Differential Gene Expression in Green Algal Competition

Charles Goodman

Marine and freshwater environments are vital resources to all life on earth and are among the biomes most severely affected by human impact; the ability to responsibly manage these ecosystems is precluded by a need for deeper understanding of their diverse networks of interactions. Green algae are nearly ubiquitous in terrestrial-aquatic, freshwater, and marine environments, and despite this, algal community dynamics are poorly understood at the molecular level. With the advent of high throughput sequencing, it is now possible to observe entire algal populations at genetic resolution, granting the potential for a novel and sophisticated *mechanistic* perspective on community dynamics. In this project, I propose a meta-transcriptomic analysis of pairwise competitions of 8 species of green algae which span more than a billion years of evolution. Though this, I aim to elucidate transcriptomic features which explain observed trends in interspecies interaction: from competition to stable coexistence to facilitation of growth.

Background & Rationale

Green algae are evolutionarily and ecologically vital, yet critically understudied. Green algae, comprising the chlorophyte and charophyte lineages, are the extant link between early single-celled eukaryotic phototrophs and terrestrial plants[1]. By carrying plastids capable of oxygenic photosynthesis from their cyanobacterial origins onto land, they were a necessary step in the development of the oxygenated atmosphere on which terrestrial life depends[1–3]. Indeed, the single largest acreage of primary productivity on earth is comprised of marine phytoplankton, which dominates net primary production in the ocean and is responsible for roughly half of Earth's annual carbon capture and oxygen production[4]. Due both to their basal position in aquatic food webs and to their relative sensitivity to environmental change, the algae are considered to be strong candidates as indicator species for overall ecosystem health[5,6]. Green algae are a vital component of the biosphere, and despite their importance, research in phycological community structure, and even in their basic genetics, has yet to benefit from the capabilities provided by next-generation sequencing and computational biology. To date, there does not exist a fully realized charophycean model system[7], and only a fraction of phytoplankton species have a fully sequenced genome[8], with green algae in particular being poorly represented in publicly available data (*Table 0*).

Historic paradigms may not adequately describe inter-species interaction. It has been shown that an ecosystem's diversity has significant positive impact on its stability[9,10], adaptability[11], resilience[12], and overall productivity[13], and a long-standing goal in community ecology has been to determine the dynamics of inter-species interaction. Within the plant and animal kingdoms predation, parasitism, competition, and mutualism are often readily observed and quantifiable phenomena. Some of these interactions are well-characterized in phytoplankton, especially competition[14] and symbiosis[1,15]. The study of inter-species competition goes back at least to Darwin, who observed that distantly related plants tended to be more successful in invasion than closely related plants, which ultimately led him to describe the "naturalization hypothesis": non-native species, being evolutionarily distant and phenotypically distinct, would be more likely to succeed when invading a new environment[16]. This observation has since been formalized into a logically appealing pair of hypotheses: "phylogenetic niche conservatism" (PNC), which suggests that recently evolved relatives tend to be more phenotypically similar than distally related species, and the "competition-relatedness hypothesis" (CRH) which posits

that closer relatives, being more phenotypically similar, will necessarily occupy more deeply overlapping niches than distal relatives, and will thus be more likely to compete[17]. This mechanism of niche partitioning forms the basis for most accepted explanations of biodiversity and interspecies competition. Support for the concept is widespread[18–25], and it has been cited in a number of applied ecological fields like invasion prediciton[18,26] and restoration ecology[27,28]. Support is not unanimous, however, and attempts to recapitulate Darwin's observations and the notion of PNC/CRH have not all been successful[29,30]. In particular, algae have been frequently claimed to defy traditional models based on phylogenetic distance (PD)[14,31–34].

Phylogenetic distance does not directly predict competition outcome. Motivated by these claims, and prior to my involvement, the Delwiche lab joined a large collaborative effort funded by an NSF DIMENSIONS of Biodiversity grant. The first component of this effort introduced and utilized an improved estimate of phylogenetic distances of all available freshwater green algae[33,35]. Venail et al. (2014) reviewed 15 empirical studies linking phylogenetic relatedness and strength of competition, finding that two thirds of studies failed to support CRH, and that more than half of these studies described instances of *facilitative* interactions between species, wherein one or both species accomplished a higher overall density in biculture as compared to monoculture[35]. To independently test whether PD has an effect on competition, Venail et al. ran a series of invasion experiments whereby an algal species was introduced to an established culture of another; a slightly negative correlation was found between PD and probability of successful invasion (*Figure 1*)[35]. In direct contrast to the expected trend, as PD increased, invasion success tended to decrease. To address the potential criticism that failure to support CRH sometimes results from the use of a poorly resolved phylogeny, the group sequenced 53 new green algal transcriptomes and produced a definitive and highly resolved green algal phylogeny (*Figure 2*)[33]. With this, they reassessed the results of three prior studies[31,32,35] and conclusively showed that it was not possible to predict interaction type or strength based on PD alone. The second component of the grant involved qualitative and observational mesocosm experiments in the field[31,33].

The third component, most relevant to this proposal, endeavored to determine the extent to which phenotypic variation at the *molecular* level affected interaction type between algal species[17]. In this study, 8 species of single celled green algae were grown as monocultures and in all possible pairwise combinations. Cultures were grown past steady-state carrying capacity, with relative biomass measured on alternate days via chlorophyll-a fluorescence and cell count. Distinguishing this study was the use of high-throughput sequencing of RNA collected at early-stage, pre-inflection log, post-inflection log, and late-stage growth for all cultures, in triplicate. Species interactions were quantified by fitting Lotka-Volterra competition models to time-series density data and ranged from competitive to facilitative. Gene-expression similarity (GES) between species was calculated via spearman-rank correlation of end time-point expression level of common genes, and phylogenetic distance was estimated from the tree produced by Alexandrou et al.[33]. It was found that, as predicted, GES and PD were negatively correlated, meaning that gene expression diverged with increasing PD (*Figure 3a*). Notably, and contrary to the CRH-based prediction, similarity in gene expression was found to be *positively* correlated with tendency to coexist (*Figure 3b*). It was also found that expression level in 15/17 *a priori* selected gene families was correlated with tendency to "over-yield" in biculture – signifying a facilitative relationship between species[17], and indicating an unmet need to go beyond phenomenological description and examine the mechanistic underpinnings of algal interactions.

An underutilized dataset presents a valuable opportunity. I believe that the transcriptome data produced by Narwani et al. (2017) has the potential to provide much deeper insight into physiological

activity at the cell level. The previous results were limited by three simplifying assumptions which could have masked significant findings. First, the group examined only the beginning and end-state timepoints of gene expression. Because the logarithmic phase of growth could reflect a critical point in the establishment of dominance in bicultures, it may be possible that DE in the early or late-log phases of growth could be significant in determining interaction type. Second, the aggregate measure of GES only accounted for homology across all 8 species and neglected to consider the presence of unique genes or distinct gene-family structures. Finally, their manual analysis was limited to 17 *a priori* selected gene families, which omitted a significant portion of the available data from closer analysis. Here, I propose an analytical pipeline that will facilitate a more complete examination of an existing dataset. I believe by widening the scope of analysis and shifting the focus from phylogenetic and phenomenological to phenetic and mechanistic, that I can elucidate evolutionarily significant trends in patterns of gene expression and provide significant insight to the observed trends in interaction among species pairs.

Experimental Plan

Aim 1. Establish baseline trends in differential gene expression (DE) across the growth curve among monocultures of eight species of green microalgae.

Specific Hypotheses:

H1. Differential gene expression results from intraspecific competition.
H2. There exist lineage-specific patterns of gene regulation and metabolic pathway usage.
H3. There are discrete patterns of gene expression that vary by phylogenetic distance.

Aim 1.1 Assemble and annotate de novo transcriptomes for each species. Sequencing was performed at four timepoints along the 46-day growth curve reflecting early (T1), pre-inflection log phase (T2), post-inflection log phase (T3), and carrying-capacity (T4) growth densities . To generate reference transcriptomes, all paired-end sequence data from T1-4 were concatenated into a single pair of fastq files per species. FastQC[36] was used to estimate sequence quality; reads below quality threshold and artifactual adapter sequences were trimmed from the set using bbduk[37]. RNAspades[38] was used for *de novo* assembly, and transcriptome completeness was assessed using BUSCO's[39] *Chlorophyte* and *Eukaryote* (pending) databases.

I obtained transcript annotations with an integrated a filtering step, using BLASTx[40,41] to compare each nucleotide reference transcriptome 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[42,43] (for the full list, see *Supplemental Table 1*). Queries that preferentially matched non-green subject sequences were separated from the green dataset for additional processing: I will blast non-green and unidentified transcripts against the NCBI non-redundant database, as well as the recently available *C. braunii*[44] and *K. nitens* (prev. *K. flaccidum)*[45] genomes, in order to identify previously overlooked green transcripts. Particular attention will be paid to non-green data that indicates culture contamination. After identifying transcripts likely to be of algal origin, protein coding regions are identified through prediction of open reading frames (ORFs) using TransDecoder[46]. This software leverages hidden-Markov models (HMMs) for all likely ORFs against the PfamA[47] database, returning only sequences which have a putative functional annotation. To supplement the annotations, I used HMMER[48] to compare the transcripts to three additional protein and gene family databases: KEGG[49], GO[50], and PANTHER[51].

