the mentioned contamination problem, which can be sorted out with blasting. Multiple operon variants per genome could also arise due to sequencing errors or chimerization between variants during assembly, including contaminant organisms. However, multiple variants may also occur for natural reasons, such as in diploid or polyploid organisms, where this is due to allelic variation. Especially in the analyzed plant genomes, we observed a high number of operon variants per genome, which could be due to their often polyploid nature. But even in haploid genomes, multiple variants can be expected if the multiple operon copies are not homogenized by concerted evolution, which has been observed in numerous organisms (Bradshaw et al., 2023). The occurrence of multiple, divergent operon variants per genome (or individual) can lead to inflation of the diversity (Thornhill et al., 2007)(ref). By considering the exhaustive ROD dataset as reference sequences in environmental sequencing studies, the problem with

multiple genomic variants can be taken into consideration and potentially corrected for during the taxonomic annotation process. This might be especially important for organismal groups with divergent operons, such as Viridiplantae. Due to the initial 99% clustering of operons within each genome (to filter out sequencing and assembly errors), we may underestimate the true variant numbers.

High variation in the operon copy number was observed across the eukaryotic tree of life, with an average high copy number present in Viridiplantae. The copy number variability will introduce PCR bias in DNA metabarcoding projects with environmental samples, where genomes and taxa with high copy number will be preferentially amplified. If the copy number variability becomes better known, measures may be taken to account for this bias in the final data (Stoddard et al., 2015)

Significant operon length differences are present between and within various taxonomic groups. Based on our data, the length variability seems especially prominent in Metazoa. One major cause for the extremely long variants is the presence of introns (Bhattacharya et al., 2000). During the amplification of longer stretches or the entirety of the rDNA operon, the length differences will introduce PCR bias, where taxa with shorter variants will be preferentially amplified. In theory, if the length variation is known across organismal groups, the bias could be modeled and taken into account.

Due to the varying rates of molecular evolution across the rDNA operon, different portions will provide different phylogenetic and taxonomic resolutions. Determining the optimal region for specific barcoding purposes remains a persistent challenge, largely due to the limited availability of reliable reference data. (Phillips, Gillis & Hanner 2022). While the ITS region is known to often provide species-level resolution for fungi (Schoch et al., 2012), more conserved markers, like 18S, may often lump related species into the same OTUs. Our clustering analyses of the different parts of the operon support that the ITS regions, as well the entire operon, provide a higher taxonomic resolution compared to 18S or the V4 region. Our results support that different markers should be treated differently during the bioinformatic processing, using e.g., different clustering levels.

The ROD database currently possesses a taxonomically highly skewed collection of reference sequences that reflect the distribution of available genomes in NCBI. Fungi are the dominating group, followed by Metazoa and Viridiplantae. Due to their small genome size and biotechnological significance, many fungal genomes have been sequenced and are available in public databases (de Vries et al., 2018). In the future, we expect more genomes from neglected branches on the Tree of Life distribution, also including more protist genomes. Due to advances in single-cell genomics, genomes from single-celled protists will expectedly become more available as references (Sieracki et al., 2019). Because of the steadily increasing pace of genome sequencing, we expect a strong increase in the number of operons that will become available for inclusion in ROD in the near future. We intend to provide regularly updated versions of ROD.

Acknowledgments

We acknowledge Ane-Rikke N. Oenting for invaluable feedback.

References

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., Campo, J. del, Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., … Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *The Journal of Eukaryotic Microbiology*, *66*(1), 4–119. https://doi.org/10.1111/jeu.12691

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bhattacharya, D., Lutzoni, F., Reeb, V., Simon, D., Nason, J., & Fernandez, F. (2000). Widespread Occurrence of Spliceosomal Introns in the rDNA Genes of Ascomycetes. *Molecular Biology and Evolution*, *17*(12), 1971–1984. https://doi.org/10.1093/oxfordjournals.molbev.a026298

Bradshaw, M. J., Aime, M. C., Rokas, A., Maust, A., Moparthi, S., Jellings, K., Pane, A. M., Hendricks, D., Pandey, B., Li, Y., & Pfister, D. H. (2023). Extensive intragenomic variation in the internal transcribed spacer region of fungi. *IScience*, *26*(8), 107317. https://doi.org/10.1016/j.isci.2023.107317

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2019). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., & Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research*, *47*(18), e103–e103. https://doi.org/10.1093/nar/gkz569

Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., Pande, N., Shang, Z., Yu, N., & Gutell, R. R. (2002). The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, *3*, 2. https://doi.org/10.1186/1471-2105-3-2

