An explorable public transcriptomics compendium for eukaryotic microalgae

Short title: microalgal transcriptomics resource

Justin Ashworth* and Peter J. Ralph
Climate Change Cluster
University of Technology Sydney, Australia

Abstract

Eukaryotic microalgae dominate primary photosynthetic productivity in fluctuating nutrient-rich environments, including coastal, estuarine and polar regions, where competition and complexity are presumably adaptive and dynamic traits. Numerous genomes and transcriptomes of these species have been carefully sequenced, providing an unprecedented view into the vast genetic repertoires and the diverse transcriptional programs operating inside these organisms. Here we collected, re-mapped, quantified and clustered publicly available transcriptome data for ten different eukaryotic microalgae in order to develop new insights into their molecular systems biology, as well as to provide a large new resource of integrated information to facilitate the efforts of others to further compare and contextualize the results of individual and new experiments within and between species. This is summarized herein and provided for public use by the eukaryotic microalgae research community.

Keywords: microeukaryotes, microalgae, transcriptomics, systems biology

Introduction

Eukaryotic microalgae dominate primary photosynthetic productivity in fluctuating nutrient-rich environments, including coastal, estuarine and polar regions, where competition and complexity are presumably adaptive and dynamic traits [1–3]. Numerous genomes and transcriptomes of these species have been carefully sequenced, providing an unprecedented view into the vast genetic repertoires and the diverse transcriptional programs operating inside these organisms [4–16]. The integration of these transcriptome data to allow wholistic, multi-experiment and system-level analysis and interpretation is important and fruitful for gaining new understandings of concerted microbial functions [17,18].

In order to facilitate new insights into the molecular systems biology of eukaryotic microalgae, we systematically collected, re-mapped, quantified and clustered publicly available transcriptome data for ten different eukaryotic microalgal species, investigated

methods and caveats of integration, performed basic clustering, functional annotation and orthology analyses within and between species, and built an interactive online resource for the further exploration of the largest single compendium of microalgal transcriptomics data to date.

Results

Data integration

The sequencing read data for nine microalgal species were obtained from the Sequence Read Archive (SRA) [19] and systematically re-mapped to current genome assemblies using HISAT2 [20], SAMtools [21], and StringTie [22] on Amazon Web Services [23] (see Materials and Methods). By-sample transcriptomic read counts for *Thalassiosira weissflogii* were estimated by mapping reads onto genome-free transcript assemblies from the Marine Microbial Eukaryote Sequencing Project [9]. Transcriptomic data for the freshwater microalga *Chlamydomonas reinhardtii* are the subject of ongoing study by other groups [24], and may be combined with these data in the future. Within-species normalization of raw transcript-per-million (TPM) [25] counts was conducted using a method developed to maximize the number of uniform genes [26].

Batch effects

The integration of independent transcriptome experiments is prone to systematic variations that are unique--but often consistent--to each individual laboratory, treatment, preparation, platform, or other unaccountable factors. It is presumed that some of these biases may be corrected by applying an appropriate normalization to the data, while others may remain [27].

To normalize RNA-seq transcriptome data, we applied a method developed to algorithmically maximize the number of "uniform genes," or those for which transcript levels are least variant over multiple experiments, to serve as internal standards for within-sample normalization factors [26]; this compared favourably in this regard to the 'Trimmed Mean of M Values' (TMM) [28] and other methods of normalization across a large RNA-seq dataset of diverse human tissues. Application to microalgal transcriptome data in this study adjusted distributions of $\log_{10}$TPM values across different experimental series (Supplementary Figure S1), resulting in improvements to the consistency of measurements for typical "housekeeping" genes, such as actins, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), tubulins, and ribosomal proteins (Supplementary Figure S2). Remaining batch effects apparent between different studies included sets of transcripts whose within-study biases may be attributable to biological and/or non-biological sources.

Previously assembled microarray data for *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* [29] were appended to this dataset to facilitate direct comparisons between the results of different platforms used to estimate transcriptome-wide expression levels, as well as between independent experiments. Microarray fold-change ratio data and RNA-seq TPM data in aggregate were distributed similarly (Supplementary Figure S3). These data are presented and provided without further normalisation to fit ideal distributions, and therefore preserve biologically faithful measurements and to facilitate further comparisons between other similarity, normalisation and clustering approaches.

In total, 1,375 samples (transcriptome measurements) were integrated from sixty-nine independent experimental studies representing ten species and four distinct clades of microalgae (Figure 1, Supplementary Table S1), providing a rich consolidated dataset for new and comparative data exploration.

