Differential Expression in Unicellular Green Algae During Self-competition

Charles A. Goodman and Charles F. Delwiche
Cell Biology and Molecular Genetics, University of Maryland

National Institutes of Health

Introduction

Marine and freshwater environments are vital resources to all life on earth and are among the biomes most severely affected by human impact, and the ability to responsibly manage these ecosystems is impeded by a need for deeper understanding of their diverse networks of interactions. Green algae are ecologically important as basal members of food webs[1] and in terms of net primary productivity[2,3]. Clade-specific trends in invasion and competition have previously been demonstrated[4,5] **(Figure 1)**, though community dynamics among primary producers are poorly understood, and historic paradigms don't adequately explain trends in competition[5,6]. As part of a larger collaborative effort to determine the potential for phylogenetic relatedness to predict outcome of competition among planktonic green algae, RNAseq data were collected to track gene expression through a series of inter- and intra-specific competition growth curves. Here, we present eight de novo transcriptome assemblies and corresponding timecourse differential gene expression analyses from species broadly distributed across the green-algal tree. Our results show distinct phases of gene expression across the monoculture growth curves, reflecting the effects of intraspecific competition. In addition, they indicate substantial differences among species in patterns of gene expression within commonly shared orthogroups. These observations provide baseline patterns of expression for a planned subsequent analysis of biculture competition data.

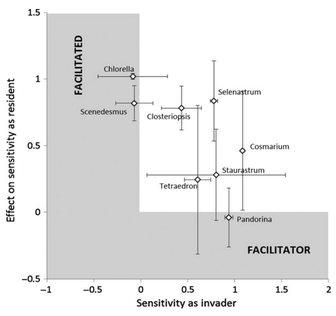


Figure 1. Sensititvity to Invasion - from Venail et al. 2014

A related study focused on the effects of invasion on growth.

The X-axis shows effect on growth rate when invading an established steady-state culture: increasing sensitivity indicates a decreasing growth rate as compared to monoculture. Values approaching 0 indicate an agressive invader, values approaching 1 indicate decreasing ability to invade.

$$S_i = (g_i[\text{mono.}] - g_i[\text{inv.}])/g_i[\text{mono.}]$$

The Y-axis is an aggregate measure of how much each species affected invader's growth rates. A high effect suggests that invader's rates were slowed, a low effect indicates that invader's rates were not affected.

$$E_i = (S_j[i] + S_k[i] + \ldots + S_p[i])/n$$

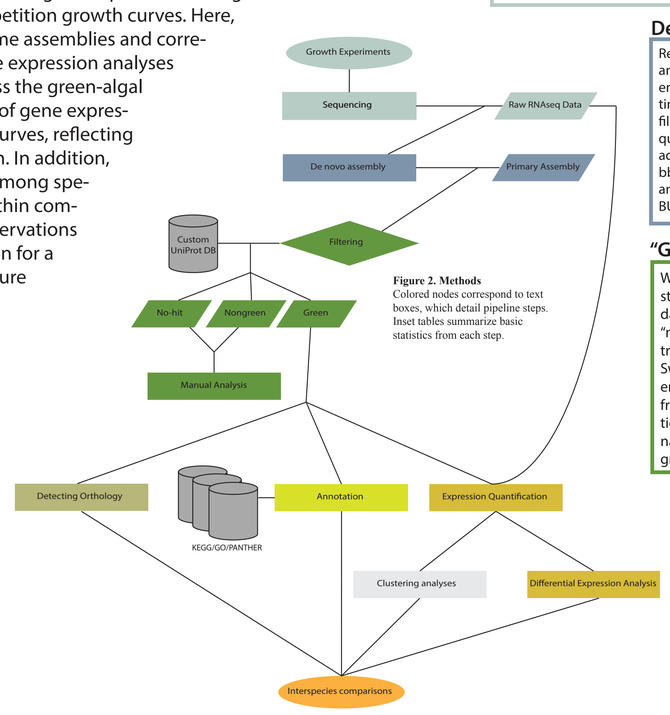


Figure 2. Methods
Colored nodes correspond to text boxes, which detail pipeline steps. Inset tables summarize basic statistics from each step.