Aim 1.2 Distinguish constitutive and differential gene expression across the growth curve. Culture densities were previously measured as described in Narwani et al. 2017. Preliminary growth curves were estimated using the original data, via GrowthCurver[52], 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[53]. To quantify gene expression, I employed Kallisto[54], 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[54] 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[55–57] 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[55], 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. By showing that there is both constitutive and differential gene expression, and that DE correlates to culture density across the growth curve, I can test my first hypothesis, that *differential gene expression results from intraspecific competition (H1).*

Aim 1.3 Perform a comparative transcriptomic analysis of the 8 species. Narwani et al. previously binned transcripts by 'gene family' via PANTHER HMMs[17]. This may be problematic, as PANTHER organizes gene families by functional pathway, which results in multiple distinct metabolic steps being grouped. In the preliminary results, I show that transcripts belonging to a single functional bin can exhibit very different expression profiles. I've thus opted to proceed with an analysis centered around single orthogroups, positing that this metric is more indicative of the actual *in vivo* protein expression profile. To that end, I inferred orthologous sequences across the 8 transcriptomes using OrthoFinder[58]. An advantage of using this software package (as opposed to a strategy relying on reciprocal best BLAST hits[59]) is that it groups divergent sequences according to their last common ancestor within the dataset, regardless of differential splicing. The significance of this is that I'm currently limited to transcriptome data containing only post-translationally modified sequence, and through this can detect and distinguish gene isoforms that might be differently grouped if relying solely on manually sorted local alignments.

Though constitutive gene expression has been described in microalgae[60], a comprehensive assessment across the major clades hasn't been made. The structure of this experiment provides the opportunity to examine a taxonomically diverse set of green algal species for the presence of such a gene set. While the scope of our data limits the analysis to the eight species in question, I'd be able to describe a rationally-identified set of genes that are constituently expressed across more than a billion years of evolution, using manual analysis of functional annotations to make inferences about conservation of constitutive cellular processes that span this evolutionary window. For differentially expressed genes, I hypothesize that *there exist lineage-specific patterns of gene regulation and metabolic pathway usage (H2).* To that end, I will employ at least two strategies. First, I'll examine the total gene set and compare expression profiles, as determined by sleuth DE analysis, among groups common to more than one species. Preliminary results indicate that there are multiple modes of expression beyond constitutive, constantly

increasing, or constantly decreasing, and it may be that these profiles are not consistent for orthologous genes across species. The second strategy will involve a manual analysis of gene families beginning with the *a priori* selected set described in Narwani et al. 2017[17]. A 2015 review of nitrate assimilation in microalgae shows that different chlorophyte species maintain distinct gene family members in the nitrate assimilation regulatory pathway[61] examined in the prior study. This suggests the possibility of differences among other families and justifies the effort in producing a species-wide comparison. Another strategy to employ is WGCNA[62], which would serve to identify co-expression networks directly from the data, thus giving insight into clusters of genes beyond "functional families" manually pulled from the annotation.

I also hypothesize that *there are discrete patterns of expression that vary by phylogenetic distance (H3).* Basing their distances on the phylogeny presented by Alexandrou et al. (2015), the group previously determined that 'gene expression similarity' (GES) – measured by spearman rank correlation of TPM values for common genes – decreased with phylogenetic distance[17]. GES, as used in the previous study, was a generalized aggregate measure of similarity across the entire transcriptome. I aim to re-evaluate the approach using more specific metrics that account both for variance in gene family structure and for more dynamic modes of expression; it's possible that GES measurements (as determined previously) do not reflect gene family structure or gene expression during logarithmic growth. I believe this approach may provide a more meaningful basis for comparison of gene expression with phylogenetic distance.

Aim 2. Characterize changes in DE profiles arising from inter-species interactions and propose a predictive model for interaction types among bicultures.

Specific Hypotheses:

H4. Responses to interspecific competition are distinct from those in intraspecific competition.
H5. Patterns of gene expression vary by inter-species interaction type.
H6. Patterns of gene expression in monoculture can predict interaction type in bicultures.

Aim 2.1 Extend interaction and DE analysis to biculture dataset. As with the monocultures, raw RNA-seq and cell-density data for bicultures remain available from the prior study where previously, Lotka-Volterra (LV) competition models were used to estimate the effects of species interactions[17]. I've since noted that culture conditions included periodic harvesting: this was not accounted for in the original analysis, and I believe that it may have significantly affected the resulting estimates. In order to obtain more accurate descriptions of species interactions, I propose to re-calculate these using a modified LV model which includes a term to account for harvesting[53].

To distinguish and quantify gene expression between species in biculture, I'll return to Kallisto, mapping each set of biculture reads back to the respective pair of previously-curated monoculture reference transcriptomes. A significant shortcoming of this strategy is the fact that I am effectively blind to transcripts that are unique to bicultures: reads that do not map to either reference transcriptome are omitted from the reported counts. To mitigate this, I propose to extract all un-mapped reads and assemble them separately. By BLASTing against the NCBI nonredundant database, I may be able to resolve ambiguous transcripts by identifying matches to near relatives – though this may be unviable in the case of species pairs that are already closely related. I also considered comparing the novel transcripts' codon-usage or kmer count profiles to those of the monocultures. To identify significant differential gene expression over time, I'll again utilize sleuth, comparing consecutive pairs of timepoints

(as with the monocultures), and define "significant DE" as FDR < 0.1 and |2LFC| > 1. I'll consider constitutive expression to include transcripts where |2LFC| = 0±0.2 across all growth-curve time points.

Aim 2.2 Identify difference in expression patterns between monoculture and bicultures. Differentially expressed genes are comprised of two subgroups. The first will include transcripts mapping to the monoculture reference transcriptomes. The second group refers to uniquely assembled transcripts identified in aim 2.1. Upon quantifying expression and mapping general profiles for both subgroups, I'll move into a largely manual stage of analysis, comparing monoculture expression to biculture expression. I hypothesize that *responses to interspecific competition are distinct from those in intraspecific competition (H4)*. By comparing expression profiles for a given gene in monoculture to that same species' expression in all 7 bicultures, I'll be able to determine the connection between the identity of the competing species and the resulting expression profile. Because direct visual comparison of expression curves for tens of thousands of genes per species isn't feasible, I'll rely on initial heuristic passes on the data to identify and filter potential genes of interest. One strategy to address this could include a differential expression analysis via sleuth, where the 'control' condition includes transcripts for a given monoculture timepoint, and 'test' conditions include expression values at the same timepoint in the 7 bicultures – effectively testing all 7 conditions against each other and simultaneously against the monoculture control at individual timepoints. In addition, any transcripts which are unique to biculture can potentially support this hypothesis.

Aim 2.3 Explore potential correlation between interaction type (compete, coexist, facilitate) and gene expression patterns, then attempt to fit a predictive model to these data. I hypothesize that *patterns of gene expression vary by inter-species interaction type (H5).* Previously, Narwani et al. identified 15 gene families which were predictive of overyielding in biculture[17]. By extending my analysis to cover the entire transcriptome across all time points, I will be more able to definitively address this hypothesis, which itself begs two questions. First, *is there a unique gene-set present among facilitating pairs vs. competing pairs?* And second, i*s there significant differential expression of common genes between facilitating vs. competing pairs?* To address these, we'll group biculture transcriptomes by interaction type as described in Aim 2.1. Though downstream strategies are TBD, I'll approach the first question by binning orthogroups by interaction type, seeking genes unique to a particular group. The second question can be addressed by once again returning to sleuth, except considering bins of common orthogroups by interaction type rather than bins of transcripts by species or timepoint. I have discussed employing machine-learning strategies with collaborator Tim Wood, of the Broad Institute, to address this question, though the concept is not yet fully conceived.

I'm most interested to test the hypothesis that *patterns of gene expression in monoculture can predict interaction types in bicultures (H6).* It would be valuable to determine a model which predicts species interaction type (in terms of the values of a pair of LV interaction coefficients) based on the individual expression profiles of the competing pair. "Expression profile" has yet to be stringently defined, however various models, including information pertaining to gene family structure, expression over time, uniquely expressed genes, etc., will be considered. Candidate models can be tested *in silico* through the use of withheld data: e.g. the model can be trained on 7/8 species, and the 8th species (representing 7 biculture interactions) can then be tested. Ideally, the withheld species would be chosen based on having shown multiple interaction types (that is, LV interaction coefficients ranging from negative to positive) depending on the co-cultured partner. DC60, *R. subcapitata*, is one such example[17].

Aim 3. Test the observed DE patterns in vivo.

Aim 3.1 Repeat competition assays between the original species and use RT-qPCR to support observations derived in my first two aims. Here, I will undertake a new series of growth assays, modeling the experiment precisely after that which was described in Narwani et al. (2017). The eight species and 28 pairwise combinations will be inoculated at 100 cells ml^{-1} in 1L enriched COMBO[63] growth medium in round culture bottles, in triplicate. Cultures will be incubated on roller-racks with a 16:8 light:dark cycle with a light intensity of ca. 80 µEinstein. To restrict growth-limiting factors to light and/or space availability, 10% of media will be replaced every other day. Using the removed volume, I will estimate community-level biomass over time via measurement of chlorophyll-a fluorescence and will extract RNA for rtQPCR. Degenerate primer sets will be designed for a rationally chosen set of candidate genes, based on the results of the second aim. Trends in differential expression will then be assessed via qPCR. Excess RNA will be stored in RNAlater for potential downstream sequencing. With this strategy, I will be able to assess the extent to which our observed trends in monoculture and biculture growth conditions are repeatable phenomena.

Aim 3.2 Using a pair of novel species, test the model in vivo. The highest goal of this project is to define a gene-based model for interaction type, whereby a group of unique or differentially expressed genes among species pairs may be used to predict the outcome of a novel species interaction assay. I've hypothesized that *patterns of gene expression in monoculture can predict interaction types in bicultures (H6).* To directly test this hypothesis *in vivo*, I will include two additional green microalgal species in the growth experiments described in the previous Aim 3.1. These will be grown as monocultures and as the single biculture combination: I will then utilize the model resulting from Aim 2.3 to form a prediction for the interaction pattern and assess the extent to which that pattern is observed.

