**Rhythm of The Night (and Day):
Predictive metabolic modeling of circadian growth in *Chlamydomonas***

Alex J. Metcalf and Nanette R. Boyle

Chemical and Biological Engineering, Colorado School of Mines, 1613 Illinois St.
Golden CO 80401

Classification: Biophysics and Computational Biology

KEY WORDS: transcriptomics, systems biology, algae, diurnal light, metabolic modeling

ABSTRACT

Algal cells experience strong circadian rhythms under diurnal light, with regular changes in both biomass composition and transcriptomic environment. However, most metabolic models – critical tools for bioengineering organisms – assume a steady state. The conflict between these assumptions and the reality of the cellular environment make such models inappropriate for algal cells, creating a significant obstacle in engineering cells that are viable under natural light. By transforming a set of discrete transcriptomic measurements from synchronized *Chlamydomonas* cells grown in a 12/12 diel light regime (1) into continuous curves, we produced a complete representation of the cell's transcriptome that can be interrogated at any arbitrary timepoint. We clustered these curves, in order to find genes that were expressed in similar patterns, and then also used it to build a metabolic model that can accumulate and catabolize different biomass components over the course of a day. This model predicts qualitative phenotypical outcomes for the *sta6* mutant, including excess lipid accumulation (2) and a failure to thrive when grown diurnally in minimal media (3), representing a qualitative prediction of phenotype from genotype even under dynamic conditions. We also extended this approach to simulate all single-knockout mutants with genes represented in the model and identified potential targets for rational engineering efforts.

SIGNIFICANCE STATEMENT

We have developed the first transient metabolic model for diurnal growth of algae based on experimental data and capable of predicting phenotype from genotype. This model enables us to evaluate the impact of genetic and environmental changes on the growth, biomass composition and intracellular fluxes of the model green alga, *Chlamydomonas reinhardtii*. The availability of this model will enable faster and more efficient design of cells for production of fuels, chemicals and pharmaceuticals.

INTRODUCTION

The dawn of a new century has also led to an awakening about our energy use; there has been a concerted push to develop more sustainable sources of energy and feedstock chemicals. One of the major challenges in developing more sustainable fuels and chemical supply chains is that the current industry is large and expensive capital has already been spent to develop the infrastructure for processing and distribution. To avoid an extensive and expensive redesign of infrastructure, renewables will have to be chemically similar to the fossil fuels we currently rely on. The fossil fuels we use today are derived from prehistoric biomass; therefore, it is a logical extension to engineer biomass to produce high concentrations and/or yields of equivalent chemical feedstocks and fuels that can be ‘dropped in’ to current infrastructure. This approach has been successfully commercialized for a handful of chemicals, such as 1,3 propanediol (4, 5), 1,4-butanediol (6-9), and succinate (10) (amongst others). One drawback of these commercialized ventures is that they utilize heterotrophic bacteria, which require a source of reduced carbon which is derived from corn or sugar cane. A more sustainable approach is the use of photosynthetic organisms that can grow on CO₂; however, the development of photosynthetically derived fuels and chemicals has fallen far short of the touted potential because the unique demands of photosynthetic organisms have not been considered in the strain development process. For example, algae and cyanobacteria have extremely strong circadian rhythms (11-13) which control the expression of cellular transcripts, proteins, and metabolic fluxes (1, 12, 14). Additionally, they are strongly entrained – cells that are removed from light/dark cycles will continue to display circadian behaviors for days afterward (15). These rhythms have a profound impact on the cell but unfortunately, they have not been studied well in the context of how they impact engineering efforts. Almost all metabolic engineering efforts in algae and cyanobacteria rely on growth in laboratory conditions with a continuous supply of light (16-29), but it is economically prohibitive to provide constant artificial light. If we are to grow these organisms outside in the most cost-effective manner, natural circadian rhythms must be considered when designing production strains.

Metabolic modeling techniques (especially constraints-based stoichiometric models) have proven extremely useful in minimizing the time it takes to develop heterotrophic production strains (30-33). One of the main assumptions in stoichiometric modeling approaches is steady state; this means that metabolite pool sizes are assumed to stay constant. While this approach works well for heterotrophic bacteria in exponential growth, it cannot be used to model the transient nature of photosynthetic growth in day/night cycles. When grown in diurnal light, 30 – 50% of genes show cyclic expression in *A. thaliana* (34, 35) and up to 80% of genes in cyanobacteria, diatoms and alga show periodic expression (36-39). These changes in gene expression impact metabolic fluxes, so to build a predictive model of photosynthetic metabolism, changing gene expression must be integrated into the model. Circadian rhythms don’t only affect the cell’s fluxes, the actual biomass composition of the cell changes over the course of the day as well; photosynthetically grown cells store starch or glycogen during the day to support metabolism at night. This also requires a different formulation of stoichiometric modeling as the steady state assumption necessitates the use of a static biomass formation equation, but diurnally grown cells alter their biomass composition significantly from day to night. The unique properties of circadian growth in photosynthetic cells means that a new modeling framework must be developed in order to be able to more accurately predict metabolism.

Here, we describe the development of a transient modeling approach capable of predicting phenotype from genotype in the model green alga *Chlamydomonas reinhardtii*. This model combines the easy to implement stoichiometric constraints-based model of *C. reinhardtii* (40) with experimentally determined transient genetic constraints on reaction bounds and a decoupled biomass formation equation that enables flexibility in the production of biomass components. By incorporating the additional complexity to account for the transient nature of diurnal growth, we are able to increase the predictive nature of our model and simulate intracellular fluxes, biomass composition and growth in diurnal light.