Growth

This study is a novel analysis of RNAseq data coming from competition growth assays done in collaboration with Narwani et al. (2017)[6]. Eight species of unicellular, planktonic, freshwater green algae (table 1) were selected for their co-occurrence in nature, compatible culture requirements, and visually distinct morphologies. Species were grown as monocultures and in their 28 possible pairwise combinations, in triplicate **(Figure 3a)**. 1L roller-bottles containing enriched COMBO growth medium were inoculated at 200 cells ml-1, and cultures were grown at 20C under a 16h:8h light:dark cycle for 46 days. 10% of each culture was removed every other day, used for density monitoring (via chlorophyll-a fluorescence), cell/colony-counts (via flow-cam), and preservation for RNA extraction. Growth curves were calculated based on cell-count data. RNA was extracted and Illumina HT sequencing was performed. For monocultures, RNA was sequenced from four time-points along the 46-day growth curve, chosen to reflect early (T1), pre-inflection log phase (T2), post-inflection log phase (T3), and carrying-capacity (T4) growth densities **(Figure 3b)**. The scope of this analysis includes only monoculture data; bicultures will be assesed in future efforts. Species images and growth curves are shown in **Box 1**.

Table 1. Experimental IDs

Species	ID
Chlorella sorokiniana	DC10
Closteriopsis acicularis	DC20
Cosmarium turpinii	DC30
Pandorina charkoweinsis	DC40
Scenedsums acuminatus	DC50
Selenastrum capricornutum	DC60
Staurastrum punctulatum	DC70
Tetraedron minimum	DC80

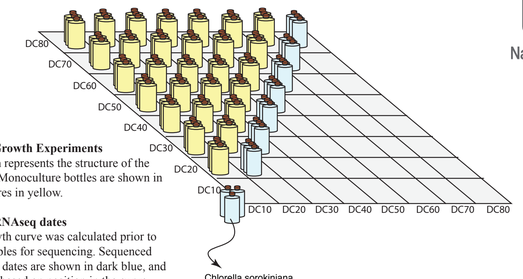
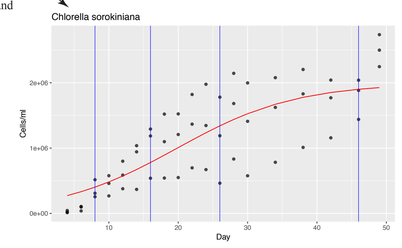


Figure 3a. Growth Experiments
This diagram represents the structure of the experiment. Monoculture bottles are shown in blue, bicultures in yellow.

Figure 3b. RNAseq dates
The red growth curve was calculated prior to chosing samples for sequencing. Sequenced RNA sample dates are shown in dark blue, and were chosen based on position in the curve.



De novo assembly

Reference genomes remain unavailable for these taxa, and so we rely on de novo assembly. To generate reference transcriptomes, paired-end sequence data from all time-points were concatenated into a single pair of fastq files per species. FastQC[7] was used to estimate sequence quality; reads below quality threshold and artifactual adapter sequences were trimmed from the set using bbduk[8]. RNAspades[9] was used for de novo assembly, and transcriptome completeness was assessed using BUSCO's[10] Chlorophyte and Eukaryote databases.

Table 2. Reference Transcriptome Assembly

ID	DC10	DC20	DC30	DC40	DC50	DC60	DC70	DC80
Reads (90bp)	2.53E+07	2.68E+07	2.85E+07	1.88E+07	2.92E+07	2.75E+07	2.42E+07	2.58E+07
Unique Contigs	80151	124149	219575	110843	118654	153536	221916	243395
Total Length (bp)	6.93E+07	8.47E+07	1.26E+08	9.68E+07	9.51E+07	7.07E+07	1.43E+08	1.35E+08
Coverage	32.84X	28.42X	20.26X	17.5X	27.67X	34.95X	15.2X	17.13X
N50 (bp)	1930	1842	1235	2309	2554	825	1090	1887
Missing BUSCOs†	8.63%	8.12%	28.18%*	8.76%	10.65%	11.35%	43.73%*	8.63%