Preliminary Results – Note: I've utilized codified terms for the eight species (*Table 1*). "DC" refers to "DIMENSIONS Competition", and the numeric value refers to species. This schema is later extended to biculture conditions: e.g. for competition between *C. sorokiniana* (DC10) and *C. acicularis* (DC20), I refer to "DC12" – "DIMENSIONS Competition, species 1 vs. species 2".

Assembly and Coverage Analysis: FastQC was used to assess overall sequence quality prior to trimming, and then I used BBDuk to remove low-quality reads and adapter sequence. FastQC was used again to verify trimmed output, and then processed reads were assembled with RNAspades. From approx. 2e7-3e7 reads, I generated between 80151 and 243395 unique contigs per species, and calculated coverages (C=LN/G) ranging from 17.1X to 34.95X (*Table 2*). To assess reference transcriptome "completeness", I ran BUSCO, comparing the assemblies to a recently-released Chlorophyte database. I observed between 8.1% and 43.73% missing BUSCOs (*Table 2*). I note that the two species missing a significantly higher percentage of sequences are Charophyte, and thus the measure of "completeness" may not accurately describe these transcriptomes. A run against the more general Eukaryote database is pending.

Initial transcript identification, "green" filtration, and sequence composition: To identify and remove non-algal sequence data I implemented the "green" filtering step described in aim 1.1. . I used blastx to compare the assembled nucleotide sequence to the combined UniProt protein data and returned the top scoring hits. I compiled a list of "green" accession numbers (containing all plant and algal IDs), and then sorted blast hits against this list using a custom Python script. In this manner, I generated "Green" transcriptomes from each reference set (*Table 3*). Analysis of non-green data is currently in process.

To assess the validity of this method, I employed two strategies. The first of which was to re-run BUSCO against the "green" filtered sequence and determine whether additional Chlorophyte BUSCOs went missing in this process. Encouragingly, I only lost between 0.1% to 5.4% of additional BUSCOs while culling from 30-75% of the original contigs. I'll note that the Charophyte species lost a significantly greater percentage of BUSCOs compared to the Chlorophytes. While missing BUSCOs may simply reflect sequences not identified in the UniProt database, I feel that having a small and measurable number of false negative IDs outweighs a large and unknown number of false positives in the context of our analysis (*Table 3*). The second strategy I employed to verify the filtering step involved comparing sequence composition for "green" vs. "non-green" contigs. The rationale was that true contaminant sequence arising from bacteria or fungi in the culture would potentially exhibit unique nucleotide usage profiles. I saw mixed success in this strategy, however for some cultures this provided a clear contrast between green and non-green populations of transcripts (*Figure 4*).

ORF prediction, Annotation, and Inter-Species Orthology: I used TransDecoder to identify likely coding sequences, leveraging the output against the PfamA database of HMMs and retaining only sequences the showed a match to an existing protein sequence having a putative annotation (*Table 4*). I found that between 77.6%-92.3% of our green transcripts were retained in this step, suggesting that the filter was successful in enriching the data for functionally annotated protein sequence. To obtain additional annotation, I followed this by running HMMER for our green transcripts against the PANTHER protein family database, finding between 74.5%-96.5% of protein sequences reported – though a significant number of these were from un-named gene families (*Table 4*). I plan to supplement these annotations with KEGG and GO data.

I next sought to establish a measure of gene orthology across species. 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[59]. 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 (*Table 5*).

Monoculture growth curves: To establish estimates of growth rate and carrying capacity by species, as well as to provide the basis for relating gene expression to overall cell density, I used an R package called GrowthCurver() to fit our cell density data to standard logistic growth curves, as was done in the previous experiment[17] (*Table 6*). I note that the carrying capacities across species vary by many orders of magnitude, however given the stark differences in cell sizes and metabolic requirements, this isn't altogether surprising. In *Figure 5*, I show the fits of my estimated curves. I've since realized that the original experiment failed to account for periodic 10% harvesting, and I've identified a logistic model which is capable of doing so through the use of an additional term[53]. This improved model will be used to calculate new single-species curves, and a similar term will be incorporated into LV competition models in my analysis of the biculture data.

Differential Expression and Time-course Analyses: I used Kallisto to map paired reads back to their respective reference transcriptomes, counting each time point individually for all biological replicates. Thus, for each species in monoculture, I had three replicates at each of four time points along the

growth curve - amounting to 12 sets of counts per species. Using the Kallisto count data, I estimated occurrences of each transcript at each timepoint.

I then sought to determine whether I could detect any signal of differential expression from simply examining each species' transcriptome-wide amino acid content across the four sequenced timepoints. To establish baseline profiles for the "green" sorted reference transcriptomes, I used a Python script to count occurrences of each amino acid in a concatenated fasta file of green-filtered ORFs. Interestingly, there appeared to be some significant differences in amino-acid composition across species – especially in usage of A, E, F, G, I, K, and P through S (*Figure 6*). Next, I calculated average amino acid usage by timepoint. Plotting the reference profile against the derived timepoint-specific profiles, I observed that while there wasn't usually a notable difference in amino-acid usage across the growth curve, that actual amino-acid usage (as when multiplied by estimated count of transcripts and then averaged) often differed significantly from the values shown by the flattened reference profile. In *Figure 7* I show DC10 as an example of this trend. A reason for this may be that a number of highly expressed genes disproportionally skew the values towards those I see reflected across the growth curve profiles, while the "flattened" reference transcriptomes are averaged across all observed transcripts (including rare or seldom-expressed genes).

To determine measures of significance in differential expression, I opted to use sleuth, an R package built specifically for the analysis of Kallisto count data. I tested two types of differential expression model for our time-course data. In the first, I considered all timepoints simultaneously in estimating a single model per species. Here, the 'reduced' linear model did not account for timepoint as a fixed effect, while the 'full' model included time as a parameter. I observed an interesting trend in the results of this first test: the reported 'significant' DE transcripts were highly enriched for those which either increased or decreased consistently throughout the entire growth curve, while genes which changed direction mid-curve were largely overlooked. Thus, any gene which might have begun to increase or decrease during log phase growth was reported as insignificant (*Figure 8*). So, while this test might be appropriate in reporting on a more predictable single-direction 'dose' response, I found that it lacked the resolving power necessary to identify changes mid-curve.

I next tested all consecutive pairs of timepoints, i.e. T1/T2, T2/T3, and T3/T4. Here, every consecutive pair of timepoints was fit to a unique model. By performing tests for DE in each timepoint, I were able to much more clearly resolve significant changes across the growth curve and were able to avoid biasing significance calls to only genes changing in a single direction. This is reflected in *Figure 9*, where values in the lower right and upper left quadrants reflect either genes that increased expression into log phase and decreased toward carrying capacity or decreased during log phase and increased at carrying capacity, respectively. I determined that testing consecutive pairs of timepoints offered higher resolving power from timepoint to timepoint, and so sleuth was used to detect DE for all monoculture species in this way using the original unfiltered assembly data.

As mentioned in aim 1.3, my initial analysis indicates that there are patterns of expression within 'gene families' that make binning expression by family invalid. This realization arose as a result of trying to recapitulate some of the trends described in Narwani et al. 2017. Therein, they examined trends in 17 candidate gene families, showing that competition uniquely affects expression based on the identity of competing species. Their metric for expression was based on TPM values for binned 'gene families'. Among these families, I arbitrarily chose two – light-harvesting complex AB, and Nitrate transporter-related genes – and plotted them to see whether general patterns of movement were consistent in their

pattern of expression (*Figure 10*). It was immediately apparent that this method of binning by family might be problematic. The left plots show genes coding light-harvesting complex AB, the upper plot shows T1/T2 and T2/T3. In the lower plot showing T2/T3 and T3/T4, we see that the entire family has shifted across the y axis, indicating a drastic change in expression between the earlier and later parts of the growth curve. On the right, the Nitrate transporter-related family spans all four quadrants in both graphs, indicating that members of this family are not following a groupable pattern of expression in any way. Clearly, the preliminary data illustrate the limitations imposed by the simplifying assumptions made in Narwani et al. (2017) and indicate the value of performing a more detailed analysis.

I am now poised to address some of our proposed first-aim hypotheses in the very near-term. My first hypothesis, that *differential gene expression results from intraspecific competition (H1),* can be tested by identifying (1) differential expression correlated with increasing cell density, and (2) constitutive gene expression persisting despite increasing cell density. Preliminary analysis indicates the presence of both. *Figure 9* shows that the total number of significant differentially expressed transcripts seems to increase from T1/T2 to T2/T3, and again from T2/T3 to T3/T4. To my understanding, this is an effect of intraspecific competition. I believe that genes centered around the origin of all four plots in *Figure 9* might be constitutively expressed – that is to say, they were not shown to have any change in expression across the three consecutive pairs of plotted timepoints.

The second hypothesis, that *there exist lineage-specific patterns of gene regulation and metabolic pathway usage (H2)* really comes as two questions. First, among common genes or pathways, are any shown to exhibit unique expression patterns in any of the species? And second, what genes or pathways are unique among species? I have the background data necessary to address both questions. Above, I described how I identified both common and unique orthogroups, and also how I identified differential expression. These designations will serve as the basis for addressing this hypothesis, which be largely an exercise in manual analysis. My third hypothesis, that *there are discrete patterns of gene expression that vary by phylogenetic distance (H3)*, is an extension of the first part of H2. By identifying common genes and pathways, as well as their patterns of differential expression, it would be very possible to frame a comparison of these patterns against a metric like branch-length to see if magnitudes of change in expression of a particular gene correlate with phylogenetic distance. I've also described that the chlorophyte nitrate assimilation pathway gene family varies with species. This suggests the same possibility for other families, and I believe it's possible that a phylogenetic signal is detectable among gene family members in these pathways.

