Authors: Kourosh Salehi-Ashtiani (ksa3@nyu.edu) and Jason A. Papin (papin@virginia.edu)

Keywords: Algae, *Chlamydomonas reinhardtii*, metabolism, ORFeome, gene annotation, network reconstruction, flux balance analysis, systems biology

I. DOE award number and information

Award Number: DE-FG02-07ER64496

Title: Experimental Definition and Validation of Protein Coding Transcripts in *Chlamydomonas reinhardtii*.

Recipient: Dana-Farber Cancer Institute

Principal Investigator: Salehi-Ashtiani, Kourosh (PI); Papin, Jason A. (subaward PI)

II. Authorized distribution limitation notices

None.

III. Executive summary

Algal fuel sources promise unsurpassed yields in a carbon neutral manner that minimizes resource competition between agriculture and fuel crops. Many challenges must be addressed before algal biofuels can be accepted as a component of the fossil fuel replacement strategy. One significant challenge is that the cost of algal fuel production must become competitive with existing fuel alternatives. Algal biofuel production presents the opportunity to fine-tune microbial metabolic machinery for an optimal blend of biomass constituents and desired fuel molecules. Genome-scale model-driven algal metabolic design promises to facilitate both goals by directing the utilization of metabolites in the complex, interconnected metabolic networks to optimize production of the compounds of interest.

Using *Chlamydomonas reinhardtii* as a model, we developed a systems-level methodology bridging metabolic network reconstruction with annotation and experimental verification of enzyme encoding open reading frames. We reconstructed a genome-scale metabolic network for this alga and devised a novel light-modeling approach that enables quantitative growth prediction for a given light source, resolving wavelength and photon flux. We experimentally verified transcripts accounted for in the network and physiologically validated model function through simulation and generation of new experimental growth data, providing high confidence in network contents and predictive applications. The network offers insight into algal metabolism and potential for genetic engineering and efficient light source design, a pioneering resource for studying light-driven metabolism and quantitative systems biology.

Our approach to generate a predictive metabolic model integrated with cloned open reading frames, provides a cost-effective platform to generate metabolic engineering resources. While the generated resources are specific to algal systems, the approach that we have developed is not specific to algae and can be readily expanded to other microbial systems as well as higher plants and animals.

IV. A comparison of the Actual accomplishments with the goals and objectives of the project

Our stated Specific Aims were: 1) Experimentally verify/define transcript structure(s) of metabolic genes; 2) identify protein-protein interaction among the metabolic gene products; 3) build protein interaction maps and metabolic networks.

During the course of the project, we have expanded some areas of the proposed work, reduced certain areas, and added additional experiments and analyses as needed. These changes are as follows: 1) Development of a functional annotation pipeline to functionally annotate *Chlamydomonas reinhardtii* proteome; 2) introduction of next-generation sequencing to more effectively sequence ORFs and verify structural annotations; 3) expansion of our ORF cloning efforts; 4) expansion of constraint-based metabolic modeling; 5) omission of yeast-two hybrid experiments from the project.

Briefly, upon intimation of the project in 2007, we recognized numerous gaps in the existing KEGG functional annotation. We therefore devised an in-house functional annotation pipeline and annotated first the JGI v3.1 proteome, JGI v4.0, and Augustus 5 proteomes. We expanded our cloning efforts. We initiated our work on JGI 3.1 ORFs and carried out cloning and RACE on a set of core metabolic ORFs. However, during the course of the project, two new annotations were released: JGI 4.0 and Augustus 5. We carried out cloning efforts on both sets of metabolic ORFs from these annotations (subsequently published in *BMC Genomics*, 2011, PMID: 21810206, and *Mol. Sys. Biol.*, 2011, PMID: 21811229). To improve structural verification of these ORFs we carried out multiple runs of 454FLX sequencing in addition to conventional high throughput Sanger sequencing.

We excluded construction of protein-protein interaction network while we expanded or metabolic modeling efforts. We first reconstructed a core metabolic network (designated as iAM303) accounting for 259 reactions using the JGI 3.1 structural annotation. This work was published in *Nature Methods* (PMID: 19597503). We then reconstructed a genome-scale network (designated as iRC1080), and experimentally verified all Augustus 5 annotated transcripts in the model. The genome-scale reconstruction of the metabolic network was published in *Mol. Syst. Biol.* in 2011 (PMID: 21811229).

V. Summary of project activities for the entire period of funding

A.1 Integration of transcript verification and central metabolic network reconstruction

Given the close relationship between gene annotation and metabolic network reconstruction, we developed a novel targeted and iterative strategy, integrating experimental transcript verification with genome-scale computational modeling. Here, an initial metabolic network reconstruction, generated based on existing literature sources and bioinformatics-generated functional annotation, serves to identify *C. reinhardtii* genes in need of experimental definition and validation. We perform Reverse-Transcription PCR (RT-PCR) and Rapid Amplification of cDNA ends (RACE) to verify existence of hypothetical transcripts, and to provide refinements to structural gene annotations. Results of transcript verification experiments are applied directly toward refining the metabolic model, with a focus on eliminating reactions associated with experimentally unverified transcripts. Gaps in pathways are filled using alternative sets of enzymes, or else further attempts and alternative approaches are used to identify transcripts associated with the necessary reactions. Further, pathways may be added and expanded to yield a more complete metabolic model which serves as the basis for another round of transcript verification experiments and network modeling. Iterative refinement continues until the network and its associated genes are fully developed and validated. As the process moves forward, the resulting network model can be used to identify targets for metabolic engineering, and the generated clone resource can be used to test these hypotheses *in vivo*.

A.2. EC assignment to JGI v3.1 transcripts

To begin this iterative process, functional annotation of the v3.1 *C. reinhardtii* genome sequence was needed. Because the Enzyme Commission (EC) annotation was only available for a previous version of the genome (JGI v3.0), we generated our own EC annotations. Using the publicly available *C. reinhardtii* version 3.1 transcripts (JGI v3.1, ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.1/Chlre3_1.fasta.gz), EC numbers were assigned by BLAST sequence comparison of *in silico* translated v3.1 transcripts against the UniProt-SwissProt database and the complete *Arabidopsis thaliana* proteome data set (http://proteomics.arabidopsis.info/). The UniProt-SwissProt database contained a set of ~120,000 proteins from over 5,000 species carrying 2,321 EC terms; the *A. thaliana* proteome data set catalogued 1,800 proteins which were assigned to 498 unique EC numbers. Our merged annotations from the two data sets yielded assignment to 929 unique EC terms for the translated JGI v3.1 transcripts, 206 of which were common to both UniProt and *Arabidopsis*. Of the EC terms common to both databases, 189 (or 91.7%) were supported by both UniProt “high confidence” values (at least 40% identity and BLAST score of 50 or higher) and *A. thaliana* orthology, and only a small portion (17 transcripts or 8.25%) showed a discrepancy. Our new annotation includes many EC terms missing from existing annotation, yielding functional differences in metabolic pathways. For example, glycerate kinase (EC 2.7.1.31) is needed for function of *Chlamydomonas* metabolism, but absent from the existing online database (JGI v3.0). In addition, five EC terms used for production of triacylglycerol, a glyceride of interest for biofuel purposes, are included in our new annotation but not in existing annotations.

A.3. Reconstruction of *Chlamydomonas* central metabolism

Having assigned EC annotation for the translated JGI v3.1 transcripts, we reconstructed the central metabolism of *C. reinhardtii*, integrating literature sources with the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/pathway.html) to establish the structure of pathways included in the metabolic network. We further drew on our newly generated EC annotation of JGI v3.1 to establish the set of enzymes to be included in the network model. Finally, KEGG, ExPASy (http://ca.expasy.org/enzyme/) and other literature sources were used to delineate stoichiometry of the included reactions. The resulting metabolic network model specifies the full stoichiometry of central carbon flow in *C. reinhardtii*, accounting for all cofactors and metabolite connections. The reconstruction accounts for reaction localization primarily in the cytosol, mitochondria and chloroplast, including the lumen as a subcompartment of the chloroplast for photosynthesis, and with additional reactions localized to the glyoxysome and the flagellum. Localization evidence was obtained mainly by literature sources and supplemented by subcellular localization predictions. Transport reactions were included to allow for the presence of metabolites in multiple compartments, inferring the correct form of transport reactions using literature evidence where possible, and supplementing with information from online databases where appropriate. Because stoichiometry of all metabolites is accounted for explicitly, our mathematical representation of the reconstruction using matrix algebra allows prediction of physiological and metabolic phenotypes based on defined environmental conditions *in silico.* Of the 69 unique EC terms contained within the initial reconstruction and used to guide transcript verification experiments (please see section A.4.), 65 were present within our annotated *C. reinhardtii* v3.1 proteome. The four missing EC terms (1.1.1.28, 1.2.7.1, 1.3.99.1, and 6.2.1.5) could be assigned to homologous *C. reinhardtii* proteins but matched better to reference proteins bearing different EC numbers; consequently, they could not be assigned unambiguously.

Working with the EC assignments generated by our group, we pooled the 174 transcripts corresponding to 65 unique EC numbers accounted for in the initial version of the metabolic network. The EC assignments for these transcripts were further confirmed by assigning enzymatic and other associated domains to their respective protein products using sensitive profile-based sequence search encoding programs like HMMER, by which we specifically assigned domain families of the Pfam database, and PSI-BLAST.

A.4. Experimental verification of central metabolic open reading frames (ORFs)

Having accurately assigned functional annotation to these transcripts, we experimentally verified them in two ways. First, RT-PCR was performed with primers corresponding to putative ORFs of the central metabolic transcripts. The successful Gateway cloning and sequencing of an ORF, as either a minipool or single colony, is evidence for the presence of the hypothesized transcript, while failure is most often due to annotation errors of the ORF termini. Second, we carried out RACE on ORFs that either could not be cloned via RT-PCR or were confirmed only at one end, with the aim of correcting ORF termini annotation errors. Using RT-PCR, we were able to confirm 78% of the tested JGI v3.1 ORF models in the central model. Analysis of the RACE results indicated confirmation of 53% and refinement of 24% of the RT-PCR failed ORFs. We were able to verify 90% and refine structural annotation of 5% of central metabolic ORFs. Altogether, these results provided experimental evidence for 172 metabolic ORFs (or 99% of the examined ORFs). Our experimental verification of ORF models guides refinement of the generated metabolic model in the next cycle of our iterative methodology. The generated ORF clones serve as a resource for downstream studies. Our iterative process includes both computational and experimental components which may be performed in parallel. Accordingly, while experiments were underway, we further expanded the metabolic network reconstruction to include more complete coverage of all pathways included in the initial model. While the initial reconstruction only included four enzymes of glyoxylate metabolism necessary for acetate uptake (EC 4.1.3.1, 2.3.3.8, 2.3.3.9, 6.2.1.1), the final reconstruction includes 19 reactions in this pathway, and reflects a more complete curation of literature and genomic evidence for this pathway.

A.5. Network validation

After updating the metabolic network reconstruction based on transcript verification, we further validated the model by comparing *in silico* predictions to literature-reported values using flux balance analysis and flux variability analysis for a variety of environmental conditions and knockouts. Validating model predictions against reported literature values allows us to assess whether the metabolic network provides qualitative and quantitative predictions consistent with what is reported in the literature. Our simulations comparing *in silico* predictions of physiological parameters under various environmental conditions predicted dark aerobic acetate growth yield within 30% of the literature derived value, indicated a dark anaerobic Formate:Ethanol:Acetate fermentation ratio of 2:1:1 which matched the literature based value for dark anaerobic conditions, supported the photosynthetic release of hydrogen gas coupled with optimal biomass production , and showed photosynthetic oxygen uptake, evolution, and net production closely paralleling experimental measurements. We also verified qualitative agreement between *in silico* predictions and literature-based characterization for three mutants with impaired use of acetate (EC 1.6.5.3, 1.9.3.1, 1.10.2.2), two mutants with restricted photosynthetic abilities (EC 4.2.1.1, 4.1.1.39), and one mutant with restricted oxygen and hydrogen photoevolution. Correspondence of our *in silico* predictions with literature-based physiological parameters, for the full network and under perturbation, suggests novel predictions made to identify targets for metabolic engineering may be viewed with more confidence.

The resulting network reconstruction, named iAM303 per established convention, accounts for 259 reactions corresponding to 106 distinct EC terms. Of the experimentally tested JGI v3.1 transcripts corresponding to EC terms in the metabolic model, only phosphofructokinase (PFK, EC 2.7.1.11) and the Rieske iron-sulfur protein of ubiquinol-cytochrome-c oxidoreductase complex (EC 1.10.2.2) were not fully verified by our RT-PCR or RACE experiments. Only one of the four transcripts corresponding to PFK in our experimental test set was left unverified in our experiments. Similarly, one of the three transcripts corresponding to ubiquinol-cytochome-c reductase complex (the Rieske iron-sulfur protein) was left unverified. As our cell samples were grown under constant light, these results suggest we have identified light/dark regulated forms of transcripts corresponding to these enzymes in *C. reinhardtii*, evidence for which has been documented with PFK in the blue-green algae *Synechocystis sp*. Although any evidence drawn from cyanobacteria is tentative, the fact that the unverified transcript for PFK was the only one mapped by subcellular localization prediction to the chloroplast further indicates that light/dark regulation may also occur in the eukaryotic *C. reinhardtii*. These findings indicate our integrative approach is flexible towards functional annotation of differentially regulated transcripts and transcript variants.

B. Functional assignment of Augustus 5 and JGI v4 transcripts

Thus far, the described work was based on the JGI v3.0 genome assembly and v3.1 transcript models. However an improved assembly of the *Chlamydomonas reinhardtii* genome became available from JGI (http://genome.jgi-psf.org/cgi-bin/searchGM?db=Chlre4). We used the new *JGI* "filtered transcript models" (Chlre4_best_transcripts and Chlre4_best_proteins), and the *Augustus 5* models released through the *JGI* portal (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) for both functional assignments and structural annotation verifications. Enzymatic functional assignments were made by associating Enzyme Classification (EC) numbers through reciprocal blast searches against UniProt enzyme database (with over 100,000 protein entries). The best match for each translated ORF was identified (with an e-value threshold of 10^{-3}) and the EC number from the UniProt best match was transferred on to the ORF. We extended the EC assignments to the respective paralogs of the ORFs by clustering ORFs using BLASTCLUST (sequence identity cut-off of 35% and sequence length cut-off of 70%) within each annotation group (i.e., *Augustu*s 5 and *JGI filtered models*). Altogether, we were able to assign 970 EC numbers to 1,427 *JGI* and to 1,874 *Augustus* models. Over 93% of the EC terms were assigned to both *JGI* and *Augustus* models. We then carried out all possible pairwise alignments between the *JGI* and *Augustus* transcripts that had been assigned the same EC numbers by the above-mentioned procedure. In contrast to the high overlap between the two models in terms of EC assignments, less than half of each set were found to be 100% identical in sequence, indicating that the structural annotations of many of the two sets differ from one another.

C. Experimental verification of the *C. reinhardtii* metabolic ORFeome

To experimentally verify annotation of the ORFs, we carried out two types of experiments: First, targeted experiments in which RT-PCR was performed with primers corresponding to putative ORFs and cloning of the resulting amplicons. Second, we carried out whole transcriptome sequencing using the 454FLX platform. The latter was done using three distinct growth conditions aiming to capture annotation information on genes that may not be expressed under the condition that the cloning experiments were carried out, and to obtain expression information on various genes under different growth conditions.

For the Augustus annotated ORFs, we synthesized primers to amplify 2,776 ORFs, including 248 transporters, 1,874 EC assigned ORFs, and 654 regulatory genes (transcription factors and chromatin associated proteins). Our EC annotation of the JGI transcript models identified 1,431 transcripts with putative enzymatic functions. Of these, 645 ORFs had identical structure as the Augustus ORFs that we had assigned enzymatic function to, we therefore did not re-synthesize primers for these overlapping ORFs; however, we synthesized primers for the remaining 786 unique JGI ORFs. Altogether we synthesized 3,421 pairs of primers for amplification and cloning of various ORFs.

We grew *C. reinhardtii* under a permissive condition by providing light, organic carbon sources and nitrogen (as ammonium chloride or other ammonium salts). RNA was isolated from cells undergoing exponential growth. The isolated RNA was reverse transcribed and used as template for amplification of the ORFs for which we had designed primers for.

The amplified ORFs were cloned using recombinational cloning into pDONR223 vector. The Augustus metabolic and transporter ORFs were end sequenced by conventional high-throughput Sanger sequencing. From 2,119 Augustus ORFs tested, we were able to verify 1,408 ORFs by cloning and sequencing, while 711 cases could not be verified confidently. In other words, based on this experiment, 66% of the Augustus ORFs could be verified (please note the chromatin associated and transcription factors ORFs were not included in this set).

As an alternative sequencing method, we carried out next generation sequencing (using the 454FLX platform). We amplified the inserts of the cloned ORFs by PCR, fragmented the amplicons, and carried out 454 sequencing. The obtained 454 reads were then aligned to the ORF reference sequences to assess annotation accuracy.

Briefly, the 454 reads could cover 90-100% of the 61.4% of the JGI ORFs, 39.16% of the Augustus ORFs, and 72.39% of ORFs common between the JGI and Augustus. These results indicate that 1) the JGI models are more accurately annotated than the Augustus models, and 2) ORFs that are common between the two annotations (i.e., JGI and Augustus) are more accurately annotated than either the unique JGI or Augustus sets.

Because the experiments described in the previous section were "targeted" i.e., relied on choice of primers used to amplify the ORFs, we carried out whole transcriptome sequencing to sample the transcriptome without primer choice bias. We grew *C. reinhardtii* under three different growth conditions: 1) permissive condition with acetate and light, light and no acetate (light autotrophic growth), and dark plus acetate. Messenger RNA was isolated from these cultures, fragmented, reverse transcribed, then prepared as a library for 454FLX sequencing. We carried out two full 454 Titanium runs for each condition. The obtained reads were then aligned to the Augustus and JGI reference sequences to assess expression and annotation accuracy.

For the Augustus models, we obtained >90% coverage for 52 to 64% the ORF models depending on the growth condition, while for the JGI, the numbers ranged from 60% to 73%. Therefore, consistent with the targeted verification results, we obtained higher verification rates for the JGI ORF models as compared to the Augustus models, indicating that the JGI models are more accurately annotated. Interestingly, between the three growth conditions examined, the light - no acetate condition produced the highest coverage rates for both JGI and Augustus models, suggesting upregulation of a significant number of genes when organic carbon source is removed from the growth medium.

D. Genome-scale reconstruction of *Chlamydomonas* metabolism network model

Beginning with our reconstruction of *C. reinhardtii* central metabolism, we added pathways to the reconstruction one-by-one according to the list of target pathways chosen for the reconstruction effort. To initiate reconstruction of each individual pathway, KEGG and other classical biochemistry references were used as a starting point, with functional EC annotation used to indicate which enzymes in the pathway were genomically present. Each pathway was then manually curated using available literature evidence from *C. reinhardtii* and related species to establish presence of particular enzymes and associated reactions, reaction directionality, and cofactors involved in particular reactions. Individual reactions were localized by experimental evidence as reported in the literature, and supplemented with PASUB localization predictions as needed.

After thorough manual curation of each pathway, we followed up with gap-filling to account for dead-ends in conversion of included intermediates and cofactors. As a general rule, enzymes absent from the EC annotation were only included in the network reconstruction if (1) literature evidence was deemed sufficient to establish presence of the enzymes, (2) only one reaction was needed to fill the gap between intermediates in the pathway and available literature evidence did not contradict presence of the associated enzyme, or (3) the reaction(s) were necessary for functionality of pathways known to be present in *C. reinhardtii*.

Reaction assignment and localization for each pathway included in the network model was followed by assignment of transporters needed for functional conversion of pathway intermediates. Literature evidence and publicly available databases (e.g. TCDB and TransportDB) were used as available to assign family and stoichiometry of transporters. In the absence of other evidence, transporters were inferred from other organisms or else assumed to take the form of passive diffusion.

Having reconstructed individual pathways of the network model, we took steps to integrate these pathways. Initial and final reactants and products of each pathway were investigated to identify potential dead-ends, and additional metabolic or transport reactions were incorporated as appropriate. In addition to these manual quality control steps for pathway integration, modeling based gap-filling was also performed in the framework of flux balance analysis, with the addition of reactions needed for *in silico* growth.

With a complete version of the metabolic network reconstruction in place, we performed global quality controls, including elemental balancing and elimination of free energy loops. Referencing the full protonated elemental composition in KEGG, we compiled an E-matrix (Elemental matrix) containing elemental composition of all included metabolites. This E-matrix was then combined with the S-matrix (Stoichiometric matrix, representing all reactions in the model), and a check of $E \cdot S = 0$ ensured elemental balance for all included reactions. Finally, our metabolic network reconstruction was evaluated with extreme pathway analysis, and all type III pathways, or internal loops corresponding to free energy consumption in the network, were removed.

The generated genome-scale reconstruction of *C. reinhardtii*, accounts for all pathways and metabolic functions indicated by the latest release of the genome (JGI v4.0) combined with our in-house generated functional annotation. The reconstruction accounts for 1,080 genes, associated with 2,190 reactions, and includes 1068 unique metabolites, and encompasses 83 subsystems distributed across10 compartments. As the most comprehensive metabolic network reconstruction of *C. reinhardtii* to date, ours is the first to account for three different wavelengths of light involved in photosynthesis and includes considerable expansion of fatty acid metabolism over previous reconstructions, with detail at the level of individual R-groups. Further, the metabolic network reconstruction presented here provides a greater level of compartmentalization than existing reconstructions of *C. reinhardtii*, with the inclusion of the lumen as a distinct component of the chloroplast for photosynthetic functionality, and the eyespot used to guide the flagella in phototaxis.

We have carried out simulations under a variety of growth conditions (e.g. acetate/no acetate, light/no light, aerobic/anaerobic), and physiological validation of *in silico* gene knockout against known mutant data for a variety of phenotypes (e.g. increased use of acetate; light; CO_2; nitrogen; and other media components, amino acid requiring, altered color). In addition, we have detailed simulations demonstrating how photon absorption and different wavelengths of light affect downstream metabolic processes, elucidating the benefits of sunlight versus artificial light conditions. Our well-validated and comprehensive genome-scale reconstruction of *C. reinhardtii* metabolism provides a valuable quantitative and predictive resource for metabolic engineering toward improved production of biofuels and other commercial targets.

VI. Products developed under the award

A. Publications

1. *Nat. Methods.* 2009 Aug;6(8):589-92. Metabolic network analysis integrated with transcript verification for sequenced genomes. Manichaikul A, Ghamsari L, Hom EF, Lin C, Murray RR, Chang RL, Balaji S, Hao T, Shen Y, Chavali AK, Thiele I, Yang X, Fan C, Mello E, Hill DE, Vidal M, Salehi-Ashtiani K, Papin JA. (PMID: 19597503)

2. *Biotechnol. J.* 2010 Jul;5(7):660-70. Metabolic systems analysis to advance algal biotechnology. Schmidt BJ, Lin-Schmidt X, Chamberlin A, Salehi-Ashtiani K, Papin, JA. (PMID: 20665641)

3. *BMC Genomics.* 2011 Jun 15;12 Suppl 1:S4. Genome-wide functional annotation and structural verification of metabolic ORFeome of Chlamydomonas reinhardtii. Ghamsari L, Balaji S, Shen Y, Yang X, Balcha D, Fan C, Hao T, Yu H, Papin JA, Salehi-Ashtiani K.

(PMID: 21810206)

4. *Mol. Syst. Biol.* 2011 Aug 2;7:518. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. Chang RL, Ghamsari L, Manichaikul A, Hom EF, Balaji S, Fu W, Shen Y, Hao T, Palsson BØ, Salehi-Ashtiani K, Papin JA. (PMID: 21811229)

B. Websites and other internet sites that provide public access to results of this project.

1. http://www.bme.virginia.edu/csbl/downloads-chlamy.php

2. http://www.biomedcentral.com/1471-2164/12/S1/S4/additional

3. http://www.nature.com/msb/journal/v7/n1/suppinfo/msb201152_S1.html

C. Network or collaborations fostered.

The following collaborations were established in support of the project:

1. Harvard University, Andrew Murray lab

2. Cornell University, Haiyuan Yu Lab

3. University of Iceland and University of California, San Diego, Bernhard Palsson lab

4. University of Iceland, Ines Thiele lab

D. Technologies/Techniques.

We developed an integrated experimental and modeling approach to facilitate our proposed work. This approach is described in section V of this document and is published (please see *Nat. Methods* 2009 in publication list above)

E. Inventions/Patents.

None.

F. Other products (physical collections, models, etc).

1. Clones: As part of our efforts under this award, we have generated a substantial number of cDNA clones. These clones are described in publications listed above (*Mol. Syst. Biol.*, 2011, PMID: 21811229; *BMC Genomics* 2011, PMID: 21810206). We intend to submit these clones to Chlamydomonas Center (http://www.chlamy.org/) for public distribution.