† BUSCO database: chlorophyta_odb10, n= 2168
* Charophyte

"Green" Filtering

We obtained transcript annotations with an integrated a filtering step, using BLASTx[11,12] to compare each transcript against a protein database containing both "green" plant and algal sequence, and "nongreen" sequence from representative taxa across the rest of the tree of life. Protein data were drawn from UniProtKB's reviewed Swiss-Prot and unreviewed TrEMBL databases[13,14]. Queries that preferentially matched non-green subject sequences were separated from the green dataset for additional processing: particular attention will be paid to non-green data that indicates culture contamination and/or fellow-tavelers. Queries that preferentially matched green subjects were carried forward in the presented analysis.

Table 3. "Green" Filtered Transcriptomes

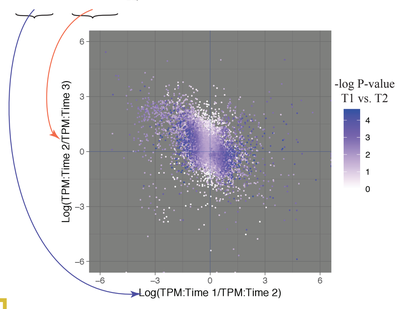
ID	DC10	DC20	DC30	DC40	DC50	DC60	DC70	DC80
Total Hits	34657	32769	62993	37852	32120	56703	61529	45492
Green Hits	27362	25820	40833	33141	25297	48508	54073	20273
Green Tot. Length (bp)	4.90E+07	5.01E+07	5.22E+07	6.74E+07	5.57E+07	3.93E+07	6.15E+07	3.65E+07
% of Reference Trans.	70.73%	59.13%	41.30%	69.59%	58.54%	55.53%	42.97%	26.97%
N50 (bp)	2471	2616	1872	3169	3539	1373	1514	2502
Green Missing BUSCOs †	9.36%	8.21%	29.38%*	8.81%	10.70%	11.35%	49.31%*	9.64%

† BUSCO database chlorophyta_odb10, n= 2168
* Charophyte

Quantification and Differential Expression Analysis

Culture densities were previously measured as described in Narwani et al. 2017. Preliminary growth curves were estimated using the original data, via GrowthCurver[23], an R package which fits cell density data to a standard logistic curve. Improved growth models are currently under consideration, including the use of a discrete (rather than continuous) curve, as well as accounting for periodic harvesting[24]. To quantify gene expression, I employed Kallisto[25], chosen for its use of a hash-based quantification approach which has benchmarked well against other common methods, running orders of magnitude more quickly than comparable software packages[25] and integrating bootstrapped estimates of expression variance. I obtained counts for each of the monocultures by counting trimmed reads against their respective reference transcriptomes, running Kallisto with 100 bootstrapped replicates.

I used sleuth[26,27,28] to identify significant differential expression of transcripts between all consecutive pairs of time points in each dataset. Sleuth is unique among DE analytical packages in that it combines biological and 'inferential' variance in its estimate, basing the latter on the bootstrapped count values produced with Kallisto. I separately compared T1/T2, T2/T3, and T3/T4 – where the early time point was taken as the 'control' value, and the later time point was the 'test' value. Sleuth has been shown to over-estimate false-discovery rate[26], though a conservative estimate is desirable as fewer genes are reported as 'significant', but those reported tend to be enriched for true differential expression. I define "significant DE" as FDR < 0.1 and |2LFC| > 1, and "constitutive expression" as transcripts for which |2LFC| = 0±0.2 across all growth-curve time points. Summary figures are shown in **Box 2**, **Fig. 3c**, **Box 3**, and via **QRC 1**.