Cavalier-Smith, T., Lewis, R., Chao, E. E., Oates, B., & Bass, D. (2009). Helkesimastix marina n. sp. (Cercozoa: Sainouroidea superfam. n.) a gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist*, *160*(3), 452–479. https://doi.org/10.1016/j.protis.2009.03.003

Chilton, N. B., Huby-Chilton, F., & Gasser, R. B. (2003). First complete large subunit ribosomal RNA sequence and secondary structure for a parasitic nematode: phylogenetic and diagnostic implications. *Molecular and Cellular Probes*, *17*(1), 33–39. https://doi.org/10.1016/s0890-8508(02)00107-x

Clark, C. G. (1987). On the evolution of ribosomal RNA. *Journal of Molecular Evolution*, *25*(4), 343–350. https://doi.org/10.1007/bf02603119

Crowley, L. M., Akinmusola, R. Y., University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2023). The genome sequence of a druid fly, Clusia tigrina (Fallén, 1820). *Wellcome Open Research*, *8*, 430. https://doi.org/10.12688/wellcomeopenres.20109.1

Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115642118. https://doi.org/10.1073/pnas.2115642118

de Vries, R. P., Grigoriev, I. V., & Tsang, A. (2018). Introduction: Overview of fungal genomics. In *Methods in Molecular Biology* (Vol. 1775, pp. 1–7). Springer New York. https://doi.org/10.1007/978-1-4939-7804-5_1

Elder, J. F., Jr, & Turner, B. J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology*, *70*(3), 297–320. https://doi.org/10.1086/419073

Estrada, R., Tantalean, J. F. C., Saldaña, C. L., Romero, Y., Vilcara, E., & Arbizu, C. I. (2024). Draft genome and SSR data mining of a Peruvian landrace of Capsicum chinense, the arnaucho chili pepper. *Genet. Resour. Crop Evol.* https://doi.org/10.1007/s10722-024-01941-4

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J.,

Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … European Reference Genome Atlas (ERGA) Consortium. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, *37*(3), 197–202. https://doi.org/10.1016/j.tree.2021.11.008

Foster, Z. S. L., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLoS Computational Biology*, *13*(2), e1005404. https://doi.org/10.1371/journal.pcbi.1005404

Ganley, A. R. D., & Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research*, *17*(2), 184–191. https://doi.org/10.1101/gr.5457707

Gaonkar, C. C., & Campbell, L. (2024). A full-length 18S ribosomal DNA metabarcoding approach for determining protist community diversity using Nanopore sequencing. *Ecology and Evolution*, *14*(4). https://doi.org/10.1002/ece3.11232

Geng, Y., Guan, Y., Qiong, L., Lu, S., An, M., Crabbe, M. J. C., Qi, J., Zhao, F., Qiao, Q., & Zhang, T. (2021). Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation. *BMC Biology*, *19*(1), 143. https://doi.org/10.1186/s12915-021-01079-0

Grigoriev, I. V., Hayes, R. D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S. J., Salamov, A., Shabalov, I., & Kuo, A. (2021). PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Research*, *49*(D1), D1004–D1011. https://doi.org/10.1093/nar/gkaa898

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., & Shabalov, I. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research*, *42*(Database issue), D699-704. https://doi.org/10.1093/nar/gkt1183

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., Vargas, C. de, Decelle, J., Campo, J. D., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H., Lara, E., Bescot, N. L., Logares, R., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(Database issue), D597-604. https://doi.org/10.1093/nar/gks1160

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E. M., & Troedsson, C. (2014). Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. *PloS One*, *9*(2), e87624. https://doi.org/10.1371/journal.pone.0087624

Hall, A. N., Morton, E., & Queitsch, C. (2022). First discovered, long out of sight, finally visible: ribosomal DNA. *Trends in Genetics: TIG*, *38*(6), 587–597. https://doi.org/10.1016/j.tig.2022.02.005

Hassouna, N., Michot, B., & Bachellerie, J. P. (1984). The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Research*, *12*(8), 3563–3583. https://doi.org/10.1093/nar/12.8.3563

Hedberg, A., & Johansen, S. D. (2013). Nuclear group I introns in self-splicing and beyond. *Mobile DNA*, *4*(1), 17. https://doi.org/10.1186/1759-8753-4-17

Hillis, D. M., & Dixon, M. T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology*, *66*(4), 411–453. https://doi.org/10.1086/417338