2. Models: We have generated two metabolic network models describing steady-state metabolic fluxes in *Chlamydomonas*. The models, iAM303 and iRC1080, describe the central metabolism and global metabolism of *C. reinhardtii* respectively. Both models are described in our peer-reviewed publications, validated, and are publically available (please see above under section VI.A. and VI.B.). These models are provided as

standard SBML format as well as ".mat" format (please see Section VI.B). Metabolic analyses on these models can be performed using the COBRA toolbox (publically available from http://opencobra.sourceforge.net/openCOBRA/Welcome.html). The COBRA toolbox runs within Matlab environment.

For background information (including methodology, assumptions, limitations, and usage) on constraint based metabolic modeling used in this project, please see:

1. *Methods Enzymol*. 2011; 500:411-433. Whole-genome metabolic network reconstruction and constraint-based modeling. Haggart CR, Bartell JA, Saucrman JJ, Papin JA. (PMID: 21943909)

2. *Nat. Protoc.* 2011 Aug 4;6(9):1290-307. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ. (PMID: 21886097)

3. *Mol. Sys. Biol.* 2009; 500:61-80. Applications of genome-scale metabolic reconstructions. Oberhardt MA, Palsson BP, Papin JA. (PMID: 19888215)

Appendix

Copies of Publications:

1. *Nat. Methods.* 2009 Aug;6(8):589-92. Metabolic network analysis integrated with transcript verification for sequenced genomes. Manichaikul A, Ghamsari L, Hom EF, Lin C, Murray RR, Chang RL, Balaji S, Hao T, Shen Y, Chavali AK, Thiele I, Yang X, Fan C, Mello E, Hill DE, Vidal M, Salehi-Ashtiani K, Papin JA. (PMID: 19597503)

2. *Biotechnol. J.* 2010 Jul;5(7):660-70. Metabolic systems analysis to advance algal biotechnology. Schmidt BJ, Lin-Schmidt X, Chamberlin A, Salehi-Ashtiani K, Papin, JA. (PMID: 20665641)

3. *BMC Genomics.* 2011 Jun 15;12 Suppl 1:S4. Genome-wide functional annotation and structural verification of metabolic ORFeome of Chlamydomonas reinhardtii. Ghamsari L, Balaji S, Shen Y, Yang X, Balcha D, Fan C, Hao T, Yu H, Papin JA, Salehi-Ashtiani K. (PMID: 21810206)

4. *Mol. Syst. Biol.* 2011 Aug 2;7:518. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. Chang RL, Ghamsari L, Manichaikul A, Hom EF, Balaji S, Fu W, Shen Y, Hao T, Palsson BØ, Salehi-Ashtiani K, Papin JA. (PMID: 21811229)

Metabolic network analysis integrated with transcript verification for sequenced genomes

Ani Manichaikul[1,6], Lila Ghamsari[2,6], Erik F Y Hom[3,6], Chenwei Lin[2,6], Ryan R Murray[2,6], Roger L Chang[4,6], S Balaji[2], Tong Hao[2], Yun Shen[2], Arvind K Chavali[1], Ines Thiele[4,5], Xinping Yang[2], Changyu Fan[2], Elizabeth Mello[2], David E Hill[2], Marc Vidal[2], Kourosh Salehi-Ashtiani[2] & Jason A Papin[1]

With sequencing of thousands of organisms completed or in progress, there is a growing need to integrate gene prediction with metabolic network analysis. Using *Chlamydomonas reinhardtii* as a model, we describe a systems-level methodology bridging metabolic network reconstruction with experimental verification of enzyme encoding open reading frames. Our quantitative and predictive metabolic model and its associated cloned open reading frames provide useful resources for metabolic engineering.

Present availability of genome sequences for diverse microorganisms brings opportunities for metabolic engineering through systems-level characterization of these organisms' metabolic networks[1]. Such efforts require both functional and structural annotation of metabolic components encoded within these genomes. Although advances have been made in defining transcribed protein coding sequences for widely studied eukaryotes, notable deficiencies in genome annotation remain[2]. These problems are evident in the genomes of less widely studied species for which comparative genomic information is scarce. Structural annotations of boundaries for many genes in newly sequenced genomes are often poorly defined because of incomplete understanding of transcriptional-initiation, termination and splicing rules, and deficiencies in gene-prediction algorithms[3]. Genes with valid structural annotations lack thorough functional annotations linking transcripts to enzymatic or regulatory activities of corresponding proteins[4].

Given the close relationship between gene annotation and metabolic network reconstruction[1,5], we propose a targeted iterative methodology, integrating experimental transcript verification with genome-scale computational modeling (**Fig. 1**). An initial metabolic network, generated using literature sources and bioinformatics-generated functional annotation, served to identify *C. reinhardtii* genes in need of experimental definition and validation. We performed reverse-transcription PCR (RT-PCR) and rapid amplification of cDNA ends (RACE) to verify existence of hypothetical transcripts and to refine structural annotations. We used the results of transcript verification experiments to refine the metabolic model, with a focus on eliminating reactions associated with experimentally unverified transcripts. We filled resulting gaps in pathways by incorporating alternative sets of enzymes and by applying more detailed functional annotation to identify transcript models associated with necessary reactions. We also added and expanded pathways to yield a more complete metabolic model, providing the basis for another round of transcript verification and network modeling. Iterative refinement continued until the network and its associated genes were fully developed and validated.

To begin our iterative process, functional annotation was needed for current *C. reinhardtii* genome sequence. Because Enzyme Commission (EC) annotation was only available for a previous version of the genome (Joint Genome Institute (JGI) v3.0), we generated our own annotations (**Supplementary Note** and **Supplementary Figs. 1**,**2**). Using the publicly available *C. reinhardtii* version 3.1 transcripts (JGI v3.1, ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.1/Chlre3_1.fasta.gz), we assigned EC numbers by basic local alignment search tool (BLAST) sequence comparison of *in silico*–translated v3.1 transcripts against UniProt-SwissProt[6] and the complete *Arabidopsis thaliana* proteome dataset. Our new annotation (**Supplementary Table 1**) included EC terms missing from existing annotation, yielding functional differences in metabolic pathways (**Fig. 2a**,**b**). For example, six EC terms used for production of triacylglycerol, a glyceride of interest for biofuel purposes, were included in our new annotation but not in existing annotations (**Supplementary Table 2**).

Having assigned EC annotation for the translated JGI v3.1 transcripts, we generated a central metabolic network reconstruction of *C. reinhardtii*, integrating literature-sourced data with our newly generated EC annotation of JGI v3.1. We used the Kyoto Encyclopedia of Genes and Genomes (KEGG), Expert Protein Analysis System (ExPASy) and literature sources to delineate pathway structure and reaction stoichiometry. The resulting metabolic network model specified the full stoichiometry of central metabolism in *C. reinhardtii*, accounting for all cofactors and metabolite connections[1], with reactions localized to the cytosol, mitochondria,

[1]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, USA. [2]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [4]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. [5]Present address: Center for Systems Biology, University of Iceland, Reykjavik, Iceland. [6]These authors contributed equally to this work. Correspondence should be addressed to K.S.-A. (kourosh_salehi-ashtiani@dfci.harvard.edu) or J.A.P. (papin@virginia.edu).

RECEIVED 29 APRIL; ACCEPTED 17 JUNE; PUBLISHED ONLINE 13 JULY 2009; CORRECTED ONLINE 20 JULY 2009 (DETAILS ONLINE); DOI:10.1038/NMETH.1348

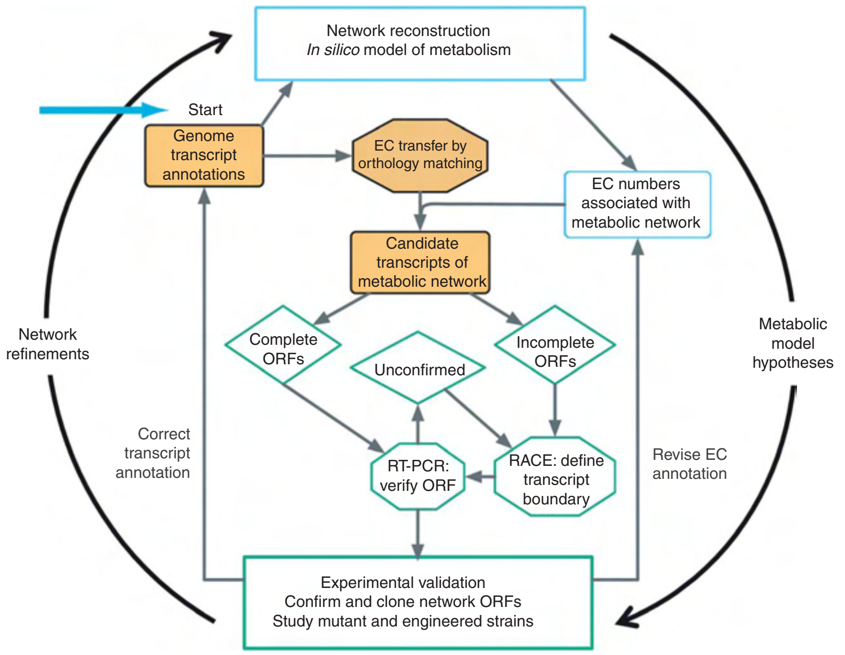

Network reconstruction
In silico model of metabolism
Start
Genome transcript annotations
EC transfer by orthology matching
EC numbers associated with metabolic network
Candidate transcripts of metabolic network
Complete ORFs
Incomplete ORFs
Unconfirmed
RT-PCR: verify ORF
RACE: define transcript boundary
Experimental validation
Confirm and clone network ORFs
Study mutant and engineered strains
Network refinements
Correct transcript annotation
Revise EC annotation
Metabolic model hypotheses

Figure 1 | Assessing and improving gene annotation for *C. reinhardtii*: iterative process integrating gene annotation experiments with metabolic network reconstruction and analysis. Starting with a draft network reconstruction, EC terms associated with model reactions are mapped to corresponding transcripts. Experimentally verified transcripts are used to propose changes in structural annotation, along with functional annotation changes that motivate refinements in the network reconstruction. The reconstructed metabolic network is then used to motivate another round of transcript verification experiments.

chloroplast (including the lumen as a subcompartment for photosynthesis) glyoxysome and flagellum. We obtained the localization evidence mainly from literature and supplemented it by subcellular localization predictions[7]. We established transport reactions using literature-sourced evidence where possible, supplementing it with information from online databases where appropriate. Of the 69 unique EC terms contained within the initial reconstruction and used to guide transcript verification experiments (**Supplementary Table 3**), all but four were annotated in the *C. reinhardtii* v3.1 proteome. The missing EC terms (1.1.1.28, 1.2.7.1, 1.3.99.1 and 6.2.1.5) could be assigned to homologous *C. reinhardtii* proteins but matched better to reference proteins bearing different EC numbers, and so could not be assigned unambiguously.

We confirmed EC assignments for 174 transcripts by assigning enzymatic domains to the protein products using hidden Markov model-based software HMMER[8] (**Supplementary Table 4**) and experimentally verified these transcripts in two ways. First, we performed RT-PCR with primers corresponding to putative open reading frames (ORFs) encoding central metabolic enzymes (**Supplementary Table 5**). The successful cloning and a matched sequence[9] of an ORF to its predicted model indicated the presence of the hypothesized transcript, whereas failure in this task was most often due to annotation errors of ORF termini[2]. Second, we carried out RACE on ORFs that either could not be cloned via RT-PCR or were confirmed only at one end, with the aim of correcting ORF termini annotation errors. Using RT-PCR, we confirmed 78% of the tested JGI v3.1 ORF models, and RACE allowed confirmation of 53% and refinement of 24% of the ORFs that we could not verify by RT-PCR. Altogether, we verified 90%, refined structural annotation of 5% and provided experimental evidence for 99% of the 174 examined ORFs encoding central metabolic enzymes (**Fig. 2c** and **Supplementary Table 4**). Our experimental verification of ORF models guided refinement of the metabolic model in the next cycle of our iterative methodology, and generated ORF clones can be used for downstream studies.

We expanded the metabolic network reconstruction to include more complete coverage of all pathways included in the initial model. For example, the glyoxylate metabolism pathway in our initial network reconstruction included only four enzymes needed for acetate uptake, but our final reconstruction included 16 enzymes, reflecting more complete curation of this pathway. After additionally updating the metabolic network reconstruction with transcript verification results, we validated the model by comparing *in silico* predictions to quantitative literature-based physiological parameters under a variety of environmental conditions and qualitative literature-based characterization of known mutants (**Supplementary Note**, **Supplementary Tables 6,7** and **Supplementary Fig. 3**). Agreement between *in silico* predictions and existing experimental data brought confidence to predictions of metabolic engineering targets (**Supplementary Fig. 4**).

The resulting network reconstruction, named iAM303 per established convention[10], accounted for 259 reactions corresponding to

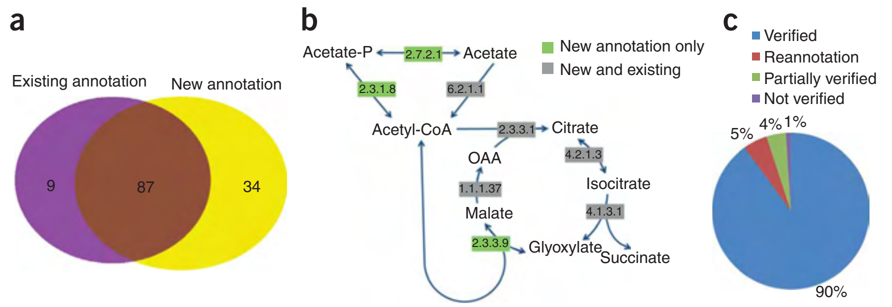


Figure 2 | Integrating the network model with transcript verification experiments. (**a**) Comparison of central metabolic EC terms annotated in existing JGI v3.0 and our annotation of JGI v3.1 (**Supplementary Note**). (**b**) Applying these two versions of EC annotation to inform the network reconstruction yielded functional differences in core metabolic pathways, as illustrated in acetate uptake pathways inferred from the two sets of annotation. As acetate is the sole carbon source used by wild-type *C. reinhardtii in vivo*, these pathway differences translate directly to measureable growth phenotypes. (**c**) Results summary for verification and structural annotation of *C. reinhardtii* central metabolic transcripts by RT-PCR and RACE. ‘Partially verified’ denotes cases for which the assembled ORF did not completely match the genome sequence or a complete sequence could not be assembled.

Table 1 | EC terms guiding reconciliation of literature, modeling and experimental evidence

				Modeling evidence[a]					
	Enzyme name (EC number)	Pathway(s) affected	Literature evidence	Dark aerobic	Dark anaerobic	Light	Light with acetate	PSI-BLAST hit(s)	Action
Absent in our annotation of JGI v3.1 translated transcripts	L-lactate dehydrogenase (1.1.1.27)	Pyruvate metabolism	Yes	WT	WT	WT	WT	estExt_fgenesh2_pg.C_190058	Perform transcript verification for functional matches identified by PSI-BLAST
	D-lactate dehydrogenase (1.1.1.28)	Pyruvate metabolism	Yes	WT	WT	WT	WT	Chlre2_kg.scaffold_1000146	
	L-lactate dehydrogenase, cytochrome (1.1.2.3)	Pyruvate metabolism	None	WT	WT	WT	WT	estExt_gwp_1H.C_90212	
	Pyruvate synthase (1.2.7.1)	Pyruvate metabolism	Yes	WT	N	WT	WT	e_gwWT.62.37.1	
	Succinate dehydrogenase (1.3.99.1)	Photosynthesis; TCA cycle	Yes	WT	WT	WT	WT	fgenesh2_pg.C_scaffold_1000904 estExt_fgenesh2_pg.C_30248	
	Limit dextrinase (3.2.1.142)	Starch metabolism	Yes	R	N	R	R	fgenesh2_pg.C_scaffold_33000007	
	Oxalate decarboxylase (4.1.1.2)	Glyoxylate metabolism	None	WT	WT	WT	WT	estExt_fgenesh2_pg.C_160183	
	Succinyl-CoA ligase (6.2.1.5)	TCA cycle	Yes	R	WT	WT	WT	estExt_GenewiseH_1.C_190100 estExt_fgenesh2_kg.C_130058	
One or more experimentally unverified transcript models	Phosphofructo-kinase (2.7.1.11)	Glycolysis	Yes	WT	N	WT	WT	Analysis not performed because transcripts were already identified for these enzymes	Perform transcript verification for cells grown in the dark
	Ubiquinol cytochrome *c* oxidoreductase (1.10.2.2)	Oxidative phosphorylation	Yes	R	WT	R	R		

[a]WT, wild-type flux; R, reduced flux; and N, no flux.

We probed these ten EC terms through *in silico* knockout experiments under the four indicated environmental conditions. We interpreted reduced or zero flux through the objective function to indicate the given enzyme was necessary or important under the stated environmental condition. Finally, we used PSI-BLAST to search more thoroughly for EC terms with no corresponding transcripts in our annotation JGI v3.1. Because PSI-BLAST identified alternative transcripts for each of these EC terms, none of the corresponding reactions were deleted from the network reconstruction.

106 distinct EC terms (**Supplementary Fig. 5**, **Supplementary Tables 8,9** and **Supplementary Data 1**). Of the experimentally tested JGI v3.1 transcripts corresponding to 65 unique EC terms from the initial metabolic model, only phosphofructokinase and the Rieske iron-sulfur protein of ubiquinol-cytochrome *c* oxidoreductase complex were not verified in our RT-PCR or RACE experiments: we left unverified one of the four transcripts corresponding to phosphofructokinase and one of the three transcripts corresponding to ubiquinol-cytochome *c* oxidoreductase complex (the Rieske iron-sulfur protein) (**Supplementary Table 4**). As we grew our cultures under constant light, these results suggest that we identified light/dark–regulated forms of transcripts corresponding to these enzymes, evidence for which has been documented for phosphofructokinase in the cyanobacteria *Synechocystis sp.*[11]. Although any parallel drawn from cyanobacteria is tentative, that the unverified phosphofructokinase transcript was the only one mapped by subcellular localization prediction[7] to the chloroplast further indicates light/dark regulation may occur in the eukaryotic *C. reinhardtii*. These findings indicate our integrative approach is flexible toward functional annotation of differentially regulated transcripts and transcript variants.

With ORF verification results for all annotated enzymes in the current version of our metabolic network reconstruction, we demonstrated a complete cycle of our iterative approach. Although not all enzymes in the model could be completely validated experimentally, we seek to recover these enzymes in the next round of experiments. For enzymes present in the network reconstruction but lacking functionally assigned transcripts in the *C. reinhardtii* genome, we performed more detailed searches using position-specific iterative BLAST (PSI-BLAST) to assign likely targets to corresponding EC numbers (**Table 1**); newly assigned transcript models can be followed up in the next iteration of experiments. EC terms annotated in JGI v3.1 which were not fully verified by our RACE and RT-PCR transcript verification experiments, but are supported by both literature and modeling evidence, suggest corresponding transcripts are present in *C. reinhardtii*, particularly under dark conditions. In the next round of experiments, we will attempt to verify these transcripts in the absence of light. Our structural reannotation of transcripts will also inform reannotation of functional enzymatic domains needed to refine and expand our metabolic network model.

Although throughput of our method is modest compared to fully automated computational approaches, we achieved higher quality structural and functional annotation for a targeted set of metabolic enzymes. Accordingly, our integrative approach produced: (i) a well-validated metabolic network reconstruction of *C. reinhardtii*, (ii) functional annotation needed to map the network reconstruction to associated transcripts and (iii) experimentally based structural annotation, providing the requisite toolset for metabolic engineering toward improved biofuel production (**Supplementary Fig. 4**). Whereas the latter does not provide direct proof of function, it establishes the necessary condition upon which functional assignments can be proposed, and targeted experiments may be performed to verify function.

With only 1% of experimentally tested transcripts left unverified, our effort provides proof of concept for the proposed approach integrating network analysis with experimental transcript verification. Because this success may be attributed in part to our focus on central metabolism, enzymes and pathways of which are generally the best characterized, our manual curation efforts will be even more important in informing high-quality transcript annotation refinement as we extend our metabolic model to the genome-wide scale. Although our work has focused on *C. reinhardtii*, integration of gene annotation experiments with network reconstruction can be applied broadly toward improved annotation of existing and emerging genome sequences. Our pipeline for functional annotation based on existing annotation of *A. thaliana* provides a computationally efficient approach to extract functional annotation for species with one or more well-annotated close relatives. For new genome sequences without availability of closely related reference sequence, more sophisticated approaches, including PSI-BLAST and hidden Markov model–based programs, may provide viable alternatives. Although existing transcriptomic technologies lag behind RT-PCR and RACE in their ability to provide well-defined ORF structure and precise definition of exon-boundaries for eukaryotic sequence data, emerging sequencing technologies[12] open possibilities to scale up the throughput of our methodology. Finally, we may look beyond metabolic network modeling toward reconstruction of regulatory[13] and signaling[14] networks as alternative systems-level frameworks to guide future efforts.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This research was supported by the Office of Science (Biological and Environmental Research), US Department of Energy, grant DE-FG02-07ER64496 (to J.A.P. and K.S.-A.), the Jane Coffin Childs Memorial Fund for Medical Research (to E.F.Y.H.) and by National Science Foundation IGERT training grant DGE0504645 (to R.L.C.).

AUTHOR CONTRIBUTIONS

A.M., A.K.C., R.L.C. and I.T. reconstructed metabolic networks; L.G., R.R.M., X.Y. and E.M. performed transcript verification experiments, E.F.Y.H. performed localization prediction; L.G., C.L., Y.S., C.F. and T.H., annotated transcripts and analyzed sequences; S.B. annotated transcripts; D.E.H. and M.V. initially developed the transcript verification pipeline; A.M., L.G., E.F.Y.H., K.S.A., J.P., development of pipeline to integrate model with experiments; A.M., L.G., E.F.Y.H., C.L., R.L.C., R.R.M., K.S.-A. and J.A.P. wrote and edited the manuscript; D.E.H. and M.V. edited the manuscript; K.S.-A. guided transcript verification experiments and transcript annotation; J.A.P. guided the metabolic network reconstruction; J.A.P. and K.S.-A. conceived the study.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L. & Palsson, B. *Nat. Rev. Microbiol.* **7**, 129–143 (2009).
2. Reboul, J. *et al. Nat. Genet.* **27**, 332–336 (2001).
3. Jones, S.J.M. *Annu. Rev. Genomics Hum. Genet.* **7**, 315–338 (2006).
4. Frishman, D. *Chem. Rev.* **107**, 3448–3466 (2007).
5. Boyle, N.R. & Morgan, J.A. *BMC Syst. Biol.* **3**, 4 (2009).
6. Apweiler, R. *et al. Nucleic Acids Res.* **32**, D115–D119 (2004).
7. Lu, Z. *et al. Bioinformatics* **20**, 547–556 (2004).
8. Zhang, Z. & Wood, W.I. *Bioinformatics* **19**, 307–308 (2003).
9. Walhout, A.J. *et al. Methods Enzymol.* **328**, 575–592 (2000).
10. Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.O. *Genome Biol.* **4**, R54 (2003).
11. Kucho, K. *et al. J. Bacteriol.* **187**, 2190–2199 (2005).
12. Shendure, J. & Ji, H. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
13. Herrgård, M.J., Covert, M.W. & Palsson, B. *Curr. Opin. Biotechnol.* **15**, 70–77 (2004).
14. Papin, J.A., Hunter, T., Palsson, B.O. & Subramaniam, S. *Nat. Rev. Mol. Cell Biol.* **6**, 99–111 (2005).

ONLINE METHODS

Metabolic network reconstruction. The metabolic network reconstruction begins with identification of key pathways to be included in the central metabolic model. The basic structure of these pathways was extracted from KEGG (http://www.genome.jp/kegg/pathway.html). Reactions were localized to specific organelles and compartments primarily using literature evidence. When no literature evidence could be identified to localize a particular reaction, we drew on subcellular localization predictions combined with localization of neighboring reactions of the same pathway to make a reasonable localization assignment. In the absence of any literature-based localization information for an entire pathway, a consensus of localization predictions for the entire pathway was taken to ensure that neighboring reactions were connected.

Pathways were initially focused to reflect specific knowledge about *C. reinhardtii* by excluding reactions for which no genes encoding the corresponding enzyme (based on our group's EC annotation) were present in v3.1 of the genome. In a second pass, pathways were supplemented with the addition of reactions deemed necessary by gap analysis, and reactions having literature evidence specifically relevant to *C. reinhardtii* were also included. Stoichiometry of metabolic reactions was extracted from KEGG or ExPASy (http://ca.expasy.org/enzyme/), and also supplemented with key literature references on metabolism of *C. reinhardtii* and related species when necessary. Because the assignment of transporters is an area of metabolic network modeling with less direct evidence available, we limited use of transport reactions to those necessary to account for metabolites appearing in more than one compartment. We then drew from literature evidence, where available, to assign the stoichiometry of transport reactions. For example, triose-phosphate transport is performed by antiport with phosphate between the cytosol and the chloroplast[15–17]. We also used predictions from online databases (TransportDB, http://www.membranetransport.org/; Transport Classification Database, http://www.tcdb.org/) as a secondary source of evidence to infer sodium-ion symport for 2-oxoglutarate and malate to the chloroplast, as well as to the mitochondria. Transporters for the remaining set of metabolites for which there was no clear evidence were assigned based on precedent from other organisms.