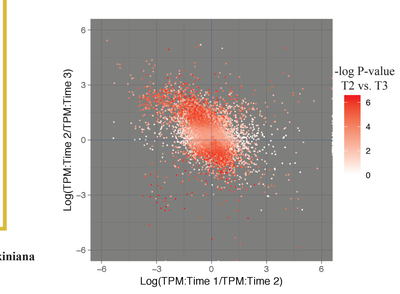


Figure 3c. Total Differential Gene Expression of C. sorokiniana in Time 1, Time 2, and Time 3

Typical DE analyses often utilize "Volcano Plots" as a method of visualizing change in expresison between two times. A normalized value such as RPKM or TPM for each gene is plotted as a point; the X axis displays either log-fold change or log-ratio of expression between conditions, and the Y-axis shows the negative log P-value - see **Box 2** for an example of this. Distance from the origin on the X-axis indicates the amount of change in expression, and in the Y-axis represents significance.

Here, in an attempt to generate a more dynamic view of gene expression over time, we visualize three timepoints in a single plot. Both the upper and lower plots contain the same exact points and axes, and they differ only by coloration.

Each point represents expression of a single transcript over time. The X-axis shows the log ratio of TPM (Transcripts per Million) between timepoint 1 and timepoint 2, the Y-axis shows the log ratio of TPM between timepoint 2 and timepoint 3. Thus: points in the upper right quadrant continually increased in expression from T1 to T2 to T3. In the lower left qudrant, expression decreased through all timepoints. In the bottom right, expression increased from T1 to T2, and decreased from T2 to T3. Note: we use Log base 2, so a value of 1 indicates that expression doubled, a value of 2 indicates expression quadrupled.

The coloration represents the negative log P-value for each point, as determined by sleuth - this is effectively a measure against total noise, and is based on the variance among repeats of a gene vs. variance of the total set. Deeper color indicates higher "significance." Purple refers to significance between T1 vs. T2, and red between T2 and T3.

The benefit of this visualiztion method is apparent when viewing individual gene families, as in **Box 3**, where we can evaluate the extent of change in groups of related genes over multiple timepoints. In an attempt to generate a visualization including ALL available timepoints, we added the log ratio of TPM between T3 and T4 on a Z-axis, highlighting genes annotated with a functional GO term, "Photosynthesis, Light Harvesting". The 3D plot isn't possible to convey here, and so we invite you to follow a link via the inset **QRC 1**:

QRC 1

Annotation

After identifying transcripts likely to be of algal origin, protein coding regions are identified through prediction of open reading frames (ORFs) using TransDecoder[17]. For each transcript, this software leverages hidden-Markov models (HMMs) for all likely ORFs against the PfamA[18] database, returning only sequences which have a putative functional annotation represented therein. To supplement these annotations, we used HMMER[19] to compare the transcripts to three additional protein and gene family databases: KEGG[20], GO[21], and PANTHER[22]. A strong majority of transcripts passing the green filter were successfully annotated.

Table 4. Annotation

ID	DC10	DC20	DC30	DC40	DC50	DC60	DC70	DC80
"Green" Trans. PfamA IDs	24660	23843	36533	28379	22872	37645	48338	18266
% "Green" w/ PfamA ID	90.12%	92.34%	89.47%	85.63%	90.41%	77.61%	89.39%	90.10%
"Green" Trans. PANTHER IDs	24142	23917	38452	24698	21990	36647	52221	18497
% "Green" w/ PANTHER ID	88.23%	92.63%	94.17%	74.52%	86.93%	75.55%	96.58%	91.24%
PANTHER IDs "family not named"	8297	8653	15485	8562	7646	12286	20721	6532

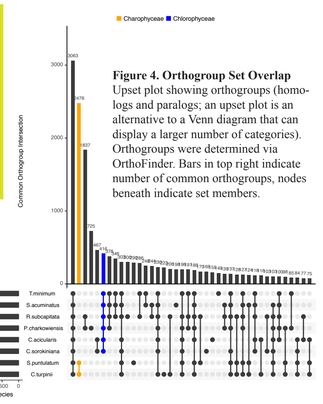


Figure 4. Orthogroup Set Overlap
Upset plot showing orthogroups (homologs and paralogs; an upset plot is an alternative to a Venn diagram that can display a larger number of categories). Orthogroups were determined via OrthoFinder. Bars in top right indicate number of common orthogroups, nodes beneath indicate set members.