Intellectual Merit

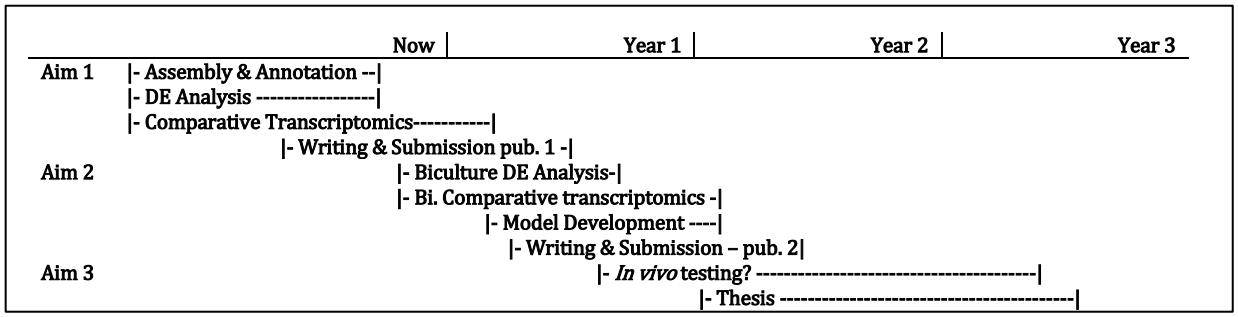
The green algae represent a poorly studied yet ecologically important set of organisms, and there exist conspicuous gaps within our understanding of species interactions at the cellular level. This project has the potential to address two pressing issues at once. My first aim will present detailed descriptions of the time-course transcriptomes of these 8 species and will provide an in-depth comparison of the resulting data. Not only will these analyses serve as the basis for my subsequent aims, but stand-alone will serve as a significant contribution to the fields of phycology and comparative transcriptomics. While this proposal does not encompass true genome analysis, the data produced in this aim would serve as valuable pilot data for future genomic efforts. The second aim will provide insight into the genetic mechanisms behind trends in community ecology, which can improve our understanding of the molecular workings of freshwater ecosystems. It might explain why these interactions, contrary to previous hypotheses, are seemingly not modeled by CRH. As near relatives of embryophytes, increasing

our knowledge of these organisms can impact our understanding of the evolution and community dynamics of land plants as well. The third aim will serve to test the second and may provide evidence toward a model for species interaction based on specific systems biology, tying genetic trends to ecological theory. Sequence data produced through this study will be deposited in GenBank, and datatypes not archived there will be made available via digital repositories such as FigShare and GitHub. I'm on schedule to present at the 2019 Ecological Society of America Meeting in New Orleans and/or the 2019 ISOP meeting in Rome, and the 12th International Phycological Congress in 2021. I hope to publish the results of the first aim during Summer 2019, and the results of the second and third aims (together or separately) within two years.

Broader Impacts

Thinking beyond the three-year window specified for this proposal, a career goal of mine is to develop and utilize algal models both for addressing questions in aquatic community dynamics, and for the advancement of applied biotech geared towards sustainable solutions in fuel and food production. Phytoplankton are underrepresented in genetic model systems: the data obtained here could be used to justify the sequencing necessary to obtain new genomes and be applied toward future annotation efforts, while developing reliable knock-in/knockout protocols will make classical reverse genetics a feasible tool for studying algal community ecology. There have been recent efforts in this, which is exciting. For example, Iben Sorenson and Jocelyn Rose have demonstrated stable transformation in Penium[64]. The direct utility is evident: if my results suggest that competition is dependent on the presence of certain genes, for example, then knockout experiments for the identified genes would serve to confirm such observations. Beyond the study of community dynamics, algae have been the subject of increasing interest in biotech. Algae-derived biofuel, among other potential applications, is a hot topic of research[65–68]. Though improvements in algal fatty-acid methyl ester (FAME) yield have been made through genetic modifications[69], there's a growing body of evidence to suggest that *community engineering* can have a significant effect on overall yield[70–72]. That is to say, the *combination* of species within the culture can greatly affect the overall yield[73,74]. Thus, my initial pursuit in understanding the genetic underpinnings of community dynamics may give rise to valuable insights for those interested in community engineering for biofuel production.

Proposed Timeline



Bibliography

1. Delwiche, C. F. & Cooper, E. D. The evolutionary origin of a terrestrial flora. *Current Biology* (2015). doi:10.1016/j.cub.2015.08.029
2. Wellman, C. H. & Strother, P. K. The terrestrial biota prior to the origin of land plants (embryophytes): a review of the evidence. *Palaeontology* **58**, 601–627 (2015).
3. Graham, L. E. The Origin of the Life Cycle of Land Plants: A simple modification in the life cycle of an extinct green alga is the likely origin of the first land plants. *Am. Sci.* **73**, 178–186
4. 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).
5. Mccormick', P. V & Cairns, J. *Algae as indicators of environmental change*. *Journal of Applied Phycology* **6**, (1994).
6. Cairns, J., Mccormick, P. V & Niederlehner, B. R. *A proposed framework for developing indicators of ecosystem health*. *Hydrobiologia* **263**, (1993).
7. Sørensen, I. *et al.* Stable transformation and reverse genetic analysis of Penium margaritaceum: A platform for studies of charophyte green algae, the immediate ancestors of land plants. *Plant J.* (2014). doi:10.1111/tpj.12375
8. de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science (80-.).* (2015). doi:10.1126/science.1261605
9. Cadotte, M. W., Dinnage, R. & Tilman, D. Phylogenetic diversity promotes ecosystem stability. *Ecology* **93**, S223–S233 (2012).
10. Ives, A. R. & Carpenter, S. R. Stability and diversity of ecosystems. *Science* **317**, 58–62 (2007).
11. Hoffmann, A. A. & Sgrò, C. M. Climate change and evolutionary adaptation. *Nature* **470**, 479–485 (2011).
12. Reusch, T. B. H., Ehlers, A., Hämmerli, A. & Worm, B. Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2826–31 (2005).
13. Cadotte, M. W., Cardinale, B. J. & Oakley, T. H. Evolutionary history and the effect of biodiversity on plant productivity. *Proc. Natl. Acad. Sci.* **105**, 17012–17017 (2008).
14. Tilman, D. Tests of Resource Competition Theory Using Four Species of Lake Michigan Algae. *Ecology* **62**, 802–815 (1981).
15. Croft, M. T., Warren, M. J. & Smith, A. G. Algae need their vitamins. *Eukaryot. Cell* **5**, 1175–83 (2006).
16. Darwin, C. *THE ORIGIN OF SPECIES BY MEANS OF NATURAL SELECTION, PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE*.
17. Narwani, A. *et al.* Ecological interactions and coexistence are predicted by gene expression similarity in freshwater green algae. *J. Ecol.* **105**, 580–591 (2017).
18. Strauss, S. Y., Webb, C. O. & Salamin, N. Exotic taxa less related to native species are more invasive. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5841–5 (2006).
19. Burns, J. H. & Strauss, S. Y. More closely related species are more ecologically similar in an experimental test. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 5302–7 (2011).

20. Maherali, H. & Klironomos, J. N. Phylogenetic and Trait-Based Assembly of Arbuscular Mycorrhizal Fungal Communities. *PLoS One* **7**, e36695 (2012).
21. Maherali, H. & Klironomos, J. N. Influence of phylogeny on fungal community assembly and ecosystem functioning. *Science* **316**, 1746–8 (2007).
22. Cavender-Bares, J., Kozak, K. H., Fine, P. V. A. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12**, 693–715 (2009).
23. Jiang, L., Tan, J. & Pu, Z. An experimental test of Darwin's naturalization hypothesis. *Am. Nat.* **175**, 415–23 (2010).
24. Violle, C., Nemergut, D. R., Pu, Z. & Jiang, L. Phylogenetic limiting similarity and competitive exclusion. *Ecol. Lett.* **14**, 782–787 (2011).
25. Macarthur, R. & Levins, R. The Limiting Similarity, Convergence, and Divergence of Coexisting Species. *Am. Nat.* **101**, 377–385 (1967).
26. Catford, J. A., Jansson, R. & Nilsson, C. Reducing redundancy in invasion ecology by integrating hypotheses into a single theoretical framework. *Divers. Distrib.* **15**, 22–40 (2009).
27. Verdú, M., Gómez-Aparicio, L. & Valiente-Banuet, A. Phylogenetic relatedness as a tool in restoration ecology: a meta-analysis. *Proceedings. Biol. Sci.* **279**, 1761–7 (2012).
28. Kettenring, K. M., Mercer, K. L., Reinhardt Adams, C. & Hines, J. EDITOR'S CHOICE: Application of genetic diversity-ecosystem function research to ecological restoration. *J. Appl. Ecol.* **51**, 339–348 (2014).
29. Cahill, J. F., Kembel, S. W., Lamb, E. G. & Keddy, P. A. Does phylogenetic relatedness influence the strength of competition among vascular plants? *Perspect. Plant Ecol. Evol. Syst.* **10**, 41–50 (2008).
30. Best, R. J., Caulk, N. C. & Stachowicz, J. J. Trait vs. phylogenetic diversity as predictors of competition and community composition in herbivorous marine amphipods. *Ecol. Lett.* **16**, 72–80 (2013).
31. Fritschie, K. J., Cardinale, B. J., Alexandrou, M. A. & Oakley, T. H. Evolutionary history and the strength of species interactions: testing the phylogenetic limiting similarity hypothesis. *Ecology* **95**, 1407–1417 (2014).
32. Narwani, A., Alexandrou, M. A., Oakley, T. H., Carroll, I. T. & Cardinale, B. J. Experimental evidence that evolutionary relatedness does not affect the ecological mechanisms of coexistence in freshwater green algae. *Ecol. Lett.* **16**, 1373–1381 (2013).
33. 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
34. Venail, P. A. & Vives, M. J. Phylogenetic distance and species richness interactively affect the productivity of bacterial communities. *Ecology* **94**, 2529–2536 (2013).
35. 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).
36. 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)
37. Bushnell, B. BBMap | SourceForge.net. *JGI* (2018). Available at: https://sourceforge.net/projects/bbmap/. (Accessed: 3rd December 2018)
38. Bushmanova, E., Antipov, D., Lapidus, A. & Przhibelskiy, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *bioRxiv* 420208 (2018). doi:10.1101/420208
39. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
40. 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).
41. Altschul, S. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
42. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
43. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
44. Nishiyama, T. *et al.* The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell* **174**, 448–464.e24 (2018).
45. Hori, K. *et al.* Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* **5**, 3978 (2014).