To develop a constraint-based model from the reconstructed network, initially no assumptions were made limiting any reaction flux in the network. Additional literature curation was performed to assemble a set of Boolean constraints for reaction activity in light or in the dark. For instance, it is known that certain plastidic enzymes are subject to either light activation or inhibition mediated via the thioredoxin system[18]. Since a major source of energy in *C. reinhardtii* is obtained through starch degradation, especially in the dark, we also determined maximal starch degradation rates from experimental values both in light and dark and under aerobic and anaerobic conditions[19]. The modeling constraints used for all simulations are reported in **Supplementary Table 9**.

We evaluated our metabolic network reconstruction with extreme pathway analysis, and all type III pathways, or internal loops corresponding to free energy consumption in the network, were removed[20]. The stoichiometry of the full set of reactions in the reconstruction was incorporated into an S-matrix, which was imported to Matlab to perform growth simulations by flux balance analysis using the COBRA toolbox[21]. Flux balance analysis[22] was used to simulate growth or survival of the organism by optimization of the precursor biomass reaction or an ATP demand reaction, as appropriate. Proposed engineering strategies for hydrogen production were achieved through flux variability analysis[23] of the full set of reaction deletion mutants grown *in silico* under light conditions and constrained to achieve a growth rate at least 95% of the optimum (**Supplementary Fig. 4**, with full results shown in **Supplementary Table 10**).

Subcellular localization prediction. The compartmentalization of network reactions was guided by subcellular localization predictions generated using PASUB, the Proteome Analyst Specialized Subcellular Localization Server[7]. cDNA sequences for the experimentally tested transcripts were translated using custom Perl scripts and subjected to PASUB analysis. Given the dual plant- and animal-like nature of the *C. reinhardtii* proteome[24], predictions were generated using both "animal" and "plant" default settings, providing localization information for all experimentally tested transcripts (**Supplementary Table 4**). Using animal settings, predictions were made with 9 possible subcellular compartments: cytoplasm, endoplasmic reticulum, extracellular, Golgi, lysosome, mitochondria, nucleus, peroxisome and plasma membrane. Using plant settings, predictions were made with 10 possible subcellular compartments: chloroplast, cytoplasm, endoplasmic reticulum, extracellular, Golgi, mitochondria, nucleus, peroxisome, plasma membrane and vacuole. Predictions involving the peroxisome or vacuole were treated as predictions to the glyoxysome. Both animal and plant predictions, along with associated enzyme reaction characteristics, were used to manually assign subcellular localization(s) for each transcript product.

***Chlamydomonas reinhardtii* strain and growth conditions.** *C. reinhardtii* strain CC-503 was used throughout our experiments. *C. reinhardtii* cells were grown in Tris-acetate-phosphate (TAP) medium containing 100 mg l^{-1} carbamicillin without agitation, at room temperature (22–25 °C) and under continuous illumination with cool white light at a photosynthetic photon flux of 60 μmol m^{-2} s^{-1}. Cells from mid-log phase were collected by centrifugation at 2,000 r.p.m. (650*g*) for 10 min for RNA isolation.

Isolation of total RNA. Total RNA was isolated by the TRIzol (Invitrogen Life Sciences) method and subsequently cleaned from DNA using 0.08 U μl^{-1} RNase-free DNase I enzyme (Ambion). The quality of RNA was assessed on a 5% TBE-urea denaturing gel (Bio-Rad Laboratories) and the concentration was measured spectrophotometrically.

RT-PCR verification experiments. We carried out RT-PCR to validate the central metabolic transcripts. The reverse transcription of the *C. reinhardtii* total RNA was performed using Superscript III reverse transcriptase (Invitrogen Life Sciences) and $dT_{(16)}$ as general primer. The reaction mixture contained 1.2 M betaine (Sigma-Aldrich) to prevent premature terminations owing to the high G+C content of *C. reinhardtii* transcriptome. The resultant cDNAs were amplified by PCR using KOD hot start DNA polymerase (Novagen). As in the reverse transcription reaction, we included 1.2 M betaine in all PCRs to optimize the yield. Forward and reverse Gateway-tailed primers were used to allow recombinational cloning[9]: The forward primers were designed using the predicted ORF

doi:10.1038/nmeth.1348

sequence starting at ATG of the annotated 5′ end exon and were 5′-tailed with the Gateway B1.1 sequence. The gene-specific part of each reverse primer was designed using the very 3′-end sequence of the annotated 3′ exon omitting stop codon and 3′-tailed with the Gateway B2.1 sequence. All primers had a melting temperature (T_m) between 55 °C and 65 °C. The sequences of the primers are available in **Supplementary Table 5**.

RACE verification experiments. We removed the cap structure from *C. reinhardtii* mRNA. Total RNA was first dephosphorylated using 1 U μl^{-1} calf intestinal phosphatase (New England Bio-Labs) to remove 5′ phosphates from truncated mRNAs and non-mRNA molecules. The dephosphorylated RNA was then treated with 0.5 U μl^{-1} tobacco acid pyrophosphorylase (Epicentre Biotechnologies) to remove the cap structure from full length mRNAs.

To generate the templates for 5′ RACE, an RNA oligo sequence (GR-RNA) was ligated to the 5′ end of the decapped RNA in a reaction catalyzed by 1 U μl^{-1} T4 RNA ligase I (New England Biolabs). The decapped-ligated RNA was then reverse transcribed by the $dT_{(24)}$ GR3 primer and random hexamers. For 3′ RACE reactions, the $dT_{(24)}$ GR3 primer was used to reverse transcribe total RNA without addition of random hexamers. Both the RNA oligo and $dT_{(24)}$ GR3 primer sequences were derived from Invitrogen GeneRace kit. cDNA synthesis was catalysed by Superscript III reverse transcriptase in a reaction mixture contained 1.2 M betaine.

RACE amplicons were generated in two PCRs. To obtain 5′ ends, we amplified the cDNA using a forward general primer that was homologous to the RNA oligo ligated to the 5′ ends (GR5S). Reverse primers were gene-specific (see **Supplementary Table 5** for sequences) and were designed antisense to the putative ORF region of the gene of interest. These primers were placed 300–350 bases 3′ to the putative start of the ORF. 3′ ends were obtained using GR3 (derived from Invitrogen GeneRacer kit) as general, reverse primer and a forward, gene-specific primer (see **Supplementary Table 5** for sequences) that was designed sense relative to the mRNA. The latter primer was placed 300–350 bases upstream of the putative stop codon. To provide these PCRs with adequate coverage of the transcriptome, the amount of reverse transcribed template was adjusted such that equivalent of ~150 ng total RNA was introduced to each reaction. PCR was performed as a 'touch-down' PCR in which the annealing temperature of the first 5 cycles was 65 °C, on average 5–10 degrees above the T_m of the gene-specific primers. We used 0.5 μl of the first PCR product as template to run the second set of PCRs, which also performed as touchdown. A set of nested, tailed and proximal primers were used in these PCR reactions. To amplify 5′ ends we used GGRn5S as forward, general nested primer. The primer was tailed with the B1.1 Gateway sequence at its 5′ end. The reverse primers were nested gene-specific and were tailed with the Gateway B2.1 sequence (**Supplementary Table 5**). The 3′ ends were amplified using GGRn3 as reverse general primer that was 3′ Gateway-tailed with the B2.1 sequence. The nested 3′ RACE gene-specific primers had the same general design as the 5′ RACE primers, except that they were in the forward orientation and contained a Gateway B1.1 tail (**Supplementary Table 5**). Nested PCR step increased sensitivity and specificity of the experiment while providing Gateway tails for cloning.

Gateway cloning and sequencing. PCR products generated in RACE or in ORF verification experiments were recombinationally cloned in a BP reaction into pDONR223 to generate Gateway Entry clones[25]. Chemically competent DH5α *E. coli* was then transformed with the BP reaction products in 96-well microtiter plates containing spectinomycin as selection marker of cells bearing entry clone. Following growth in liquid media, the transformed bacteria were used as a source of template in PCR reactions, containing 1.2 M betaine and KOD hot start DNA polymerase (Novagen) to amplify the clones. Vector primers were used to generate the final DNA template for sequencing. PCR products were sequenced bidirectionally using conventional automated cycle sequencing to generate ORF sequence tags (OSTs)[2] or RACE sequence tags (RSTs). 3′ RACE products were sequenced unidirectionally from 5′ ends owing to the presence of poly(A) tails. Sequencing was carried out by Agencourt Bioscience Corp.

Trace analysis: ORF sequence tags (OSTs). Forward and reverse sequences were vector-clipped (using Cross_match), quality-trimmed, then assembled. For quality trimming, we kept the longest continuous sequence with average Phred score above 15 in a window of 20 nucleotides. We used Phrap (http://www.phrap.org/) to assemble the forward and reverse sequences. Both assembled contigs and singlets were aligned against the coding sequences (CDSs) of corresponding predicted transcripts from *C. reinhardtii* assembly v.3.1 (JGI v3.1, ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.1/Chlre3_1.fasta.gz) using T-Coffee[26] or MUSCLE[27]. The alignment files were then used to verify the CDSs of the predicted transcripts.

Trace analysis: RACE sequence tags (RSTs). We obtained both forward and reverse reads for 5′ RSTs, whereas only forward reads were generated for 3′ RSTs (owing to difficulties associated with sequencing through poly(A) tails). For 5′ RSTs, we assembled the forward and reverse reads using Phrap. The 5′ RST contigs and singlets, as well as 3′ RSTs, were aligned against CDSs of JGI v.3.1 predicted transcripts using T-Coffee or MUSCLE and evaluated (**Supplementary Note**).

15. Belknap, W.R. & Togasaki, R.K. *Proc. Natl. Acad. Sci. USA* **78**, 2310–2314 (1981).
16. Klein, U., Chen, C. & Gibbs, M. *Plant Physiol.* **72**, 488–491 (1983).
17. Clemetson, J.M., Boschetti, A. & Clemetson, K.J. *J. Biol. Chem.* **267**, 19773–19779 (1992).
18. Lemaire, S.D. *et al. Proc. Natl. Acad. Sci. USA* **101**, 7475–7480 (2004).
19. Gfeller, R.P. & Gibbs, M. *Plant Physiol.* **75**, 212–218 (1984).
20. Price, N.D., Famili, I., Beard, D.A. & Palsson, B.O. *Biophys. J.* **83**, 2879–2882 (2002).
21. Becker, S.A. *et al. Nat. Protocols* **2**, 727–738 (2007).
22. Lee, J.M., Gianchandani, E.P. & Papin, J.A. *Brief. Bioinform.* **7**, 140–150 (2006).
23. Mahadevan, R. & Schilling, C.H. *Metab. Eng.* **5**, 264–276 (2003).
24. Merchant, S.S. *et al. Science* **318**, 245–250 (2007).
25. Rual, J.F., Hill, D.E. & Vidal, M. *Curr. Opin. Chem. Biol.* **8**, 20–25 (2004).
26. Notredame, C., Higgins, D.G. & Heringa, J. *J. Mol. Biol.* **302**, 205–217 (2000).
27. Edgar, R.C. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

doi:10.1038/nmeth.1348

Review

Metabolic systems analysis to advance algal biotechnology

Brian J. Schmidt[1]**, *Xiefan Lin-Schmidt*[1], *Austin Chamberlin*[1], *Kourosh Salehi-Ashtiani*[2]* and *Jason A. Papin*[1]

[1] Department of Biomedical Engineering, University of Virginia, Health System, Charlottesville, VA, USA
[2] Center for Cancer Systems Biology (CCSB), Department of Cancer Biology and Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA, USA

Algal fuel sources promise unsurpassed yields in a carbon neutral manner that minimizes resource competition between agriculture and fuel crops. Many challenges must be addressed before algal biofuels can be accepted as a component of the fossil fuel replacement strategy. One significant challenge is that the cost of algal fuel production must become competitive with existing fuel alternatives. Algal biofuel production presents the opportunity to fine-tune microbial metabolic machinery for an optimal blend of biomass constituents and desired fuel molecules. Genome-scale model-driven algal metabolic design promises to facilitate both goals by directing the utilization of metabolites in the complex, interconnected metabolic networks to optimize production of the compounds of interest. Network analysis can direct microbial development efforts towards successful strategies and enable quantitative fine-tuning of the network for optimal product yields while maintaining the robustness of the production microbe. Metabolic modeling yields insights into microbial function, guides experiments by generating testable hypotheses, and enables the refinement of knowledge on the specific organism. While the application of such analytical approaches to algal systems is limited to date, metabolic network analysis can improve understanding of algal metabolic systems and play an important role in expediting the adoption of new biofuel technologies.

Received 12 April 2010
Revised 9 June 2010
Accepted 10 June 2010

Supporting information available online

Keywords: Algae · Biofuels · *Chlamydomonas reinhardtii* · Flux balance analysis · Metabolic network

1 Introduction

The use of microorganisms to produce compounds of commercial value enjoys a rich history. Of recent interest is the use of algae for the synthesis of nutraceuticals and biofuels. For example, high-value molecules are extracted from microalgae, such as carotenoid pigments and docosahexaenoic acid (DHA), an ω3 fatty acid [1]. Polysaccharides, sterols, and polyunsaturated fatty acids are all nutraceutical compounds extracted from algae [2]. Large-scale commercial culture of strains of *Chlorella* and *Arthospira* as a nutritious food date back to the 1960s and 1970s, respectively [1]. Table 1 lists a few of the well-defined molecules of commercial value that are purified from algal sources.

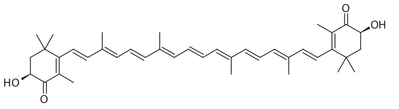
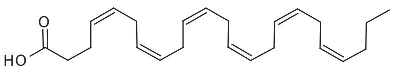
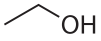
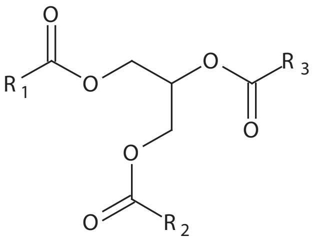
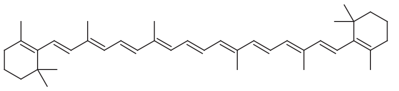
Microalgae hold promise as a source of renewable energy. Algae-derived hydrogen, methane, triacylglycerols, and ethanol all serve as potential materials for biofuels [3–6]. For example, depending on production conditions, *Schizochytrium* sp. and *Botryococcus braunii* may yield 50–77% and 25–75% oil by mass, respectively [3]. Algae oils are rich in the triacylglycerols that serve as material for conversion to biodiesel [3]. Some species of microalgae, such as *Chlamydomonas reinhardtii*, may pro-

Correspondence: Dr. Jason A. Papin, Department of Biomedical Engineering, University of Virginia, Box 800759, Health System, Charlottesville, VA 22908, USA
E-mail: papin@virginia.edu

Abbreviation: FBA, flux balance analysis

* Additional corresponding author: Dr. Kourosh Salehi-Ashtiani
E-mail: kourosh_salehi-ashtiani@dfci.harvard.edu
** *Current address*: Entelos, Foster City, CA 94404, USA

Table 1. Selected molecularly defined products currently isolated from microalgae

Name or family	Structure	Companies	Commercialized species
Astaxanthin (food colorant, antioxidant [1])		Cyanotech [1] Mera Pharmaceuticals [1] Bioreal [1] Parry's Pharmaceuticals [1] Algatech [1]	*Haematococcus pluvialis* [1]
Docosahexaenoic acid (ω3 fatty acid, cardiovascular health, brain development [1])		Seambiotic Martek Biosciences Corporation [1] OmegaTech [1] Nutrinova [1]	*Crypthecodinium conhii* [1] *Shizochtrium* sp. [1] *Ulkenia* sp. [1]
Ethanol (biofuel)		Algenol Biofuels [70] Seambiotic Inventure Chemical	Various cyanobacteria
Hydrogen (biofuel)	H_2	Solarvest BioEnergy [70]	*Chlamydomonas reinhardtii* [71]
Triacylglycerols (biodiesel precursor)		Aurora Biofuels [70] Solarvest BioEnergy Seambiotic Inventure Chemical Solazyme [72]	*Haematococcus pluvialis* [10, 73][a)]
β-Carotene (food colorant, provitamin A, antioxidant [1])		Western Biotechnology [1] Betatene [1]	*Dunaliella salina* [1]

a) Species data from published trial results by Aquasearch [10]. Additional species are likely suitable for biodiesel production [74], but the identity of algae employed in new commercial ventures are not usually publicized.

duce hydrogen directly [4, 7]. Additionally, the doubling time of microalgae in the exponential growth phase is as short as 3.5 h, and they are efficient at utilizing light to produce biomass, facilitating rapid fuel production [4]. Although some algae may be capable of utilizing biomass feedstocks as other microbes do, utilizing the photosynthetic route will arguably be the most efficient means of biofuel production [5].

Microalgal biofuel cultivation promises to be highly sustainable. Importantly, microalgae are much more distant from the human food chain than plant crops, avoiding competition between agricultural and biofuel resources [3]. As shown in Table 2, biodiesel produced from photosynthetic microalgae have a much higher yield than current biofuels and can be cultured on marginal land, further reducing the diversion of agricultural resources. Additionally, some algae can be cultured with saltwater or wastewater, avoiding use of freshwater resources [5]. Since microalgal fuel yields on an area basis are higher than currently possible with crops, they are more capable of meeting fuel demand [3]. Furthermore, microalgae cultures have been demonstrated to fix carbon dioxide, and may be utilized in the bioremediation of industrial flue gases [8–10]. Algal fuels are therefore carbon neutral, or carbon negative in the case of hydrogen.

Despite the advantages of algae as a source of biofuels, there are still significant challenges that must be addressed before algal biofuels can be widely adopted. Although compatible with the existing fuel infrastructure, biodiesel from algae is not yet economically competitive with fossil fuels or corn ethanol (Table 2). For algae biodiesel production, an additional challenge will be altering the selected algae to produce triacylglycerol fatty acid constituents with the optimal length and hydrocarbon saturation [5]. In this review article, we describe a systems level metabolic modeling approach that enables the generation of hypotheses to modify algal metabolism towards more efficient

Table 2. Comparison between gasoline, corn ethanol, microalgae biodiesel, and microalgae hydrogen as fuel sources

	Gasoline	Corn ethanol		Microalgae biodiesel		Microalgae hydrogen	
Energy	[75]	[75]		[75]		[7]	
(BTU/gal)	118 170	76 300		116 090		0.0458[a]	
(kJ/kg)	46 000	27 000		39 000		142 000	
		Low	High	Low	High	Low	High
Yield	N/A	[76]	[76]	[5]	[3]	[7]	[7]
(L/Ha/yr)		3970	5590	12 000[b]	136 900[c]	160 000[a]	830 000[a]
(g/m^2/day)		1.2	1.7	3.7	43	1.4[d]	4.5[e]
Cost to produce	2009 est.	2007 est.[77]		2009 est.[78]		2004 est. [79]	
\$/gal	1.86[f]	1.69		2.5–25[g]		\$0.57/kg–\$13.53/kg	
\$/10^6 BTU	15.7	22.1		21.5–215		4.2–100	

[a] Reported values for hydrogen on a volumetric basis assume standard temperature and pressure.
[b] Algae harvest 10 g/m^2/day, 30% oil in biomass.
[c] Algae harvest 60 g/m^2/day, 70% oil in biomass.
[d] Assumes the demonstrated 2% photoconversion efficiency.
[e] Assumes the theoretical limit of 10.6% photoconversion efficiency.
[f] Assumes the 2009 national average retail price of \$2.31/gal (www.eia.doe.gov), corrected for a tax of \$0.45/gal (www.api.org).
[g] Range depends on algae productivity.

production of desired compounds. We describe how such network models are constructed and present a number of case studies in which network modeling has been carried out.

2 *In silico* directed metabolic engineering approaches facilitate economical production schemes

Theoretically, the yield and synthesis rate of any metabolite could be optimized through the process of metabolic engineering. Metabolic engineering can be described as the optimization of entire metabolic or biosynthetic pathways through the manipulation of the genetic content or environmental context [11]. The advantages of utilizing *in silico* directed metabolic engineering to optimize microbial production processes over traditional strain improvement methods have already been demonstrated for commercially important microbes such as *Saccharomyces cerevisiae* [12]. Traditional methods of strain improvement include many rounds of selection, mutagenesis, mating, and hybridization [12]. Modeling approaches obviate this labor-intensive process and further minimize the potential for the introduction and accumulation of undesired mutations that may compromise production conditions [12]. Metabolic engineering can also exploit quantitative fine-tuning of gene expression to optimize product yields [12]. Due to the required efficiency of the production process and necessity to achieve high yields, metabolic models may play an essential part in making microalgal biofuel production commercially viable. Several examples where metabolic engineering guided by large-scale mathematical models have optimized the production of a desired metabolite are presented in Table 3. Many of the models now employed capture metabolic processes at a genome scale [13].

The approaches taken for the production of pharmaceutical compounds, especially biologics, offer a significant contrast to what would be expected for successful methodology for biofuel production. Pharmaceutical production often employs genetic engineering approaches to overexpress a single recombinant protein [11]. This approach cannot be used to optimize production of small molecule metabolites, such as triacylgycerols. The interconnectivity of metabolic pathways, with many metabolites feeding into multiple reactions, can make the optimization process counter-intuitive; a greater knowledge of metabolic network properties and mathematical modeling of these networks are needed to optimize bioproduction processes [14]. The production of some small molecule therapeutics may also take advantage of mathematical modeling of metabolic networks in the future. For example, metabolic modeling has been applied to investigate ways to increase penicillin production [15], and metabolic models will likely play a role in optimizing strains for the production of new antibiotics [16]. Notably, *C. reinhardtii* is being developed for the production of therapeutic proteins [17].

Mathematical modeling of metabolism can elucidate metabolic network properties and facilitates optimization. At the simplest level, metabolic modeling can supplement high-throughput data generation technologies, such as transcrip-

Table 3. Examples where genome-scale mathematical models have demonstrated the potential to optimize the microbial production of commercially important compounds. A compendium of genome-scale metabolic models can be found elsewhere [80]

Microorganism	Metabolite	Application	Reference(s)
Saccharomyces cerevisiae	Ethanol	Biofuel	[18]
Clostridium thermocellum	Ethanol	Biofuel	[54]
Lactococcus lactis	Diacetyl	Food (dairy flavor)	[81]
Pseudomonas putida	Polyhydroxyalkonoates	Plastics	[82, 83]
Corynebacterium glutamicum	L-Lysine	Food and animal feed (essential amino acid)	[84]
Clostridium acetobutylicum	Butanol	Biofuel	[85]
Escherichia coli	L-Threonine	Food, animal feed, pharmaceutical and cosmetic	[86]
Escherichia coli	Lycopene	Nutraceutical	[50]
Escherichia coli	Succinic acid	Polymers and many others	[87]
Mannheimia succiniciproducens	Succinic acid	Polymers and many others	[88–91]

tional profiling, to develop a meaningful visual representation of network function [14]. Furthermore, optimizing individual pathways can impact the utilization of global cofactors, such as NADH, NADPH, and ATP [18]. The utilization of pooled resources by different pathways is one reason a genome-scale mathematical model can be necessary to interpret phenotypic changes in metabolically engineered organisms [14]. Additionally, mathematical optimization focuses development efforts on the engineering strategies most likely to yield improvements in yield, titer, productivity, and robustness [19]. The required resources and time to commercialization can be greatly reduced compared to purely experimental development methods [19].

3 Selection of microalgae for biofuels production

In general, two approaches might be utilized to develop a metabolically engineered organism. Novel strains with perhaps less well-defined metabolisms but with unique, advantageous characteristics (*e.g.*, ability to process a particular substrate) may be utilized and subjected to targeted genetic modification as needed [11]. Alternatively, a model organism with relatively well-defined metabolic machinery already in place could be utilized.

The advantage to utilizing microalgae strains that already produce a desired metabolite is that it may be possible to find wild-type strains that give good yields. The search for such microorganisms is called bioprospecting [11]. The disadvantage is that molecular techniques may not exist for efficiently introducing and obtaining expression of genes in novel microorganisms [11]. For example, difficulties encountered trying to engineer *Clostridium acetobutylicum* to increase butanol production have led some researchers to develop new butanol production microbes in place of *C. acetobutylicum* [20, 21]. Notably, aside from *Chlamydomonas reinhardtii*, methods for the genetic manipulation of algal species are not well established [5].