Detecting Orthology

While looking at reciprocal best blast hits is a known strategy in establishing orthology, evidence suggests it isn't universally reliable in predicting nearest-neighbor relationships[29]. I therefore used OrthoFinder to identify and bin orthologous sequence across our eight species – this has the added advantage of considering same-species orthologs and gene isoforms. OrthoFinder's basic output bins genes by "orthogroups", which does not automatically distinguish ortholog from paralog, but does provide a basis for doing so with additional phylogenetic analysis. I've identified 3156 orthogroups which are common to all 8 species, with between 38.4-50.1% of genes belonging to these groups, and also identified a number of species-unique orthogroups. Groupings of shared OrthoGroups are shown in **Figure 4.**

Table 5. OrthoFinder Results

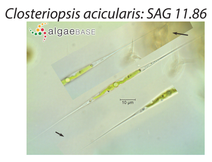
ID	DC10	DC20	DC30	DC40	DC50	DC60	DC70	DC80
OrthoGroups	7672	8219	10905	10939	87722	14044	10081	9191
Genes Identified †	26738	25519	40367	32492	24292	46862	53560	19914
Genes Assigned	19654	20788	30058	22503	18260	33913	39815	16221
% Genes Unassigned	26.49%	18.54%	25.54%	30.74%	24.83%	27.63%	25.66%	18.54%
Common Genes ††	9935	9921	13090	8643	8301	13177	17837	6758
% Common	50.55%	47.72%	43.55%	38.41%	45.46%	38.86%	44.80%	41.66%
Unique Orthogroups	74	54	39	31	44	52	102	22

† Incl. all transcript isoforms
†† Genes mapped to 3156 orthogroups common to all 8 species