RESULTS AND DISCUSSION

Fitting and Clustering of Transcriptomic Profiles. Prior to using the experimental data for constraining the model, it first has to be converted from discrete data to continuous so that we can model smaller time steps than the experimental data (2-hour time steps). This way, the transcript abundance of any transcript can be modeled at any time step. By fitting transcripts to best fit functions, we can also easily compare the profile of expression across several genes (Figure 1). The transcript expression profiles are clustered with other transcripts of similar expression patterns. We have chosen to highlight 4 specific genes and the most closely associated expression profiles in Figure 1. As previously discussed, up to 80% of genes in cyanobacteria, diatoms and alga show periodic expression (36-39) in diurnal light, which makes it difficult to identify housekeeping genes to normalize gene expression to. Therefore, the first gene we investigated further was *RACK1* (*CBLP*), which is commonly used as a housekeeping gene (41), which shows very little variation in expression over time. We also found 19 named genes that are clustered closely with *RACK1*; these may be good candidates for housekeeping genes (see supplemental file 2). We noticed that the expression profile of many genes was most accurately modeled by the Kronecker Delta function, when we investigated this cluster, we saw that they cluster with known stress response genes, such as *LHCSR3.1* (42) (see supplemental file 3). This is likely more an artifact of the experimental design, where the light is turned on immediately at the onset of day instead of ramping up as it would do in natural conditions. It is well known that Ribulose-1,5-bisphosphatase carboxylase-oxygenase (RuBisCO) is a light regulated enzyme (43); our expression profile of *RBCS1*, the gene encoding the small subunit of RuBisCO, shows that the cell anticipates the onset of light by starting transcription of this gene prior to the onset of light and the peak transcript abundance occurs 6 hours after the onset of light. This profile is also clustered with multiple other named genes, including *FUOI*, associated with the electron transport chain, and *PGK1*, which encodes for phosphoglycerate kinase, a critical enzyme within the Calvin cycle. (see supplemental file 4 for more details). We also looked into genes known to be expressed at night, such as *FAP85*, a gene encoding a flagellar protein; genes associated with the synthesis of cilia are coordinately expressed at night following cell division (44). Other genes closely clustered with *FAP85* include *CPC1*, which is associated with the central pair microtubule complex (45), as well as ten other genes specifically identified as encoding flagella associated proteins (see Supplemental File 5). Fitting transcript profiles to continuous functions not only aides in the development of transient constraints for the model, but it also helps to identify similar expression profiles and potentially gene function.

Figure 1. Clustering time dependent transcriptomic data provides insight into gene expression patterns. A) Heat map of all expression patterns within the dataset. Every pixel in the main square of this image represents the degree of closeness between a pair of best-fit transcript functions, while the column on the right side indicates the relative weight of each of the best-fit models to the transcript in question; the larger amount of a given color in the column, the higher probability that that model fits the actual expression pattern better than any other of the evaluated choices. The multicolored column on the right of the heat map indicates the weight of each proposed curve fit to the gene expression pattern for a given gene; for example, if the majority of the column is pink, then that gene is best fit by the Kronecker Delta function. B) Examples of select genes of interest (shown in red) and genes which are most closely co-expressed (dark grey lines). The genes included in each cluster shown above are provided in supplemental files 2-5.

Dynamic Diel Modeling. By dynamically adjusting the model constraints according to transcriptomic abundance, we constructed a metabolic model that is capable of accurately describing the shifts in metabolism due to circadian rhythms and changing light conditions. In our simulations, we had an on/off onset of light but due to the addition of gene expression constraints we see a ramp up of carbon flux through the Calvin Benson Bassham Cycle over the course of the day; it peaks at 5 hours, after which carbon flux is ramped back down as the simulation moves toward the onset of night (Figure 2, animated in Supplemental File 6). Despite the simulated cell experiencing the same amount of light at both timepoints, it utilizes less of it early in the day because the transcripts associated with the light reactions are less present, thereby reflecting transient physical limitations within the constraints of a dynamic model. These daytime ramps are driven by changes in expression of the photosystems and associated photosynthetic light reactions, but they end up reflected in the carbon flux throughout the cell as well (see Supplemental Files 5-18). The cell shuttles large amounts of carbon between the chloroplast and the mitochondria, likely to make use of the more efficient mitochondria ATP synthase (46). The cell also experiences a night slowdown, during which time it simply attempts to produce the maintenance ATP and does not encounter any significant reaction bounds. This phenomena is supported by the finding of Strenkert et al. that the cell does not utilize its full respiratory capacity at night (1).

Figure 2. Flux maps of central carbon metabolism and light harvesting reactions over a 24 hour period with 12 hour light 12 hour dark conditions. Labels above the flux maps give time since the onset of light, at $t = 12$ hours the light turns off. The thickness of the arrows indicates the amount of flux in mmol/gDW-hr through the given reaction, and the color indicates inclusion – black lines carry significant flux, gray lines do not, and red lines are not in the model of this organism. The model includes compartmentalization of reactions as shown in the figure: orange represents the mitochondria and green represents the plastid; dotted lines indicate transport flux between these different compartments. It is clear from these select time points that the cell experiences dramatic shifts in carbon and light reactions during the day.

Modeling Mutant Phenotypes. The model is capable of predicting phenotype based solely on genomic and transcriptomic data. As a validation, we simulated the phenotype of a starchless mutant (*sta6*) (47) in diurnal light both with acetate and on minimal media and a wild type strain in the same conditions (Figure 3A). Based on our simulations, the wild type strain can grow in diurnal light in both conditions, though as expected, in the presence of a reduced carbon source (acetate), growth rate is higher. The *sta6* mutant however, is only capable of growth in the presence of acetate; this is the expected result as a cell that is unable to store starch during the day cannot sustain itself at night and agrees with published data by Davey et al. (3). This also highlights the power of our modeling approach because a normal flux balance analysis (FBA) model would not predict that this gene is essential. Due to the way FBA is formulated, it can only model steady state conditions, not the transient changes that occur during the switch back and forth. This can result in knockouts that are not fatal under steady state becoming so, such as Cre02.g145800.t1.2, a gene that encodes for mitochondrial malate dehydrogenase. As might be expected from a cell that heavily relies upon shuttling carbon to the mitochondrial ATP synthase, removing the cell's ability to do so efficiently results in long-term failure to thrive. There are also some knockouts that are predicted to overproduce in a circadian environment, despite being fatal in a steady state one – but these are universally the result of the dynamic's model changing biomass constraints. Because the model does not strictly constrain the ratios of biomass components, the model can suggest that some knockouts are feasible despite their inability to produce biomass components; a good example of this is Cre17.g728950.t1.1, which controls production of flavin adenine dinucleotide (FAD). This redox coenzyme is only required in small amounts in the biomass equation, but is large and therefore metabolically costly to produce. Mutants that cannot make it are predicted to grow faster, but this result is unlikely to be experimentally borne out. A full summary of these knockouts is available in Supplemental File 19.