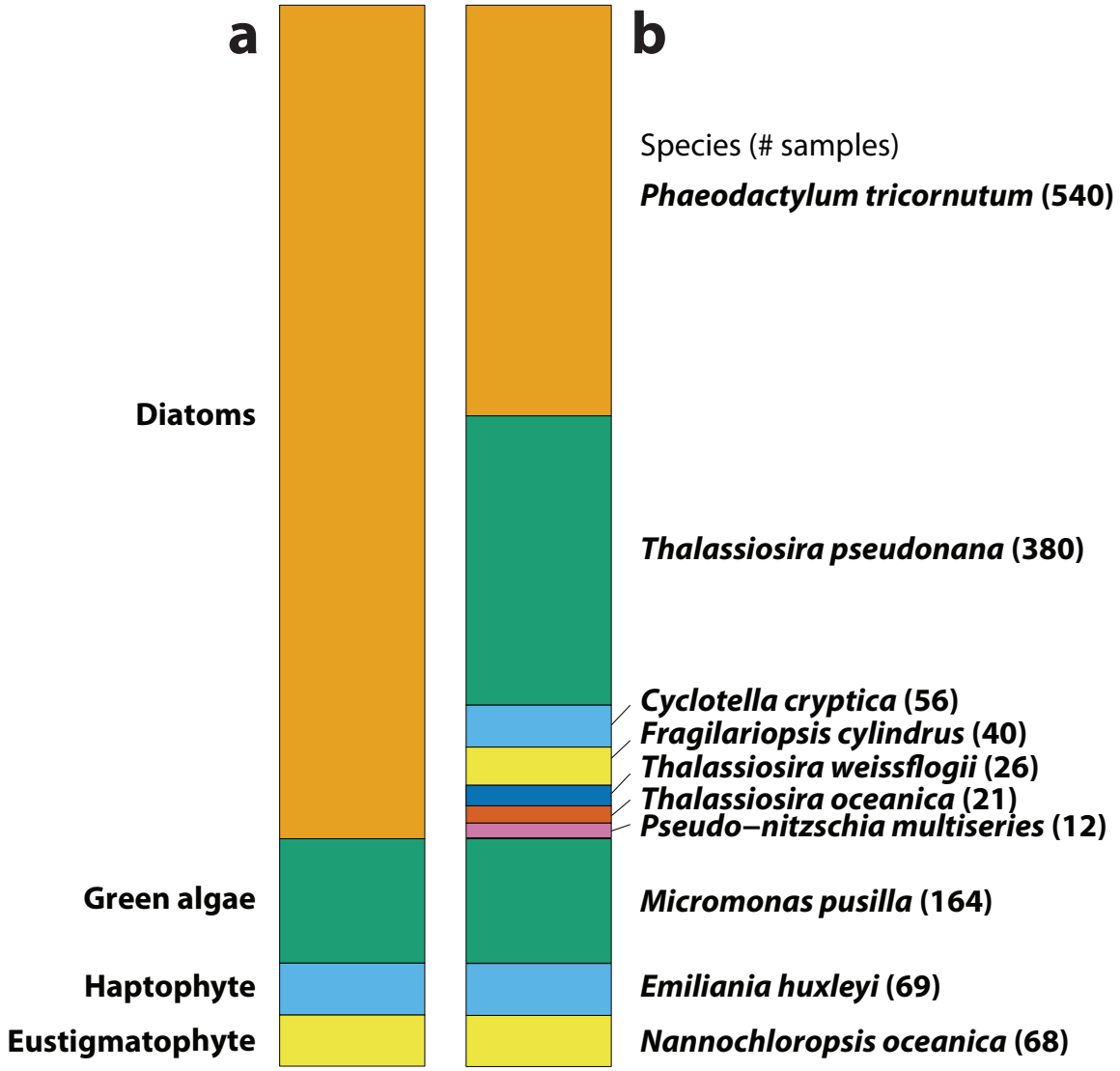

Figure 1. Summary of sample counts by a) microalgal clade and b) species.

Clustering with efficient empirical estimates of reproducibility

Agglomerative hierarchical clustering based on aggregate within-sample correlations [30] was performed for each species, resulting in high numbers of putative clusters with distinct conditional and experimental patterns of expression in all ten microalgal species. In order to assess the robustness, reproducibility and statistical boundaries of apparent clusters of co-expressed transcripts, we employed deep bootstrapping of hierarchical trees with multi-scale resampling using a version of Pvclust [31] refactored to run more efficiently on large matrices [29]. This produced data-supported, fine-grained clusters of transcripts (Supplementary Table S1) whose expression patterns over all conditions imply--but not do prove--uniquely shared responses, regulation, activities, or functions that appear to be linked to biological or environmental change. Large, robust clusters of co-expressed and putatively related genes include major aspects of microalgal biology, such as the regulation of light harvesting and ribosomes, while smaller clusters of condition-specific genes abound, such as those for nutrient scavenging (Figure 2).