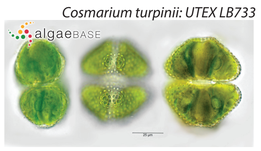
Box 1. Species & Growth Curves

Chlorella sorokiniana: UTEX 2805

Closteriopsis acicularis: SAG 11.86



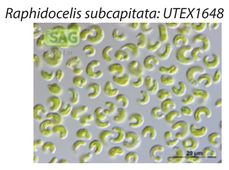
Cosmarium turpinii: UTEX LB733



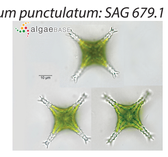
Pandorina charkoweinsis: UTEX LB840

Scenedesmus acuminatus: SAG 38.81

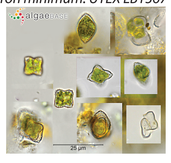
Raphidocelis subcapitata: UTEX1648

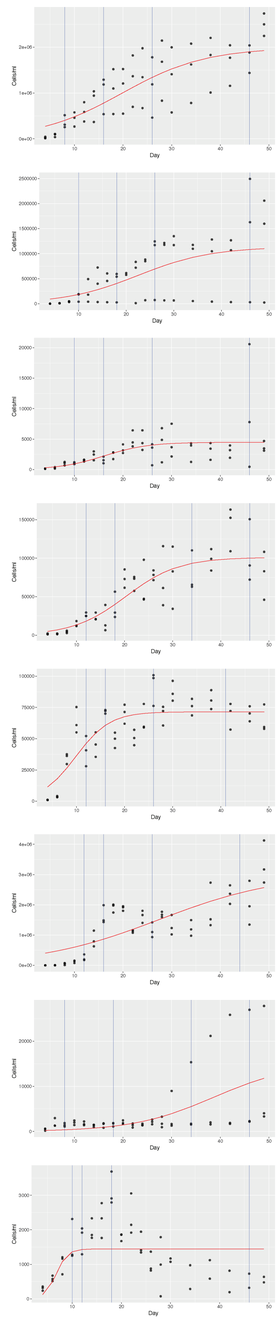


Staurastrum punctulatum: SAG 679.1



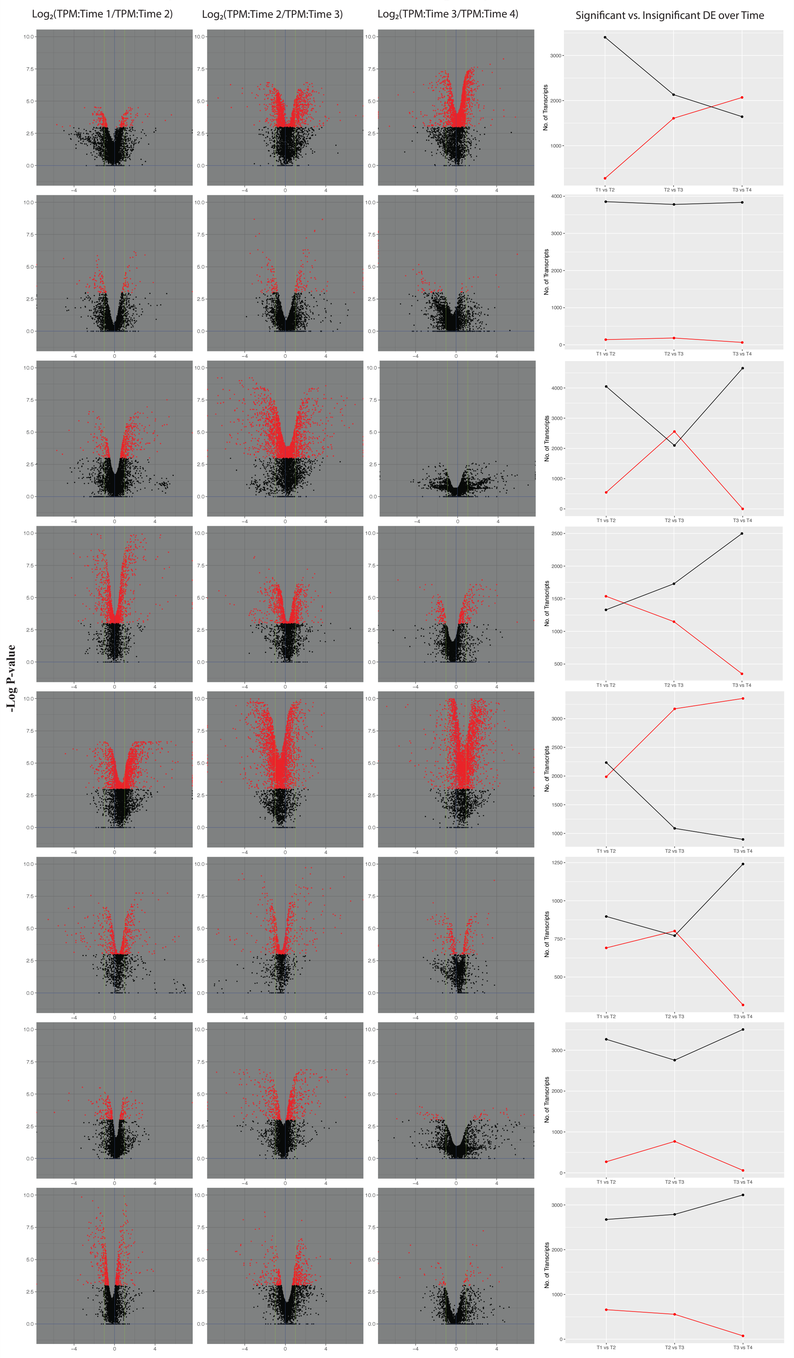
Tetraedron minimum: UTEX LB1367





Cell Count Value
Derived Growth Curve
RNA sample dates

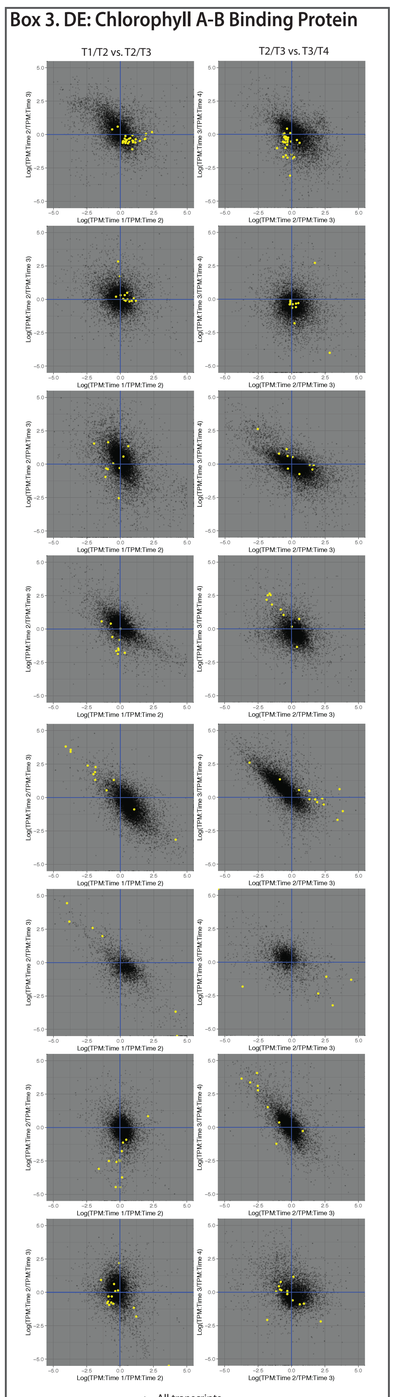
Box 2. Differential Epression (DE) of Common Orthogroups



Mean TPM, P>0.05
Mean TPM, P<=0.05
2LR +/- 1
No. Transcripts P>0.05
No. Transcripts P<=0.05

This is a series of volcano plots demonstrating differential expression of orthogroups common to all species. Each plot shows a pair of sequential timepoints; T1 to T2, T2 to T3, or T3 to T4. Individual plot points represent, for each annotated transcript, the log ratio of average TPM between two timepoints. TPM, or transcripts per million, are an estimate of gene expression normalized by the length of the transcriptome. Per gene, estimated counts are divided by the length of the gene in kilobases, giving "reads per kilobase", or RPK. Total RPK across the transcriptome are summed and divided by 1,000,000 - giving the "per million" scaling factor, SF. RPK/SF = TPM. A log base-2 ratio of TPM indicates proportional change over time, e.g. a value of 1 would indicate a twofold increase in expression, and a value of -1 would indicate a twofold decrease. Each row of plots corresponds to the species growth curve to the left, and reading the plots from left to right indicates a general profile of change in total (common) gene expression over time. Points in red indicate a P-value <0.05. The line graphs in the right column simply compare total numbers of "significant" vs "non-significant" log ratios of expression, as a way to visualize the effect of self-competition (i.e. growing to capacity) on gene expression.