The ideal selection for *de novo* metabolic design would be a laboratory model organism, such as *Escherichia coli*, *S. cerevisiae*, or *C. reinhardtii* due to the availability of laboratory techniques for genetic manipulation, a sequenced genome, and availability of genome-scale metabolic models. Two advantages of utilizing model organisms are that the tools for genetic manipulation are present and mathematical descriptions of the metabolic pathways may already exist [11]. A potential disadvantage of utilizing model organisms is that introducing entire metabolic pathways may present a substantial challenge in itself.

A more extreme case of truly *de novo* metabolic design would be to build an organism from scratch for the optimal production of the metabolite of interest [22]. Indeed, recent advances in synthetic biology techniques include the construction of full *Mycoplasma* genomes and their introduction into an organism [23, 24]. A fully synthetic approach would facilitate the design of microbial factories that would use a minimal mixture of inexpensive feedstock for growth and the optimized conversion to the desired metabolite [25]. However, although synthetic approaches have been successfully utilized to add pathways and gene networks to organisms [26], these approaches have not yet been utilized to make a minimal, fully engineered microbe capable of producing compounds of economic value. There are also additional fundamental challenges to constructing a vi-

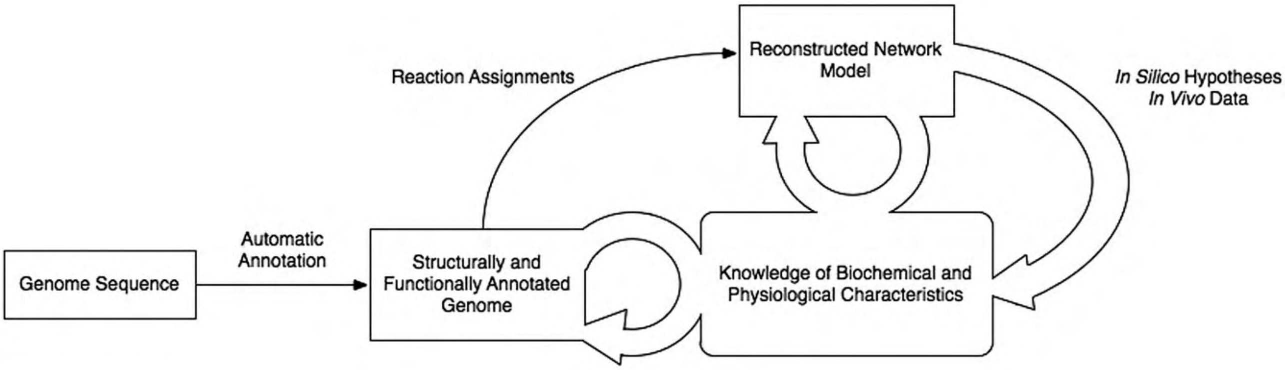

Reconstructed Network Model
Reaction Assignments
In Silico Hypotheses
In Vivo Data
Automatic Annotation
Genome Sequence
Structurally and Functionally Annotated Genome
Knowledge of Biochemical and Physiological Characteristics

Figure 1. The process of developing a metabolic model.

able synthetic organism as a commercial production platform, such as making the associated regulatory networks sufficiently robust to environment perturbations, mutations, and noise in gene expression [27, 28].

Each approach has advantages, and all hold potential for microalgal biofuels production. There is substantial interest in both algal bioprospecting and developing laboratory algae such as *C. reinhardtii*. Metabolic systems analysis can play an important role in both approaches.

4 Developing systems biology of algae through metabolic network modeling

Systems biology provides the means to understand the emergent properties of biological systems and predict systems behavior under different physiological conditions. Metabolic network modeling, as a systems approach, integrates different large-scale datasets, genomic information, and mathematical equations, to model and predict the metabolic fluxes of an organism. As described in more detail below, network reconstruction is an iterative process that starts with building a draft metabolic network using the available literature and genomic evidence, the incorporation of reaction stoichiometry, gene-reaction association, and cellular localization of reactions. The next step is the conversion of the reconstructed network into a computable format. The final step is the evaluation and refinement of the network model through comparison with experimental data [29]. The iteration of these steps can improve the accuracy of the model.

4.1 Metabolic network reconstruction and analysis

The workflow for the development and refinement of a metabolic network model is illustrated in Fig. 1. Sequenced genomes serve as a starting point for the reconstruction. In addition to the *C. reinhardtii* genome [30], complete genome assemblies for several algae-related species are available (e.g., *Acaryochloris marina* [31], *Anabaena* sp., *Cyanidioshyzon merolae* [32, 33], *Ostreococcus tauri* [34], and *Synechococcus* sp. [35–37]). As additional high-quality metabolic network reconstructions emerge, metabolism in multiple algal species can be compared *in silico*. Their reconciliation may serve as an additional validation of their reconstruction and enhance understanding of microbial specialization. Comparisons will facilitate selecting an optimal species as a starting point for biofuel production. Furthermore, analysis of several metabolic networks may help to identify ideal species for modification based on the best production potential rather than optimal production in the starting strain [38].

After sequencing, the genome is structurally annotated to define genes and transcribed elements. Once open reading frames (ORFs) are delineated, molecular function can be assigned through comparison with genes associated with proteins of known functions. Functional assignments can be made through profile-based domain assignments [39] or, as a first draft, by predicting protein function based on sequence similarity with proteins of previously annotated function in a database such as Uniprot (http://www.uniprot.org/). The automated annotation pipeline results in a genome annotated with Enzyme Commission (EC) numbers which designate the putative catalytic function of the gene product [40]. The reliability of this process is improved by the availability of accurate annotation data for related organisms.

With an annotated genome in hand, a reconstruction can be generated in a structured format such as a stoichiometric matrix. The stoichiometric matrix accounts for compounds (as rows) and corresponding chemical transformations (as columns) in which the elements of the matrix correspond to the stoichiometric coefficients. While the stoichiometry of metabolic reactions is fixed, the annotated genome enables the identification of which reactions are to be included in a given network. Reactions are assigned to the annotated genes using a metabolic database such as the Kyoto Encyclopedia of Genes and Genomes (KEGG). Reaction properties such as reversibility or localization to specific cellular compartments are also built into the network model [41]. The resulting reaction network may contain incomplete pathways or lack metabolic functions for which there is empirical evidence. In such cases, the network is curated to make it consistent with the known physiological and biochemical characteristics of the organism [42]. The model is then converted to a computable format to allow for quantitative analysis [43]. SBML formats facilitate the exchange of models between research groups and compatibility with software tools (http://sbml.org).

It is likely that the model will lack reactions that are present in the organism, as many gene functions are undetermined. It is also possible the model may include reactions which are not present [43]. Developing a metabolic network model is an iterative process in which the model is refined as hypotheses based on simulations are tested against experimental results [44]. Metabolomic and transcriptomic data from high-throughput experiments can be used to evaluate and refine the model, iteratively improving its capacity to predict phenotypes.

With a mathematically defined model, analysis can be performed to optimize or characterize the network. Because metabolic reactions occur on a fast time scale relative to other cellular processes, a reasonable assumption that enables the application of several analytical approaches is that the metabolic network operates at steady state. The steady-state assumption is inherent to flux balance analysis (FBA), a widely used metabolic modeling strategy. To analyze the network, constraints are placed on reaction fluxes, such as on the exchange reactions responsible for taking in nutrients, and the network is optimized with respect to a goal, frequently taken to be the growth of the organism (biomass production). The maximization of the objective function subject to constraints makes the linear programming problem a cornerstone of metabolic FBA. However, metabolic systems models are most frequently underdetermined: there are more reactions than metabolites, and there are frequently many solutions that give the same maximum objective. Software tools to perform constraint-based analysis on stoichiometric metabolic models are freely available (for example, the COBRA toolbox [45]). Genome-scale constraint-based models and FBA have been reviewed in more depth elsewhere [13, 46].

The constraint-based analysis approach can be applied to predict flux through metabolic pathways, optimal growth media, product yields, and other factors relevant to bioprocess design and optimization. In the context of metabolic engineering, gene knockouts are simulated by removing the corresponding reactions from the model. While the wild-type system is typically assumed to be optimized for biomass production, techniques have been developed to explore knockout combinations and gene additions that will maximize the production of a target metabolite by coupling it to cell growth [47, 48]. Interestingly, knockout phenotypes may no longer have the same biological objectives as their wild-type parents. It has been noted that the metabolic networks of mutants behave suboptimally with respect to growth, and instead more closely resemble the unperturbed network [49]. Thus, mutant phenotypes may be modeled more accurately through Minimization of Metabolic Adjustment (MOMA) rather than optimization of biomass production [49, 50]. These analytical tools may be useful for metabolic engineering strategies.

5 Case studies

As described, the process of metabolic network reconstruction naturally lends itself to an iterative approach. Subsequent rounds of model refinement facilitate the testing of hypotheses *in vivo*. The metabolic network model becomes a tool not just for finding the optimal solution to industrially relevant metabolic engineering challenges, but an integral part of conducting genome-scale research into the fundamental operating principels and mechanisms of organisms. To truly exploit the power of the metabolic network modeling approach, *in silico* research can be directly coupled to experimental verification, improving knowledge of the network components, annotation of the genome, and confidence in model predictions. We discuss three such examples where metabolic modeling has demonstrated encouraging results in the development of engineered microbial strains. A more extensive listing of model-driven metabolic engineering is shown in Table 3. While the application of these metabolic network analyses to algal systems is relatively limited to date, these examples provide an overview of the status of the field and some of the opportunities available.

5.1 Metabolic network reconstruction of *C. reinhardtii* with transcript verification

Manichaikul *et al.* [51] have described an iterative methodology for building a high-confidence, experimentally verified model of central metabolism and have applied the method to an updated *C. reinhardtii* genome sequence. Interestingly, the first round of automated functional annotation found six new enzymatic reactions involved in the production of triacylglycerols that were not present in the previous annotated genome, an enhancement potentially very relevant for future studies into biodiesel production. The reconstructed metabolic network model was initially focused on central metabolism. The network structure and reaction stoichiometry were identified by coupling the automated functional annotation with a manual review of the literature, KEGG, and the Expert Protein Analysis System proteomics server (ExPASy). Postulated transcripts encoding the enzymes mediating the network reactions were verified *in vivo* utilizing RT-PCR and rapid amplification of cDNA ends (RACE). The experimental verification improved the original structural annotation of the sequence, refining 5% of the ORFs. An additional round of expansion and verification was then applied to the network. Interestingly, two of the transcripts could not be verified experimentally, and literature evidence showed that one of the unverified transcripts is regulated by light in a genus of cyanobacteria. It is, therefore, likely the approach can be applied to account for the differential regulation of transcript expression, and thus network structure, based on growth conditions. Boyle and Morgan [52] also constructed a model of *C. reinhardtii* central metabolism and also demonstrated the utility of such network models for refining genome annotation and predicting phenotypes of the alga under defined environments. These network reconstructions can serve as a platform and starting point for more detailed metabolic engineering programs (as described below).

5.2 Optimization of ethanol production in *C. thermocellum*

Recently, a genome-scale metabolic model of *C. thermocellum* was constructed to investigate the production of ethanol from the alkaline cellulose degradation product, cellobiose [53]. The model identified several important knowledge gaps related to central metabolism. None of the existing genome annotations contained a gene for pyruvate kinase, and BLASTP identified several candidate genes that could encode the enzyme. The analysis also identified a gap in the citric acid cycle. The genome does not appear to encode for succinate dehydrogenase and enzymatic activity could not be detected. However, small amounts of succinate were detected in *C. thermocellum* culture, so it is likely there is an alternate pathway utilizing the metabolite [53, 54]. It will be important to determine the metabolic fate of succinate in future experiments and refine the model for improved predictions.

Strategies for increasing ethanol production through genetic modifications and altering the feedstock were identified. Metabolic reactions can exhibit a range of theoretical flux values while meeting the biological objective, and, notably, Roberts *et al.* [53] found this to be the case for ethanol production. Alternative solutions that result in the same optimal objective were sought in flux variability analysis (FVA). FVA predicted strains missing ferroredoxin hydrogenase and growth in media supplemented with lactate and malate results in a maximal 35-fold increase in the maximum theoretical ethanol yield, to about 140 mmol/gDW/h.

5.3 Optimization of lycopene production in *E. coli*

Alper *et al.* [50] investigated gene knockout methods to further optimize an industrial *E. coli* strain for the production of lycopene. A significant difficulty for *in silico* metabolic knockout design is that exhaustive search strategies are combinatorially complex and, therefore, not practical for designs exploiting multiple knockouts. Sequential strategies are not theoretically guaranteed to find the global optimum in the gene knockout space, especially if synergistic interactions are critical to the optimal solution. However, their sequential search method, as validated by an exhaustive pairwise search, performed excellently in identifying the best knockout combinations. Overall, *in vivo* verification of changes in microbial growth and ethanol production agreed well with predictions. However, the accuracy of the *in silico* prediction was compromised when the knockout resulted in the accumulation of 3-phosphoglycerate, a metabolite with known regulatory functions. The genome-scale stoichiometric model utilized did not incorporate regulatory effects, which may explain the discrepancy. Utilizing the sequential search strategy, a triple knockout mutant along the optimal *in silico* path was verified *in vivo* to produce 37% more lycopene than the parent industrial strain.

6 Ready for application: Algal metabolic systems analysis

To date, efforts aimed at genome-scale metabolic modeling have been primarily directed at bacterial networks. Model bacteria, such as *E. coli*, are the among the best characterized organisms, simplifying the substantial task of building a high-quality, curated model [55]. The small genomes of bacteria such as *E. coli* and *H. pylori* have also facilitated the expansion of the scope of the metabolic models to the genome scale [55], which have been iteratively tested and refined [55–57]. Additionally, industries of commercial scale, where modeling and optimization approaches have demonstrated value for other products, have a critical interest in also optimizing the production processes for products derived from microbes [19]. Notably, models of a much more ambitious scale have recently been constructed, such as multicompartmental genome-scale models of human metabolism [58] and the plant *Arabidopsis thaliana* [59].

These advances are being employed in the field of algal biotechnology, and arguably the field of algal systems biotechnology is still in an early stage of development. The sequencing of *C. reinhardtii*'s nuclear [30, 60], mitochondrial [61, 62], and chloroplast genomes [63] has enabled the few published large-scale computational models of algal metabolism. Three computational models of *C. reinhardtii* metabolism have been published. The first constraint-based model featured 484 reactions and 458 metabolites located in the cytosol and mitochondria [52]. Shortly thereafter, an independent model was published with 259 reactions and 267 metabolites localized to the cytosol, mitochondria, chloroplast, glyoxysome, and flagellum [51]. Additionally, a relatively large kinetic model of algal metabolism has been constructed that includes 95 reactions with 38 metabolites localized to the cytosol and mitochondria [64].

Construction and validation of accurate algal models is certainly more challenging than prokaryotic organisms given the multiple organelles and genomes. However, there is some guidance available from efforts with another complex, photosynthetic organism, *A. thaliana*, and *C. reinhardtii* should be an easier organism to work with [65]. One of the fundamental difficulties with complex multicompartmental models is determining the compartments to which specific metabolic reactions are localized, as duplication of portions of biochemical pathways occurs. A recent study employing three pentose phosphate model alternatives in *A. thaliana* was not able to distinguish between the possibilities using steady-state isotope labeling data [66]. This result emphasizes the need for additional biochemical evidence to develop accurate metabolic models, especially if a metabolic design situation requires manipulating compartment-specific reaction fluxes. However, it is worth noting that networks as large as that of the human [67] and *A. thaliana* [68] have been modeled with less accounting for compartmentalization. It is accepted that model construction is an iterative process [69], and the algae field is well-situated to begin applying and refining these models to guide experimental methods to produce products of commercial value from *C. reinhardtii*.

7 Summary and conclusions

Systems-based metabolic engineering holds promise for algal bioprocess design. Genome-scale models will generate testable hypotheses that may increase understanding of algal metabolism and lead to non-intuitive optimization strategies that traditional methods are unlikely to produce. The adoption of systems-based approaches to metabolic engineering of algae may be a critical step towards making algae-derived biofuels economically competitive. Several sequencing projects are underway, and the subsequent development of *in silico* models will cooperatively reinforce the utility of systems analysis for the algal biotechnology industry.

We would like to thank Ani Manichaikul for insightful discussions into the applications of microalgae-derived biotechnologies. This work was supported by a grant from the Office of Science (Biological and Environmental Research), US Department of Energy, DE-FG02-07ER64496, to J.P. and K.S.-A.

The authors have declared no conflict of interest.

8 References

[1] Spolaore, P., Joannis-Cassan, C., Duran, E., Isambert, A., Commercial applications of microalgae. *J. Biosci. Bioeng.* 2006, *101*, 87–96.

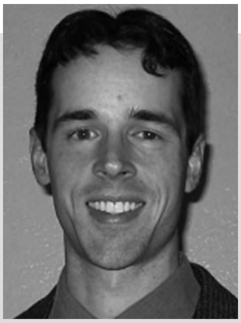
[2] Barrow, C., Shahidi, C., *Marine Nutraceuticals and Functional Foods*. Taylor & Francis Group, LLC, Boca Raton 2008.

[3] Chisti, Y., Biodiesel from microalgae. *Biotechnol. Adv.* 2007, *25*, 294–306.

[4] Hu, Q., Sommerfeld, M., Jarvis, E., Ghirardi, M. *et al.*, Microalgal triacylglycerols as feedstocks for biofuel production: Perspectives and advances. *Plant. J.* 2008, *54*, 621–639.

[5] Schenk, P., Thomas-Hall, S., Stephens, E., Marx, U. *et al.*, Second generation biofuels: High-efficiency microalgae for biodiesel production. *Bioenerg. Res.* 2008, *1*, 20–43.

Dr. Jason A. Papin is an Assistant Professor of Biomedical Engineering at the University of Virginia. He completed his undergraduate and graduate work in Bioengineering at the University of California, San Diego. His current research interests include the development of systems biology methods and the application of these methods to a diverse set of problems. In collaboration with Dr. Kourosh Salehi-Ashtiani, he is working on algal metabolic engineering with a focus on the model organism *Chlamydomonas reinhardtii*.



Dr. Brian J. Schmidt completed his BS in chemical engineering summa cum laude at the University of Pittsburgh in 2001. After working in the medical device industry, he joined Dr. Michael Lawrence's laboratory in the Department of Biomedical Engineering at the University of Virginia. There, he developed computational methods and novel assays based on microfluidic technology to study the mechanokinetic properties of leukocyte adhesion molecules and their impact on dynamic function. After completing his PhD in 2009, he joined Dr. Jason Papin's Computational Systems Biology Laboratory. His postdoctoral work focused on identifying predictive markers of tumor metastasis, invasiveness, and responsiveness to therapeutic agents and also linear systems modeling of cardiomyocyte metabolic dysfunction in diabetes. He currently works at Entelos, Inc., developing and applying dynamic models of disease pathophysiology to understand therapeutic mechanisms of action and investigating the clinical efficacy of new compounds.

[6] Rupprecht, J., From systems biology to fuel – *Chlamydomonas reinhardtii* as a model for a systems biology approach to improve biohydrogen production. *J. Biotechnol.* 2009, *142*, 10–20.

[7] Hankamer, B., Lehr, F., Rupprecht, J., Mussgnug, J. H. *et al.*, Photosynthetic biomass and H_2 production by green algae: From bioengineering to bioreactor scale-up. *Physiol. Plant* 2007, *131*, 10–21.

[8] Zeiler, K., Heacox, D., Toon, S., Kadam, K. *et al.*, The use of microalgae for assimilation and utilization of carbon dioxide from fossil fuel-fired power plant flue gas. *Energy Convers. Mgmt* 1995, *36*, 707–712.

[9] Brown, L. M., Uptake of carbon dioxide from flue gas by microalgae. *Energy Convers. Mgmt* 1996, *37*, 1363–1367.

[10] Huntley, M., Redalje, R., CO_2 mitigation and renewable oil from photosynthetic microbes: A new appraisal. *Mitigation Adapt. Strat. Global Change* 2007, *12*, 573–608.

[11] Alper, H., Stephanopoulos, G., Engineering for biofuels: Exploiting innate microbial capacity or importing biosynthetic potential? *Nat. Rev. Microbiol.* 2009, *7*, 715–723.

[12] Nevoigt, E., Progress in metabolic engineering of *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* 2008, *72*, 379–412.

[13] Price, N. D., Reed, J. L., Palsson, B. O., Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2004, *2*, 886–897.

[14] Smid, E. J., Molenaar, D., Hugenholtz, J., de Vos, W. M. *et al.*, Functional ingredient production: Application of global metabolic models. *Curr. Opin. Biotechnol.* 2005, *16*, 190–197.

[15] Jorgensen, H., Nielsen, J., Villadsen, J., Mollgaard, H., Metabolic flux distributions in *Penicillium chrysogenum* during fed-batch cultivations. *Biotechnol. Bioeng.* 1995, *46*, 117–131.

[16] Rokem, J. S., Lantz, A. E., Nielsen, J., Systems biology of antibiotic production by microorganisms. *Nat. Prod. Rep.* 2007, *24*, 1262–1287.

[17] Mayfield, S. P., Manuell, A. L., Chen, S., Wu, J. *et al.*, *Chlamydomonas reinhardtii* chloroplasts as protein factories. *Curr. Opin. Biotechnol.* 2007, *18*, 126–133.

[18] Bro, C., Regenberg, B., Forster, J., Nielsen, J., *In silico* aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metab. Eng.* 2006, *8*, 102–111.

[19] Otero, J. M., Nielsen, J., Industrial systems biology. *Biotechnol. Bioeng.* 2010, *105*, 439–460.

[20] Inui, M., Suda, M., Kimura, S., Yasuda, K. *et al.*, Expression of *Clostridium acetobutylicum* butanol synthetic genes in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* 2008, *77*, 1305–1316.

[21] Tang, W. L., Zhao, H., Industrial biotechnology: Tools and applications. *Biotechnol. J.* 2009, *4*, 1725–1739.

[22] Lee, S. K., Chou, H., Ham, T. S., Lee, T. S. *et al.*, Metabolic engineering of microorganisms for biofuels production: From bugs to synthetic biology to fuels. *Curr. Opin. Biotechnol.* 2008, *19*, 556–563.

[23] Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A. *et al.*, Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 2008, *319*, 1215–1220.

[24] Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N. *et al.*, Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 2010, in press. DOI: 101126/science. 1190719.

[25] Forster, A. C., Church, G. M., Towards synthesis of a minimal cell. *Mol. Syst. Biol.* 2006, *2*, 45.

[26] Heinemann, M., Panke, S., Synthetic biology – Putting engineering into biology. *Bioinformatics* 2006, *22*, 2790–2799.

[27] Chen, B. S., Chang, C. H., Lee, H. C., Robust synthetic biology design: Stochastic game theory approach. *Bioinformatics* 2009, *25*, 1822–1830.

[28] Chen, B. S., Wu, C. H., A systematic design method for robust synthetic biology to satisfy design specifications. *BMC Syst. Biol.* 2009, *3*, 66.

[29] Thiele, I., Palsson, B. O., A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 2010, *5*, 93–121.

[30] Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H. *et al.*, The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 2007, *318*, 245–250.

[31] Swingley, W. D., Chen, M., Cheung, P. C., Conrad, A. L. *et al.*, Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 2005–2010.

[32] Matsuzaki, M., Misumi, O., Shin, I. T., Maruyama, S. *et al.*, Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 2004, *428*, 653–657.

[33] Nozaki, H., Takano, H., Misumi, O., Terasawa, K. *et al.*, A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC Biol.* 2007, *5*, 28.

[34] Derelle, E., Ferraz, C., Rombauts, S., Rouze, P. *et al.*, Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 11647–1152.

[35] Sugita, C., Ogata, K., Shikata, M., Jikuya, H. *et al.*, Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: Gene content and organization. *Photosynth. Res.* 2007, *93*, 55–67.

[36] Palenik, B., Ren, Q., Dupont, C. L., Myers, G. S. *et al.*, Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 13555–13559.

[37] Palenik, B., Brahamsha, B., Larimer, F. W., Land, M. *et al.*, The genome of a motile marine *Synechococcus*. *Nature* 2003, *424*, 1037–1042.

[38] Trinh, C. T., Unrean, P., Srienc, F., Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Appl. Environ. Microbiol.* 2008, *74*, 3634–3643.

[39] Krogh, A., Brown, M., Mian, I. S., Sjolander, K. *et al.*, Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* 1994, *235*, 1501–1531.

[40] Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W. *et al.*, IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* 2004, *32*, Database issue, D434–437.

[41] Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. *et al.*, Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 2009, *7*, 129–143.

[42] Borodina, I., Nielsen, J., From genomes to *in silico* cells via metabolic networks. *Curr. Opin. Biotechnol.* 2005, *16*, 350–355.

[43] Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S. *et al.*, Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.* 2001, *26*, 179–186.

[44] Palsson, B., The challenges of *in silico* biology. *Nat. Biotechnol.* 2000, *18*, 1147–1150.

[45] Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G. *et al.*, Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protoc.* 2007, *2*, 727–738.

[46] Oberhardt, M. A., Chavali, A. K., Papin, J. A., Flux balance analysis: Interrogating genome-scale metabolic networks. *Methods Mol. Biol.* 2009, *500*, 61–80.