Our model can predict not only growth and metabolic fluxes, but also changes in biomass composition as well. This feature is incredibly important as most efforts to engineer algae for biofuels have focused on the accumulation of lipids or fatty acids and our model is capable of predicting the effect of environmental or genetic changes on biomass. We used the model to predict the effect of nitrogen limitation on the accumulation of lipids in the wild type and *sta6* mutant in both continuous light and diurnal light (Figure 3B). In continuous light, both strains accumulate more lipids with decreasing nitrogen in the medium and we see the phenotype that has been reported in a number of papers on *sta6* (27, 47-50): that it is capable of accumulating more lipid than the wild type. When we repeated the simulations for growth in diurnal light, the wild type strain loses its ability to respond to nitrogen stress with the accumulation of lipids, it has almost the same lipid content regardless of the nitrogen limitation imposed on the model. As we described earlier, the *sta6* mutant is incapable of growth in diurnal light and this is repeated (and likely exacerbated) with nitrogen limitation. This feature of the model lets us evaluate how external or internal changes in the cell impact biomass composition.

Figure 3. Predicting phenotype from genotype. A) Growth and biomass composition simulations for WT and *sta6* in diurnal light (light indicated in white and night indicated in grey) with and without an organic carbon source. In the presence of acetate, both strains grow well, with the wild type strain outperforming the *sta6* mutant (overall biomass in black line). Without acetate (only CO₂), the model predicts much slower growth for the WT strain and no growth for the *sta6*. B) Predicted lipid content for WT and *sta6* strains in constant light (solid bars) and in diurnal light (lined bars). C) Phenotype of individual gene knockouts for every gene in the metabolic model. Each vertical line in the graph represents a single gene knockout. The parent strain (WT) doubles once in 24 hours and thus its mass ratio is 2. The majority of gene knockouts have the same final biomass as the WT, but there are many that result in severe defects in growth (on left side of the graph) and several that result in higher biomass.

Genotype to Phenotype Prediction. The model was then used to predict phenotypes of other genetic knockouts (see Figure 3C). A key parameter for economical production of algal based biofuels is productivity; for growth associated products, higher growth rates lead to higher productivities. Therefore, we first focused on the impact these gene knockouts had on growth. Our simulations are based on experimental data that was specifically designed to have one cell doubling in a 24-hour period; therefore, any knockouts that impair growth have a biomass less than 2 and those that result in higher biomass have biomass larger than 2 (Figure 4). Not surprisingly, many of the underperforming mutants are knockouts of critical enzymes involved in pigment biosynthesis, chlorophyll biosynthesis and light harvesting (see supplemental Table S1; supplemental file 20). Table S1 provides more detail for all mutants with a growth defect of 25% or more; of the 146 knockouts in the table, 23 encode enzymes which are part of the GreenCut2 (51), a functional classification of enzymes specific to the plant lineage. We investigated the functionality of these genes with Phtyozome (52), and then compared these mutants to the genome-wide knockout library constructed and tested for growth rate in minimal medium photoautotrophically by Li et al.; given the major growth defects predicted by the model, it is not surprising that most were not found (53). Of the 1355 genes in the model, only 39 were predicted to result in a 10% or higher increase in biomass (see supplemental Table S2, supplemental file 21). The simulation predicts higher growth in diurnal conditions, but closer inspection of the list can throw out several possibilities prior to constructing the mutant. For example, *nic2*, *nic7* and *nic13* are auxotrophic for nicotinamide (54) and two strains are deficient in genes encoding enzymes

present in the photorespiratory pathway (*GYDI* and *AGT2*), thus requiring high CO₂ for growth and one strain is lacking a key glyoxylate shunt enzyme (*MASI*), which is a known carbon conserving mechanism. The other mutants predicted to have better growth have not yet been characterized experimentally for growth in diurnal light.

CONCLUSION

While steady state metabolic modeling is an important tool for biological engineering, photoautotrophic species exhibit strong circadian rhythms that are evident both in the cyclic changes in transcript abundance and biomass composition. These transient changes require a modeling approach that allows for the dynamic changes occurring in the cell. Here, we leverage the ease of implementation of dynamic FBA models with gene expression and biomass composition data to develop a model that is capable of predicting phenotype in diurnal light. This model simulates growth in both day and night continuously, accounting for catabolism, changing objective functions, and transcriptomic constraints. We have shown its ability to accurately predict phenotype from genotype, making it a useful tool both for physiological interrogation of biology and metabolic engineering efforts. That predictive power, in combination with the inherently parallelizable nature of *in silico* modeling, means that it can be used to quickly assess potential bioengineering approaches for feasibility, as demonstrated by individual simulations of a broad swath of single knockout mutants.