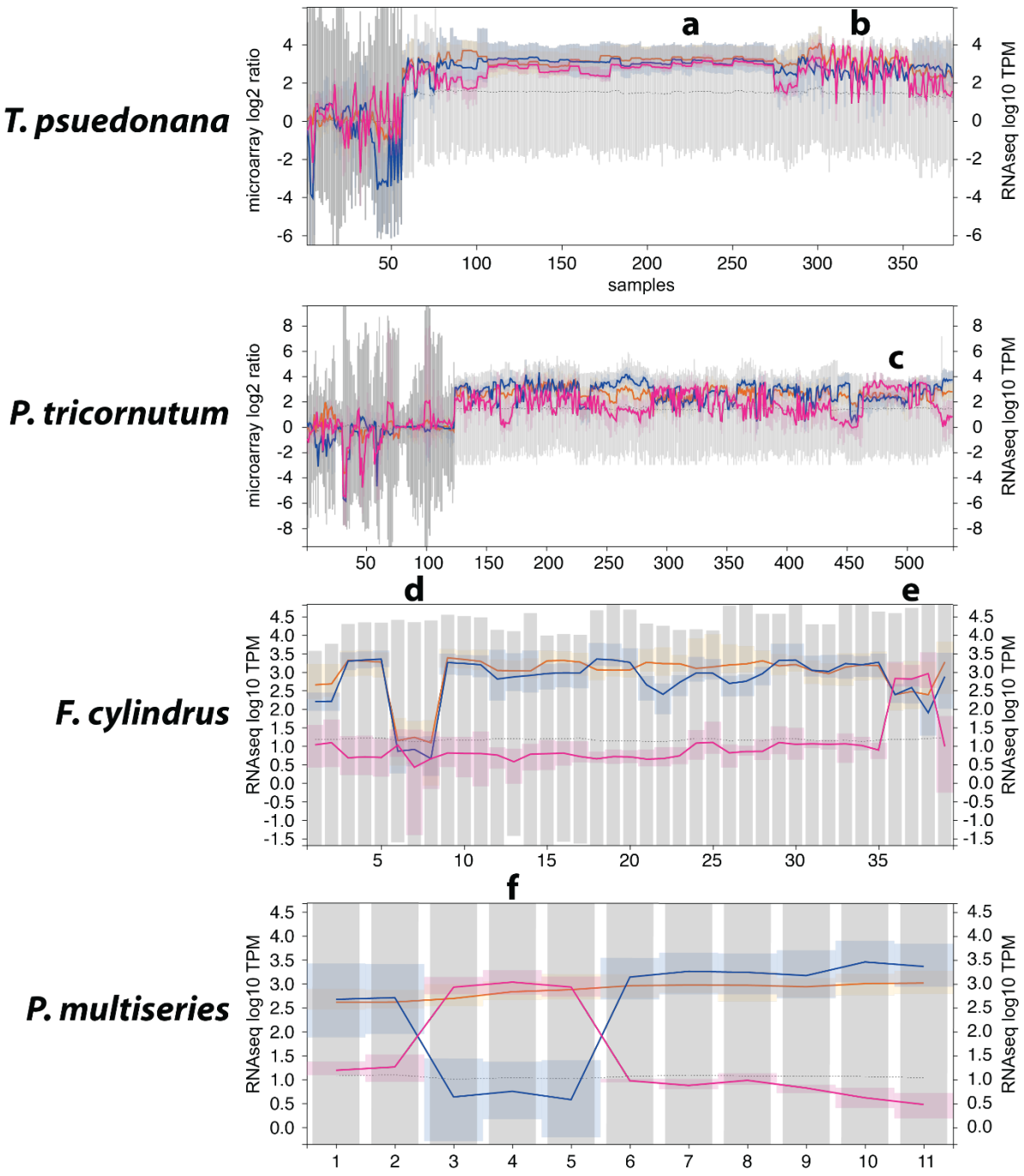

Figure 2. Distinct bootstrap-supported clusters of co-expressed genes in four diatom species, which include ribosomal proteins (orange), photosynthesis and light harvesting genes (blue), and nitrate uptake genes (magenta). Notable conditions affecting the transcript levels of these gene groups include a) nitrate-limited chemostat cultures of *T. pseudonana* [32], b) batch cultures of *T. pseudonana* experiencing periodic nutrient stress [33], c) *P. tricornutum* cultures subject to nitrate depletion [34], *F. cylindrus* subject to d) prolonged darkness [12] and e) nitrate limitation [35], and f) *Pseudo-nitzschia multiseries* subject to nitrate limitation [35].

The occurrence of orthologically related clusters of co-expressed genes across all included species reflects a conservation of concerted transcriptional regulation of fundamental microalgal processes. In the case of nitrate uptake mechanisms in diatoms (Figure 2, magenta lines), the transcript levels of putatively orthologous nitrate uptake transporters correlate with organic nitrogen-limited conditions in multiple species. An abundance of smaller apparent clusters, comprised of large proportions of novel and under-characterized genes, implies a wealth of divergently regulated functions whose unique expression patterns might contribute to species specificity and environmentally linked biological programs.

Public exploration via online resource

To further increase the public utility and accessibility of this data resource, we created a lightweight database that includes basic protein homolog predictions using BLASTp [36], protein sequences, functional annotations, putative promoter region sequences, as well as an interactive online interface to this compendium using Django [37] and d3.js [38]. This resource (Figure 3) is currently available at http://115.146.92.162/Clusterview/ in cooperation with the Australian National eResearch Collaboration Tools (NeCTAR) [39]. The transcriptomic datasets assembled in this work are also available for download, independent analysis and use.