HMMER 3.4 (Aug 2023). (n.d.). HMMER. http://hmmer.org/

Hooper, C., Ward, G. M., Foster, R., Skujina, I., Ironside, J. E., Berney, C., & Bass, D. (2023). Long amplicons as a tool to identify variable regions of ribosomal RNA for improved taxonomic resolution and diagnostic assay design in microeukaryotes: using ascetosporea as a case study. *Frontiers in Ecology and Evolution*, *11*. https://doi.org/10.3389/fevo.2023.1266151

Ihrmark, K., Bödeker, I. T. M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K. E., & Lindahl, B. D. (2012). New primers to amplify the fungal ITS2 region - evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology*, *82*(3), 666–677. https://doi.org/10.1111/j.1574-6941.2012.01437.x

Jamy, M., Biwer, C., Vaulot, D., Obiol, A., Jing, H., Peura, S., Massana, R., & Burki, F. (2022). Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nature Ecology & Evolution*, 1–13. https://doi.org/10.1038/s41559-022-01838-4

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., & Burki, F. (2020). Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, *20*(2), 429–443. https://doi.org/10.1111/1755-0998.13117

Karpov, S. A., Mamanazarova, K. S., Popova, O. V., Aleoshin, V. V., James, T. Y., Mamkaeva, M. A., Tcvetkova, V. S., Vishnyakov, A. E., & Longcore, J. E. (2017). Monoblepharidomycetes diversity includes new parasitic and saprotrophic species with highly intronized rDNA. *Fungal Biology*, *121*(8), 729–741. https://doi.org/10.1016/j.funbio.2017.05.002

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kauserud, H. (2023). ITS alchemy: On the use of ITS as a DNA marker in fungal ecology. *Fungal Ecology*, *65*(101274), 101274. https://doi.org/10.1016/j.funeco.2023.101274

Krabberød, A. K., Deutschmann, I. M., Bjorbækmo, M. F. M., Balagué, V., Giner, C. R., Ferrera, I., Garcés, E., Massana, R., Gasol, J. M., & Logares, R. (2022). Long-term patterns of an interconnected core marine microbiota. *Environmental Microbiome*, *17*(1), 2021.03.18.435965. https://doi.org/10.1186/s40793-022-00417-1

Lawniczak, M. K. N., Crowley, L. M., McAlister, E., University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2023). The genome sequence of the dark-edged bee fly, Bombylius major (Linnaeus, 1758). *Wellcome Open Research*, *8*, 379. https://doi.org/10.12688/wellcomeopenres.19804.1

Lewin, H. A., Richards, S., Aiden, E. L., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello, N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S.-M., Chen, L., Childers, A. K., Coddington, J. A., … Zhang, G. (2022). The Earth BioGenome Project 2020: Starting the clock. *Proceedings of the National Academy of Sciences*, *119*(4), e2115635118. https://doi.org/10.1073/pnas.2115635118

Lindner, D. L., Carlsen, T., Henrik Nilsson, R., Davey, M., Schumacher, T., & Kauserud, H. (2013). Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution*, *3*(6), 1751–1764. https://doi.org/10.1002/ece3.586

Logares, R., Haverkamp, T. H. A., Kumar, S., Lanzén, A., Nederbragt, A. J., Quince, C., & Kauserud, H. (2012). Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods*, *91*(1), 106–113. https://doi.org/10.1016/j.mimet.2012.07.017

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Jr, Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., & Tiedje, J. M. (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Research*, *29*(1), 173–174. https://doi.org/10.1093/nar/29.1.173

Medlin, L., Elwood, H. J., Stickel, S., & Sogin, M. L. (1988). The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene*, *71*(2), 491–499. https://doi.org/10.1016/0378-1119(88)90066-2

Nelles, L., Fang, B.-L., Volckaert, G., Vandenberghe, A., & Wachter, R. D. (1984). Nucleotide sequence of a crustacean 18S ribosomal RNA gene and secondary structure of eukaryotic small subunit ribosoraal RNAs. *Nucleic Acids Research*, *12*(23), 8749–8768. https://doi.org/10.1093/nar/12.23.8749

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajilić-Stojanović, M., Heilig, H. G. H. J., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., & De Vos, W. M. (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environmental Microbiology*, *11*(7), 1736–1751. https://doi.org/10.1111/j.1462-2920.2009.01900.x

Ram, K., & Wickham, H. (2023). *wesanderson: A Wes Anderson Palette Generator*. https://CRAN.R-project.org/package=wesanderson

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium, & Fungal Barcoding Consortium Author List. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Seemann, T. (n.d.). *barrnap: Bacterial ribosomal RNA predictor*. Retrieved February 23, 2024, from https://github.com/tseemann/barrnap