[47] Burgard, A. P., Pharkya, P., Maranas, C. D., Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 2003, *84*, 647–657.

[48] Pharkya, P., Burgard, A. P., Maranas, C. D., OptStrain: A computational framework for redesign of microbial production systems. *Genome Res.* 2004, *14*, 2367–2376.

[49] Segre, D., Vitkup, D., Church, G. M., Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 15112–15117.

[50] Alper, H., Jin, Y. S., Moxley, J. F., Stephanopoulos, G., Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab. Eng.* 2005, *7*, 155–164.

[51] Manichaikul, A., Ghamsari, L., Hom, E. F., Lin, C. *et al.*, Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat. Methods* 2009, *6*, 589–592.

[52] Boyle, N. R., Morgan, J. A., Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Syst. Biol.* 2009, *3*, 4.

[53] Roberts, S. B., Gowen, C. M., Brooks, J. P., Fong, S. S., Genome-scale metabolic analysis of *Clostridium thermocellum* for bioethanol production. *BMC Syst. Biol.* 2010, *4*, 31.

[54] Chinn, M. S., Nokes, S. E., Strobel, H. J., Influence of process conditions on end product formation from *Clostridium thermocellum* 27405 in solid substrate cultivation on paper pulp sludge. *Bioresour. Technol.* 2007, *98*, 2184–2193.

[55] Kim, H. U., Kim, T. Y., Lee, S. Y., Metabolic flux analysis and metabolic engineering of microorganisms. *Mol. Biosyst.* 2008, *4*, 113–120.

[56] Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., *et al.*, A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 2007, *3*, 121.

[57] Thiele, I., Vo, T. D., Price, N. D., Palsson, B. O., Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): An *in silico* genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol.* 2005, *187*, 5818–5830.

[58] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I. *et al.*, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 1777–1782.

[59] de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M. *et al.*, AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol*. *152*, 579–589.

[60] Shrager, J., Hauser, C., Chang, C. W., Harris, E. H. *et al.*, *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.* 2003, *131*, 401–408.

[61] Gray, M. W., Boer, P. H., Organization and expression of algal (*Chlamydomonas reinhardtii*) mitochondrial DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 1988, *319*, 135–147.

[62] Michaelis, G., Vahrenholz, C., Pratje, E., Mitochondrial DNA of *Chlamydomonas reinhardtii*: The gene for apocytochrome b and the complete functional map of the 15.8 kb DNA. *Mol. Gen. Genet.* 1990, *223*, 211–216.

[63] Maul, J. E., Lilly, J. W., Cui, L., dePamphilis, C. W. *et al.*, The *Chlamydomonas reinhardtii* plastid chromosome: Islands of genes in a sea of repeats. *Plant Cell* 2002, *14*, 2659–2679.

[64] Chang, C., Alber, D., Graf, P., Kwison, K. *et al.*, Addressing unknown constants and metabolic network behaviors through petascale computing: Understanding H2 production in green algae. *J. Phys. Conf. Ser.* 2007, *78*, 012011.

[65] Stitt, M., Lunn, J., Usadel, B., *Arabidopsis* and primary photosynthetic metabolism – More than the icing on the cake. *Plant J.* 2010, *61*, 1067–1091.

[66] Masakapalli, S. K., Le Lay, P., Huddleston, J. E., Pollock, N. L. *et al.*, Subcellular flux analysis of central metabolism in a heterotrophic *Arabidopsis* cell suspension using steady-state stable isotope labeling. *Plant Physiol.* 2010, *152*, 602–619.

[67] Ma, H., Sorokin, A., Mazein, A., Selkov, A. *et al.*, The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* 2007, *3*, 135.

[68] Poolman, M. G., Miguet, L., Sweetlove, L. J., Fell, D. A., A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiol.* 2009, *151*, 1570–1581.

[69] Kim, T. Y., Sohn, S. B., Kim, H. U., Lee, S. Y., Strategies for systems-level metabolic engineering. *Biotechnol. J.* 2008, *3*, 612–623.

[70] Waltz, E., Biotech's green gold? *Nat. Biotechnol.* 2009, *27*, 15–18.

[71] Surzycki, R., Cournac, L., Peltier, G., Rochaix, J. D., Potential for hydrogen production with inducible chloroplast gene expression in *Chlamydomonas*. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 17548–17553.

[72] Grant, B., Future oil. *Scientist* 2009, *23*, 2.

[73] Li, Y., Horsman, M., Wu, N., Lan, C. Q. *et al.*, Biofuels from microalgae. *Biotechnol. Prog.* 2008, *24*, 815–820.

[74] Mata, T., Martins, A., Caetano, N., Microalgae for biodiesel production and other applications: A review. *Renew. Sustain. Energy Rev.* 2010, *14*, 217–232.

[75] Darzins, A., Algal Biofuel Technologies. *States Biomass/Clean Cities Web Conference* 2008.

[76] Liska, A., Yang, H., Bremer, V., Klopfenstein, T. *et al.*, Improvements in life cycle energy efficiency and greenhouse gas emissions of corn-ethanol. *J. Ind. Ecol.* 2009, *13*, 58–74.

[77] Curtis, B., U. S. Ethanol Industry: The Next Inflection Point. BCurtis Energies & Resource Group 2008.

[78] Pienkos, P. T., Darzins, A., The promise and challenges of microalgal-derived biofuels. *Biofuels Bioprod. Bioref.* 2009, *3*, 431–440.

[79] Amos, W., Updated cost analysis of photobiological hydrogen production from *Chlamydomonas reinhardtii* green algae. National Renewable Energy Laboratories 2004.

[80] Milne, C. B., Kim, P. J., Eddy, J. A., Price, N. D., Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. *Biotechnol. J.* 2009, *4*, 1653–1670.

[81] Oliveira, A. P., Nielsen, J., Forster, J., Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* 2005, *5*, 39.

[82] Nogales, J., Palsson, B. O., Thiele, I., A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Syst. Biol.* 2008, *2*, 79.

[83] Puchalka, J., Oberhardt, M. A., Godinho, M., Bielecka, A. *et al.*, Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. *PLoS Comput. Biol.* 2008, *4*, e1000210.

[84] Kjeldsen, K. R., Nielsen, J., *In silico* genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.* 2009, *102*, 583–597.

[85] Lee, J., Yun, H., Feist, A. M., Palsson, B. O. *et al.*, Genome-scale reconstruction and *in silico* analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. *Appl. Microbiol. Biotechnol.* 2008, *80*, 849–862.

[86] Lee, K. H., Park, J. H., Kim, T. Y., Kim, H. U. *et al.*, Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol. Syst. Biol.* 2007, *3*, 149.

[87] Lee, S. Y., Hong, S. H., Moon, S. Y., *In silico* metabolic pathway analysis and design: Succinic acid production by metabolically engineered *Escherichia coli* as an example. *Genome Inform.* 2002, *13*, 214–223.

[88] Hong, S. H., Kim, J. S., Lee, S. Y., In, Y. H. *et al.*, The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat. Biotechnol.* 2004, *22*, 1275–1281.

[89] Lee, S. J., Song, H., Lee, S. Y., Genome-based metabolic engineering of *Mannheimia succiniciproducens* for succinic acid production. *Appl. Environ. Microbiol.* 2006, *72*, 1939–1948.

[90] Kim, T. Y., Kim, H. U., Park, J. M., Song, H. *et al.*, Genome-scale analysis of *Mannheimia succiniciproducens* metabolism. *Biotechnol. Bioeng.* 2007, *97*, 657–671.

[91] Kim, T. Y., Kim, H. U., Song, H., Lee, S. Y., *In silico* analysis of the effects of H_2 and CO_2 on the metabolism of a capnophilic bacterium *Mannheimia succiniciproducens*. *J. Biotechnol.* 2009, *144*, 184–189.

BMC Genomics

RESEARCH Open Access

Genome-wide functional annotation and structural verification of metabolic ORFeome of *Chlamydomonas reinhardtii*

Lila Ghamsari[1,2†], Santhanam Balaji[1,2†], Yun Shen[1,2], Xinping Yang[1,2], Dawit Balcha[1,2], Changyu Fan[1,2], Tong Hao[1,2], Haiyuan Yu[3*], Jason A Papin[4*], Kourosh Salehi-Ashtiani[1,2,5*]

Abstract

Background: Recent advances in the field of metabolic engineering have been expedited by the availability of genome sequences and metabolic modelling approaches. The complete sequencing of the *C. reinhardtii* genome has made this unicellular alga a good candidate for metabolic engineering studies; however, the annotation of the relevant genes has not been validated and the much-needed metabolic ORFeome is currently unavailable. We describe our efforts on the functional annotation of the ORF models released by the Joint Genome Institute (JGI), prediction of their subcellular localizations, and experimental verification of their structural annotation at the genome scale.

Results: We assigned enzymatic functions to the translated JGI ORF models of *C. reinhardtii* by reciprocal BLAST searches of the putative proteome against the UniProt and AraCyc enzyme databases. The best match for each translated ORF was identified and the EC numbers were transferred onto the ORF models. Enzymatic functional assignment was extended to the paralogs of the ORFs by clustering ORFs using BLASTCLUST.
In total, we assigned 911 enzymatic functions, including 886 EC numbers, to 1,427 transcripts. We further annotated the enzymatic ORFs by prediction of their subcellular localization. The majority of the ORFs are predicted to be compartmentalized in the cytosol and chloroplast. We verified the structure of the metabolism-related ORF models by reverse transcription-PCR of the functionally annotated ORFs. Following amplification and cloning, we carried out 454FLX and Sanger sequencing of the ORFs. Based on alignment of the 454FLX reads to the ORF predicted sequences, we obtained more than 90% coverage for more than 80% of the ORFs. In total, 1,087 ORF models were verified by 454 and Sanger sequencing methods. We obtained expression evidence for 98% of the metabolic ORFs in the algal cells grown under constant light in the presence of acetate.

Conclusions: We functionally annotated approximately 1,400 JGI predicted metabolic ORFs that can facilitate the reconstruction and refinement of a genome-scale metabolic network. The unveiling of the metabolic potential of this organism, along with structural verification of the relevant ORFs, facilitates the selection of metabolic engineering targets with applications in bioenergy and biopharmaceuticals. The ORF clones are a resource for downstream studies.

* Correspondence: Haiyuan.Yu@cornell.edu; papin@virginia.edu; ksa3@nyu.edu
† Contributed equally
[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[3]Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA
Full list of author information is available at the end of the article

Background

Recent advances in sequencing genomes of prokaryotes and eukaryotes [1] and the explosion of the development and use of genome-scale metabolic network reconstructions [2] are expected to facilitate the selection of targets for metabolic engineering [3,4]] . The unicellular green alga *Chlamydomonas reinhardtii* has been an attractive organism for exploration of metabolic engineering hypotheses due to its capability to flexibly regulate alternative biochemical pathways to produce biofuels [6-9]. However, the optimal selection of the enzymatic targets has been so far hindered by the lack of a comprehensive knowledge of the encoded genes that carry out the metabolic activities of the organism. Although the released genome sequence of *C. renihardtii* by the Joint Genome Institute (JGI) [10] provided the needed resource to predict nearly 17,000 genes in this organism, it alone does not reveal the underlying principles of metabolic network function, nor does it disclose the functions of the predicted “parts-list” of the organism. To define genes and map their products to function, computational algorithms have been extensively applied to annotate the accumulated genomic data from many organisms including *C. reinhartii*[11,12]. Most of these approaches are unable to predict the transcript structures precisely and accurately in a uniform manner due to 1) the incompleteness of the EST data, 2) the lack of comparative genomic information, particularly in less widely studied species, and 3) the complexity of the rules governing transcription initiation, termination and splicing events. Even for the well-studied nematode *C. elegans*, for which a high quality genome sequence has been available for over 10 years, inconsistencies still remain in defining the ORF structures [13,14]]. Previous large-scale studies on *C. reinhardtii*, have included microarray [15,16]], proteomics [17], and, more recently, RNAseq experiments [18] which have provided valuable expression data based on earlier releases of JGI annotations. Currently, the JGI v4.0 predicted *C. reinhardtii* ORFeome remains for the most part unverified; therefore, the functional annotation and experimental structural verification of the encoded ORFs are urgently needed prior to use in functional studies including metabolic engineering experiments.

We previously reported the functional annotation of the gene products involved in central metabolism of *C. reinhardtii* using JGI v3.0 gene models [19] in which we improved the existing functional and structural annotations of the ORF models. In the re-evaluation of the central metabolic ORFs, for which the ORFs are generally the best characterized in the proteome, we observed that as much as 10% of the ORFs were annotated with structuralerrors. The errors included incorrect 5’ or 3’ boundary annotations, which we identified through RACE [19].

In this study, we computationally assigned enzyme functions to the predicted and newly released JGI v4.0 protein-coding ORF models and targeted the enzymatic ORFeome for structural verification. Our results, in addition to structural verification, provide expression evidence for the enzymatic gene products, predict their subcellular localization, and identify the ORF models that may need to be re-annotated.

Results and discussion

Functional annotation of JGI v4.0 transcripts

We used the new JGI “filtered transcript models” released through the JGI portal (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) for both functional assignments and structural annotation verifications. Enzymatic functional assignments to the *C. reinhardtii* ORFs were made by associating Enzyme Commission (EC) numbers through reciprocal BLAST searches against the UniProt enzyme database [20] (http://www.uniprot.org/, with over 100,000 protein entries) (Figure 1A) supplemented with AraCyc database entries [21] . The best match for each translated ORF was identified (with an e-value threshold of 10^{-3}) and the EC number from the UniProt best match (or enzyme annotation from AraCyc) was transferred on to the JGI predicted ORF. We extended the EC assignments to the respective paralogs of the ORFs by clustering ORFs for the JGI filtered models. Altogether, we were able to assign 886 EC numbers to 1,427 JGI ORFs (Figure 1B, Additional file 1). KEGG currently provides 603 enzymatic annotations for the JGI v4.0 transcripts, of which there are 441 shared with our annotation. Theassignments given in this study provide an additional 445 EC numbers not present in KEGG. The list of the enzymatic JGI v4.0 gene models with their assigned EC numbers are provided in Additional file 1.

In order to provide additional functional information, WoLF PSORT [22] was implemented to assign subcellular localizations to each translated JGI v4.0 enzymatic ORF. WoLF PSORT is a high-performance localization prediction algorithm evolved from PSORT [23] , PSORT II [24] and iPSORT [25]; it combines localization features from these algorithms together with amino acid composition in a weighted *k*-nearest neighbors framework. Based on the cross-validation results, WoLF PSORT makes reliable predictions for nucleus, mitochondria, cytosol, plasma membrane, extracellular and (in plants) chloroplast. For other subcellular compartments, the performance is not as good, but still informative [22] . Compared to other methods, WoLF PSORT has been shown to have good performance for most

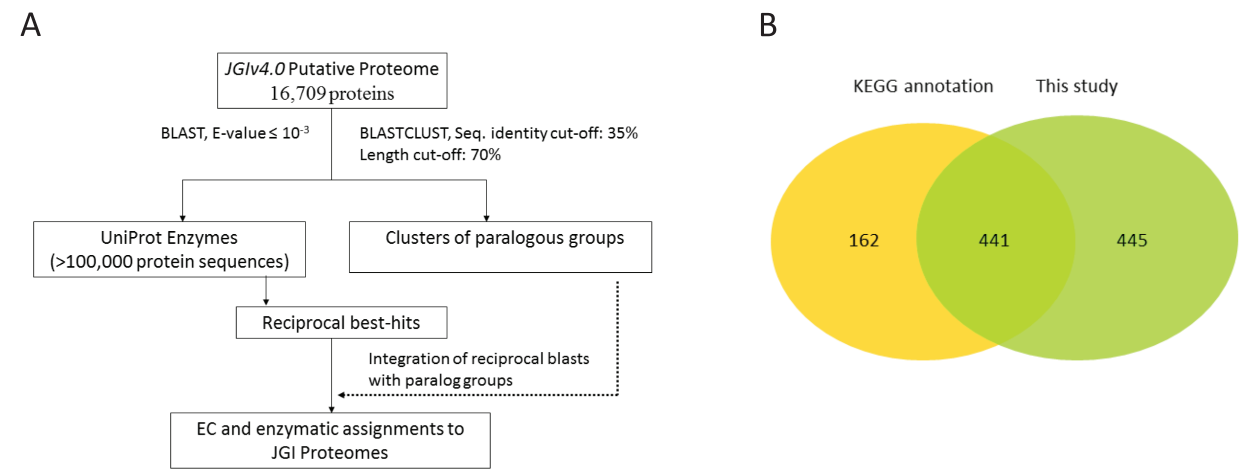

A
JGIv4.0 Putative Proteome
16,709 proteins
BLAST, E-value ≤ 10^{-3}
BLASTCLUST, Seq. identity cut-off: 35%
Length cut-off: 70%
UniProt Enzymes
(>100,000 protein sequences)
Clusters of paralogous groups
Reciprocal best-hits
Integration of reciprocal blasts
with paralog groups
EC and enzymatic assignments to
JGI Proteomes
B
KEGG annotation
This study
162
441
445

Figure 1 Functional annotation of *C. reinhardtii* JGI v4.0 translated ORFs. Enzymatic functions were assigned to the JGI v4.0 translated ORFs by comparing the sequences with the UniProt and AraCyc enzyme databases. The computational pipeline (A) entailed transfer of enzyme annotation to JGI ORFs identified through reciprocal BLAST, then establishing paralog groups to extend enzyme annotation to paralogs. Our functional annotation identified 886 EC numbers, of which only ~50% are currently annotated by KEGG (B).

subcellular localizations [26]. Importantly, predictions are not made on the basis of signal sequences that can introduce vulnerability to errors in sequence and/or annotations on the 5' end of the gene [27]. Furthermore, due to the unique phylogenetic position of *C. reinhardtii* and a lack of extensive GO annotation, alternative methods such as MultiLoc2 [28], which use GO annotation for refinement of predictions, would not be applicable here.

The results (Additional file 2) are presented as the number of nearest neighbors in different subcellular compartments for each protein. The default value for the total number of nearest neighbors (i.e., *k*) is 32. Even though *C. reinhardtii* is in the plant lineage, it has retained key animal genes [10] and is a unicellular organism that shares ancestry at the branching point of plants and animals. We therefore performed two WoLF PSORT runs in which *C. reinhardtii* was considered either as a plant or animal. Because *C. reinhardtii* is closer to plants than animals [10], predictions made when considering it as a plant are likely to be more accurate. However, because WoLF PSORT uses homology to known proteins, and some *C. reinhardtii* proteins may be closer to those in animals than plants [10], the predictions assuming an animal lineage provide alternative assignments, particularly for cases where ambiguous predictions are made for the proteins assuming plant origins. To summarize the obtained results (Fig. 2), we have binned the encoded proteins based on the assigned probability values for each protein, such that, if more than 50% of the nearest neighbors of the protein belong to a given compartment, that protein is assigned to a single compartment as its primary localization site. In cases where different localization predictions made based on animal and plant assumptions both meet an 85% cutoff, we took the higher confidence prediction as the final localization assignment (Additional file 3). Using this integration scheme, the largest compartment is the chloroplast when *C. reinhardtii* is considered a plant, and the second largest is the mitochondrion (Fig. 2C). These localization predictions agree with the fact that these genes are all related to metabolism. To verify the performance of our predictions, we manually curated a number of experimentally derived *C. reinhardtii* subcellular protein localizations recently reported by Weinkoop et al.[29]. Due to the limited number of localizations that could be transferred to v4.0 annotations from this study, we were only able to evaluate 9 ORFs in our set. Our predicted localizations of all 9 ORFs agreed with the experimentally determined localizations. Although the number is too small for adequate statistical analysis, it still shows the high quality of the predictions.

Experimental verification of *C. reinhardtii* enzymatic ORFeome

Our EC annotation of the JGI v4.0 transcript models identified 1,427 predicted transcripts with putative enzymatic functions. To experimentally verify structural annotation of the enzymatic ORFs, we carried out targeted transcriptome sequencing experiments after we amplified the ORFs by reverse transcription-PCR (RT-PCR) (Figure 3A). The generated amplicons were sequenced using the 454FLX platform before and after

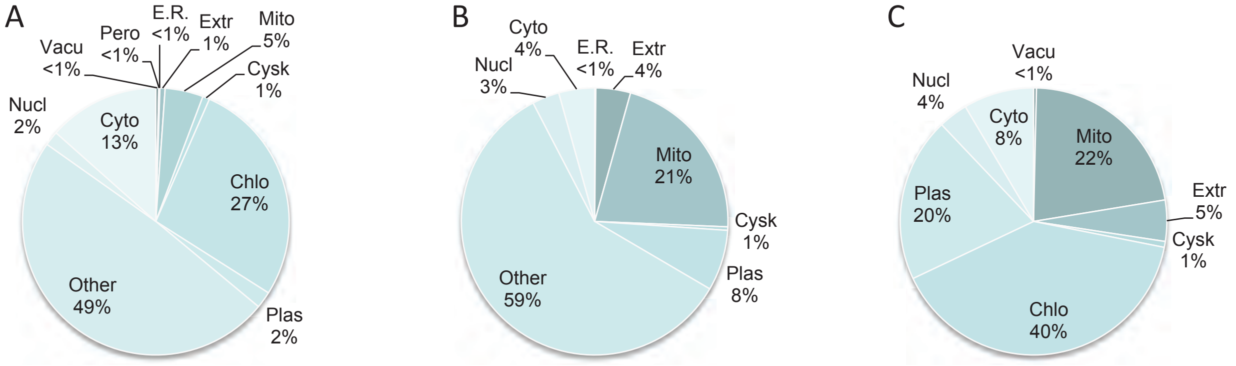


Figure 2 Subcellular localization prediction of JGI v4.0 enzymes. Following enzyme classification assignments to JGI v4.0 translated ORFs, subcellular localization of the proteins were predicted by WoLF PSORT [22] as plant (A) or animal (B) proteins. Based on the obtained probability values, each protein was assigned a compartment when 50% or higher percentage of the nearest neighbors for the protein belonged to a given compartment. When the 50% threshold is not reached, the protein, or its encoding ORF are assigned to "other" category to designate multiple compartments or ambiguous predictions. In (C), the predictions made as animal and plant were consolidated into a single set by increasing the threshold to 85%, then reporting the predicted assignment with the higher value. Abbreviations are: Chlo: chloroplast, Cyto: cytosol, Cysk: cytoskeleton, E.R.: endoplasmic reticulum, Extr: extracellular, Mito: mitochondrion, Nuc: nucleus, Pero: peroxisome, Plas: plasma membrane, Vacu: vacuolar membrane.

cloning of the amplicons into a Gateway vector. The sequences of the clones were further verified by conventional Sanger sequencing.

In order to perform the verification experiments, we grew *C. reinhardtii* under permissive condition by providing light, organic carbon sources and other nutrients (Methods). Total RNA from cells undergoing exponential growth was isolated and reverse transcribed to serve as a template for amplification of the ORFs for which we designed Gateway-tailed primers. Following

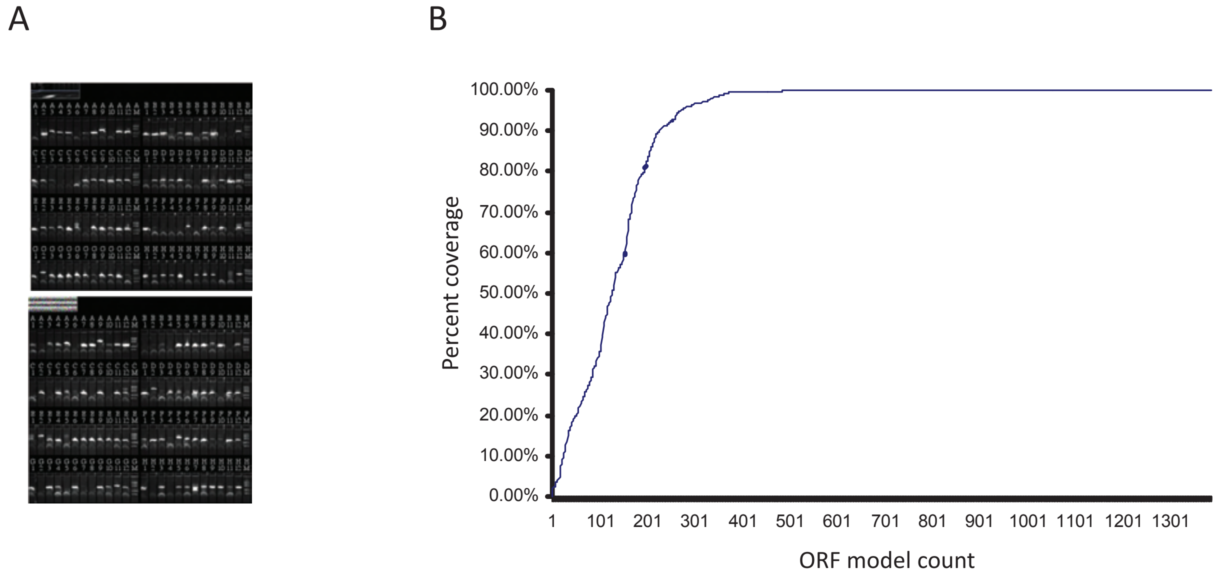


Figure 3 ORF verification by RT-PCR followed by multiplex sequencing. RNA isolated from *C. reinhardtii* grown under a permissive condition (continuous illumination and acetate as a source of carbon) was reverse transcribed, then used as template for PCR in which ORF-specific primers were used to amplify the JGI annotated ORFs. The amplicons were then sequenced directly using the 454FLX platform, or cloned, then sequenced by 454. (A) Amplification of representative metabolic ORFs are shown after electrophoresis (192 amplicons analyzed in two 96 well E-gels). (B) Percent coverage of 1,427 enzymatic ORF reference sequences by the obtained reads from 454 sequencing. The 454 reads were aligned to the JGI ORF reference sequences and percent coverage of the length of each reference sequence was determined (100% denotes all bases of the reference sequences could be covered by one or more 454 read). The entire lengths of 699 ORFs were 100% verified.

amplification, we carried out next generation sequencing (using the 454FLX platform) of the amplicons. The obtained 454 reads were then aligned to the JGI v4.0 ORF reference sequences to assess annotation accuracy. The aligned ORFs were binned according to their percent coverage; i.e., based on the percentage of the entire length of the ORF reference sequence that could be covered by the contigs assembled from the 454 reads.