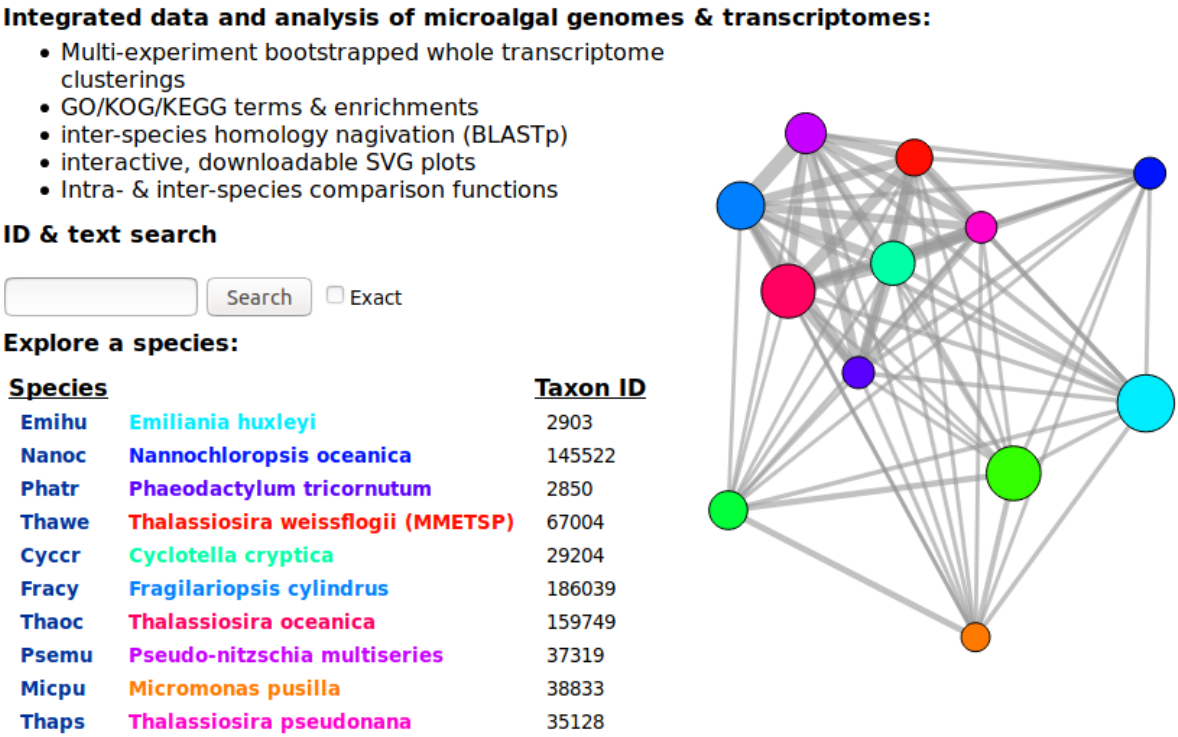

Figure 3. The “Alganaut” public interactive web resource for the exploration of microalgal transcriptomes.

Discussion

Assembly and informed analysis of ever-expanding integrated datasets are crucial for organismal and environmental sciences. While dozens of microalgal genomes have now been sequenced, assembled and translated into large and diverse proteomes, the identities and functions of the majority of putative proteins discovered in these organisms remain a mystery. Approximately half or more of each new microalgal proteome bears no reliable similarity to any other species, nor to any known proteins or functional domains in existing gene databases [4–13]. Furthermore, mechanistically accurate gene networks, molecular "biomarkers," robust environmental covariates, and intra- and inter-species dependencies are all yet to be discovered for eukaryotic microalgal species. "Data-driven" large-scale prediction and discovery of putative links between known and unidentified proteins will assist in further modeling, characterization and understanding of algal species.

Data science potential

In most species, the transcript levels of nearly all functional gene products depend on the physiological state of the cell or cell population, the biological and regulatory programs in operation, and environmental conditions. It is presumed that groups of transcripts that exhibit uniquely similar or correlated expression patterns over various conditions are more likely to be functionally related or co-regulated than those for which expression patterns diverge.

Mining of transcriptome data can be performed in numerous ways. Most approaches yield a finite number of large clusters with similar memberships based on various algorithms, as well as increasingly divergent sub-clusters based on various arbitrary, heuristic, or statistical choices. Additional types of clustering and significance estimation can be fruitfully applied to these data. Both a) empirically appropriate statistical metrics and b) new and orthogonal experimental measurements will be required to agnostically compare the results of different "data-driven" predictions, as has been accomplished for established bacterial, metazoan and human systems [40]. Sources of validation and extension of models include fresh transcriptional data series, gene knockdowns, knockouts, overexpression, genome-wide binding studies, comparative genomics, proteomics, and subcellular localization atlases. Increasing amounts of these orthogonal experimental measurements will be particularly valuable to provide a sufficient scientific basis upon which to judge the accuracy and applicability of biological and statistical models.

Limitations of transcriptome data

Transcriptomics alone is insufficient to understand the complex biology and inner workings of microalgae. However, it is currently the easiest approach to gather comprehensive data to describe the intracellular behaviour, cellular programming, environmental responses, and

comparative molecular biology. The ease of obtaining, handling and interpreting sequence read data continues to outpace other types of measurements, including proteomics, phenomics, fluorometry, cytometry, and imaging. But the value of genomic and transcriptomic data are limited without adequate context--thus the relative shortage of other data types such as those mentioned above make them increasingly valuable to collect in parallel with transcript data.

Post-transcriptional regulation and post-translational features including RNA modifications, binding and trafficking, degradation, protein modifications, subcellular localization, protein and chemical signalling, and allostery all crucially contribute to the cellular holo-program. Protein levels do, as an apparent rule, correlate closely with their corresponding transcript levels [42,43], but it is possible to identify interesting exceptions. For example, peptide signalling is prevalent in eukaryotic microalgae [44,45], and the complex subcellular locations of variously acting proteins is crucial to correctly understand their physiological roles across a large number of different membranes and compartments [46]. The measurement of protein phospho-states [47,48], which can provide data to model protein signalling and the post-translational activities of many transcription factors, may also be possible and relevant to advance systems biology studies in microalgae.

Integrated datasets for environmental biology

Biology is governed by detailed regulatory and metabolic programs operating throughout various environmental conditions. While rapid, high-throughput reverse genetics and phenotyping approaches are under development eukaryotic microalgal species, multi-omics and experimental systems biology are the fastest and broadest initial means to observe and predict the roles, functions and importance of new genes in new and non-model organisms. Molecular systems biology consists of collecting detailed and comprehensive measurements of numerous observable molecular features and parameters, and then using computing, statistics, quantitative hypotheses and biological models to apply these data to new questions [49]. The synthesis of community-wide datasets will continue to deepen our understanding of microalgae and help to inform further efforts including biological oceanography, direct genetics, and comparative microbial biology, and biotechnology.

