**TITLE:** Protist.guru: gene expression, co-expression network and comparative transcriptomics database for protists

**AUTHORS:** Erielle Marie Fajardo Villanueva1, Peng Ken Lim1, Jolyn Jia Jia Lim1, Shan Chun Lim1, Pei Yi Lau1, Kenny Ting Sween Koh1, Emmanuel Tan1, Ryanjit Singh Kairon1, Wei An See1, Jian Xiang Liao1, Ker Min Hee1, Varsheni Vijay1, Ishani Maitra1, Chong Jun Boon1, Kevin Fo1, Yee Tat Wang1, Ryan Jaya1, Li Anne Hew1, Yong Yee Lim1, Wei Quan Lee1, Zhi Qi Lee1, Herman Foo1, Adriana Lopes dos Santos2, Marek Mutwil1\*

1School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

2Asian School of the Environment, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

\*Corresponding author:

Marek Mutwil

School of Biological Sciences,

Nanyang Technological University, 60 Nanyang Drive,

637551, Singapore,

Singapore

Email: [mutwil@ntu.edu.sg](mailto:mutwil@ntu.edu.sg)

**SUPPLEMENTARY METHODS**

*Download of genomic and transcriptomic data*

The protist.guru database is based on the CoNekT framework, an open-source platform that facilitates comparative genomic and transcriptome analysis (Proost and Mutwil, 2018). For the 15 protist species, RNA-seq data was sourced from the publicly available Sequence Read Archive (SRA) from NCBI. IDs of Illumina-based sequencing runs and corresponding experimental metadata were generated from the SRA to stream fastq files and annotate runs respectively.

For each species, coding sequence (CDS) files (Table S1) were downloaded either from Ensembl Genomes or NCBI. The CDS files were subsequently used to generate Kallisto index files using Kallisto v0.46.0 (Bray et al., 2016) with default parameters. By applying the LSTrAP-Cloud (Qiao et al., 2020) pipeline, each experiment was streamed as a fastq file from the European Nucleotide Archive (ENA) (Leinonen et al., 2010). In total, 2,482 experiments were downloaded (Table S2). TPM (transcripts per million) expression values were extracted from the files generated from Kallisto’s quant function, which was utilized with default parameters. To annotate the RNA-seq experiments, the aforementioned metadata from the run tables was used to include information such as culture medium, genotype and other experimental variables. Annotation data was also supplemented by existing publications that are associated with the runs.

*Quality Control*

For quality control, thresholds were established: > 1 million reads for the number of processed reads (NPR) (with the exception of *Cladocopium sp.* clade C, *Micromonas pusilla*, *Porphyridium purpureum,* and *Thalassiosira pseudonana*), as well as a species-specific percentage for the percentage of pseudoaligned reads (PPR). The NPR and PPR values were obtained from Kallisto index files. The PPR threshold was set by manual observation of scatter-plots which displayed graphs of NPR values on the x-axis against PPR values on the y-axis (Supplementary Figure). These thresholds were set to remove samples that either had an insufficient NPR or PPR value. The 1,651 samples which passed the thresholds were subsequently used to generate expression matrices for all 15 protists (Tables S3–S17).

*Functional analysis of proteins*

To predict gene function, protein IDs in the form of pep files were obtained using the conversion feature onboard the CoNekT framework. For each protein, the Pfam domains and Gene Ontology (GO) terms were obtained via Interproscan-5.51-85.0 (Jones et al., 2014). Orthogroups were identified and inferred phylogenetic trees were obtained via the use of Orthofinder v2.3.12 (Emms and Kelly, 2015) and Diamond (Buchfink et al., 2014) with default settings. However, certain species (*O. trifallax, P. tetraurelia and T. thermophila*) had CDS files which contained a disproportionate number of stop codons within coding sequences, resulting in truncated peptide sequences after conversion. Hence, these species were omitted from the database.

*Construction of protist.guru database*

The database was constructed based on the CoNekT framework and substantiated by protist data. The coexpression networks were constructed using the Highest Reciprocal Rank metric (Mutwil et al., 2009). Coexpression clusters for each species were generated via Heuristic Cluster Chiseling Algorithm (HCCA) (Mutwil et al., 2009) where cluster sizes were limited to 100 genes.

**SUPPLEMENTARY REFERENCES**

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology, 34(5), 525–527.

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59–60.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology, 16(1), 157.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236–1240.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., ten Hoopen, P., Vaughan, R., Zalunin, V., Cochrane, G. (2010). The European Nucleotide Archive. Nucleic Acids Research, 39(Database), D28–D31.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., & Persson, S. (2009). Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm. Plant Physiology, 152(1), 29–43.

Proost, S., & Mutwil, M. (2018). CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. Nucleic Acids Research, 46(W1), W133–W140.

Qiao, WT., Goh, W., & Mutwil, M. (2020). LSTrAP-Cloud: A User-Friendly Cloud Computing Pipeline to Infer Coexpression Networks. Genes, 11(4), 428.