For 78% of the JGI v4.0 ORF reference sequences, the 454 reads provided 95-100% coverage (Fig. 3B; Additional file 1), of this set approximately 92% had a coverage rate of 99-100%, demonstrating high verification rates. Approximately 10% of the ORF models showed coverage of 50-95%. The remaining 12% were covered less than 50% and of this set, 7% of the ORF models had less than 20% of their length verified by 454-reads.

As an alternative method of verifying the ORFs, we end-sequenced the cloned PCR products by conventional high-throughput Sanger sequencing. From 1,427 JGI v4.0 ORFs tested, we were able to obtain 661 ORF sequence tags (OSTs) that were aligned to the 5' end of the ORF models, and 631 OSTs that could be aligned to the 3' ends. Altogether, 42% (602) ORFs had OSTs that verified both ends of the ORF models. We could assemble full-length contigs for 242 ORFs (Additional file 1).

Overall, we obtained expression evidence for 1,401 of 1,427 ORF models with assigned enzymatic functions based on targeted transcriptome sequencing results and sequencing of the clones, though clearly not all of these ORF models can be considered verified. We consider an ORF model to be verified if 98 to 100% of its reference sequence could be covered by 454-reads, or if a full-length contig generated from Sanger sequencing of an obtained clone completely matched the reference sequence. For 73% of the ORF models, the 454-reads give confirmation at the 98-100% level. Sanger sequencing of the clones could verify an additional 36 ORF models (for which we could assemble contigs using 3' and 5' end reads). These models can therefore be considered verified, though it should be noted that even 100% coverage of an ORF model does not exclude the possibility of the presence of exons that were not annotated. The high coverage rates do, however, guarantee that the annotated exons are expressed. Furthermore, incomplete coverage by 454-reads does not necessarily imply inaccurate annotation; in some cases, less than 100% coverage could be the result of low expression level of the transcript and consequently low sequencing depth. We note that due to the amplification of the transcripts, the targeted transcriptome method that we have used is expected to normalize the abundance of the amplicons to a degree.

While end verification by Sanger sequencing can confidently verify the 5' and 3' ends, this method provides no information on the internal exon structure of long ORFs (unless internal primer walking [30] is carried out). We also find that the overall success rate of sequencing clones using the Sanger method is significantly lower than the 454 sequencing of amplicons. Cloning bottlenecks, failure to generate contigs due to end reads not covering the internal segments, and random sequencing failures could be among the contributing factors. Direct sequencing of amplicons through 454 or other parallel sequencing methods clearly bypasses these limitations.

Conclusions

A central challenge in the post-genomic era is the mapping of the genotype-phenotype relationship. For biochemical networks, the functional connections between genotype and phenotype are deciphered through the use of the available high-throughput experimental and computational platforms. Each technology can be used to generate a vast amount of data particular to some aspects of a given biochemical network. Ultimately the gathered data could be used to manipulate the biochemical systems for biotechnological and medical purposes. However, such efforts rest upon the availability of accurate structural and functional annotations, as well as the availability of the biological resources, such as ORF clones. In this study, we have carried out both computational functional annotation and direct experimental verification of structural annotations of JGI v4.0 enzymatic ORFs, which include both metabolic and non-metabolic enzymes. We carried out targeted amplification of the ORFs by RT-PCR and sequenced the products (before and after the cloning) to verify the ORF structures. The approach of using targeted amplification of ORFs offers several advantages over other high-throughput approaches that are not targeted; importantly, it establishes the *cis*-connectivity between the 5' and 3' ends of the ORF. Such *cis*-connectivity cannot be established from whole transcriptome sequencing, tiling array analysis or other high-throughput transcriptome survey methodologies (e.g., [18,31-34]). In addition, the generated amplicons can be cloned, as we have done so here, to provide reagents for downstream large- or small-scale experiments, which can be used to define genotype to phenotype maps as well as accomplishing bio-engineering tasks. With an ever-increasing number of organisms whose genome sequences are becoming available (e.g., the diatom *Phaeodactylum tricornutum* [35], the algae *Ostreococcus* Sp. [36] and *Volvox carteri* [37]), the need for structural and functional annotation and their verification is clear. The approach and experiments carried out in this study can be readily extended to other species to facilitate functional annotation and structural verification of their gene models.

Methods

Enzyme annotation of JGI v4 Proteome

We assigned Enzyme classification (EC) to the translated JGI v4.0 filtered ORF models (Chlre4_best_transcripts and Chlre4_best_proteins) using UniProt [20] and AraCyc [21] enzyme protein sequences and their EC annotations as the basis. The transfer of enzyme annotations to ORF models involved two main steps: (1) Carrying out and deciphering reciprocal best-hits, if any, for each of the translated JGI ORF models to the UniProt and AraCyc sequences, then transferring the EC from the best-hits UniProt/AraCyc sequences to the corresponding ORF models. This transfer was done using BLASTP with an e-value threshold 0.001 [38,39]]; (2) Identification of paralogs, in the entire collection of translated JGI models, of already EC assigned translated ORF models and then transferring their EC annotations to their paralogs as well. This transfer was done using BLASTCLUST (http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/doc/blast/blastclust.html) with a sequence identity cut-off of 35% and length cut-off of 70%. BLASTCLUST can cluster protein sequences (using BLAST) systematically through pair wise alignments when statistically significant matches are found. Importantly, BLASTCLUST uses "single-linkage" clustering, which allows linkage of clusters through their "best matching" components. This aspect of the algorithm allows for clustering of sequences, which otherwise may lie below a set similarity threshold among themselves, but are linked through a sequence that has an above threshold similarity.

Subcellular localization predictions

WoLF PSORT [22] was used to assign subcellular localizations to each translated JGI v4.0 enzymatic ORFs. The output for each ORF provides the number of nearest neighbors in different subcellular compartments for each protein. The default value for total number of nearest neighbors (i.e., *k*) is 32. For each protein, the result can be transformed into a probability model:

$$P(c_i) = \frac{N(c_i)}{\sum_{i=1}^{m} N(c_i)}$$

where c_i is the ith subcellular compartment; $N(c_i)$ is the number of nearest neighbors the protein has for the ith subcellular compartment, and m is the total number of subcellular compartments predicted for the protein. We carried out the localization assignments of *C. reinhardtii* ORFs considering it as a plant and animal.

C. reinhardtii strain and growth condition

C. reinhardtii strain CC-503 was used for our experiments. *C. reinhardtii* cells were grown in Tris-acetate-phosphate (TAP) medium containing 100 mg l^{-1} carbamicillin without agitation, at room temperature (22–25 ° C) and under continuous illumination with cool white light at a photosynthetic photon flux of 60 μmol m^{-2} s^{-1}.

RNA isolation and quality assessment

Total RNA was isolated from *C. reinhardtii* cells grown in TAP medium and under constant light. Cells from mid-log phase were collected by centrifugation at 2,000 rpm (650g) for 10 min. Total RNA was isolated using TRIzol reagent (Invitrogen). The quality of the isolated RNA was improved by digesting the remainder of the cellular DNA using 0.08 U μl^{-1} RNase-free DNase I enzyme (Ambion). The integrity and quality of the total RNA was assessed by Agilent 2100 Bioanalyzer (Agilent) using RNA pico 6000 kit and by following the manufacturer's instruction. The fraction of RNA with RNA Integrity Number (RIN) of more than 7.5 was used for cDNA synthesis. The concentration of the RNA was measured spectrophotometrically.

Structural verification of the JGI v4.0 transcripts: Reverse transcription-PCR of the metabolic ORFs

The annotated metabolic ORFs were subjected to reverse transcription followed by PCR to verify their predicted sequences. Reverse transcription of RNA was carried out using Superscript III reverse transcriptase (Invitrogen) following the manufacturer's instructions using random N6 and dT(16) (Ambion) as universal primers. The reaction mixture contained 1.2 M betaine (Sigma-Aldrich) to prevent premature terminations owing to the high G+C content of the *C. renhardtii* transcriptome. The synthesized cDNAs were used as templates in PCR reactions. ORF-specific primers tailed with Gateway compatible sequences were designed automatically using the OSP program [40] The forward primer starts from nucleotide A of the ATG start codon and was flanked with the Gateway B1.1 sequence at its 5' end. The reverse primer starts from the codon immediately before the termination codon and carried the Gateway B2.1 sequence at its 5' end. All primers had a melting temperature (Tm) between 55 °C and 65 °C. KOD hot start DNA polymerase (Novagen) catalyzed the amplification of ~1,430 ORFs individually in separate 50 μl reaction mixtures containing 1.2 M betaine and 0.25 μg/μl cDNA.

Gateway cloning of the metabolic ORFs, their transformation and amplicon generation for sequencing

The generated amplicons were recombinationally cloned into the pDONR223 Gateway vector to generate Entry clones [41]. The recombinational cloning was performed using BP clonase (Invitrogen) following the manufacturer's instructions. The Entry clones were subsequently transformed into chemically competent *E. coli* DH5α. The positive transformants were selected and grown in 96-well format plates containing LB and 100 mg/l spectinomycin. Following growth in liquid media, the transformed bacteria were used as a source of template in PCR reactions containing 1.2 M betaine and KOD hot start DNA polymerase (Novagen) to amplify the clones. Vector primers were used to generate the final DNA templates for sequencing.

Generation of ORF sequence tags (OSTs) by Sanger sequencing

PCR products were sequenced bi-directionally using conventional automated cycle sequencing to generate ORF sequence tags (OSTs) [42]. Sequencing was carried out by Agencourt Bioscience Corp.

Forward and reverse sequences were vector-clipped (using Cross_match, http://www.phrap.org/phredphrap/general.html), then assembled. We used Phrap (http://www.phrap.org/) to assemble the forward and reverse sequences. Both assembled contigs and singlets were aligned against the coding sequences (CDSs) of corresponding predicted transcripts from *C. reinhardtii* assembly v4.0 (http://genome.jgi-psf.org/Chlre4/Chlre4.home.html) using MUSCLE [43,44]]. The alignment files were then used to verify the CDSs of the predicted transcripts. An ORF model was considered verified if a contig could be assembled from both end reads and if the contig verifies the predicted sequence.

ORF model verification by 454FLX sequencing

The generated ORF amplicons were sequenced using the 454FLX Titanium sequencing system (454 Life Sciences Corp., Roche). For targeted transcriptome sequencing, the amplicons generated in RT-PCR reactions were pooled in equimolar ratios. For verification of cloned ORFs, the PCR products of the entry clones were pooled in equimolar quantities. The resulting mixes were partially purified using Qiagen MinElute PCR purification kit following the manufacturer's instruction. Five micrograms of DNA from each sample was subjected to nebulization for 90 seconds under nitrogen gas pressure of 30 psi(2.1 bar). After purification of the sheared DNA using the MinElute PCR purification kit, the DNA fragments were size-selected using AMPure beads (Agencourt). DNA fragments with the size range of 300-800 bp were end repaired and the adaptors were ligated to the ends. After melting into single stranded DNA molecules, the quality of the DNA library was assessed on a BioAnalyzer RNA Pico 6000 LabChip (Agilent). The resulting single stranded DNA libraries were then purified and used to set up emulsion PCR reactions according to the manufacturer's instruction (454 Life Sciences Corp., Roche). After the amplification step, the emulsions were chemically broken and the beads carrying the amplified DNA library were recovered and enriched. The sequencing was performed on the Roche 454 Genome Sequencer Instrument with the GS FLX Titanium Sequencing Kit XLR70. Approximately 800,000 DNA-carrying beads along with enzyme and packing beads were loaded onto a PicoTitrePlate device. The sequencing was operated and monitored for ~9 hrs during which 200 flow cycles were completed. The generated data were processed using the GS FLX data analysis software v2.3. The vector sequences and Gateway tail sequences were trimmed from the raw reads and the reads shorter than 20 nt were filtered out. The trimmed and filtered reads were aligned against JGI v4.0 reference sequences using the GS Reference Mapper application (*gsMapper* v2.3). A minimum overlap length of 40 nt and minimum overlap identity of 90% were used to align the reads against the JGI v4.0 reference sequences. An ORF model was called verified if more than 98% of its entire length was covered by (matched to) the assembled contigs from the 454 reads.

Additional material

Additional File 1: JGIv4.0 gene model names, their predicted sequence, EC annotation, and verification status of their structural annotation.

Additional File 2: Subcellular localization prediction of JGI v4.0 enzymes predicted by WoLF PSORT as plant or animal proteins.

Additional File 3: A consolidated set of high confidence subcellular localization predictions made by WoLF PSORT. Subcellular compartments predicted for JGI v4.0 as plant or animal at 0.85 or higher ratio relative to other compartments were selected then consolidated by reporting the prediction with the higher value.

List of abbreviations used

ORF: Open Reading Frame; OST: ORF Sequence Tag; JGI: Joint Genome Institute

Acknowledgements

This research was supported by the Office of Science (Biological and Environmental Research), US Department of Energy, grant No. DE-FG02-07ER64496 (to J.P. and K.S.-A.), and additionally by New York University Abu Dhabi Research funds (to K.S.-A.), and Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative (to CCSB).
This article has been published as part of *BMC Genomics* Volume 12 Supplement 1, 2011: Validation methods for functional genome annotation. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S1.

Author details

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [3]Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA.

[4]Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908, USA. [5]New York University Abu Dhabi, Abu Dhabi, UAE, and Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA.

Authors' contributions

LG designed the cloning experiments, carried out molecular cloning, 454 sequencing, sequence analysis and drafted the manuscript. SB designed the functional annotation pipeline and carried out functional annotations of the ORFs; DB contributed to cloning; XY contributed to 454 sequencing. YS, CF, and TH carried out primer design and sequence alignments. HY carried out localization prediction of the ORFs. HY, JP, and KSA conceived the study, participated in its design and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 15 June 2011

References

1. Galperin MY, Koonin EV: **From complete genome sequence to 'complete' understanding?** *Trends Biotechnol* 2010, **28(8)**:398-406.
2. Oberhardt MA, Palsson BØ, Papin JA: **Applications of genome-scale metabolic reconstructions.** *Mol Syst Biol* 2009, **5**:320.
3. Park JH, Lee SY: **Towards systems metabolic engineering of microorganisms for amino acid production.** *Curr Opin Biotechnol* 2008, **19(5)**:454-460.
4. Schmidt BJ, Lin-Schmidt X, Chamberlin A, Salehi-Ashtiani K, Papin JA: **Metabolic systems analysis to advance algal biotechnology.** *Biotechnol J* 2010, **5(7)**:660-670.
5. Li Y, Han D, Hu G, Sommerfeld M, Hu Q: **Inhibition of starch synthesis results in overproduction of lipids in Chlamydomonas reinhardtii.** *Biotechnol Bioeng* 2010, **107(2)**:258-268.
6. Boyle NR, Morgan JA: **Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii.** *BMC Syst Biol* 2009, **3**:4.
7. Rupprecht J: **From systems biology to fuel–Chlamydomonas reinhardtii as a model for a systems biology approach to improve biohydrogen production.** *J Biotechnol* 2009, **142(1)**:10-20.
8. Kruse O, Rupprecht J, Bader KP, Thomas-Hall S, Schenk PM, Finazzi G, Hankamer B: **Improved photobiological H2 production in engineered green algal cells.** *J Biol Chem* 2005, **280(40)**:34170-34177.
9. Jans F, Mignolet E, Houyoux PA, Cardol P, Ghysels B, Cuine S, Cournac L, Peltier G, Remacle C, Franck F: **A type II NAD(P)H dehydrogenase mediates light-independent plastoquinone reduction in the chloroplast of Chlamydomonas.** *Proc Natl Acad Sci U S A* 2008, **105(51)**:20546-20551.
10. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernández E, Fukuzawa H, González-Ballester D, González-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riaño-Pachón DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martínez D, Ngau WC, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS, Grossman AR: **The Chlamydomonas genome reveals the evolution of key animal and plant functions.** *Science* 2007, **318(5848)**:245-250.
11. Mao X, Cai T, Olyarchuk JG, Wei L: **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 2005, **21(19)**:3787-3793.
12. Wortman JR, Haas BJ, Hannick LI, Smith RK Jr., Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, White OR, Town CD: **Annotation of the Arabidopsis genome.** *Plant Physiol* 2003, **132(2)**:461-468.
13. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH: **Massively parallel sequencing of the polyadenylated transcriptome of C. elegans.** *Genome Res* 2009, **19(4)**:657-666.
14. Salehi-Ashtiani K, Lin C, Hao T, Shen Y, Szeto D, Yang X, Ghamsari L, Lee H, Fan C, Murray RR, Milstein S, Svrzikapa N, Cusick ME, Roth FP, Hill DE, Vidal M: **Large-scale RACE approach for proactive experimental definition of C. elegans ORFeome.** *Genome Res* 2009, **19(12)**:2334-2342.
15. Eberhard S, Jain M, Im CS, Pollock S, Shrager J, Lin Y, Peek AS, Grossman AR: **Generation of an oligonucleotide array for analysis of gene expression in Chlamydomonas reinhardtii.** *Curr Genet* 2006, **49(2)**:106-124.
16. Nguyen AV, Thomas-Hall SR, Malnoë A, Timmins M, Mussgnug JH, Rupprecht J, Kruse O, Hankamer B, Schenk PM: **Transcriptome for photobiological hydrogen production induced by sulfur deprivation in the green alga Chlamydomonas reinhardtii.** *Eukaryot Cell* 2008, **7(11)**:1965-1979.
17. May P, Wienkoop S, Kempa S, Usadel B, Christian N, Rupprecht J, Weiss J, Recuenco-Munoz L, Ebenhöh O, Weckwerth W, Walther D: **Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii.** *Genetics* 2008, **179(1)**:157-166.
18. González-Ballester D, Casero D, Cokus S, Pellegrini M, Merchant SS, Grossman AR: **RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival.** *Plant Cell* 2010, **22(6)**:2058-2084.
19. Manichaikul A, Ghamsari L, Hom EF, Lin C, Murray RR, Chang RL, Balaji S, Hao T, Shen Y, Chavali AK, Thiele I, Yang X, Fan C, Mello E, Hill DE, Vidal M, Salehi-Ashtiani K, Papin JA: **Metabolic network analysis integrated with transcript verification for sequenced genomes.** *Nat Methods* 2009, **6(8)**:589-592.
20. Apweiler R, Bairoch A, Wu CH: **Protein sequence databases.** *Chem Biol* 2004, **8(1)**:76-80.
21. Mueller L, Zhang P, Rhee SY: **AraCyc: a biochemical pathway database for Arabidopsis.** *Plant Physiol* 2003, **132(2)**:453-460.
22. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic Acids Res* 2007, **35(Web Server issue)**:W585-587.
23. Nakai K, Horton P: **PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.** *Trends Biochem Sci* 1999, **24(1)**:34-36.
24. Nakai K, Kanehisa M: **A knowledge base for predicting protein localization sites in eukaryotic cells.** *Genomics* 1992, **14(4)**:897-911.
25. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: **Extensive feature detection of N-terminal protein sorting signals.** *Bioinformatics* 2002, **18(2)**:298-305.
26. Casadio R, Martelli PL, Pierleoni A: **The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation.** *Brief Funct Genomic Proteomic* 2008, **7(1)**:63-73.
27. Reinhardt A, Hubbard T: **Using neural networks for prediction of the subcellular location of proteins.** *Nucleic Acids Res* 1998, **26(9)**:2230-2236.
28. Blum T, Briesemeister S, Kohlbacher O: **MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction.** *BMC Bioinformatics* 2009, **10**:274.
29. Wienkoop S, Weiss J, May P, Kempa S, Irgang S, Recuenco-Munoz L, Pietzke M, Schwemmer T, Rupprecht J, Egelhofer V, Weckwerth W: **Targeted proteomics for Chlamydomonas reinhardtii combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses.** *Mol Biosyst* 2010, **6(6)**:1018-1031.
30. Voss H, Schwager C, Wiemann S, Zimmermann J, Stegemann J, Erfle H, Voie AM, Drzonek H, Ansorge W: **Efficient low redundancy large-scale DNA sequencing at EMBL.** *J Biotechnol* 1995, **41(2-3)**:121-129.
31. Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, Lin C, Szeto D, Denoeud F, Calvo M, Frankish A, Harrow J, Makrythanasis P, Vidal M, Salehi-Ashtiani K, Antonarakis SE, Gingeras TR, Guigó R: **Efficient targeted transcript discovery via array-based normalization of RACE libraries.** *Nat Methods* 2008, **5(7)**:629-635.

32. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5(7)**:621-628.
33. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proc Natl Acad Sci USA* 2003, **100(26)**:15776-15781.
34. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270(5235)**:484-487.
35. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kröger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jézéquel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV: **The Phaeodactylum genome reveals the evolutionary history of diatom genomes.** *Nature* 2008, **456(7219)**:239-234.
36. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev IV: **The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation.** *Proc Natl Acad Sci USA* 2007, **104(18)**:7705-7710.
37. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, Hellsten U, Chapman J, Simakov O, Rensing SA, Terry A, Pangilinan J, Kapitonov V, Jurka J, Salamov A, Shapiro H, Schmutz J, Grimwood J, Lindquist E, Lucas S, Grigoriev IV, Schmitt R, Kirk D, Rokhsar DS: **Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri.** *Science* 2010, **329(5988)**:223-226.
38. Madan Babu M, Balaji S, Aravind L: **General trends in the evolution of prokaryotic transcriptional regulatory networks.** *Genome Dyn* 2007, **3**:66-80.
39. Balaji S, Babu MM, Aravind L: **Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E. coli.** *J Mol Biol* 2007, **372(4)**:1108-1122.
40. Hillier L, Green P: **OSP: a computer program for choosing PCR and DNA sequencing primers.** *PCR Methods Appl* 1991, **1(2)**:124-128.
41. Walhout AJ, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S, Vidal M: **GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes.** *Methods Enzymol* 2000, **328**:575-592.
42. Reboul J, Vaglio P, Tzellas N, Thierry-Mieg N, Moore T, Jackson C, Shin-i T, Kohara Y, Thierry-Mieg D, Thierry-Mieg J, Lee H, Hitti J, Doucette-Stamm L, Hartley JL, Temple GF, Brasch MA, Vandenhaute J, Lamesch PE, Hill DE, Vidal M: **Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans.** *Nat Genet* 2001, **27(3)**:332-336.
43. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
44. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5)**:1792-1797.

doi:10.1186/1471-2164-12-S1-S4
Cite this article as: Ghamsari *et al.*: Genome-wide functional annotation and structural verification of metabolic ORFeome of *Chlamydomonas reinhardtii*. *BMC Genomics* 2011 **12**(Suppl 1):S4.