Materials and Methods

RNA-seq data integration and normalization

Sequencing read data for eukaryotic microalgae with sufficient numbers of samples for integration (≥10, Supplementary Table S2) were obtained from the Sequence Read Archive (SRA) [19] as of February 2018 and re-mapped to current genome assemblies using

HISAT2 v2.1.0 [20], SAMtools 1.7 [21], and StringTie 1.3.4b [22] on Amazon Web Services [23]. Genome and RNA-seq data for *C. cryptica* were obtained from Traller et. al [11]. By-sample transcriptomic read counts for *T. weissflogii* were estimated by mapping reads onto genome-free transcript assemblies from the Marine Microbial Eukaryote Sequencing Project [9]. The ordering and sample names of data series were taken as-is in accordance with source annotations. Within-species normalization of raw transcript-per-million (TPM) [25] counts was conducted using a method developed to maximize the number of uniform genes [26]. Previously assembled microarray data for *T. pseudonana* and *P. tricornutum* [29] were appended as $\log_2$ ratios of changes in expression to within-experiment internal reference controls. Various scripts used and intermediate datasets are available as Supplementary Information.

Bootstrapped hierarchical clustering

Agglomerative hierarchical clustering was performed using a c++ wrapper to call the Fastcluster [30] library directly and repeatedly, with in-memory multi-scale resampling of the data to perform efficient bootstrapping [29]. A partially refactored version of Pvclust [31] was used to assign bootstrap P-values to branch nodes in the hierarchical tree, providing a robust empirical estimate of reproducible and bifurcating sub-cluster memberships.

Database and resource implementation

An SQL database was used to combine transcriptomic data with pairwise reciprocal inter-proteome homolog predictions (BLASTp 2.6.0 [36]), GO, KEGG, and KOG functional annotations where publicly available from draft genomes, and putative upstream promoter region sequences taken from public genome assemblies. These data were rendered explorable using Django 2.0.3 [37] and d3.js version 4 [38]. The data resource was made available on-line using a remote server instance on the Australian NeCTAR Research Cloud [39].

Acknowledgements

J.A. is the recipient of an Australian Research Council Discovery Early Career Award (DE160100615) funded by the Australian Government. This research was supported by use of the NeCTAR Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy.

References

1. Armbrust EV. The life of diatoms in the world's oceans. Nature. 2009;459: 185–192. doi:10.1038/nature08057

2. Karsenti E, Acinas SG, Bork P, Bowler C, Vargas CD, Raes J, et al. A Holistic Approach to Marine Eco-Systems Biology. PLOS Biol. 2011;9: e1001177. doi:10.1371/journal.pbio.1001177

3. Vargas C de, Audic S, Henry N, Decelle J, Mahé F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348: 1261605. doi:10.1126/science.1261605

4. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, et al. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science. 2004;306: 79–86. doi:10.1126/science.1101156

5. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008;456: 239–244. doi:10.1038/nature07410

6. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science. 2009;324: 268–272. doi:10.1126/science.1167222

7. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature. 2013;499: 209–213. doi:10.1038/nature12221

8. Vieler A, Wu G, Tsai C-H, Bullard B, Cornish AJ, Harvey C, et al. Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga Nannochloropsis oceanica CCMP1779. PLOS Genet. 2012;8: e1003064. doi:10.1371/journal.pgen.1003064

9. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12: e1001889. doi:10.1371/journal.pbio.1001889

10. Lommer M, Specht M, Roy A-S, Kraemer L, Andreson R, Gutowska MA, et al. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol. 2012;13: R66. doi:10.1186/gb-2012-13-7-r66

11. Traller JC, Cokus SJ, Lopez DA, Gaidarenko O, Smith SR, McCrow JP, et al. Genome and methylome of the oleaginous diatom Cyclotella cryptica reveal genetic flexibility toward a high lipid phenotype. Biotechnol Biofuels. 2016;9: 258. doi:10.1186/s13068-016-0670-3

12. Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, et al.

Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature. 2017;541: 536–540. doi:10.1038/nature20803

13. Basu S, Patil S, Mapleson D, Russo MT, Vitale L, Fevola C, et al. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. New Phytol. 2017;215: 140–156. doi:10.1111/nph.14557

14. Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, et al. Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. Proc Natl Acad Sci U S A. 2008;105: 1579–1584. doi:10.1073/pnas.0707946105

15. Allen AE, Laroche J, Maheswari U, Lommer M, Schauer N, Lopez PJ, et al. Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. Proc Natl Acad Sci U S A. 2008;105: 10438–10443. doi:10.1073/pnas.0711370105

16. Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, et al. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. Proc Natl Acad Sci. 2012;109: E317–E325. doi:10.1073/pnas.1118408109

17. Ashworth J, Coesel S, Lee A, Armbrust EV, Orellana MV, Baliga NS. Genome-wide diel growth state transitions in the diatom Thalassiosira pseudonana. Proc Natl Acad Sci U S A. 2013;110: 7518–7523. doi:10.1073/pnas.1300962110

18. Levering J, Dupont CL, Allen AE, Palsson BO, Zengler K. Integrated Regulatory and Metabolic Networks of the Marine Diatom Phaeodactylum tricornutum Predict the Response to Rising CO2 Levels. mSystems. 2017;2: e00142-16. doi:10.1128/mSystems.00142-16

19. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011;39: D19–D21. doi:10.1093/nar/gkq1019

20. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12: 357–360. doi:10.1038/nmeth.3317

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

22. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11: 1650–1667. doi:10.1038/nprot.2016.095