MATERIALS AND METHODS

Informational Pipeline. In order to utilize a large dataset of circadian transcriptomic expression patterns, we built a pipeline for mass-fitting simple expression curves to every single individual transcript. These curves are simple enough to reduce overfitting, but also have enough parameters to preserve critical cyclical information. Once these curves were fit, we used them in two ways. The first was a clustering approach that enabled us to compare how every single transcript's expression pattern compares with every other transcript within the cell. This produced a heatmap where several clear correlations could be identified, including the collection of genes with a strong response to sudden light onset. The second was the integration of these curves into a genome-scale metabolic model; by querying the transcriptomic curves at every time point, the boundaries of the metabolic model could be dynamically adjusted over the course of the day (Figure 4).

Figure 4. The data pipeline used to create a dynamic metabolic model and a set of correlations. This can be utilized for any algal cell that has these data sets. Transcriptomic correlations aid in the discovery of new interactions between genes; the mechanistic metabolic model enables for the testing of knockouts and media adjustments under dynamic circadian conditions.

Fitting Discrete Transcriptomic Data with Continuous Functions. To build this data-driven model of *Chlamydomonas reinhardtii*, we used published transcriptomic data from cells grown in 12:12 day night cycles (1). This data set provided two hour resolution (or less) of changes in gene expression across the entire genome of *Chlamydomonas*. To use this data as constraints for a model with smaller time steps, we first needed to fit the discrete data points with continuous functions. After a quick visual inspection of the data, we chose 8 different types of transcript expression models that would fit the different expression profiles well: One term cosine model, two term

cosine model, hat model, fixed decay model, variable decay model, constant model, Kronkecker delta model, and cosine multiplicative model (see supplemental figure S1). For every gene in the transcriptome, we first estimated the parameters for each of the 8 proposed expression profiles which provided the best fit (see supplemental methods). Then, we used the corrected Akaike Information Criterion (AICc) (55, 56) to identify the profile that provided the best fit with the least loss of information.

The parameter estimation for each transcript expression model was performed with `scipy.optimize.curve_fit`. As there was no guarantee of global optima, the initial conditions for the fitting protocol were randomly perturbed and re-fit 100 times, with the lowest value for the AICc fit being selected for that particular model. Then, all the transcript expression models were compared against each other on the basis of AICc; the lowest overall AICc model was assumed to be the best model for this specific transcript. This protocol was repeated for every single transcript, thereby transforming thousands of sets of discrete data points into a collection of curves that represented the cell's circadian rhythms.

Clustering Genes by Expression Profile. The curves assigned to specific transcripts can also be used to evaluate the similarity between each transcript. This is typically performed in two ways: score and integral similarity. As described below, score is based off of vectorizations of the parameters for each curves – by calculating the angle and distance between two fits, a quick measurement of curve similarity can be assessed. Integral similarity between curves is found by sampling the best fit curve for each transcript at small intervals, linearly mapping the sampled values such that the minimum sampled value is zero and the maximum sampled value is one in order to normalize each, and finally comparing the differences between the normalized curves at each sample point. Those with a high integral of difference are very dissimilar curves, while those with a very small integral have highly similar expression shapes (though they may have large differences in expression magnitudes). By calculating these measurements for every possible pair of transcriptomic interactions and clustering the results, we generate a heatmap that summarizes the relationships between the circadian expressions of every gene. This map can be built for either method of comparison, either score or integral; the clustering in this paper was performed with the integral method.

Scoring. The score method is faster than the integral method at examining correlations between transcripts, and provides a good “first look” at potential clusters. However, it is also much more abstracted than the integral method. By vectorizing the parameters of the curves associated with each transcript, we can compare two vectors with relatively few calculations. Minimizing calculations is critical, as the number of comparisons scales as $O(n^2)$.

In order to vectorized a given expression curve, we begin by calculating the weight of the transcript expression models associated with the curve. A model's weight is calculated by comparing the AICc to the AICc's of all models summed together in the following fashion, where Δ_i is the difference between the AICc of model i and the lowest AICc within the set of transcript expression models associated with the curve (here represented by R):

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2}\Delta_r\right)} \quad [1]$$

Weight is roughly conceptually equivalent to the probability that a specific model is the best of all given models at fitting the data. Repeating the above equation for all models within the set R shows that the weights of all model fits for a particular set of data must sum to 1 (57). When we multiply the weight of a given transcript expression model by the parameter values that provide the best fit for that same model, we transform the fit into a vector that represents both the fit's parameters and the utility of said fit; repeating this calculation for every model and concatenating all the resulting vectors into one gives us a vector that encapsulates all of the transcript expression models simultaneously.

These vectors enable inexpensive comparison between different transcripts. Because vector operations are, computationally speaking relatively cheap, we can quickly run a number of calculations for each vector comparison. We first calculate the distance between each pair of vectors, using the Euclidian norm. The distance between a pair of vectors can indicate the similarity of fit parameters – even if two curves heavily favor the same model, the difference in parameters will result in a high distance. However, two curves that weight different models strongly may still have a small Euclidian distance between them. To account for this, we also calculate the angle between the vectors. If one transcript heavily weights a particular model while another places a significant weight on another, it is likely that the shape of the transcript expression is quite different. When the angle is calculated between them, it will reflect that information and approach orthogonality. We then combine these pieces of information in a multiplicative fashion, so that if either distance or angle indicates dissimilarity, the score will reflect that fact. While the cosine of the angle is already conveniently normalized between -1 and 1, it's desirable to also map the score onto a similar space. To do so, we generated a distance value:

$$\min\left(1 - \frac{\log(d)}{\log(d_m)}, 1\right) \quad [2]$$

where d is the distance between the vectors, and d_m is the maximum distance for all gene ID vector pairs. This is neatly locked between 0, when the distance of the pair is the maximum distance in the entire data set, and 1, when the distance is so close as to be negligible. Moreover, the log operations shift the sensitivity of the curve so that most of the 0-1 range of distance values are mapped onto relatively small distance values, thereby highlight small differences in distance even when maximum distance values are large. Multiplying the cosine of the angle by the distance value gives the score expression:

$$S = \cos \theta * \min\left(1 - \frac{\log(d)}{\log(d_m)}, 1\right) \quad [3]$$

It is immediately clear that the only way to approach the maximum score of 1 is to have vectors that are closely aligned and of similar magnitudes. Should two vectors not be closely aligned, they cannot approach 1, regardless of how distant they are from each other – and should two vectors have near-identical direction but different magnitudes, their score will also be low.