Molecular Systems Biology 7; Article number 518; doi:10.1038/msb.2011.52
Citation: ***Molecular Systems Biology*** **7:518**
© 2011 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/11
www.molecularsystemsbiology.com

Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism

Roger L Chang[1,9], Lila Ghamsari[2,3,9], Ani Manichaikul[4], Erik FY Hom[5], Santhanam Balaji[2,3], Weiqi Fu[6], Yun Shen[2,3], Tong Hao[2,3], Bernhard Ø Palsson[1], Kourosh Salehi-Ashtiani[2,3,7,8,*] and Jason A Papin[4,*]

[1] Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA, [2] Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA, [3] Department of Genetics, Harvard Medical School, Boston, MA, USA, [4] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA, [5] Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA, [6] Center for Systems Biology, University of Iceland, Reykjavik, Iceland, [7] New York University Abu Dhabi, Abu Dhabi, UAE and [8] Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA
[9] RL Chang led the network reconstruction and computational modeling; L Ghamsari led the transcript verification
* Corresponding author. K Salehi-Ashtiani, New York University Abu Dhabi, Abu Dhabi, UAE, and Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA. Tel.: +1 212 992 6964; Fax: +1 212 995 4015; E-mail: ksa3@nyu.edu or JA Papin, Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908, USA. Tel.: +1 434 924 8195; Fax: +1 434 982 3870; E-mail: papin@virginia.edu

Received 15.2.11; accepted 18.6.11

Metabolic network reconstruction encompasses existing knowledge about an organism's metabolism and genome annotation, providing a platform for omics data analysis and phenotype prediction. The model alga *Chlamydomonas reinhardtii* is employed to study diverse biological processes from photosynthesis to phototaxis. Recent heightened interest in this species results from an international movement to develop algal biofuels. Integrating biological and optical data, we reconstructed a genome-scale metabolic network for this alga and devised a novel light-modeling approach that enables quantitative growth prediction for a given light source, resolving wavelength and photon flux. We experimentally verified transcripts accounted for in the network and physiologically validated model function through simulation and generation of new experimental growth data, providing high confidence in network contents and predictive applications. The network offers insight into algal metabolism and potential for genetic engineering and efficient light source design, a pioneering resource for studying light-driven metabolism and quantitative systems biology.

Molecular Systems Biology **7**: 518; published online 2 August 2011; doi:10.1038/msb.2011.52
Subject Categories: metabolic and regulatory networks; plant biology
Keywords: *Chlamydomonas reinhardtii*; lipid metabolism; metabolic engineering; photobioreactor

Introduction

Algae have garnered significant interest in recent years for their potential commercial applications in biofuels (Hu *et al*, 2008; Hemschemeier *et al*, 2009) and nutritional supplements (Spolaore *et al*, 2006). Among eukaryotic microalgae, *Chlamydomonas reinhardtii* has arisen as the hallmark, model organism (Harris, 2001). *C. reinhardtii* has been widely used to study photosynthesis, cell motility and phototaxis, cell wall biogenesis, and other fundamental cellular processes (Harris, 2001).

Commercial use and basic scientific research of photosynthetic organisms could benefit from better understanding of how light is absorbed and affects cellular systems. The quality of light sources implemented in photobioreactors largely determines the efficiency of energy usage in industrial algal farming (Fernandes *et al*, 2010). Light spectral quality also affects how photon absorption induces various metabolic processes: photosynthesis, pigment and vitamin synthesis, and the retinol pathway required for phototaxis.

Metabolic network reconstruction provides a framework to integrate diverse experimental data for investigation of global properties of metabolism, and as such, can provide clear advantages as a mode of studying the effects of light upon a photosynthetic biological system if light input is accounted for explicitly. The standardized reconstruction process (Thiele and Palsson, 2010) yields a biochemically and genomically structured knowledgebase and, coupled with the standard simulation approach of flux balance analysis (FBA) (Orth *et al*, 2010), provides a basis for predictive phenotype modeling; both contexts have been used for a variety of applications (Durot *et al*, 2009; Oberhardt *et al*, 2009; Gianchandani *et al*, 2010), among them the design of genetic engineering strategies for production strains (Bro *et al*, 2006; Park *et al*, 2011). To date, however, photon flux, with associated spectral constraints, has not been integrated into a metabolic network reconstruction.

Characterizing algal metabolism is key to engineering production strains and framing the study of photosynthesis. Extensive literature on *C. reinhardtii* metabolism, reviewed in Stern *et al* (2008), and multiple metabolic mutants (Harris *et al*, 2008) provide a solid foundation for detailed characterization of its metabolic functions. The availability of complete genome sequence data for *C. reinhardtii* (Merchant *et al*, 2007) and its functional annotation have enabled bioinformatic approaches to inform the presence of genome-encoded enzymes (Grossman *et al*, 2007; Boyle and Morgan, 2009; Manichaikul *et al*, 2009). We have employed these resources to reconstruct and experimentally validate a genome-scale metabolic network of *C. reinhardtii*, the first network to account for detailed photon absorption permitting growth simulations under different light sources. This network accounts for the activity of substantially more genes with metabolic functions than existing reconstructions (Boyle and Morgan, 2009; Manichaikul *et al*, 2009).

Results

Reconstruction contents and advances

The genome-scale *C. reinhardtii* metabolic network (Figure 1A; Supplementary Figure S1; Supplementary Table S1; Supplementary Table S2; Supplementary Model S1) accounts for 1080 genes, associated with 2190 reactions and 1068 unique metabolites, and encompasses 83 subsystems distributed across 10 compartments. As per convention (Reed *et al*, 2003), we call this network *i*RC1080 based on the primary reconstructionist and the scope of genomic content. Of the putative protein-coding genes in the *C. reinhardtii* genome (http://augustus.gobics.de/predictions/chlamydomonas/augustus.u5.aa), an estimated 20% function in metabolism (Supplementary Table S3). *i*RC1080 accounts for the activity of >32% of the estimated genes with metabolic functions, a significant expansion over existing reconstructions (Boyle and Morgan, 2009; Manichaikul *et al*, 2009). *i*RC1080 is the most comprehensive metabolic network reconstruction of *C. reinhardtii* to date based on inclusion of pathways and a level of detail absent from previous reconstructions.

A major emergent feature of *C. reinhardtii* metabolism, apparent in Figure 1A, is the relative centrality of the chloroplast and its importance in light-driven metabolism. The chloroplast, including the thylakoid and eyespot subcompartments, accounts for >30% of the total reactions in the network and 9 of the 10 photon-utilizing reactions. The thylakoid contains essential pathways for photoautotrophic growth including photosynthesis, chlorophyll synthesis, and carotenoid synthesis, producing photoprotective pigments also valuable as fish feed additives and nutritional supplements for human consumption. The eyespot accounts for retinol metabolism, the mechanistic basis for phototaxis. Several pathways are partially duplicated across the chloroplast and other cellular compartments, in agreement with known biochemistry. A few crucial pathways are divided between the chloroplast and cytosol, including glycolysis and glycerolipid metabolism. Among the glycerolipids, triacylglycerides carrying high energy, long-chain fatty acids relevant for biofuel production accumulate substantially in microalgae. *i*RC1080 provides a thorough resource for studying these and other metabolic products and a basis for strain design for genetic engineering.

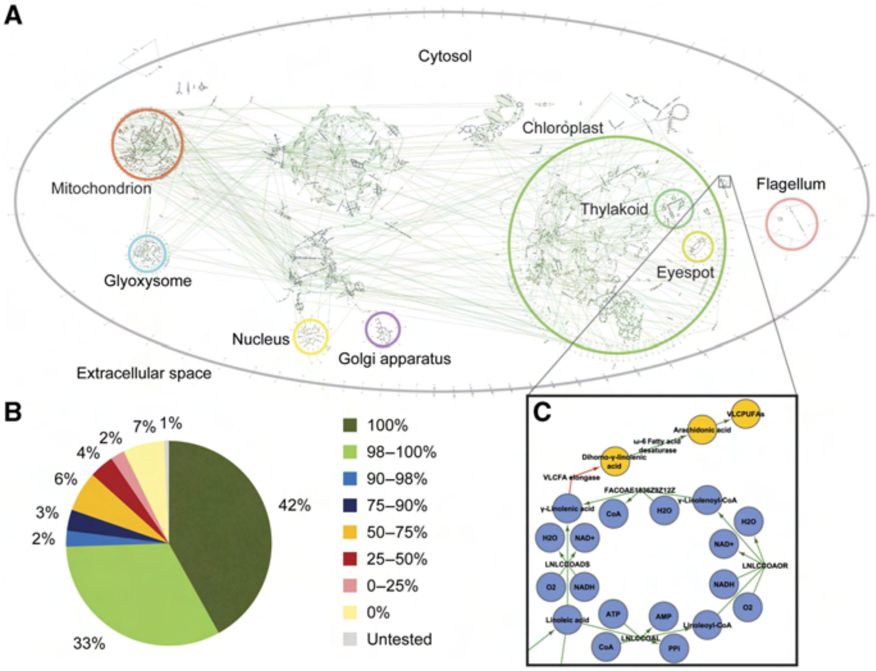


Figure 1 Contents of the *i*RC1080 metabolic network reconstruction. (**A**) Compartmentalized network diagram. The full genome-scale metabolic network is depicted, denoting compartments. A high-resolution diagram without compartment labels is also available (Supplementary Figure S1). (**B**) Global transcript verification status. The graph shows the distribution of transcripts accounted for in the network categorized by their verification status. Color codes correspond to the noted percentage of transcript sequence verified experimentally. For example, 42% of transcripts in the network were verified experimentally by 100% sequence coverage. (**C**) Latent VLCPUFA pathway diagram. Blue nodes represent metabolites included in *i*RC1080, and orange nodes represent metabolites not included in *i*RC1080, hypothesized to be absent in *C. reinhardtii*. Green edges represent enzyme activities accounted for in our functional annotation, and the red edge represents the VLCFA elongase missing from our annotation and hypothesized to have been lost in *C. reinhardtii*'s evolution. This pathway diagram also demonstrates the detail of the high-resolution network diagram (Supplementary Figure S1).

*i*RC1080 considerably expands lipid metabolic pathways over previous reconstructions. We compared the lipid pathways of *i*RC1080 with several previously published metabolic reconstructions (Duarte *et al*, 2007; Feist *et al*, 2007; Boyle and Morgan, 2009; Mo *et al*, 2009; Montagud *et al*, 2010) counting the number of genes, reactions, and chemically distinct lipid molecules included in pathways for each lipid class (Table I). The extent of gene, reaction, and metabolite content of lipid pathways in *i*R1080 is, in general, greater than previous reconstructions. The coverage of ketoacyl lipid chemical properties represented in each network was also analyzed for all metabolites in fatty acyl, glycerolipid, glycerophospholipid, and sphingolipid classes; the fraction of lipid metabolites in the networks that account for a given applicable property was determined (Table I). Lower coverage signifies incompletely specified molecular species and often lumped lipid reactions and metabolites. *i*RC1080 accounts explicitly for all metabolites in these pathways, providing sufficient detail to completely specify all individual molecular species: backbone molecule and its stereochemical numbering of acyl-chain positions; acyl-chain length; and number, position, and *cis–trans* stereoisomerism of carbon–carbon double bonds. This level of detail enables a significantly higher degree of precision in lipid studies and in metabolic engineering design involving these pathways.

Experimental transcript verification

We have analyzed *i*RC1080 via experimental transcript verification under permissive growth conditions (Supplementary Table S4), representing the largest genome-scale transcript validation effort to date. More than 75% of included

Table I Lipid pathway reconstruction properties in *i*RC1080 in comparison to other metabolic network reconstructions

	Reconstructions					
	*i*RC1080 *C. reinhardtii*	[*i*NB305] *C. reinhardtii*	*i*Syn669 *Synechocystis*	*i*MM904 *S. cerevisiae*	*i*AF1260 *E. coli*	Recon 1 *Homo sapiens*
Ketoacyl lipid chemical properties[a]						
Backbone molecule	1.00	0.94	1.00	1.00	1.00	1.00
Stereochemical numbering	1.00	0.00	0.60	0.85	1.00	0.00
Acyl-chain length	1.00	0.72	0.90	0.91	1.00	0.70
C=C number	1.00	0.72	0.75	0.91	1.00	0.70
C=C positions	1.00	0.00	0.80	0.42	0.91	0.60
E–Z stereoisomerism	1.00	0.00	0.80	0.50	0.42	0.53
Fatty acyls						
G[b]	64	7	13	32	26	91
R[c]	167	41	71	108	139	233
M[d]	104	21	55	55	95	137
Glycerolipids						
G[b]	40	0	0	18	0	27
R[c]	292	4	0	12	0	13
M[d]	135	4	2	4	7	4
Glycerophospholipids						
G[b]	47	0	8	46	22	87
R[c]	126	5	7	52	227	51
M[d]	56	4	3	4	102	22
Sphingolipids						
G[b]	8	0	0	21	0	54
R[c]	10	0	0	63	0	79
M[d]	6	0	0	31	0	59
Sterol lipids						
G[b]	22	0	1	32	0	87
R[c]	34	0	3	49	0	156
M[d]	26	0	4	22	0	105
Prenol lipids						
G[b]	37	4	15	9	16	21
R[c]	59	5	53	17	20	50
M[d]	43	4	42	15	17	41
Total lipids						
G[b]	218	11	37	158	64	367
R[c]	688	55	134	301	386	582
M[d]	370	33	106	131	221	368

[a] Values are the fraction of lipid metabolites in each network that account for each property, when applicable.
[b] Gene transcripts (can be duplicated across lipid classes).
[c] Lipid pathway reactions (non-transport).
[d] Lipid metabolites (unique lipids).

transcripts were verified at >90% sequence coverage, and 92% of tested transcripts were at least partially validated experimentally (i.e. a portion of the sequence was recovered in the sequenced transcripts) (Figure 1B). We also analyzed the strength of transcript verification by specific metabolic subsystems (Figure 2, a representative subset; Supplementary Figure S2, the full set). The full lengths of all transcripts associated with 10 subsystems were verified, notably including biosynthesis of unsaturated fatty acids, histidine metabolism, and phenylalanine, tyrosine and tryptophan biosynthesis, with 12, 12, and 24 transcripts, respectively. Many more subsystems were also well verified, 61 out of 76 gene-associated subsystems with >90% of associated transcripts at least partially validated. It should be noted that only sequencing reads that uniquely map to reference transcript sequences were counted toward the percentage of length validation; thus, sequencing reads unique enough to unambiguously specify the corresponding reference transcript were detected for every transcript with >0% validation. A few subsystems stood out as being more poorly verified, including chloroplast and mitochondrial transport systems and sphingolipid metabolism, all of which exhibited <80% of transcripts validated to any extent. This may reflect low expression level or imperfect structural annotation of these genes, particularly compartmental transporters. Low expression levels or complete deactivation of these genes is consistent with a hypothesized evolutionary trend (see below) in the case of sphingolipid metabolism.

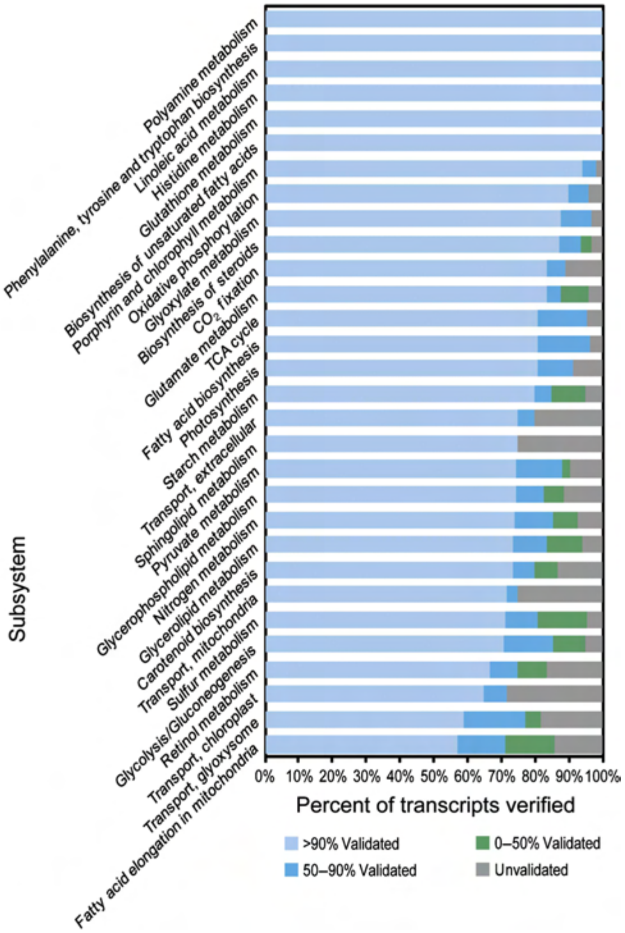


Figure 2 Experimental transcript verification by subsystem. The graph summarizes transcript verification status (see Materials and methods and Supplementary information for details) for 30 of the 76 gene-associated subsystems of *i*RC1080. Identical analysis for the full complement of 76 subsystems is also available (Supplementary Figure S2). The *x* axis corresponds to the percentage of subsystem-associated transcripts that were experimentally verified to the extent noted by the color code.

Evolution of latent lipid pathways

The comprehensive reconstruction of lipid metabolism in *i*RC1080 revealed hypothetical latent pathways, the functions of which have likely been lost through evolution. Previous studies established that *C. reinhardtii* lacks the practically ubiquitous membrane lipids phosphatidylcholine (Giroud *et al*, 1988) and phosphatidylserine (Riekhof *et al*, 2005). Similarly, our reconstruction suggests that *C. reinhardtii* also lacks very long-chain fatty acids (VLCFAs), their polyunsaturated analogs (VLCPUFAs) (Figure 1C), and ceramides.

Surveys of *C. reinhardtii* lipid species have not detected VLCFAs (Giroud *et al*, 1988; Giroud and Eichenberger, 1989; Tatsuzawa *et al*, 1996; Dubertret *et al*, 2002; Kajikawa *et al*, 2006; Lang, 2007), likely due to a lack of functional VLCFA elongase (Weers and Gulati, 1997; Guschina and Harwood, 2006; Kajikawa *et al*, 2006). No candidate VLCFA elongase was identified in our comprehensive functional annotation (Supplementary Table S3), and our annotation suggests several downstream gaps in arachidonic acid metabolism as well, corroborating this hypothesis. Arachidonic acid, the 20-carbon parent fatty acid of all VLCFAs and VLCPUFAs, is synthesized by a VLCFA elongase-catalyzed extension of γ-linolenic acid, which is present in *C. reinhardtii* (Griffiths *et al*, 2000). Notably, *C. reinhardtii* does encode a fatty acid desaturase that accepts arachidonic acid as substrate (Kajikawa *et al*, 2006) and, based on our functional annotation, encodes several other enzymes that act upon this substrate, indicating that algal ancestors likely had a functional VLCFA elongase.

Multiple lines of evidence uncovered during the reconstruction also support the absence of ceramides in *C. reinhardtii*. Our functional annotation did not uncover a convincing candidate for ceramide synthetase (EC:2.3.1.24), a required enzyme for ceramide synthesis, nor, to our knowledge, has one been discovered by previous efforts, including *C. reinhardtii* enzyme annotations of the Kyoto Encyclopedia of Genes and Genomes. Similarly, our functional annotation suggests substantial gaps downstream in the sphingolipid metabolic pathway. As aforementioned, *C. reinhardtii* also lacks VLCFAs, and VLCFA-CoA is a required substrate for the ceramide synthetase reaction (Hills and Roscoe, 2006). Finally, our experimental transcript analysis failed to verify 2 out of 8 transcripts associated with sphingolipid metabolism (Figure 2) that were included in *i*RC1080, 1 of 2 serine C-palmitoyltransferases and a putative sphingosine 1-phosphate aldolase. This result may reflect still further gene function loss in this pathway, perhaps occurring more recently in evolutionary time given that our functional annotation actually detected candidate sequences for these enzymes. Considering this evidence, we suggest that the evolutionary history of *C. reinhardtii* includes the loss of ceramide metabolism, although this hypothesis remains to be verified. Annotated enzymes in this pathway separated from the broader network by gaps may represent multifunctional proteins or proteins that have evolved to function in a pathway distinct from ceramide synthesis. These gaps in *C. reinhardtii* metabolism not only increase understanding of the evolution of algal lipid pathways but also represent potential targets for genetic engineering in an effort to expand the diversity of lipids this alga can synthesize. Such engineering efforts serve as valuable test cases for engineering industrial strains and could improve *C. reinhardtii* as a model alga for biofuel development.

Modeling metabolic light usage

Our reconstruction accounted for effective light spectral ranges by analyzing biochemical activity spectra (Figure 3A), either reaction activity or absorbance at varying light wavelengths. Defining effective spectral bandwidths associated with each photon-utilizing reaction enabled our network to model growth under different light sources via stoichiometric representation of the spectral composition of emitted light, which we term prism reactions. The coefficients for different photon wavelengths in prism reactions correspond to the ratios of photon flux in the defined effective spectral ranges to the total photon flux in the visible spectrum emitted by a given light source (Figure 3A and B). In this manner, it is possible to distinguish the amount of emitted photons that drive different metabolic reactions. We created prism reactions for 11 distinct light sources (Supplementary Figure S3), covering most sources that have been used in published studies for algal and plant growth including solar light, various light bulbs, and LEDs.

The network reconstruction provides a detailed account of metabolic photon absorption by light-driven reactions. Photosystems I and II in *i*RC1080 stoichiometrically absorb photons according to the *Z*-scheme (Berg *et al*, 2007). The light-dependent protochlorophyllide oxidoreductases require a single photon per catalysis as demonstrated in wheat (Griffiths *et al*, 1996). Extrapolation of UVB energy requirements for spontaneous provitamin D_3 conversion to vitamin D_3 (Bjorn, 2007) based on the average photon energy in the UVB range suggests a stoichiometric ratio of approximately one. Two phototactic rhodopsins, reactants of the rhodopsin photoisomerase reaction, are encoded by *C. reinhardtii*, one

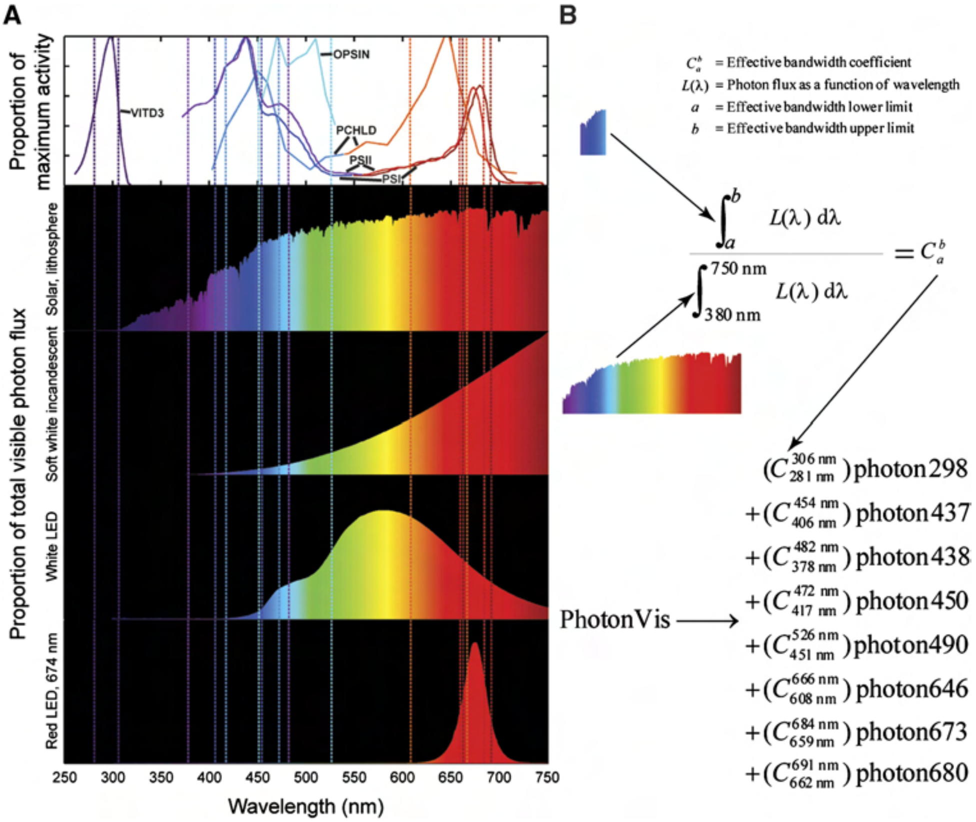


Figure 3 Analysis of light spectra. (**A**) Activity and irradiance spectra. The top graph displays activity spectra for photon-utilizing reactions included in *i*RC1080. The abbreviated reactions are defined as follows: VITD3, vitamin D_3 synthesis; OPSIN, rhodopsin photoisomerase; PCHLD, both protochlorophyllide photoreductase and divinylprotochlorophyllide photoreductase; PSI, photosystem I; PSII, photosystem II. The *y* axis for the activity spectra is the fraction of maximum-measured activity with respect to each noted reaction. Four of the eleven sample irradiance spectra (Supplementary Figure S3) are depicted with *y* axes set as the percentage of total visible photon flux at each wavelength (*x* axis). Effective spectral bandwidths are denoted by vertical dashed lines color coded to match the activity spectra for each reaction. (**B**) Prism reaction derivation. The photon flux from wavelengths *a* to *b* is normalized by the total visible photon flux from 380 to 750 nm to yield the effective spectral bandwidth coefficient *C*. The coefficients for each range are compiled into a single prism reaction for a given light source, representing the composition of emitted light as defined by photon-utilizing metabolic reactions. Equation variables are defined at top.

requiring a single photon and one requiring two photons for activation; the average effective stoichiometric photon count was measured to be 1.6 (Hegemann and Marwan, 1988).