23. Amazon Web Services (AWS) - Cloud Computing Services. In: Amazon Web Services, Inc. [Internet]. [cited 31 Jul 2018]. Available: https://aws.amazon.com/

24. López García de Lomana A, Schäuble S, Valenzuela J, Imam S, Carter W, Bilgin DD, et al. Transcriptional program for nitrogen starvation-induced lipid accumulation in Chlamydomonas reinhardtii. Biotechnol Biofuels. 2015;8: 207. doi:10.1186/s13068-015-0391-z

25. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010;26: 493–500. doi:10.1093/bioinformatics/btp692

26. Glusman G, Caballero J, Robinson M, Kutlu B, Hood L. Optimal Scaling of Digital Transcriptomes. PLOS ONE. 2013;8: e77885. doi:10.1371/journal.pone.0077885

27. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17: 13. doi:10.1186/s13059-016-0881-8

28. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11: R25. doi:10.1186/gb-2010-11-3-r25

29. Ashworth J, Turkarslan S, Harris M, Orellana MV, Baliga NS. Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum. Mar Genomics. 2016;26: 21–28. doi:10.1016/j.margen.2015.10.011

30. Müllner D. Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. J Stat Softw. 2013;53: 1–18.

31. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22: 1540–1542. doi:10.1093/bioinformatics/btl117

32. Hennon GMM, Ashworth J, Groussman RD, Berthiaume C, Morales RL, Baliga NS, et al. Diatom acclimation to elevated CO2 via cAMP signalling and coordinated gene expression. Nat Clim Change. 2015;5: 761–765. doi:10.1038/nclimate2683

33. Valenzuela JJ, Lomana ALG de, Lee A, Armbrust EV, Orellana MV, Baliga NS. Ocean acidification conditions increase resilience of marine diatoms. Nat Commun. 2018;9: 2328. doi:10.1038/s41467-018-04742-3

34. McCarthy JK, Smith SR, McCrow JP, Tan M, Zheng H, Beeri K, et al. Nitrate Reductase Knockout Uncouples Nitrate Transport from Nitrate Assimilation and Drives Repartitioning of Carbon Flux in a Model Pennate Diatom. Plant Cell. 2017;29: 2047–2070. doi:10.1105/tpc.16.00910

35. Bender SJ, Durkin CA, Berthiaume CT, Morales RL, Armbrust EV. Transcriptional

responses of three model diatoms to nitrate limitation of growth. Front Mar Sci. 2014;1. doi:10.3389/fmars.2014.00003

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

37. The Web framework for perfectionists with deadlines | Django [Internet]. [cited 31 Jul 2018]. Available: https://www.djangoproject.com/

38. Bostock M. D3.js - Data-Driven Documents [Internet]. [cited 31 Jul 2018]. Available: https://d3js.org/

39. Science Clouds. In: Nectar [Internet]. [cited 31 Jul 2018]. Available: https://nectar.org.au/science-clouds/

40. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci. 2010;107: 6286–6291. doi:10.1073/pnas.0913357107

41. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

42. Dyhrman ST, Jenkins BD, Rynearson TA, Saito MA, Mercier ML, Alexander H, et al. The Transcriptome and Proteome of the Diatom Thalassiosira pseudonana Reveal a Diverse Phosphorus Stress Response. PLOS ONE. 2012;7: e33768. doi:10.1371/journal.pone.0033768

43. Wang D, Eraslan B, Wieland T, Hallstrom BM, Hopf T, Zolg DP, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. bioRxiv. 2018; 357137. doi:10.1101/357137

44. Gschloessl B, Guermeur Y, Cock JM. HECTAR: A method to predict subcellular targeting in heterokonts. BMC Bioinformatics. 2008;9: 393. doi:10.1186/1471-2105-9-393

45. Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. Plant J. 2015;81: 519–528. doi:10.1111/tpj.12734

46. Matsuda Y, Hopkinson BM, Nakajima K, Dupont CL, Tsuji Y. Mechanisms of carbon dioxide acquisition and CO2 sensing in marine diatoms: a gateway to carbon metabolism. Phil Trans R Soc B. 2017;372: 20160403. doi:10.1098/rstb.2016.0403

47. Saleem RA, Rogers RS, Ratushny AV, Dilworth DJ, Shannon PT, Shteynberg D, et al. Integrated Phosphoproteomics Analysis of a Signaling Network Governing Nutrient Response and Peroxisome Induction. Mol Cell Proteomics. 2010;9: 2076–2088.

doi:10.1074/mcp.M000116-MCP201

48. Chen Z, Yang M, Li C, Wang Y, Zhang J, Wang D, et al. Phosphoproteomic Analysis Provides Novel Insights into Stress Responses in Phaeodactylum tricornutum, a Model Diatom. J Proteome Res. 2014;13: 2511–2523. doi:10.1021/pr401290u

49. Ashworth J. Marine Microalgae: Systems Biology from 'Omics.' Systems Biology of Marine Ecosystems. Springer, Cham; 2017. pp. 207–221. doi:10.1007/978-3-319-62094-7_10

Supporting Information

Table S1. Dataset summary.