Image construction. For every clustering map, the dataframe containing the relevant data (either score or integral values for each pair) is loaded. As this file is generally around 6GB, it's infeasible to generate the entire map at once. Instead, a set of linkages is created using scipy's

cluster.hierarchy.linkage with the ward map. The linkages are then used to produce a truncated dendrogram with a set number of leaves. Each leaf is then processed into a non-square heatmap, where every pixel represents one pair-wise comparison between transcripts. The maps are made iteratively, with each being cleared out of the memory before the next is added. Once all the maps are complete, the dataframe is no longer required and can be cleared out as well, leaving space for each map to be loaded back in and stitched together into a final square map. For the weight sidebar, each model weight is included as a section of a stacked bar graph, which is saved separately and then added to the square map. Additional information can be added if required, such as the highlighting of specific genes.

Using transcriptomic data to constrain solution space with E-Flux2. Simulating a cell with a constraint-based approach requires constraints, which ideally reflect actual physiological or environmental phenomena. Transcripts can be integrated into these models in order to inform these constraints, using approaches like the E-Flux2 methodology (58); this method multiplies the level of transcriptomic expression, normalized to a fraction of the maximum, by the maximum allowable flux for all reactions, and sets the result as the maximum allowable flux for that specific transcript. In this simulation, the maximum allowable flux through any reaction was set by matching the maximum light uptake through the transcriptomic constraints to the maximum amount of light in the environment.

We then extended this method by dynamically shifting the constraints of all reactions with transcriptomic linkages at every time point. At each time point, each curve was interrogated and normalized against the maximum for that curve over the day, and the bounds lowered to that normalized fraction. When reactions involved more than one gene, the bound for each was calculated and then combined according to the structure of the gene-reaction rule itself (AND relationships took the minimum of the involved fluxes, while OR relationships summed together the fluxes). This is depicted mathematically below; Equation 4 defines the normalization function, Equations 5 and 6 show the equations for AND and OR relationships, respectively, while Equations 7, 8, and 9 show an example.

$$v_a^t = \frac{a(t)}{\max(a(t): t = [0,24))} \quad [4]$$

Equation 4's resulting value, v_a^t , represents the normalized flux fraction for a given gene a at time t , where $a(t)$ represents the value of the best-fit curve for gene a at time t .

$$v_{\wedge} = \min(v_a^t, v_b^t) \quad [5]$$

$$v_{\vee} = v_a^t + v_b^t \quad [6]$$

$$R_A = a \vee (b \wedge c) \quad [7]$$

$$v_a(t) = v_a^t + \min(v_b^t, v_c^t) \quad [8]$$

Equation 7 is a mathematical representation of the gene-reaction rule a OR (b AND c); Equation 8 is the corresponding representation of how to calculate the flux fraction for the same rule.

$$u_A(t) = u_{max} * \min(1, v_a(t)) \quad [9]$$

Finally, the flux fraction is multiplied by the maximum reaction flux of the system to calculate the bound for the reaction A at time t . This value is used for the upper bound of the reaction. If the reaction is reversible, the negative of it is used for the lower as well. Otherwise, the lower bound of the reaction is zero. The maximum reaction flux of the system was calculated such that the maximum rate of light uptake over the course of the day matched the light availability of the system. This approach is feasible, despite the discreet nature of the transcriptomic data set, because the process of fitting the curves to the point creates a continuous representation of the cell's transcriptome that can be queried at any arbitrary time.

Decoupling Biomass and Transient Objectives. A typical constraint based model formulation uses a static biomass formation equation to represent metabolites being used for the creation of biomass. Such a fixed equation is inappropriate for a circadian model, both because the composition of the cell changes over the diel cycle and because the cell's capability to make individual components of the broader equation may change over time. In order to accommodate for this difference, we split the cell's biomass equation into sub-groups and laid a control system over the top of it, such that the model would always attempt to aim for a long-term stable solution that approached the known biomass composition at different times of the day. The decoupled biomass equation consists of 10 components: Carotenoid, Chlorophyll A, Chlorophyll B, Protein, Lipid, Lipid Droplet, Carbohydrate, Nucleotide, Redox, and Starch; details of these are in supplemental file 22. This division of biomass allows the model to produce different components of biomass when it can, without compromising the cell's ability to grow if a specific biomass metabolite cannot be made.

In order to properly treat lipid production, we separated the lipid components of biomass into both a general lipid component and a lipid droplet component. The lipid droplet contained all triacylglycerides in the original biomass equation that had tails measured in Wang, et al., while the general lipid class contained all other lipids (2). This separation of lipids allows the lipid droplet fraction to rise and fall independently of the other lipid components. This behavior is pertinent for modeling lipid fraction under nitrogen stress, as the general lipid component contains lipids that contain nitrogen, and therefore will be constrained in times of low nitrogen availability.

Setting catabolism and anabolism rates. After the model is constrained to a timepoint, the production possibilities need to be evaluated. Critically, this means that the production space for biomass components must be defined. In this modeling approach, the biomass equation of the original FBA model – iCre1355 (40) in this implementation, though this approach is not limited to only this model – is split into components, so that not all biologically relevant metabolites need to be available in the correct ratios for any of them to be consumed. For every timepoint in the diel cycle, the production of each component is tested: first, for production from just the exchange reactions available, and then for production from consumption of each of the other biomass components.