46. Haas, B. J. & Papanicolaou, A. TransDecoder.
47. 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).
48. 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
49. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
50. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, 258D–261 (2004).
51. 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).
52. Sprouffske, K. GrowthCurver.
53. Tschirhart, J. Integrated Ecological-Economic Models. *Annu. Rev. Resour. Econ.* **1**, 381–407 (2009).
54. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
55. 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).
56. Yi, L., Pimentel, H., Bray, N. L. & Pachter, L. Gene-level differential analysis at transcript-level resolution. *Genome Biol.* **19**, 53 (2018).
57. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
58. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
59. Koski, L. B. & Golding, G. B. The Closest BLAST Hit Is Often Not the Nearest Neighbor. *J. Mol. Evol.* **52**, 540–542 (2001).
60. Adler-Agnon, Z., Leu, S., Zarka, A., Boussiba, S. & Khozin-Goldberg, I. Novel promoters for constitutive and inducible expression of transgenes in the diatom Phaeodactylum tricornutum under varied nitrate availability. *J. Appl. Phycol.* **30**, 2763–2772 (2018).
61. Sanz-Luque, E., Chamizo-Ampudia, A., Llamas, A., Galvan, A. & Fernandez, E. Understanding nitrate assimilation and its regulation in microalgae. *Front. Plant Sci.* **6**, 899 (2015).
62. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
63. Kilham, S. S., Kreeger, D. A., Lynn, S. G., Goulden, C. E. & Herrera, L. COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia* **377**, 147–159 (1998).
64. Sørensen, I. *et al.* Stable transformation and reverse genetic analysis of *Penium margaritaceum* : a platform for studies of charophyte green algae, the immediate ancestors of land plants. *Plant J.* **77**, 339–351 (2014).
65. Smith, V. H., Sturm, B. S. M., deNoyelles, F. J. & Billings, S. A. The ecology of algal biodiesel production. *Trends Ecol. Evol.* **25**, 301–309 (2010).
66. Shurin, J. B. *et al.* Industrial-strength ecology: trade-offs and opportunities in algal biofuel production. *Ecol. Lett.* **16**, 1393–1404 (2013).
67. Schenk, P. M. *et al.* Second Generation Biofuels: High-Efficiency Microalgae for Biodiesel Production. *BioEnergy Res.* **1**, 20–43 (2008).
68. Rodolfi, L. *et al.* Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. *Biotechnol. Bioeng.* **102**, 100–112 (2009).
69. Ajjawi, I. *et al.* Lipid production in Nannochloropsis gaditana is doubled by decreasing expression of a single transcriptional regulator. *Nat. Biotechnol.* **35**, 647–652 (2017).
70. Kazamia, E., Riseley, A. S., Howe, C. J. & Smith, A. G. An Engineered Community Approach for Industrial Cultivation of Microalgae. *Ind. Biotechnol.* **10**, 184–190 (2014).
71. Kazamia, E., Aldridge, D. C. & Smith, A. G. Synthetic ecology – A way forward for sustainable algal biofuel production? *J. Biotechnol.* **162**, 163–169 (2012).
72. Shurin, J. B., Mandal, S. & Abbott, R. L. Trait diversity enhances yield in algal biofuel assemblages. *J. Appl.*

Ecol. **51**, 603–611 (2014).
73. Jackrel, S. L. *et al.* Ecological engineering helps maximize function in algal oil production. *Appl. Environ. Microbiol.* **84**, e00953-18 (2018).
74. Lian, J., Wijffels, R. H., Smidt, H. & Sipkema, D. The effect of the algal microbiome on industrial production of microalgae. *Microb. Biotechnol.* **11**, 806–818 (2018).

Tables & Figures

Table 0. NCBI Available Genomes†

Search Term	#
(ALL)	40824
"Animalia"	1128
"Land Plants"	339
"Chlorophyceae"	18
"Charophyceae"	1

† NCBI database, Dec. 2018

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

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

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

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

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

Table 6. GrowthCurver

ID	Growth Rate (r)	Carrying Capacity (K)
DC10	0.117	1.99e6
DC20	0.13	1.13e6
DC30	0.22	4.48e3
DC40	0.183	1.01e5
DC50	0.282	7.14e4
DC60	0.08	3.02e6
DC70	0.121	1.51e4
DC80	0.861	1.44e3

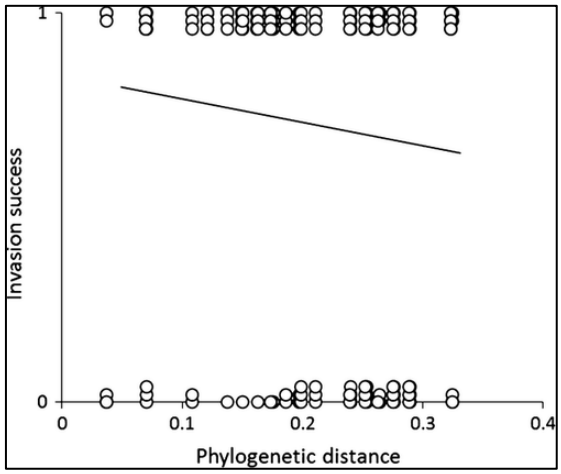
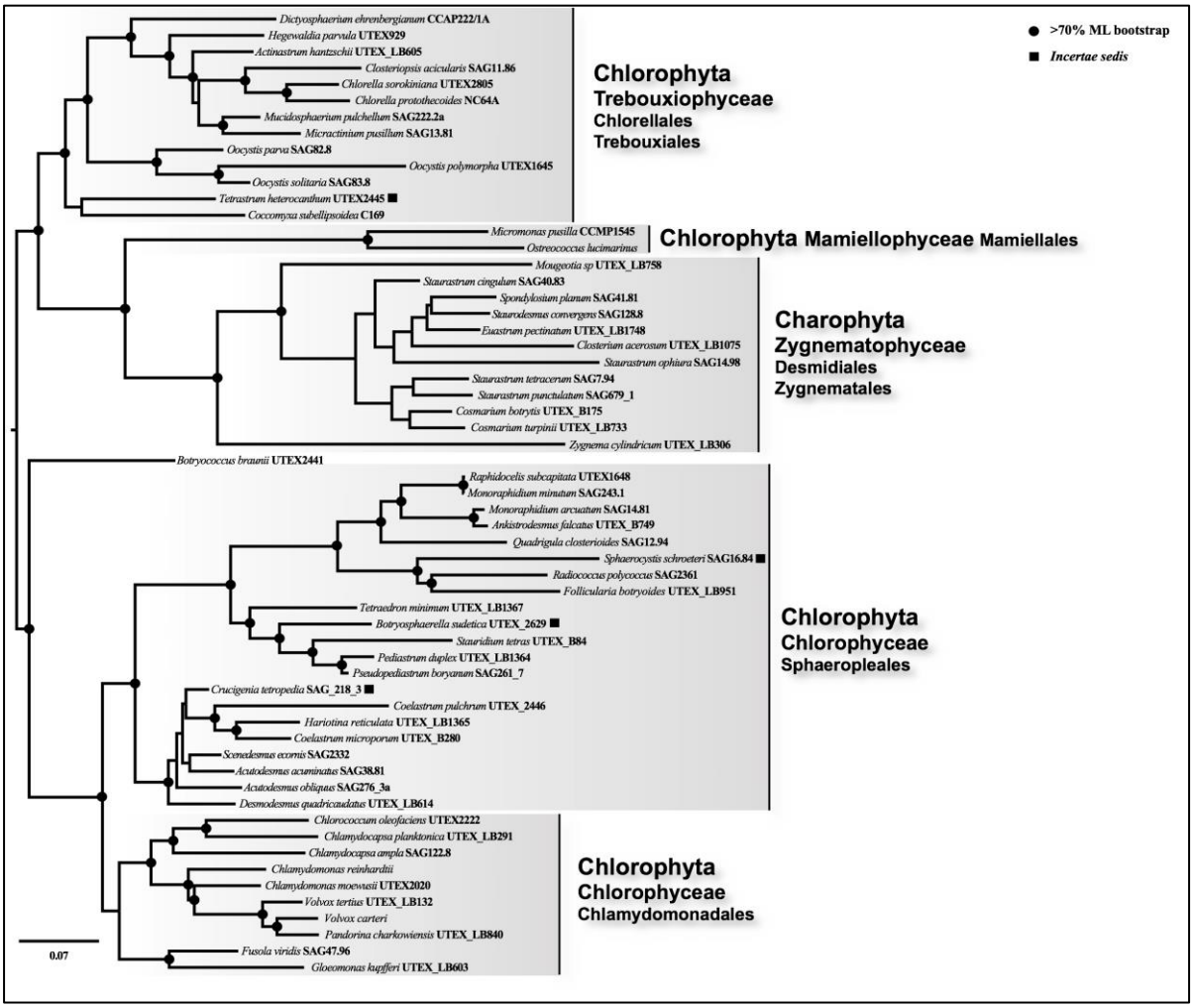