Box 3. DE: Chlorophyll A-B Binding Protein



All transcripts
"Chlorophyll A-B Binding"

Here we show a series of plots modeled after **Figure 3c**, which instead of P-value, highlight expression of transcripts coding for Chlorophyll A-B Binding protein. X-axes show the log ratio of TPM for the first two time points, Y-axes show the log ratio of TPM of the second and third.The left colum effectively shows expression from T1-T2-T3, and the right column shows T2-T3-T4. Compellingly, we see distinct differences in grouping and orientation when comparing the 8 species (rows), which suggests that they show unique responses to self-competition, in terms of expression of this gene family.

Conclusions

For eight previously unsequenced species of unicellular green algae, we present annotated reference transcriptomes and differential gene expression analyses representing the effects of self competition - or growth to carrying capacity in monoculture.

We generated matching gene expression profiles with between 20273-50730 verified "green" gene transcripts per species. Of these, from 77.6%-92.3% were succesfully annotated based on sequence similarity to available public data, and are pending functional analysis via classical molecular techniques. Within each taxon, between 38%-50% of total transcripts belong to 3083 orthogroups common to all eight taxa, which provides a wide basis for interspecies comparisons. In addition, we have identified all sets of common orthogroups across smaller subsets of species. Of particular interest are the potential differences between the chlorophyte and charophyte-specific gene sets, shown in blue and orange in **Figure 4**.