Once the timepoints of the daily cycle have been evaluated, the model begins to run. At each time point, the model is constrained according to the transcriptomic curves, and the production goals and consumption preferences are set by the error terms for the components. These error terms are calculated using a pair of PI control systems – one for the production terms, and one for the consumption terms – thereby abstracting the complex cellular regulatory networks that govern

anabolic and catabolic processes into a simpler parameter space. After the production and consumption error terms for a given time point have been calculated, the model tries to produce the non-growth associated maintenance ATP (NGAM) with the media – if it cannot, it will then consume one or more components of biomass to stay alive.

Component consumption is prioritized by the consumption error terms – those with the highest positive error (i.e. the component in question is likely above its setpoint and has been for some time) are consumed first. Once the cell has made NGAM, it evaluates production potential. If the media can be successfully consumed (i.e. it can utilize light to produce NGAM), or if any components with positive degradation error remain after making ATP, the cell will attempt to produce components from either the media or the positive error components, respectively. The production possibilities are precalculated for each time point and possible energy source, either media or specific components. In the event that metabolites in a component cannot be catabolized by the cell, they are assumed to be dumped out of the cell entirely, resulting in an overall mass and energy loss. The cell uses the production error terms and the production possibilities to calculate which components are produced at what ratios. These components and ratios form the initial objective function for the cell's simulation step.

During the production phase, the cell attempts to produce as much of the initial objective as possible. Once it reaches a limitation, the model fixes the fluxes, then checks which component is limiting production and attempts to make more of the non-limiting desired ones. This iterative expansion of the production window continues until there are no more desired components, or – more commonly – the constraints of the cell limit overall production. Once components have been consumed and produced, the rates of production and consumption are multiplied by the total cell mass and used to update the amounts of each component.

Universal Single-Knockout Evaluation. In order to demonstrate the utility of this modeling approach, we conducted a survey of all gene knockouts within the model. Using high performance computing (HPC), we were able to simultaneously evaluate large numbers of genes. To simulate each gene knockout, the transcriptomic fits for the gene in question were set to zero, thereby simulating a complete knockout. The knockout mutants were simulated for 168 hours, a full week, at light and non-growth-associated maintenance ATP light levels that produced approximately one doubling in the wild type cell over the length of the simulation. For every run, the full biomass composition was recorded at the end of the simulation, and the impacts of each gene on the overall growth and biomass composition can be clearly seen.

ACKNOWLEDGEMENTS

We would like to thank Dr. Sabeeha Merchant and Dr. Daniela Strenkert for their valuable discussion during the development of the model. This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0019171

REFERENCES

1. D. Strenkert *et al.*, Multiomics resolution of molecular events during a day in the life of *Chlamydomonas*. *Proceedings of the National Academy of Sciences* **116**, 2374-2383 (2019).
2. Z. T. Wang, N. Ullrich, S. Joo, S. Waffenschmidt, U. Goodenough, Algal Lipid Bodies: Stress Induction, Purification, and Biochemical Characterization in Wild-Type and Starchless *Chlamydomonas reinhardtii*. *Eukaryotic Cell* **8**, 1856-1868 (2009).
3. M. P. Davey *et al.*, Triacylglyceride production and autophagous responses in *Chlamydomonas reinhardtii* depend on resource allocation and carbon source. *Eukaryotic Cell* **13**, 392-400 (2014).
4. C. E. Nakamura, G. M. Whited, Metabolic engineering for the microbial production of 1,3-propanediol. *Current Opinion in Biotechnology* **14**, 454-459 (2003).
5. J. Kurian, A New Polymer Platform for the Future — Sorona® from Corn Derived 1,3-Propanediol. *J Polym Environ* **13**, 159-167 (2005).
6. H. Yim *et al.*, Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat Chem Biol* **7**, 445-452 (2011).
7. M. Eichner, S. A. Kranz, B. Rost, Combined effects of different CO₂ levels and N sources on the diazotrophic cyanobacterium *Trichodesmium*. *Physiologia plantarum* **152**, 316-330 (2014).
8. S. A. Kranz, S. Dieter, K.-U. Richter, B. Rost, Carbon acquisition by *Trichodesmium*: the effect of pCO₂ and diurnal changes. *Limnology and Oceanography* **54**, 548-559 (2009).
9. X. Cai, K. Gao, Levels of daily light doses under changed day-night cycles regulate temporal segregation of photosynthesis and N₂ fixation in the Cyanobacterium *Trichodesmium erythraeum* IMS101. *PloS one* **10**, e0135401 (2015).
10. J. G. Zeikus, M. K. Jain, P. Elankovan, Biotechnology of succinic acid production and markets for derived industrial products. *Applied Microbiology and Biotechnology* **51**, 545-552 (1999).
11. M. Mittag, The Circadian Clock in *Chlamydomonas reinhardtii*. What Is It For? What Is It Similar To? *PLANT PHYSIOLOGY* **137**, 399-409 (2005).
12. F. Goulard, K. Lüning, S. Jacobsen, Circadian rhythm of photosynthesis and concurrent oscillations of transcript abundance of photosynthetic genes in the marine red alga *Grateloupia turuturu*. *European Journal of Phycology* **39**, 431-437 (2004).
13. T. Matsuo, M. Ishiura, *Chlamydomonas reinhardtii* as a new model system for studying the molecular basis of the circadian clock. *FEBS Letters* **585**, 1495-1502 (2011).
14. V. Wagner, M. Fiedler, C. Markert, M. Hippler, M. Mittag, Functional proteomics of circadian expressed proteins from *Chlamydomonas reinhardtii*. *FEBS Letters* **559**, 129-135 (2004).
15. N. Müller *et al.*, A Plant Cryptochrome Controls Key Features of the *Chlamydomonas* Circadian Clock and Its Life Cycle. *Plant Physiology* **174**, 185-201 (2017).
16. X. Y. Liu, J. Sheng, R. Curtiss, Fatty acid production in genetically modified cyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 6899-6904 (2011).
17. Y. E. Cheah *et al.*, Diel light:dark cycles significantly reduce FFA accumulation in FFA producing mutants of *Synechocystis* sp. PCC 6803 compared to continuous light. *Algal Research* **12**, 487-496 (2015).