Sherrill-Mix, S. (2023). *taxonomizr: Functions to Work with NCBI Accessions and Taxonomy*. https://CRAN.R-project.org/package=taxonomizr

Shim, J., Hong, S. Y., Han, J.-H., Yu, Y., Yoo, E., Sung, J., Chin, J. H., & Lee, O. N. (2023). A Genomic Evaluation of Six Selected Inbred Lines of the Naturalized Plants of Milk Thistle (Silybum marianum L. Gaertn.) in Korea. *Plants*, *12*(14). https://doi.org/10.3390/plants12142702

Sieracki, M. E., Poulton, N. J., Jaillon, O., Wincker, P., de Vargas, C., Rubinat-Ripoll, L., Stepanauskas, R., Logares, R., & Massana, R. (2019). Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Scientific Reports*, *9*(1), 6025. https://doi.org/10.1038/s41598-019-42487-1

Simon, U. K., & Weiss, M. (2008). Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Molecular Biology and Evolution*, *25*(11), 2251–2254. https://doi.org/10.1093/molbev/msn188

Sonnenberg, R., Nolte, A. W., & Tautz, D. (2007). An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. *Frontiers in Zoology*, *4*(1), 6. https://doi.org/10.1186/1742-9994-4-6

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K., & Schmidt, T. M. (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*, *43*(Database issue), D593-8. https://doi.org/10.1093/nar/gku1201

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, *19 Suppl 1*, 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA*. Oxford University Press.

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

Tedersoo, L., Anslan, S., Bahram, M., Kõljalg, U., & Abarenkov, K. (2020). Identifying the 'unidentified' fungi: a global-scale long-read third-generation sequencing approach. *Fungal Diversity*, *103*(1), 273–293. https://doi.org/10.1007/s13225-020-00456-4

Thornhill, D. J., Lajeunesse, T. C., & Santos, S. R. (2007). Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Molecular Ecology*, *16*(24), 5326–5340. https://doi.org/10.1111/j.1365-294X.2007.03576.x

Tian, M., Hao, L., Li, R., & Zhang, L. (2024). Genome transposon analysis of blackleg causal pathogen. In *Research Square*. https://doi.org/10.21203/rs.3.rs-3987713/v1

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In *PCR Protocols* (pp. 315–322). Elsevier. https://doi.org/10.1016/b978-0-12-372180-8.50042-1

Wickham, H. (2016). *Ggplot2* (2nd ed.) [PDF]. Springer International Publishing.

Wilke, C. O. (2024). *ggridges: Ridgeline Plots in "ggplot2."* https://CRAN.R-project.org/package=ggridges

Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., Kagami, M., & Nilsson, R. H. (2019). Introducing ribosomal tandem repeat barcoding for fungi. *Molecular Ecology Resources*, *19*(1), 118–127. https://doi.org/10.1111/1755-0998.12944

Zhao, Y., Wang, Y., Ma, D., Feng, G., Huo, Y., Liu, Z., Zhou, L., Zhang, Y., Xu, L., Wang, L., Zhao, H., Zhao, J., & Wang, F. (2022). A chromosome-level genome assembly and annotation of the maize elite breeding line Dan340. *GigaByte (Hong Kong, China)*, *2022*, gigabyte63. https://doi.org/10.46471/gigabyte.63

Data Accessibility and Benefit-Sharing

The exhaustive set (69,480) and the reduced set of operon variants (11,867) are available in the ROD database (https://github.com/krabberod/ROD). The release used in this study is ROD v1.0 (GoldenRod), which is also deposited on Zenodo (https://zenodo.org/records/10951604).

Author Contributions

AKK conceived the idea of establishing ROD and assembled the database. AKK and EES did the analyses, HK and AKK drafted the manuscript while all co-authors commented and edited.

Figures

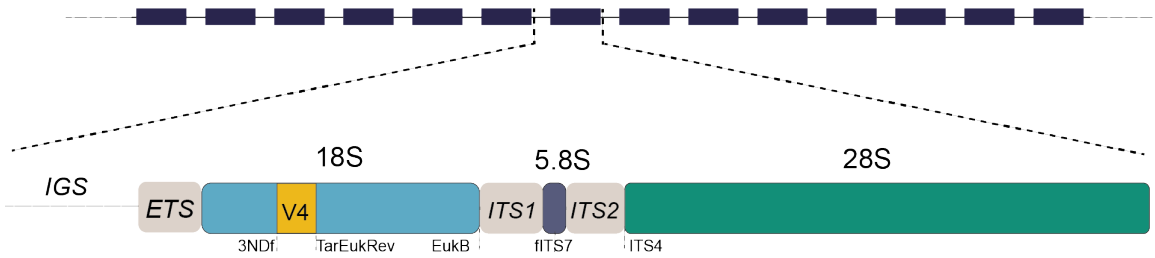

Figure 1. Typical arrangement of the ribosomal operon in eukaryotes. The ribosomal operon is commonly arranged in tandem repeats with intergenic spacers (IGS) between copies and an external transcribed spacer before the 18S gene. The genes 18S, 5.8S, and 28S are separated with internally transcribed spacers (ITS1 and ITS2), and the binding sites of primers used to separate regions of interest are annotated underneath the operon.