Figure 1: Effect of PD on Invasion Success[19]

From Venail et al. 2014 – Phylogeny based on 18s and rbcl data for 37 species of North American freshwater Algae. $\chi^2 = 0.02$, $P = 0.06$, $n = 168$

Figure 2: Improved Green Algal Phylogeny[20]

From Alexandrou et al. 2015



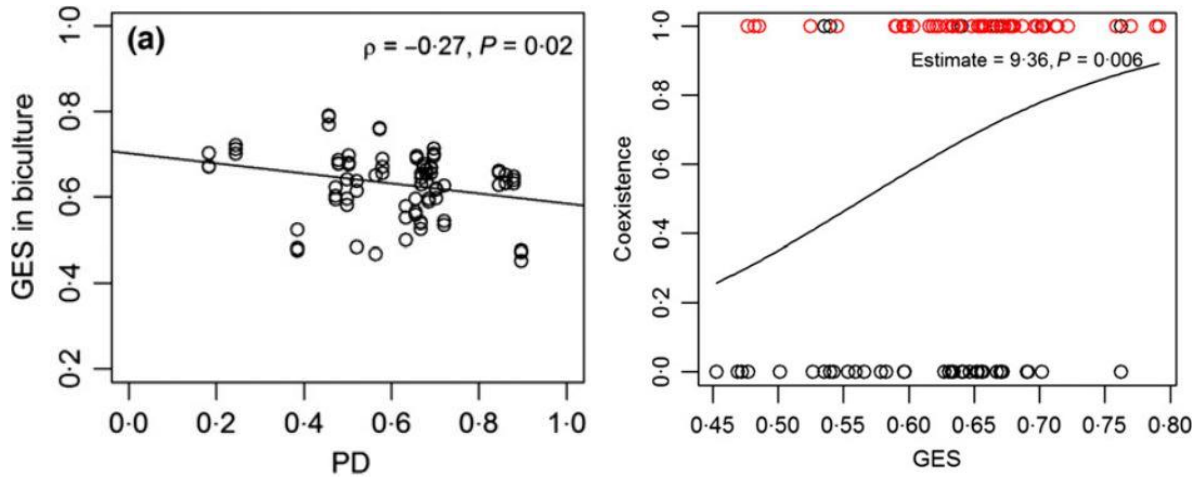


Figure 3: Gene Expression Similarity, Phylogenetic Distance, and Tendency to Coexist[21]

Fig 3a *(left) shows Fig 1 from Narwani et al. 2017, plotting phylogenetic distance of species in biculture against gene expression similarity among common genes. GES was calculated via spearman rank correlation of expression level (in TPM values from the end timepoint) for genes common to all species.*

Fig 3b *(right) shows Fig 3 from Narwani et al. 2017, plotting logistic regression of GES against "coexistence", which was estimated by extending fitted Lotka-Volterra models 100 days forward. 1 = both species present after 100 days, 0 = one or both species dead.*

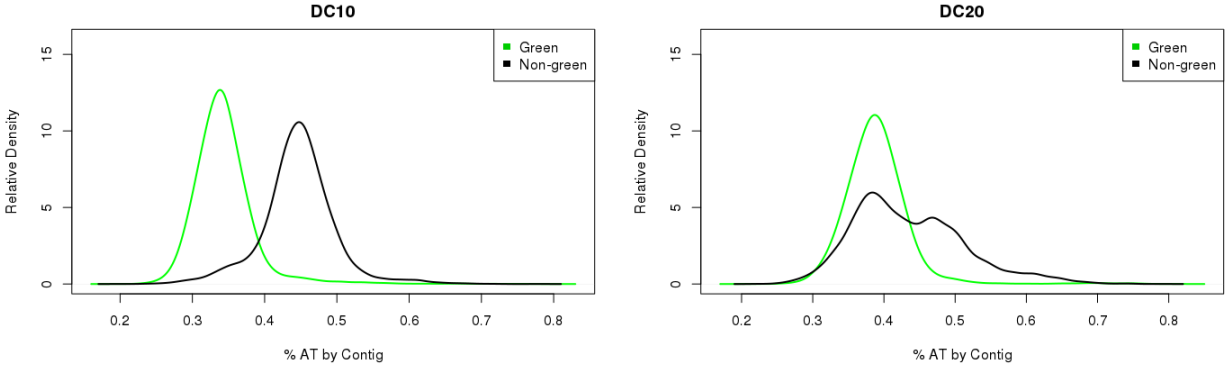


Figure 4: %AT by Contig

Plotting %AT for Green vs. Non-green contigs in DC10 shows a distinct bimodal population, suggesting that "Non-green" identified species may be derived from true culture contamination. DC20 shows a shows clear overlap of populations, and notably a bimodal population of "Non-green" with the left local maximum appearing to align with the "Green" average %AT. This could indicate that "Non-green" sequence in this case is potentially also comprised of algal sequences that were bypassed by the filtering step.

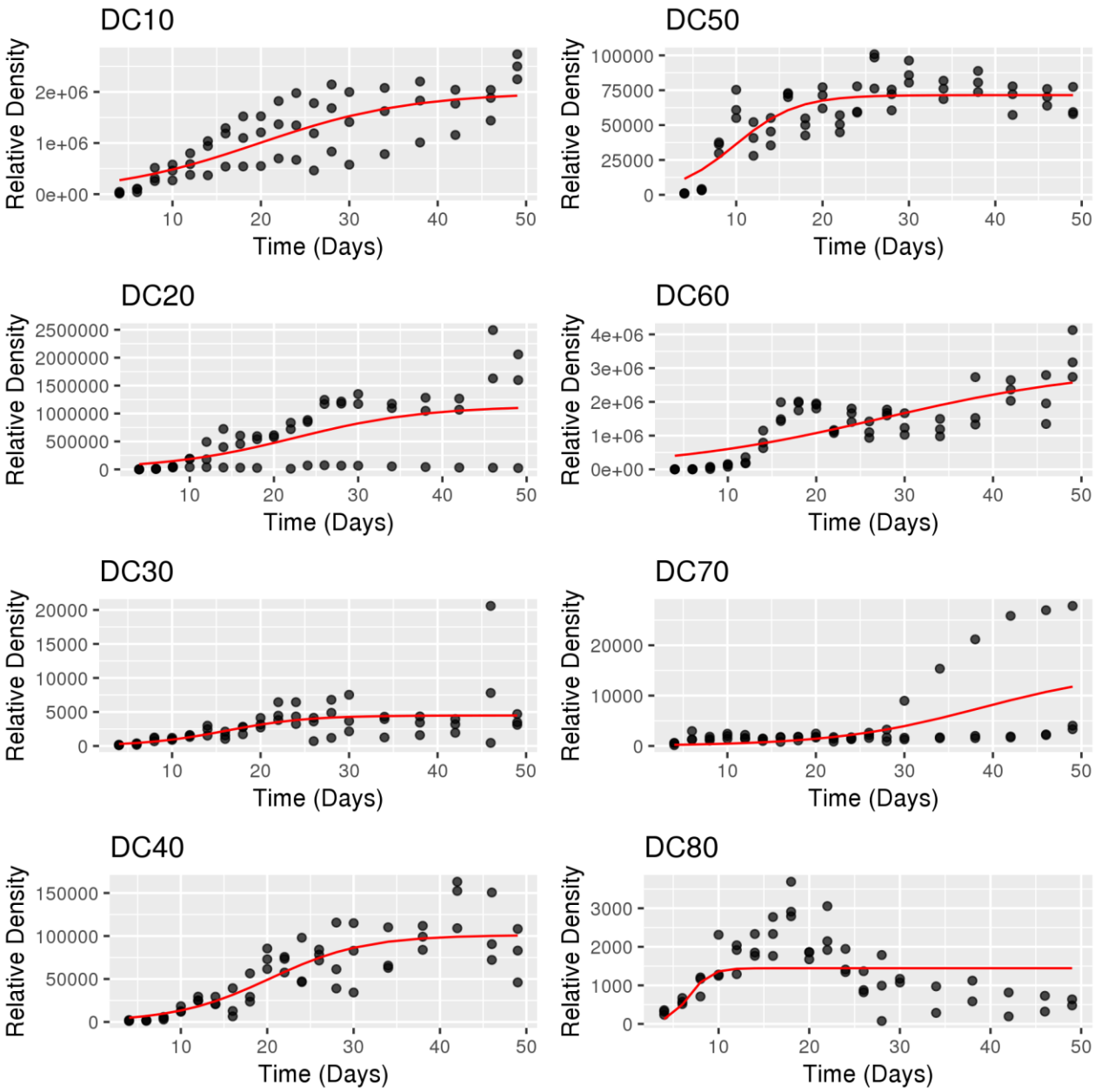


Figure 5: Growth Curves

Density measurements were made every other day, counting algal natural units (single cells or colonies) via FlowCam™ (Fluid Imaging Technologies Inc., Scarborough ME, USA). Curves were fit via standard logistic regression, outliers were not removed in this iteration. Fitting and parameter estimates (Tab. 6) were made using the GrowthCurver() package in R.