In **Box 2**, we show that among common orthogoups, overall changes in expression are evident throughout the growth curve, indicating that self-competition drives differential gene expression. In addition, we see that the patterns of total DE are distinct among species, presumably reflecting how each species physiologically responds to competition: this may provide insight into observed patterns of competitive success as reported by Venail et al.[4]

In **Box 3**, we demonstrate our ability to narrow our analysis to gene-family resolution by pulling expression profiles for any of the thousands of identified gene families in the dataset. Within the plotted family, "Chlorophyll A-B Binding Protein" (Orthogroup 4), we see clear differences in the expression profiles among our taxa. We are now at the point of determining how to validate such observations: a challenge with *de novo* assembly arises in distinguishing true gene copies from assembly artefacts. We will use qPCR to validate our observations at the bench.

Our long-term goal is to determine whether competitiveness can be explained in terms of differential gene expression. The monoculture data presented here will serve as a baseline for comparison to a much larger RNAseq dataset representing the biculture competitions, which will allow us to examine the similarites and differences between intra- and extra-specific competition, and to study the degree to which gene expression predicts competition.

Acknowledgements:
Thanks to our prior collaborators A. Narwani, B. Bentlage, B. Cardinale, P. Venail, M. Alexandrou, T. Oakley, K. Fritschie, and J. Thierer for their efforts in carrying out the growth and sequencing experiments. Thanks to Tim Wood for his guidance in the early stages of developing the bioinformatic pipeline, and to the current and past members of Delwiche lab for their support. CAG was supported by the Cell and Molecular Biology Training Grant T32 GM080201 to the University of Maryland. CFD was supported in part by the University of Maryland Agricultural Experiment Station.

1. Cadotte, M. W., Dinnage, R. & Tilman, D. Phylogenetic diversity promotes ecosystem stability. Ecology 93, S223–S233 (2012).
2. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281, 237–40 (1998).
3. Delwiche, C. F. & Cooper, E. D. The evolutionary origin of a terrestrial flora. Current Biology (2015). doi:10.1016/j.cub.2015.08.029
4. Venail, P. A. et al. The influence of phylogenetic relatedness on species interactions among freshwater green algae in a mesocosm experiment. J. Ecol. 102, 1288–1299 (2014).
5. Alexandrou, M. A. et al. Evolutionary relatedness does not predict competition and co-occurrence in natural or experimental communities of green algae. doi:10.1098/rspb.2014.1745
6. Narwani, A. et al. Ecological interactions and coexistence are predicted by gene expression similarity in freshwater green algae. J. Ecol. 105, 580–591 (2017).
7. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (2010). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 3rd December 2018)
8. Bushnell, B. BBMap | SourceForge.net. JGI (2018). Available at: https://sourceforge.net/projects/bbmap/. (Accessed: 3rd December 2018)
9. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. bioRxiv 420208 (2018). doi:10.1101/420208
10. Waterhouse, R. M. et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548 (2018).
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
12. Altschul, S. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997).
13. Bateman, A. et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169 (2017).
14. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212 (2015).
15. Nishiyama, T. et al. The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. Cell 174, 448-464.e24 (2018).
16. Hori, K. et al. Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nat. Commun. 5, 3978 (2014).
17. Haas, B. J. & Papanicolaou, A. TransDecoder.
18. Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins Struct. Funct. Genet. 28, 405–420 (1997).
19. EDDY, S. R. A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE. in Genome Informatics 2009 205–211 (PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO., 2009). doi:10.1142/9781848165632_0019
20. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30 (2000).
21. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, 258D – 261 (2004).
22. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41, D377-86 (2013).
23. Sprouffske, K. GrowthCurver.
24. Tschirhart, J. Integrated Ecological-Economic Models. Annu. Rev. Resour. Econ. 1, 381–407 (2009).
25. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).
26. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods 14, 687–690 (2017).
27. Yi, L., Pimentel, H., Bray, N. L. & Pachter, L. Gene-level differential analysis at transcript-level resolution. Genome Biol. 19, 53 (2018).
28. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31, 46–53 (2013).
29. Koski, L. B. & Golding, G. B. The Closest BLAST Hit Is Often Not the Nearest Neighbor. J. Mol. Evol. 52, 540–542 (2001)

cagood@umd.edu, delwiche@umd.edu

Link: C. Goodman GitHub repositories

Currently, data and pipeline steps are privately held, pending publication.