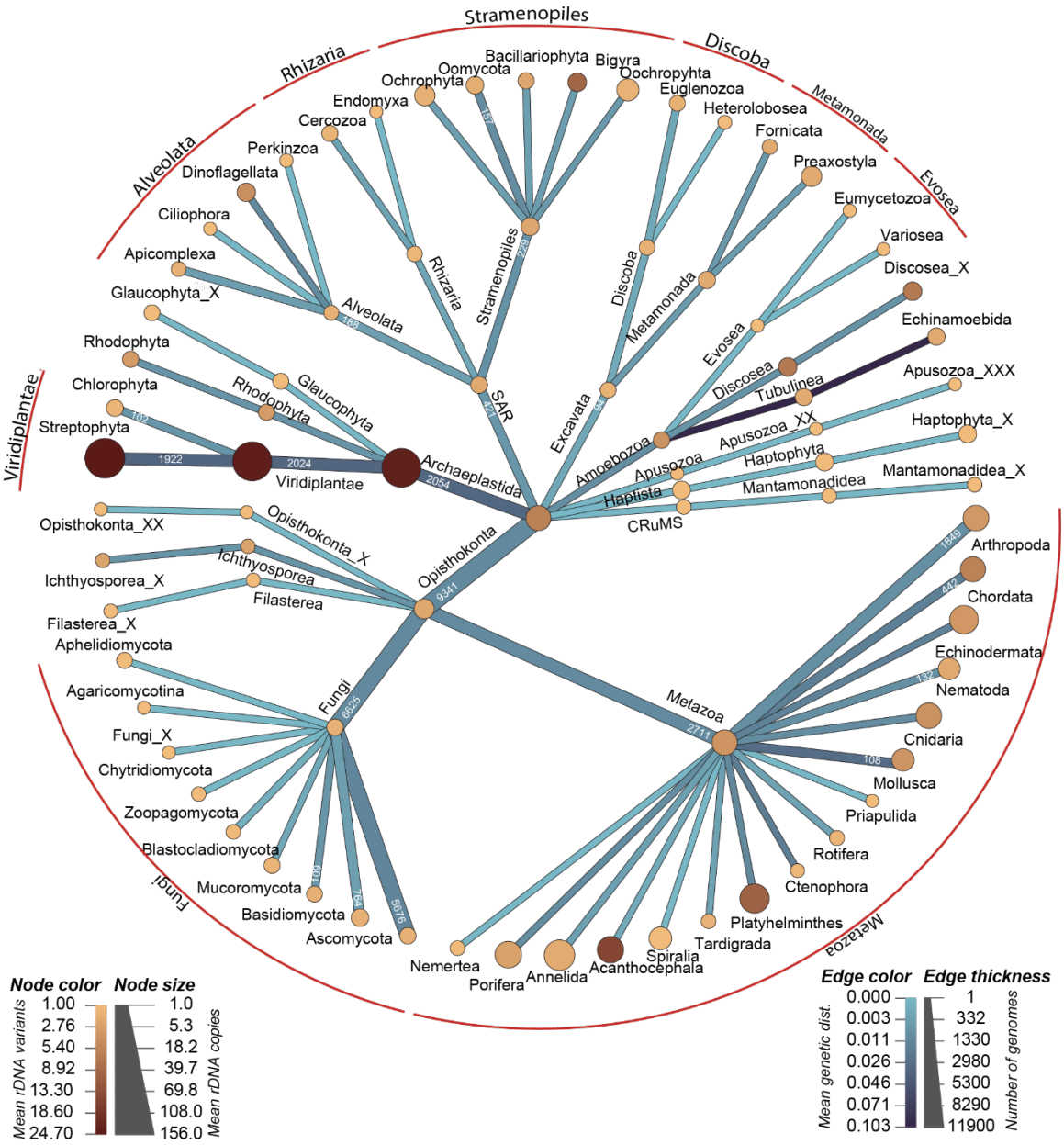

Figure 2. The Distribution of the 69,480 analyzed rDNA operons across the eukaryotic tree of life in the Ribosomal Operon Database (ROD). The unrooted tree is a schematic representation of eukaryotes produced using Metacoder (Foster et al., 2017). The thickness of the branches represents the number of genomes for the clade, and the color of the branches is the weighted genetic distance for the variants. The size of the nodes represents the Mean rDNA copy number for the genomes of that node and the color of the nodes represents the mean number of variants.