18. P. Lindberg, S. Park, A. Melis, Engineering a platform for photosynthetic isoprene production in cyanobacteria, using *Synechocystis* as the model organism. *Metabolic Engineering* **12**, 70-79 (2010).
19. D. Lagarde, L. Beuf, M. Vermaas, Increased production of zeaxanthin and other pigments by application of genetic engineering techniques to *Synechocystis* sp strain PCC 6803. *Appl. Environ. Microbiol.* **66**, 64-72 (2000).
20. J. Zhou, H. Zhang, Y. Zhang, Y. Li, Y. Ma, Designing and creating a modularized synthetic pathway in cyanobacterium *Synechocystis* enables production of acetone from carbon dioxide. *Metab Eng* **14**, 394-400 (2012).
21. J. Ungerer *et al.*, Sustained photosynthetic conversion of CO₂ to ethylene in recombinant cyanobacterium *Synechocystis* 6803. *Energ Environ Sci* **5**, 8998-9006 (2012).
22. S. Atsumi, W. Higashide, J. C. Liao, Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nat Biotechnol* **27**, 1177-1180 (2009).
23. J. W. Oliver, I. M. Machado, H. Yoneda, S. Atsumi, Cyanobacterial conversion of carbon dioxide to 2,3-butanediol. *Proc Natl Acad Sci U S A* **110**, 1249-1254 (2013).
24. H. Li, J. C. Liao, Engineering a cyanobacterium as the catalyst for the photosynthetic conversion of CO₂ to 1,2-propanediol. *Microb Cell Fact* **12**, 4 (2013).
25. M.-D. Deng, J. R. Coleman, Ethanol synthesis by genetic engineering in cyanobacteria. *Appl. Envir. Microbiol.* **65**, 523-528 (1999).
26. T. S. Pinto *et al.*, Rubisco mutants of *Chlamydomonas reinhardtii* enhance photosynthetic hydrogen production. *Applied Microbiology and Biotechnology* **97**, 5635-5643 (2013).
27. U. Goodenough *et al.*, The Path to Triacylglyceride Obesity in the sta6 Strain of *Chlamydomonas reinhardtii*. *Eukaryotic Cell* **13**, 591-613 (2014).
28. X. Tan *et al.*, Photosynthesis driven conversion of carbon dioxide to fatty alcohols and hydrocarbons in cyanobacteria. *Metab Eng* **13**, 169-176 (2011).
29. F. K. Davies, V. H. Work, A. S. Beliaev, M. C. Posewitz, Engineering Limonene and Bisabolene Production in Wild Type and a Glycogen-Deficient Mutant of *Synechococcus* sp. PCC 7002. *Frontiers in Bioengineering and Biotechnology* **2**, 21 (2014).
30. J. I. Hendry, A. Bandyopadhyay, S. Srinivasan, H. B. Pakrasi, C. D. Maranas, Metabolic model guided strain design of cyanobacteria. *Current Opinion in Biotechnology* **64**, 17-23 (2020).
31. M. B.S., D. A., "Genome-Scale Metabolic Modeling of Escherichia coli and Its Chassis Design for Synthetic Biology Applications" in Computational Methods in Synthetic Biology, M. M.A., Ed. (Humana, New York, NY, 2021), vol. 2189.
32. M. Kim, B. G. Park, E.-J. Kim, J. Kim, B.-G. Kim, In silico identification of metabolic engineering strategies for improved lipid production in *Yarrowia lipolytica* by genome-scale metabolic modeling. *Biotechnology for Biofuels* **12**, 187 (2019).
33. A. Alsiyabi, N. B. Chowdhury, D. Long, R. Saha, Enhancing in silico strain design predictions through next generation metabolic modeling approaches. *Biotechnology Advances* **54**, 107806 (2022).
34. T. P. Michael *et al.*, Network Discovery Pipeline Elucidates Conserved Time-of-Day–Specific cis-Regulatory Modules. *PLOS Genetics* **4**, e14 (2008).
35. M. F. Covington, J. N. Maloof, M. Straume, S. A. Kay, S. L. Harmer, Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biology* **9**, R130-R130 (2008).