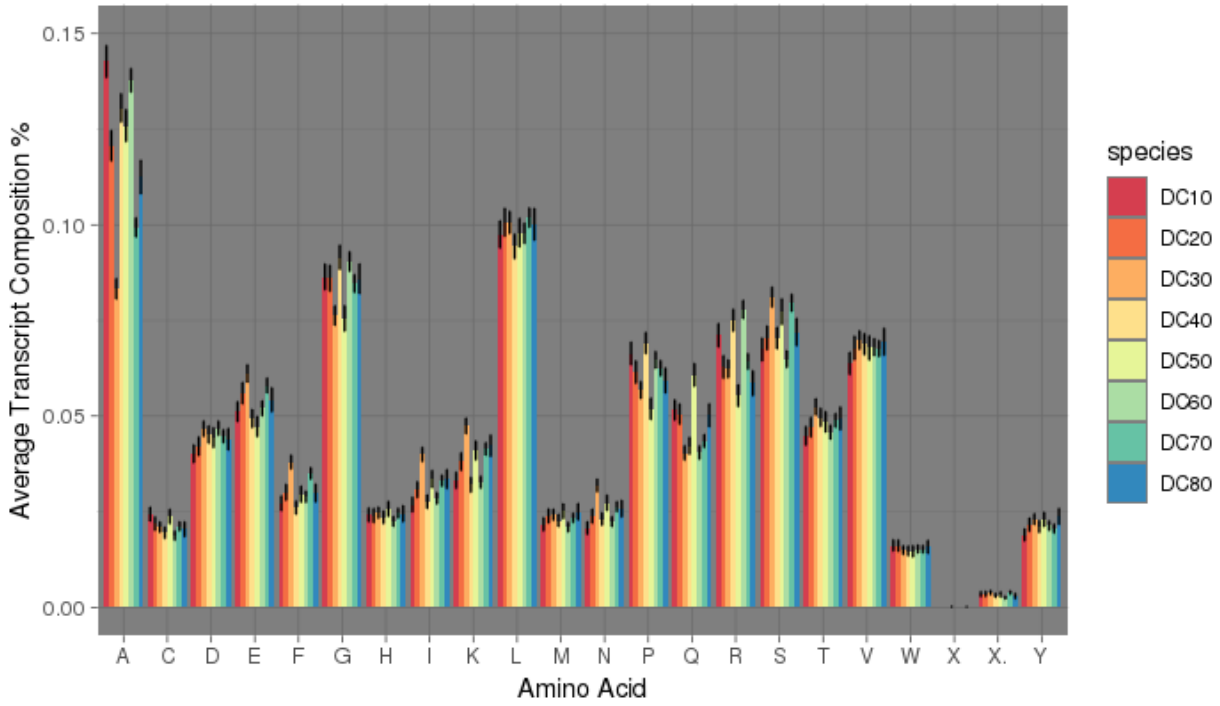


Figure 6: Mean Amino Acid Composition by Species

Amino acid counts averaged across all transcripts in each species. "X" is unidentified, "X." refers to stop codons. Confidence intervals are based on the SE of Proportion for each amino acid (SEp = sqrt[p(1-p)/n]), calculated individually by species.

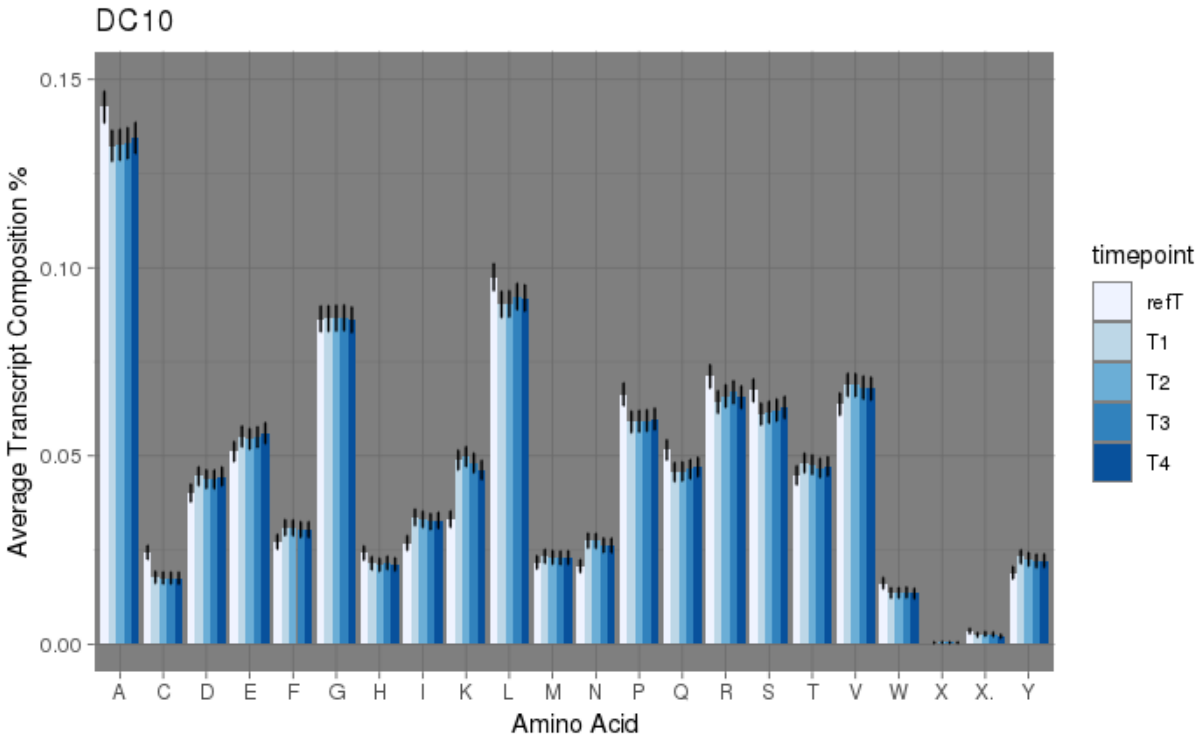


Figure 7: Mean Timecourse Amino Acid Composition of DC10

Amino acid counts normalized by estimated transcript expression for DC10. "X" is unidentified, "X." refers to stop codons. Confidence intervals are based on the SE of Proportion for each amino acid (SEp = sqrt[p(1-p)/n]), calculated individually by species.

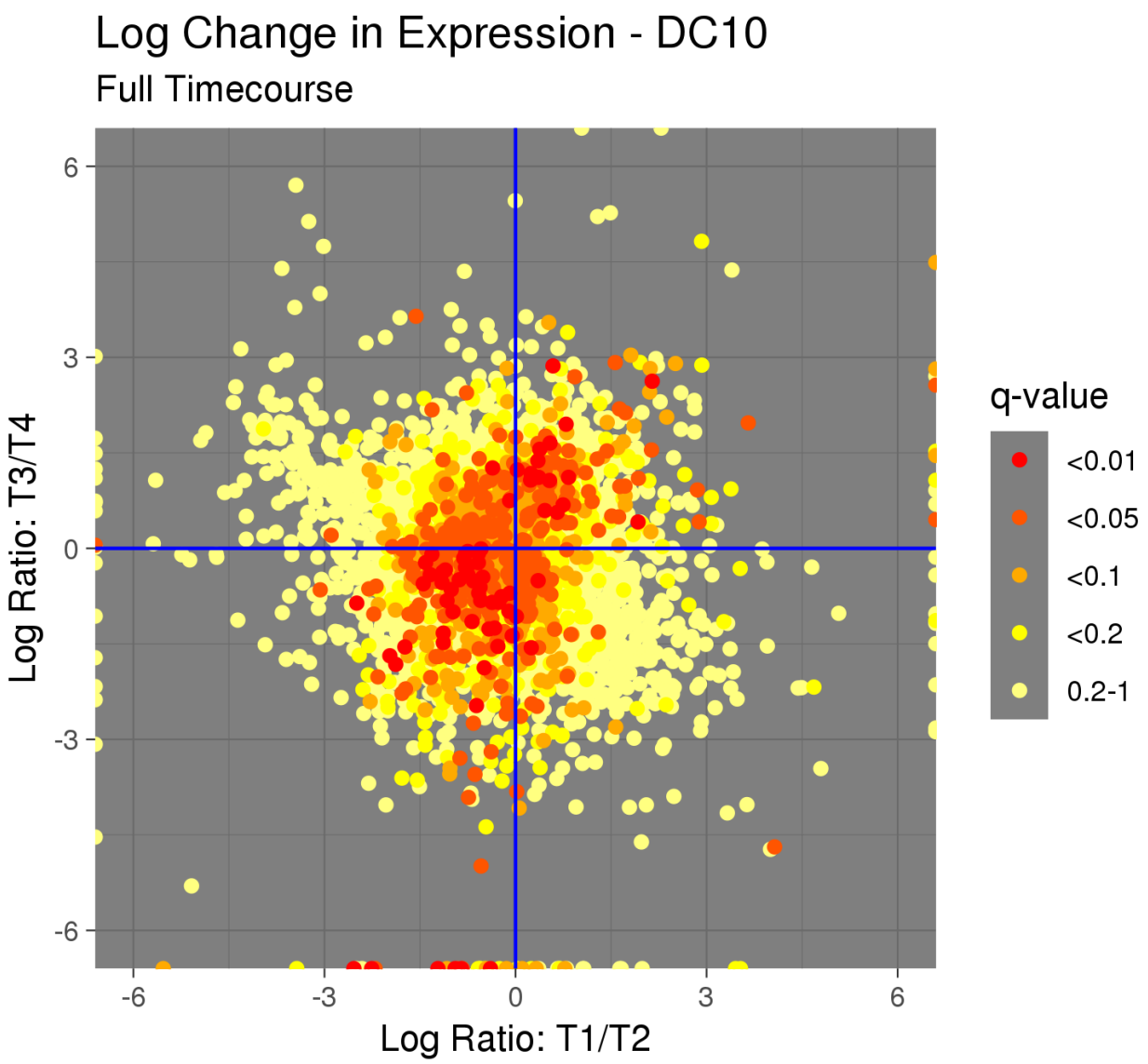


Figure 8: Logfold Change in Expression - DC10 Full Timecourse

Log-fold change was calculated by taking the log ratios of mean Kallisto-reported TPM values for transcripts at consecutive pairs of timepoints (T). q-values were determined via sleuth, testing final timepoint TPM against a model built from all available expression data. Points reflect single transcripts, where x = 2log(tpm(T1)/tpm(T2)), and y = 2log(tpm(T3)/(tpm(T4)). Only genes belonging to orthogroups common to all 8 species are plotted. +/- coordinate values reflect +/- log-fold change, respectively.