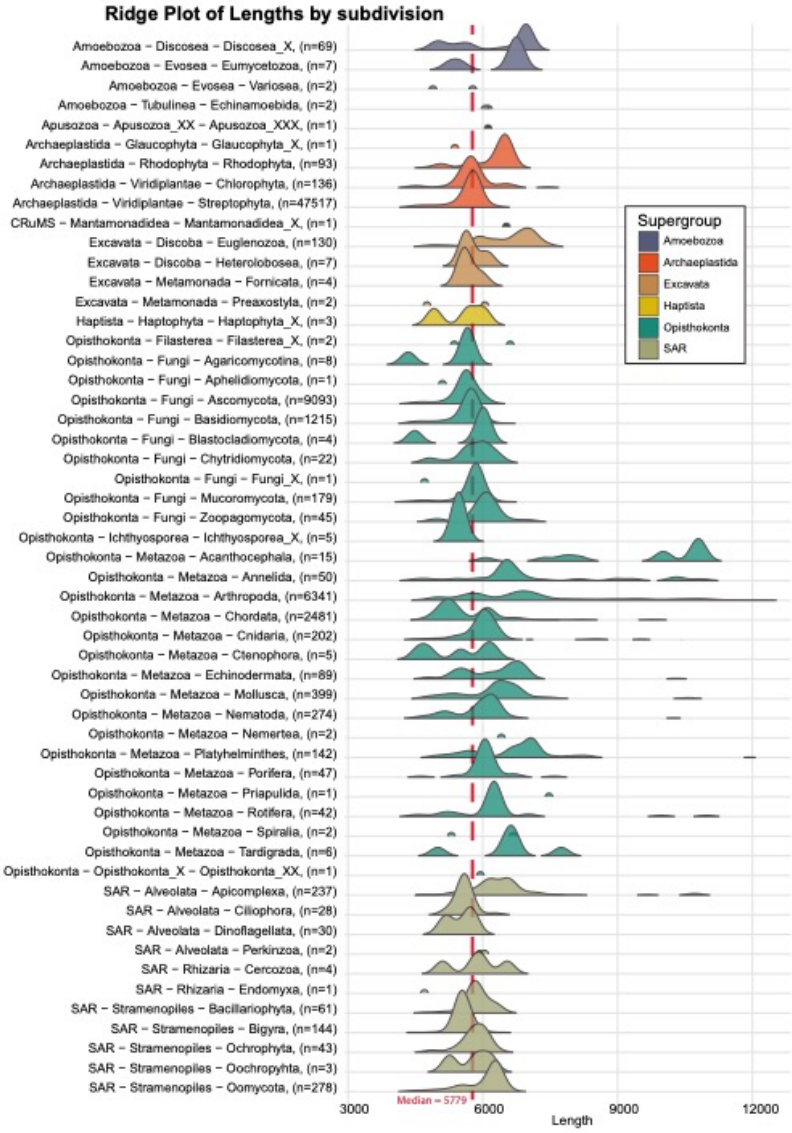

Figure 3. Length distribution of the recorded rDNA operons across the eukaryotic subdivisions in the Ribosomal Operon Database (ROD). The ridge plot represents the proportional distribution within each group. The number of genomes for each subdivision is written in parentheses. The median length is recorded as a red dashed line.