		Dataset		Bootstrapped clustering		
Species	**Abbr.**	**Samples**	**Projects**	**No. clusters**	**Max. size**	**Median size**
Phaeodactylum tricornutum	(Phatr)	539 (123*)	26 (7*)	2010	68	4
Thalassiosira pseudonana	(Thaps)	380 (56*)	17 (6*)	2172	64	4
Micromonas pusilla	(Micpu)	164	4	1439	165	5
Emiliania huxleyi	(Emihu)	69	6	9435	26	3
Nannochloropsis oceanica	(Nanoc)	68	5	2408	60	4
Cyclotella cryptica	(Cyccr)	56	1	4062	38	4
Fragilariopsis cylindrus	(Fracy)	40	4	5040	42	3
Thalassiosira weissflogii	(Thawe)	26	1	4121	44	3
Thalassiosira oceanica	(Thaoc)	21	3	8560	20	3
Pseudo-nitzschia multiseries	(Psemu)	12	2	3436	36	4

*microarray samples/studies

Table S2. RNA-seq samples included in dataset (separate file).

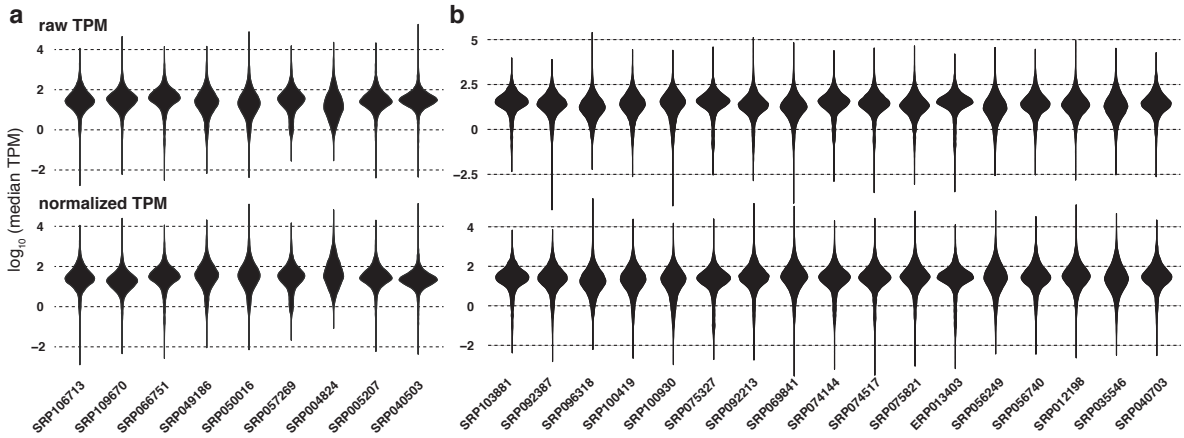

Figure S1. Log_{10} by-transcript median TPM distributions (top: raw, bottom: normalized) for different RNA-seq experimental project series, for a) *T. pseudonana* and b) *P. tricornutum*.