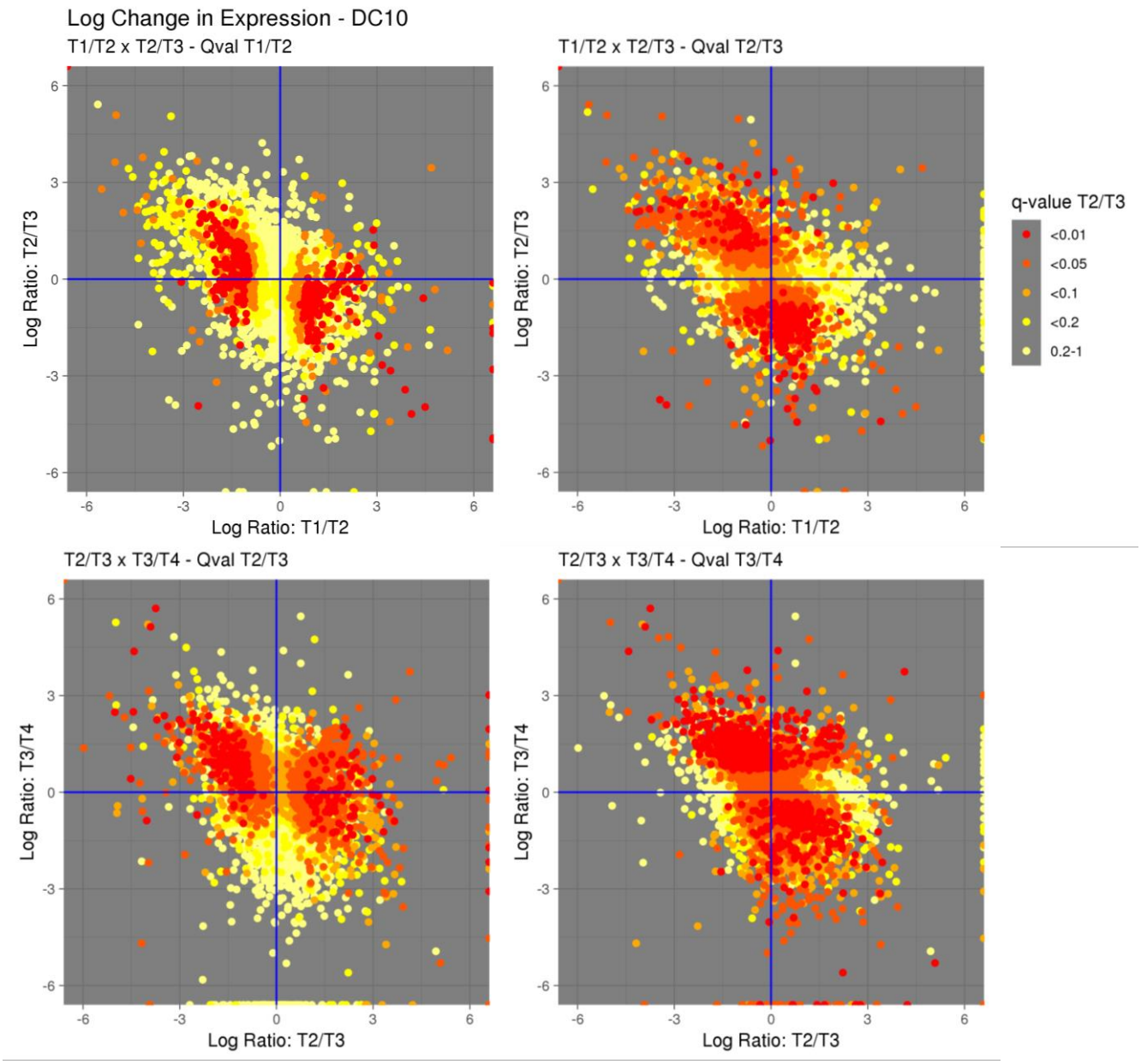


Figure 9: LFC in expression – DC10, consecutive pairwise tests

Log-fold change in expression was calculated by taking the log ratios of mean Kallisto-reported TPM values for transcripts at consecutive pairs of timepoints (T). q-values were determined via sleuth, testing consecutive pairs of timepoints individually. Points reflect single transcripts. Upper plots show 2log(tpm(T1)/tpm(T2)) by 2log(tpm(T2)/tpm(T3)), and the lower plots show 2log(tpm(T2)/tpm(T3)) by 2log(tpm(T3)/tpm(T4)). The left plots show significance values from the time-pair reflected in the x-axis, and the right plots show significance values from the time-pair reflected in the y-axis. The notable difference between basing our model on the full timecourse (Fig. 8) and using individually-fit models for each pair of timepoints is that here, we see distinctly significant expression change occurring in the bottom right and top left quadrants, which indicates directional change in expression during log phase. When using the full-timecourse model, much of this change is overlooked as 'noise'.

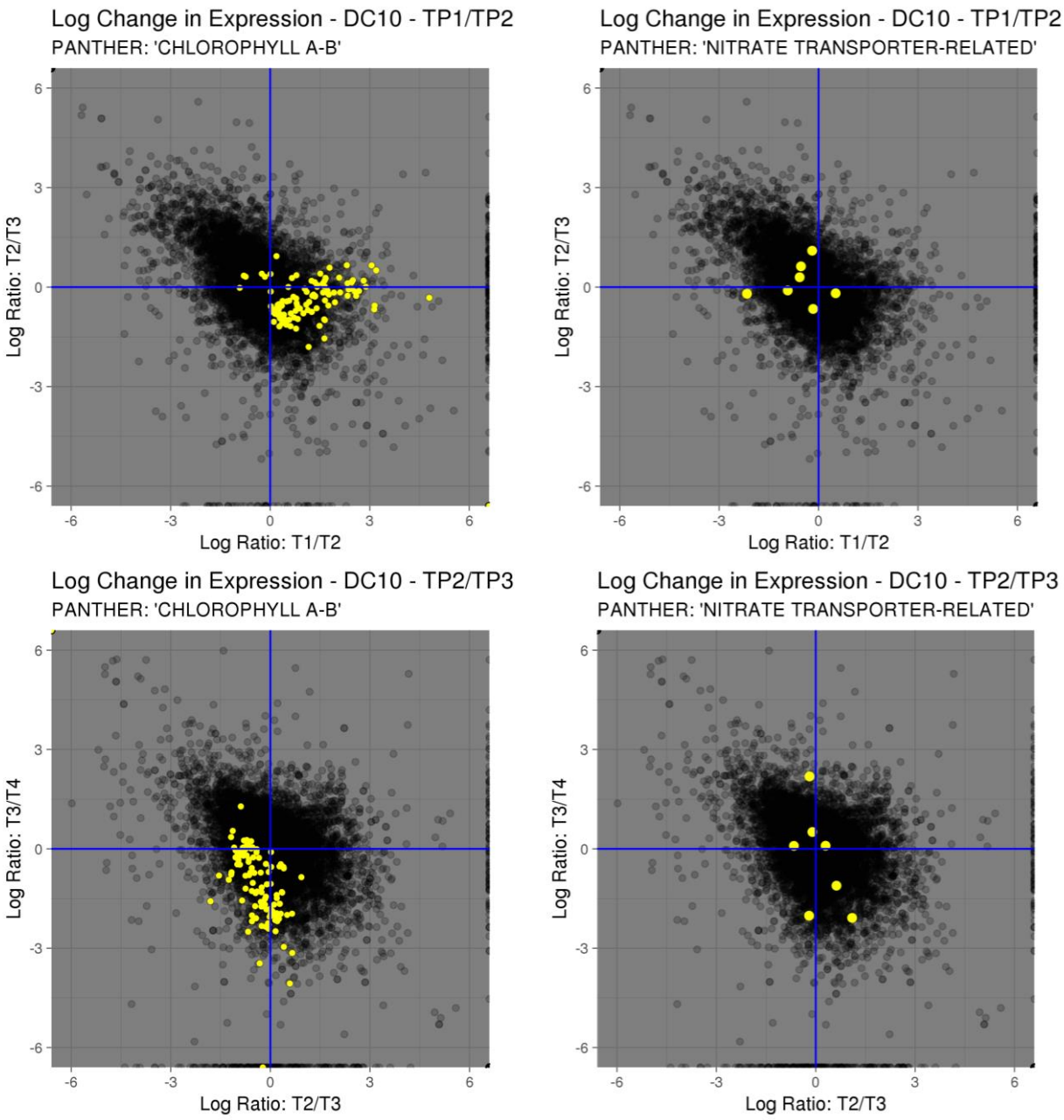


Figure 10: LFC in expression – DC10, specific families highlighted

Log-fold change was calculated and plotted as in figures 8 and 9, points reflect individual transcripts, and only transcripts from orthogroups common to the eight species are shown. Gene families were identified via HMM search of the PANTHER database, and genes belonging to said families are shown here in red. Left plots show genes from light-harvesting complex AB, right plots show nitrate transporter-related genes. Upper plots show T1/T2 and T2/T3, and lower plots show T2/T3 and T3/T4. We show that expression patterns in PANTHER-defined gene families are not consistent across the growth curve, nor at a given time point.