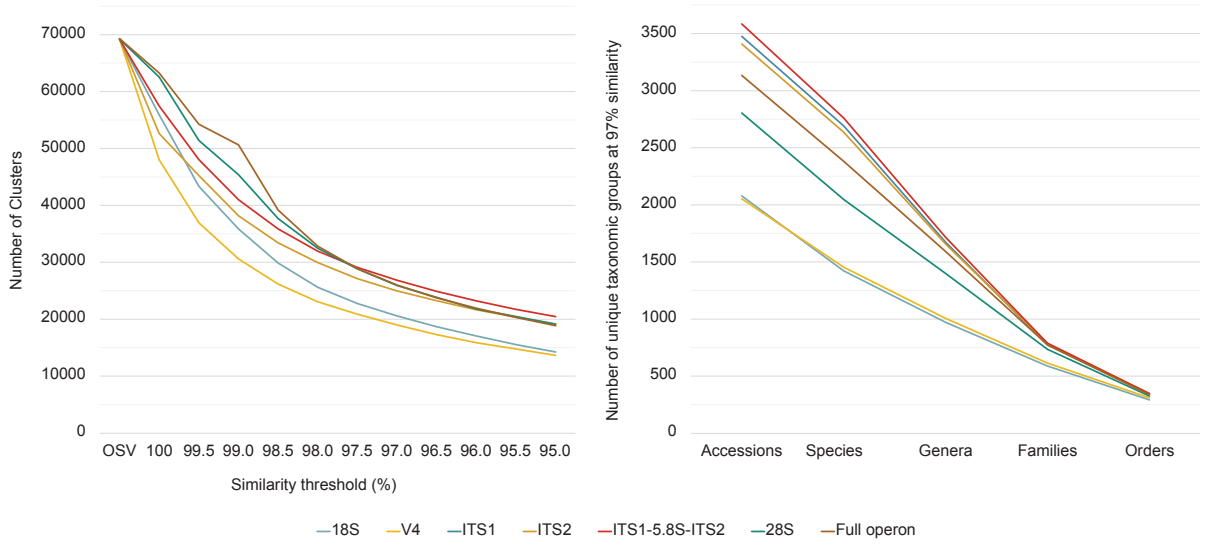

Figure 4. **Cluster analysis of 69,480 operon variants in ROD.).** (a) Clustering of genetic markers commonly used in DNA metabarcoding studies. The number of clusters (y-axis) is plotted against the clustering threshold for the entire operon as well as different subparts. (b) The number of unique taxonomic annotations for low taxonomic levels when clustered at 97% similarity, given by centroid annotations

Table 1. Summary statistics. The number of genomes/accessions for each Subdivision. The maximum number of copies found in a given subdivision and the number of variants (defined as at least 1% dissimilar). The average genetic distance is calculated based on the variants and weighted by the number of copies per variant. The minimum, maximum, and mean operon length for each subdivision. For more detailed information per accession see supplementary table 1.

									Operon length		
Supergroup	***Division***	***Subdivision***	***Genomes***	***Mean copy number***	***Max copies***	***Variants, mean***	***Max variants***	***Avg. genetic distance***	***Min.***	***Max.***	***Mean.***
Amoebozoa	Evosea	Eumycetozoa	7	1.1	2	1.0	1	-	5,260	6,837	6,362
Amoebozoa	Evosea	Variosea	2	1.0	1	1.0	1	-	4,891	5,778	5,335
Amoebozoa	Discosea	Discosea	10	11.3	59	6.9	36	0.006	4,814	7,027	5,765
Amoebozoa	Tubulinea	Echinamoebida	1	8.0	8	2.0	2	0.103	6,055	6,125	6,125
Apusozoa	Apusozoa_XX	Apusozoa_XXX	1	1.0	1	1.0	1	-	6,118	6,118	6,118
Archaeplastida	Viridiplantae	Streptophyta	1,922	155.8	5,947	24.7	2,097	0.033	4,136	13,249	5,645
Archaeplastida	Viridiplantae	Chlorophyta	102	7.1	112	1.3	6	0.005	4,347	7,448	5,695
Archaeplastida	Rhodophyta	Rhodophyta	29	5.7	21	3.2	10	0.006	4,694	6,634	6,073
Archaeplastida	Glaucophyta	Glaucophyta	1	5.0	5	1.0	1	-	5,381	5,381	5,381
CRuMS	Mantamonadidea	Mantamonadidea	1	3.0	3	1.0	1	-	6,527	6,527	6,527
Excavata	Discoba	Euglenozoa	84	4.8	36	1.5	11	0.001	4,684	7,448	6,627
Excavata	Discoba	Heterolobosea	7	1.9	6	1.0	1	-	5,605	6,105	5,766
Excavata	Metamonada	Fornicata	2	3.5	6	2.0	3	0.004	5,580	5,975	5,778
Excavata	Metamonada	Preaxostyla	1	17.0	17	2.0	2	0.003	4,757	6,047	4,757
Haptista	Haptophyta	Haptophyta	3	9.7	20	1.0	1	-	4,915	6,020	5,535
Opisthokonta	Metazoa	Arthropoda	1,849	42.1	4,691	3.4	139	0.008	4,395	16,463	6,685
Opisthokonta	Metazoa	Chordata	442	38.3	1,007	5.6	665	0.015	4,162	13,305	5,370
Opisthokonta	Metazoa	Mollusca	108	24.9	505	3.7	96	0.027	4,640	11,390	6,384
Opisthokonta	Metazoa	Echinodermata	26	58.9	710	3.4	45	0.014	4,772	10,310	5,865
Opisthokonta	Metazoa	Nematoda	132	23.3	448	2.1	15	0.007	4,415	10,248	5,972
Opisthokonta	Metazoa	Platyhelminthes	14	63.7	679	10.1	76	0.011	4,910	11,943	6,426
Opisthokonta	Metazoa	Cnidaria	49	41.2	880	4.1	40	0.008	4,409	10,417	5,936
Opisthokonta	Metazoa	Ctenophora	4	2.0	4	1.3	2	0.016	4,558	6,236	5,275
Opisthokonta	Metazoa	Annelida	26	72.5	646	1.9	6	0.003	4,440	10,955	6,612
Opisthokonta	Metazoa	Priapulida	1	1.0	1	1.0	1	-	7,473	7,473	7,473
Opisthokonta	Metazoa	Rotifera	33	3.3	25	1.3	4	0.002	4,453	10,979	6,228
Opisthokonta	Metazoa	Tardigrada	4	3.0	3	1.5	3	0.002	5,000	7,748	6,246
Opisthokonta	Metazoa	Spiralia	2	26.5	49	1.0	1	-	5,300	6,675	5,988
Opisthokonta	Metazoa	Acanthocephala	1	46.0	46	15.0	15	0.001	6,064	10,883	10,776
Opisthokonta	Metazoa	Nemertea	2	3.0	4	1.0	1	-	6,413	6,415	6,414
Opisthokonta	Metazoa	Porifera	18	47.0	258	2.6	8	0.009	4,624	7,588	5,960
Opisthokonta	Fungi	Ascomycota	5,676	6.1	1,078	1.6	81	0.012	4,185	6,905	5,595
Opisthokonta	Fungi	Basidiomycota	764	8.9	463	1.6	26	0.003	4,272	6,691	5,652
Opisthokonta	Fungi	Mucoromycota	109	4.9	68	1.6	5	0.005	4,263	6,646	5,692