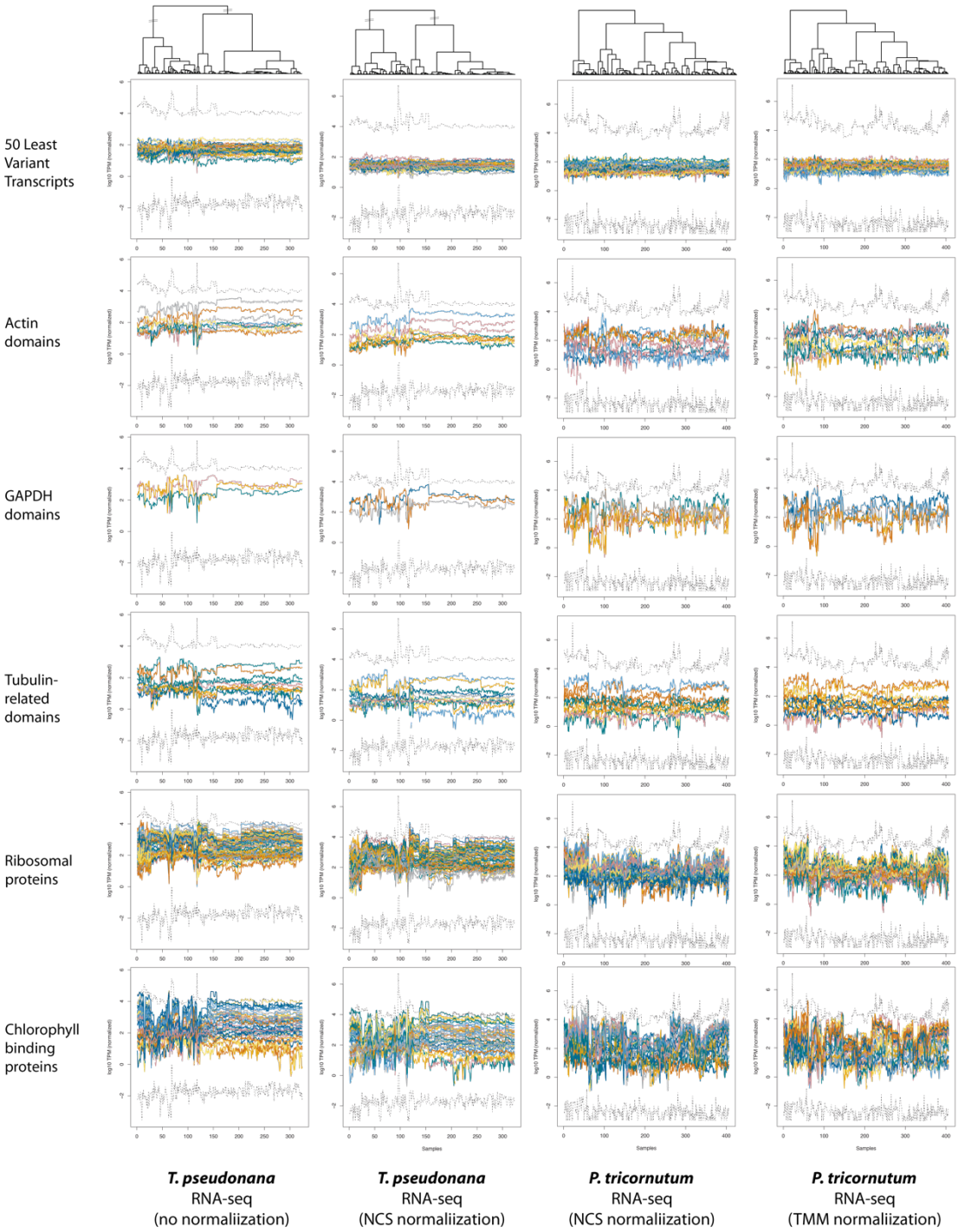

Figure S2. Combined RNA-seq datasets for *T. pseudonana* and *P. tricornutum*, illustrating characteristic results of two routine batch normalization algorithms: “network centrality scaling” (NCS) and “trimmed mean of means” (TMM). The transcripts per million (TPM, y-axis) for individual transcripts over all samples (x-axis) are shown as arbitrarily colored lines. Dashed lines indicate the within-sample minimum and maximum range of TPM values over all transcripts.

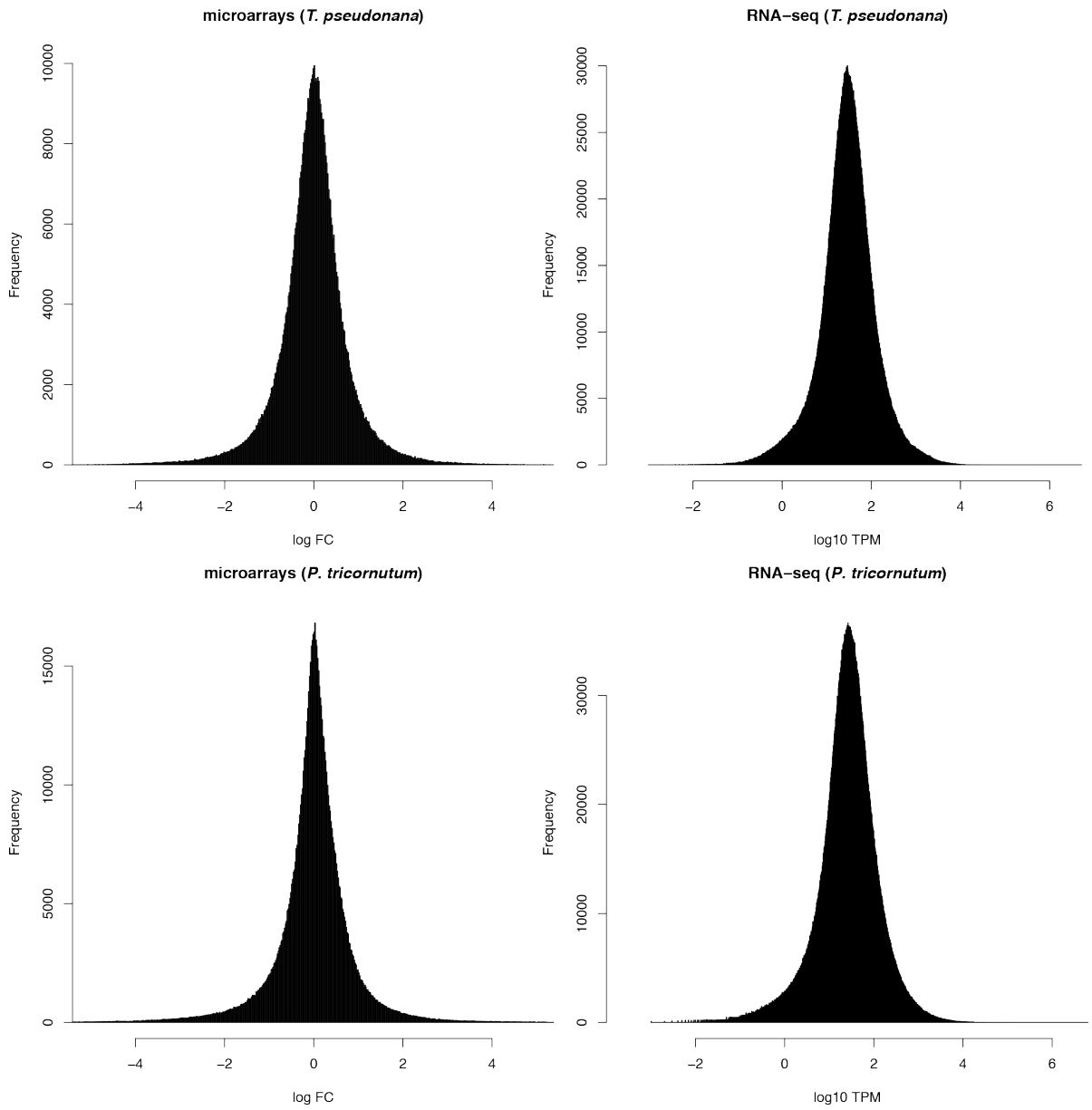

Figure S3. Distributions of microarray fold change (FC) (left) and RNA-seq normalized transcripts per million (TPM) values (right) included in the integrated datasets for *T. pseudonana* (top) and *P. tricornutum* (bottom).