36. A. Monnier *et al.*, Orchestrated transcription of biological processes in the marine picoeukaryote *Ostreococcus* exposed to light/dark cycles. *BMC Genomics* **11**, 192 (2010).
37. J. M. Zones, I. K. Blaby, S. S. Merchant, J. G. Umen, High-Resolution Profiling of a Synchronized Diurnal Transcriptome from *Chlamydomonas reinhardtii* Reveals Continuous Cell and Metabolic Differentiation. *The Plant Cell* **27**, 2743-2769 (2015).
38. C. H. Johnson, P. L. Stewart, M. Egli, The Cyanobacterial Circadian System: From Biophysics to Bioevolution. *Annual review of biophysics* **40**, 143-167 (2011).
39. J. Ashworth *et al.*, Genome-wide diel growth state transitions in the diatom *Thalassiosira pseudonana*. *Proceedings of the National Academy of Sciences* **110**, 7518-7523 (2013).
40. S. Imam *et al.*, A refined genome-scale reconstruction of *Chlamydomonas* metabolism provides a platform for systems-level analyses. *The Plant Journal* **84**, 1239-1256 (2015).
41. F. Mus, A. Dubini, M. Seibert, M. C. Posewitz, A. R. Grossman, Anaerobic Acclimation in *Chlamydomonas reinhardtii*. *Journal of Biological Chemistry* **282**, 25475-25486 (2007).
42. N. Liguori, V. Novoderezhkin, L. M. Roy, R. van Grondelle, R. Croce, Excitation dynamics and structural implication of the stress-related complex LHCSR3 from the green alga *Chlamydomonas reinhardtii*. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1857**, 1514-1523 (2016).
43. I. Andersson, Catalysis and regulation in Rubisco. *Journal of Experimental Botany* **59**, 1555-1568 (2008).
44. P. A. Salomé, S. S. Merchant, Co-Expression Networks in the Green Alga *Chlamydomonas reinhardtii* Empower Gene Discovery and Functional Exploration. *bioRxiv* 10.1101/2020.10.05.326611, 2020.2010.2005.326611 (2020).
45. H. Zhang, D. R. Mitchell, Cpc1, a *Chlamydomonas* central pair protein with an adenylate kinase domain. *J Cell Sci* **117**, 4179-4188 (2004).
46. J. Petersen, K. Forster, P. Turina, P. Graber, Comparison of the H⁺/ATP ratios of the H⁺-ATP synthases from yeast and from chloroplast. *Proceedings of the National Academy of Sciences* **109**, 11150-11155 (2012).
47. A. Krishnan *et al.*, Metabolic and photosynthetic consequences of blocking starch biosynthesis in the green alga *Chlamydomonas reinhardtii* *sta6* mutant. *The Plant Journal* **81**, 947-960 (2015).
48. I. K. Blaby *et al.*, Systems-Level Analysis of Nitrogen Starvation-Induced Modifications of Carbon Metabolism in a *Chlamydomonas reinhardtii* Starchless Mutant. *The Plant Cell Online* **25**, 4305-4323 (2013).
49. Y. Li *et al.*, *Chlamydomonas* starchless mutant defective in ADP-glucose pyrophosphorylase hyper-accumulates triacylglycerol. *Metabolic Engineering* **12**, 387-391 (2010).
50. V. H. Work *et al.*, Increased Lipid Accumulation in the *Chlamydomonas reinhardtii* *sta7-10* Starchless Isoamylase Mutant and Increased Carbohydrate Synthesis in Complemented Strains. *Eukaryotic Cell* **9**, 1251-1261 (2010).
51. S. J. Karpowicz, S. E. Prochnik, A. R. Grossman, S. S. Merchant, The GreenCut2 Resource, a Phylogenomically Derived Inventory of Proteins Specific to the Plant Lineage. *Journal of Biological Chemistry* **286**, 21427-21439 (2011).
52. D. M. Goodstein *et al.*, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**, D1178-D1186 (2012).

53. X. Li *et al.*, A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis. *Nat Genet* **51**, 627-635 (2019).
54. H. Lin, A. L. Kwan, S. K. Dutcher, Synthesizing and Salvaging NAD⁺: Lessons Learned from *Chlamydomonas reinhardtii*. *PLOS Genetics* **6**, e1001105 (2010).
55. K. P. Burnham, D. R. Anderson, K. P. Huyvaert, AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology* **65**, 23-35 (2011).
56. J. E. Cavanaugh, Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters* **33**, 201-208 (1997).
57. K. P. Burnham, D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach* (Springer, New York, NY, ed. 2. ed., [4. printing], 2010), pp. 488.
58. M. K. Kim, A. Lane, J. J. Kelley, D. S. Lun, E-Flux2 and SPOT: Validated Methods for Inferring Intracellular Metabolic Flux Distributions from Transcriptomic Data. *PLOS ONE* **11**, e0157101 (2016).

FIGURE LEGENDS

Figure 1. Clustering time dependent transcriptomic data provides insight into gene expression patterns. A) Heat map of all expression patterns within the dataset. Every pixel in the main square of this image represents the degree of closeness between a pair of best-fit transcript functions, while the column on the right side indicates the relative weight of each of the best-fit models to the transcript in question; the larger amount of a given color in the column, the higher probability that that model fits the actual expression pattern better than any other of the evaluated choices. The multicolored column on the right of the heat map indicates the weight of each proposed curve fit to the gene expression pattern for a given gene; for example, if the majority of the column is pink, then that gene is best fit by the Kronecker Delta function. B) Examples of select genes of interest (shown in red) and genes which are most closely co-expressed (dark grey lines). The genes included in each cluster shown above are provided in supplemental files 2-5.

Figure 2. Flux maps of central carbon metabolism and light harvesting reactions over a 24 hour period with 12 hour light 12 hour dark conditions. Labels above the flux maps give time since the onset of light, at $t = 12$ hours the light turns off. The thickness of the arrows indicates the amount of flux in mmol/gDW·hr through the given reaction, and the color indicates inclusion – black lines carry significant flux, gray lines do not, and red lines are not in the model of this organism. The model includes compartmentalization of reactions as shown in the figure: orange represents the mitochondria and green represents the plastid; dotted lines indicate transport flux between these different compartments. It is clear from these select time points that the cell experiences dramatic shifts in carbon and light reactions during the day.

Figure 3. Predicting phenotype from genotype. A) Growth and biomass composition simulations for WT and *sta6* in diurnal light (light indicated in white and night indicated in grey) with and without an organic carbon source. In the presence of acetate, both strains grow well, with the wild type strain outperforming the *sta6* mutant (overall biomass in black line). Without acetate (only CO₂), the model predicts much slower growth for the WT strain and no growth for the *sta6*. B) Predicted lipid content for WT and *sta6* strains in constant light (solid bars) and in diurnal light (lined bars). C) Phenotype of individual gene knockouts for every gene in the metabolic model. Each vertical line in the graph represents a single gene knockout. The parent strain (WT) doubles once in 24 hours and thus its mass ratio is 2. The majority of gene knockouts have the same final biomass as the WT, but there are many that result in severe defects in growth (on left side of the graph) and several that result in higher biomass.

Figure 4. The data pipeline used to create a dynamic metabolic model and a set of correlations. This can be utilized for any algal cell that has these data sets. Transcriptomic correlations aid in the discovery of new interactions between genes; the mechanistic metabolic model enables for the testing of knockouts and media adjustments under dynamic circadian conditions.