Opisthokonta	Fungi	Chytridiomycota	22	2.6	11	1.0	1	-	4,787	6,409	5,803
Opisthokonta	Fungi	Blastocladiomycota	3	5.3	8	1.3	2	0.001	4,486	6,058	6,013
Opisthokonta	Fungi	Zoopagomycota	41	3.0	49	1.1	4	0.000	4,855	7,105	6,090
Opisthokonta	Fungi	Fungi_X	1	1.0	1	1.0	1	-	4,706	4,706	4,706
Opisthokonta	Fungi	Aphelidiomycota	1	5.0	5	1.0	1	-	5,102	5,102	5,102
Opisthokonta	Fungi	Agaricomycotina	8	1.0	1	1.0	1	-	4,323	5,678	5,331
Opisthokonta	Filasterea	Filasterea_X	2	2.0	3	1.0	1	-	5,367	6,611	5,989
Opisthokonta	Ichthyosporea	Ichthyosporea	2	2.5	4	2.5	4	0.006	5,407	5,486	5,446
Opisthokonta	Opisthokonta_X	Opisthokonta i.s.	1	1.0	1	1.0	1	-	5,946	5,946	5,946
SAR	Alveolata	Apicomplexa	151	3.3	59	1.6	30	0.003	4,647	10,794	6,338
SAR	Alveolata	Ciliophora	28	1.1	3	1.0	1	-	5,209	6,274	5,584
SAR	Alveolata	Dinoflagellata	7	11.0	33	4.3	13	0.010	5,078	5,844	5,509
SAR	Alveolata	Perkinzoa	2	1.5	2	1.0	1	-	5,936	6,044	5,990
SAR	Rhizaria	Cercozoa	3	5.3	6	1.3	2	0.002	5,107	6,545	6,118
SAR	Rhizaria	Endomyxa	1	1.0	1	1.0	1	-	4,702	4,702	4,702
SAR	Stramenopiles	Ochrophyta	28	17.6	376	1.5	7	0.003	4,799	6,383	5,897
SAR	Stramenopiles	Oomycota	157	9.0	72	1.8	15	0.008	4,261	7,317	6,054
SAR	Stramenopiles	Bacillariophyta	27	5.7	24	2.3	13	0.003	4,713	6,401	5,795
SAR	Stramenopiles	Bigyra	15	11.3	102	9.6	87	0.003	4,436	6,489	5,664
SAR	Stramenopiles	Oochropyhta	2	24.5	48	1.5	2	0.004	5,262	6,164	5,554