A prism reaction is the intermediate step between light input and the specific photon-utilizing metabolic reactions mentioned above. Flux through the photon exchange reaction ‘EX_photonVis(e)’ represents the total metabolically active photon flux incident upon the cell. Flux passing through this exchange reaction then passes through a single user-specified prism reaction, for example ‘PRISM_solar_litho,’ and is distributed across specific spectral ranges. These ranges are specified explicitly in the photon-dependent metabolic reaction formulas (Supplementary Table S2), thereby making these reactions wavelength specific. Flux through the photon-dependent metabolic reactions is then propagated through the network. Excess wavelength-specific photon fluxes that are not absorbed metabolically leave the system via demand reactions, for example ‘DM_photon298(c),’ completing the pathway of light through the network.

To accurately model metabolic activity of a photosynthetic organism, it is also important to consider regulatory effects resulting from lighting conditions. Indeed, light and dark conditions have been shown to affect metabolic enzyme activity in *C. reinhardtii* at multiple levels: transcriptional regulation (Bohne and Linden, 2002), chloroplast RNA degradation (Salvador *et al*, 1993), translational regulation (Cahoon and Timko, 2000), and thioredoxin-mediated enzyme regulation (Lemaire *et al*, 2004). As a preliminary attempt to incorporate light and dark regulatory effects, literature was reviewed to identify such regulation upon enzymes in *i*RC1080 (Supplementary Table S5), focusing mainly on thioredoxin regulation of chloroplast enzymes since most published data relate to this mode. In the absence of activity spectra for these effects, it is not yet possible to represent these effects via prism reactions. Therefore, we modeled regulation with Boolean reaction flux constraints following published approaches (Covert *et al*, 2001).

Environmental and genetic validation of *i*RC1080

Implementing light-regulated constraints and basic environmental exchange constraints (Supplementary Table S6) yielded photoautotrophic, heterotrophic, and mixotrophic models from *i*RC1080. We simulated various growth conditions (Supplementary Table S7) and all gene knockouts for which phenotypes have been published and are assessable in our network (Supplementary Table S8) to validate the predictive ability of the models. All 30 validations involving environmental parameters displayed very close agreement with experimental results (Supplementary Table S7). Of particular note is the ability of our photosynthetic model in sunlight to accurately recapitulate O_2-PAR (photosynthetically active radiation) energy conversion efficiency, predicting an efficiency of 2% compared with the experimental result (Greenbaum, 1988) of 1.3–4.5%. Of the 14 gene knockouts simulated, 7 were partially or completely validated relative to experimental results (Supplementary Table S8). The unconfirmed gene knockout phenotypes may result from network errors or an incomplete set of constraints in the model (e.g. enzyme capacity, regulatory, thermodynamic, or other constraints). No internal model reactions were constrained in the models except indirectly via constraints on the input exchanges and the few explicitly noted Boolean regulatory constraints imposed (Supplementary Table S5). The unconfirmed knockout phenotypes were investigated through model analysis and literature search, although in most cases, current literature evidence could not completely explain these discrepancies, leaving them to be fully accounted for by future studies.

Two discrepancies may result from incomplete genome functional annotation or missing constraints. Knockout of mitochondrial NADH:ubiquinone oxidoreductase complex I (EC:1.6.5.3) in the model fails to recapitulate a reduced heterotrophic growth phenotype (Remacle *et al*, 2001a). The NDA2 and NDA3 genes can substitute completely for this activity in the current model. Sequence-based localization analysis places both proteins in the mitochondria, but this may be incorrect as a recent study suggests that both may be plastid localized (Desplats *et al*, 2009). Two other network reactions can also substitute for the reduction of ubiquinone, succinate dehydrogenase (ubiquinone) (EC:1.3.5.1) and electron transfer flavoprotein-ubiquinone oxidoreductase (EC:1.5.5.1). The cytochrome c oxidase complex IV (EC:1.9.3.1) knockout does not result in an obligate photoautotrophic phenotype (Remacle *et al*, 2001b) in the model because the cytochrome c peroxidase (EC:1.11.1.5) reaction is capable of compensating. The *C. reinhardtii* CCPR1 protein is homologous to mitochondrial cytochrome c peroxidases from a number of species, but no focused studies have been carried out to provide further evidence for this enzyme. In the model, the complex IV and CCPR1 double knockout is an obligate photoautotroph. These discrepancies point out important genes that should be the focus of subsequent experimentation in order to more clearly understand these metabolic phenotypes.

Another discrepancy may result from missing thermodynamic constraints. The zeaxanthin epoxidase (EC:1.14.13.90) knockout does not preclude antheraxanthin, violaxanthin, or neoxanthin production (Baroli *et al*, 2003) in the model because violaxanthin de-epoxidase (EC:1.10.99.3) reactions compensate. This substitution depends on the reversibility of these de-epoxidase reactions and may point to missing thermodynamic constraints or to undiscovered regulation under this condition.

Two discrepancies result from the lack of accounting for kinetics of the reactions of ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) from the model. Both phosphoglycolate phosphatase (EC:3.1.3.18) (Suzuki *et al*, 1990) and glycolate dehydrogenase (EC:1.1.99.14) (Nakamura *et al*, 2005) deficient mutants require high CO_2 for photoautotrophic growth *in vivo*, not recapitulated in simulations. This phenotype results from dominance of the oxygenase over carboxylase activity of RuBisCO under lower CO_2 conditions, both reactions sharing the same catalytic site. *In vivo*, these two mutants are deficient in the salvage of carbon from 2-phosphoglycolate, a product of the oxygenase activity of RuBisCO. Although these two reactions are carried out by the same enzyme in the model, their fluxes are treated as independent and not competitive; due to an absence of kinetic parameters in the model, the effect of relative CO_2 and O_2 concentrations upon RuBisCO activity cannot be explicitly expressed. Because the carboxylase activity more efficiently promotes growth, both high and low CO_2 conditions drive only this reaction and not the oxygenase reaction in the model; therefore, the salvage pathway is unnecessary in the model to achieve wild-type growth rates.

Finally, two mutant phenotype discrepancies in the model result from complex compensatory pathways that convert an input carbon source to the mutant-required carbon source. The high CO_2 requirement for photoautotrophic growth due to knockout of the chloroplast carbonic anhydrase (EC:4.2.1.1) (Spalding *et al*, 1983; Funke *et al*, 1997) can be compensated for in the model by activity of a six-reaction pathway of pyrimidine metabolism leading from bicarbonate incorporation via carbamoyl-phosphate synthase (EC:6.3.5.5) to conversion to CO_2 via orotidine-5′-phosphate decarboxylase (EC:4.1.1.23). The chloroplast ATP synthase (EC:3.6.3.14) deficient mutant (Smart and Selman, 1991; Dent *et al*, 2005; Drapier *et al*, 2007) with an acetate-requiring phenotype can be compensated for in the model by a complex pathway consisting of >15 reactions by which CO_2 is converted to acetate, which is then used in pathways similar to those supporting heterotrophic growth. Although this complex pathway has many branch points, it is notable that chloroplast malate dehydrogenase (EC:1.1.1.40) and the diffusion of pyruvate between the cytosol and chloroplast are essential to coupling the CO_2 fixation reactions to pyruvate metabolism and ultimate conversion to acetate but are not essential to the wild-type photoautotrophic or heterotrophic models. Loss of either of these conditionally essential reactions prevents the CO_2-to-acetate conversion and recapitulates the acetate-requiring phenotype. Given the complexity of these compensatory pathways, a number of possible missing constraints could explain their inactivity *in vivo* under photosynthetic conditions, and the model offers a starting point to explore possible targets of regulation under these conditions.

Gene essentiality analysis

To demonstrate the prospective use of *i*RC1080 in predicting phenotypic outcomes of genetic manipulations of *C. reinhardtii*, comprehensive essentiality analysis of all simulated single-gene knockouts was performed in models under four basic environmental conditions: growth in sunlight with and without acetate, aerobic growth in dark on acetate, and anaerobic subsistence in dark on starch. Phenotypes were defined as growth equivalent to wild-type, reduced growth relative to wild-type, or lethal based on the comparative objective fluxes of the mutant and wild-type models (Supplementary Table S9). A lethal phenotype was defined as no flux through the biomass reaction (defined as the objective function) in the mutant. Simulation results exhibited distinct metabolic system dependencies under each condition. There were 201 and 144 lethal knockouts in the model with sunlight and with and without acetate, respectively. There were 147 and only 3 lethal knockouts in the aerobic and anaerobic dark model, respectively. The metabolic processes associated with essential genes were ranked, and the three subsystems associated with the essential genes were compared under each condition. Photosynthesis, porphyrin and chlorophyll metabolism, and phenylalanine, tyrosine, and tryptophan biosynthesis were the most essential subsystems in light without acetate. Phenylalanine, tyrosine, and tryptophan biosynthesis, porphyrin and chlorophyll metabolism, and purine metabolism were the most essential subsystems in light with acetate. Expectedly, photosynthesis is most crucial for photoautotrophic growth and not required in the presence of acetate. The dark, aerobic condition had the same top ranked essential subsystems as in the mixotrophic condition, which is also expected as amino acids, chlorophyll, and nucleotides make up a high proportion of the required biomass components under both conditions. For subsistence in dark on starch, glycolysis/gluconeogenesis, starch metabolism, and starch and sucrose metabolism were the most essential subsystems, paralleling the expected core pathways for ATP maintenance with starch breakdown. While these predicted genotype–phenotype relationships demonstrate a compelling prospective use of the network, the majority of the mutant phenotypes remain to be validated experimentally; however, these predictions could be used to help define the scope and expected outcomes of such future studies.

Light-source-specific model validations

Next, we performed more extensive validations of light models grown under specific light sources at varying intensities. Varying sunlight intensity in our model and evaluating photosynthetic O_2 evolution, we observed that the model reached photosynthetic saturation at light intensity consistent with experimental measurement (Polle *et al*, 2003) (Figure 4A). Our model under red LED light (653 nm) also showed fair agreement with our experimentally measured maximum growth rate across the range of unsaturated photon flux (Figure 4B), despite divergence above the experimental saturation point. The principal explanation for this divergence lies in the relative CO_2 supplies of the experimental setup and the model. All reported photoautotrophic model simulations utilize the same maximum CO_2 exchange constraint corresponding to the maximum-measured cellular uptake rate under non-CO_2-limiting conditions (Supplementary Table S6), while the CO_2 supply in our bioreactor setup was clearly growth-limiting given that the light-saturated maximum growth rate was 0.01 gDW/h, much lower than the maximum growth rate of 0.14 gDW/h under non-CO_2-limiting conditions (Janssen *et al*, 2000). It should also be noted that the linearity of the simulation trends is a property of steady-state system modeling, which is incapable of kinetic representation of growth shifts observable in the *in vivo* experiments. For further validation, we present that the maximum biomass yield under incandescent white light is 5.7×10^{-5} gDW/mE (Janssen *et al*, 2000), in close agreement with our analogous prediction of 2.6×10^{-5} gDW/mE (Figure 4C). Similarly, our predicted biomass yield on 674 nm peak LED light of 1.1×10^{-4} gDW/mE is on the same order of magnitude as our experimental results for *C. reinhardtii* under 660 nm peak LED light near growth-saturating photon flux, 4.3×10^{-4} gDW/mE. This agreement is striking given that the network explicitly accounts for the spectral photon flux of these light sources and the subsequent processing of this energy to generate all of the constituents of biomass without any parameter fitting to the experimental data. Together, these results constitute an

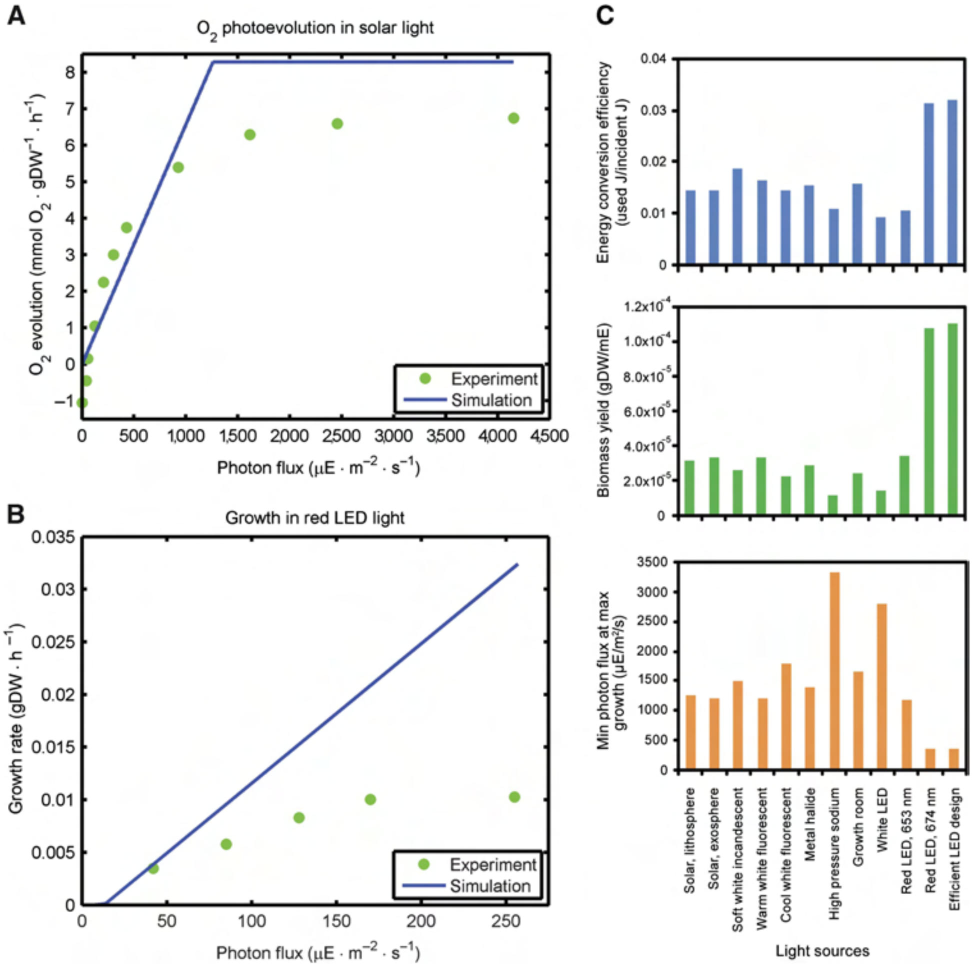


Figure 4 Photosynthetic model simulation results. (**A**) O_2 photoevolution under solar light. Simulated (blue line) and experimentally measured (green dots) O_2 evolution are compared. (**B**) Photosynthetic growth under red LED light. Simulations were performed using the 653-nm prism reaction, and experimentally grown culture was exposed to 660 nm LED light. Simulated (blue line) and experimentally measured (green dots) growth are compared. (**C**) Efficiency of light utilization. The minimum photon flux required for maximum-simulated growth (bottom), biomass yield (middle), and energy conversion efficiency (top) are presented for 11 light sources derived from measured spectra and for the designed growth-efficient LED.

important validation of our models using three different light sources.

To quantitatively evaluate the significance of the agreement between our reported model simulations using prism reactions derived through analysis of irradiance spectra and experimental measurements under the three light sources reported above, we compared the reported simulation results for each of these light sources with an unbiased sample of results representative of potential solutions achievable using our network. We sampled the space of possible light models by generating random prism reactions with the same total metabolically active photon flux. To obtain stoichiometric coefficients for a random prism reaction, a set of random fractions of the sum of stoichiometric coefficients of the prism reaction representing the evaluated light source was generated, contingent upon resulting in the same sum of coefficients. The simulations as reported above for sunlight, red LED, and white incandescent light were repeated using such random prism reactions. The Euclidean distance between the simulated and experimental results was compared with the distribution of distances for 10 000 randomly sampled results (Figure 5). The probability of randomly achieving experimental agreement closer than seen in our simulations was determined empirically based on these distributions of distances. Only 77 of 10 000 randomized simulations (Figure 5A) had experimental agreement better than the simulated oxygen photoevolution under sunlight (Figure 4A), yielding an empirical *P*-value of 0.0077, and indicating our model had experimental agreement statistically significantly better than a random model constrained to have the same total metabolically active photon flux. Simulated growth under 665 nm peak LED (Figure 4B) had a suggestive *P*-value of 0.1947 (Figure 5B), although the reported simulation was still closer to experiment than the mean of randomized simulations. Our simulated growth under white incandescent light was statistically significantly closer to experiment (Janssen *et al*, 2000) than random (Figure 5C) with a *P*-value of 0.0285. This analysis shows that the reported model for each of these

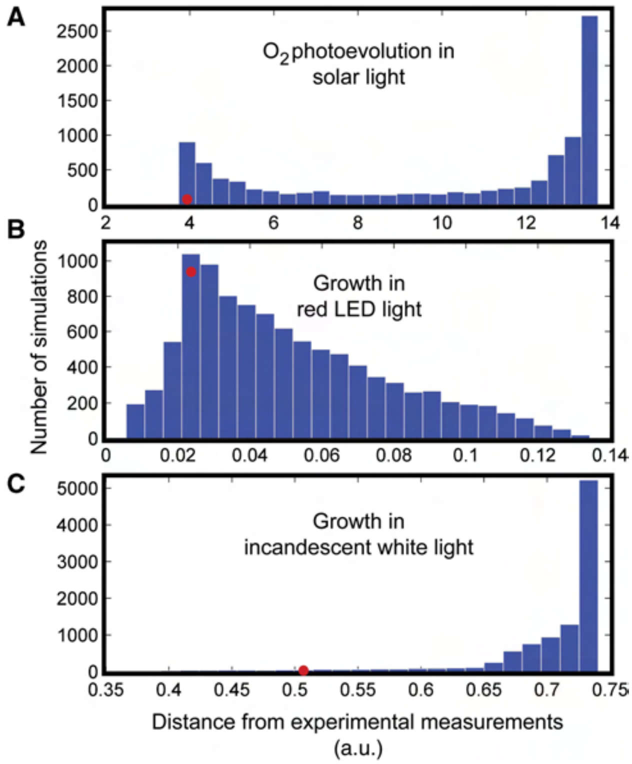


Figure 5 Distributions of randomly sampled distances from experimental measurements. (**A**) O_2 photoevolution under solar light. (**B**) Photosynthetic growth under red LED light. (**C**) Photosynthetic growth under white incandescent light. All three distance distributions result from 10 000 unbiased sampling results in which random prism reactions were generated with the same total metabolically active photon flux as the given light source. Each distribution is depicted in 25 equal-sized bins. The red dot in each plot is placed over the bin in which the distance of the reported simulation result for the given light source falls; the vertical placement of each red dot indicates the number of randomly sampled distances within the same bin that are less than that of the reported result.

light sources is exceptionally close to recapitulating experimental results and thus serves as an excellent validation. These results indicate that the network has the capacity to broadly differentiate light-dependent growth based on spectral properties and that the formulation of a prism reaction serves to accurately narrow the space of possible flux distributions relevant to a specific light source.

Application of *i*RC1080 to evaluate light source efficiency and design

Our photosynthetic model was applied prospectively to evaluate the efficiency of light utilization under different light sources. The photon energy conversion efficiency (Supplementary Equation 1) and biomass yield on light (Supplementary Equation 2) were computed for each light source given the minimum incident photon flux required to achieve maximum growth rate (Figure 4C); the minimum photon flux for maximum growth rate is the growth-saturating photon flux value for a given light source. One clear result is that red LEDs provide the greatest efficiency in terms of both absorbed photon energy and biomass yield, about two and three times as efficient as can be optimally achieved in sunlight by these respective measures. Although experimental growth data for validation is only presented for three light sources, simulation results are presented for all 11 light sources for which irradiance spectra were obtainable (Figure 4C). This analysis demonstrates the prospective extensibility of the network and modeling approach to any possible lighting condition, natural or manmade, for which an irradiance spectrum can be measured.

Given the capability of our photosynthetic model to evaluate light source efficiency, we applied it to design an LED spectrum providing maximum photon utilization efficiency for growth (Supplementary Figure S3). The result was a 677-nm peak LED spectrum with a total incident photon flux of 360 $\mu E/m^2/s$ (Figure 4C; Supplementary Figure S3), which is quite close to the 674-nm LED with a minimum incident photon flux of 362 $\mu E/m^2/s$ for maximum growth. This result suggests that for the simple objective of maximizing growth efficiency, LED technology has already reached an effective theoretical optimum, which is further supported by experimental measurements of the spectral peak of light absorption for green algae (Akkerman *et al*, 2002) and the quantum action spectrum of land plants (Barta *et al*, 1992) (Supplementary Table S7).

Discussion

We have presented a genome-scale network reconstruction of *C. reinhardtii* metabolism, well validated in content and function, and its application for detailed modeling of diverse light sources. Initial model validations also highlight the need for more experimental studies to uncover regulatory mechanisms in order to expand understanding of the complexity of light regulation of algal metabolism. This open research topic presents important challenges and opportunities in enumerating such effects on a genome scale.

Given the importance of lipid metabolism in biofuel production, *i*RC1080 was reconstructed enumerating all lipids supported by evidence in the literature and genome functional annotation. The capacity of *i*RC1080 as a knowledgebase was demonstrated through analysis of lipid metabolism to generate novel hypotheses about latent metabolic pathways resulting from algal evolution. In particular, the exclusion of certain enzymatic reactions in VLCFA and sphingolipid pathways from *i*RC1080 suggests evolutionary recession of these pathways in *C. reinhardtii*, a hypothesis supported by undetected lipids in experimental measurements, gaps in genome functional annotation for these enzymes, and incomplete transcript verification for other enzymes included in these pathways. Not only do these network gaps reflect the relatively simple lipid biosynthetic capabilities of *C. reinhardtii* among microalgae, but their identification suggests gene insertions that could expand its lipid metabolic repertoire, relevant for industrial and scientific purposes. Of particular interest may be the potential for enabling algal synthesis of essential fatty acids for human health such as docosahexaenoic acid (Yashodhara *et al*, 2009). Candidate enzymes for the conversion of arachidonic acid to essential fatty acids downstream of the apparently absent VLCFA elongase reaction are present in our functional annotation.

The models developed from *i*RC1080 provide a platform for prediction of phenotypic outcomes of system perturbations, light source evaluation and design, and genetic engineering design for production of biofuels and other commodity chemicals. We demonstrated an approach applying *i*RC1080 to the design of an energetically efficient light source for growth, a novel application of metabolic networks. Other light sources may be more efficient for other metabolic objectives or under other environmental conditions or genetic backgrounds. This result could be of significant interest to the metabolic engineering and bioreactor-design communities because it demonstrates that our network and light-modeling approach are capable of accurately predicting light source efficiencies in terms of a metabolic objective.

The prism reactions developed and applied in this study to quantitatively integrate spectral quality with biological activity represent a significant integration of diverse data types for biological system modeling, which hopefully will encourage a new paradigm for systems biology. This modeling approach could be used for applications beyond light source design, including as a metabolic basis for studying and simulating phototaxis. Given the acquisition of appropriate biological spectral activity data, this approach could be extended to other biological light-response phenomena and other organisms. The importance of understanding how light parameters affect biological systems may also extend beyond natural phenomena with recent progress in protein engineering leading to chimeric light-inducible proteins (Shimizu-Sato *et al*, 2002; Levskaya *et al*, 2005).

The *i*RC1080 network and presented metabolic modeling represent a milestone in systems biology. Our network provides a broad knowledgebase of the biochemistry and genomics underlying global metabolism of a photoautotroph, and our modeling of light-driven metabolism exemplifies how integration of largely unvisited data types, such as physicochemical environmental parameters, can expand the diversity of applications of metabolic networks.

Materials and methods

Metabolic network reconstruction

Building from our previously published reconstruction of *C. reinhardtii* central metabolism (Manichaikul *et al*, 2009), *i*AM303, the *i*RC1080 network was reconstructed in a bottom–up manner according to current standards (Thiele and Palsson, 2010) on a pathway-by-pathway basis, drawing biochemical, genomic, and physiological evidence from >250 publications (Supplementary Table S2). The genomic evidence was derived from our own functional annotation (Supplementary Table S3) of metabolic enzymes, coenzymes, and transport proteins. Network gap-filling was performed to make pathways functional and account for dead-end metabolites. Global quality control checks were then performed, including elemental balancing and elimination of as many internal thermodynamically infeasible loops and new photon-driven, input-only pathways as possible (Supplementary Figure S4; Supplementary information). We also accounted for subcellular compartment pH in the protonation states of metabolites as much as possible.

*i*RC1080 is publicly available at http://www.ebi.ac.uk/biomodels (Accession: MODEL1106200000) and as Supplementary Model S1.

Functional annotation of transcripts

Functional annotation for *i*RC1080 was performed using a consensus of two separate approaches. In the first approach, gene models (http://augustus.gobics.de/predictions/chlamydomonas/augustus.u5.aa) from the Augustus update 5 (Au5) of *C. reinhardtii* genome assembly version JGI v4.0 were functionally annotated by assigning enzyme classification (EC) terms using BLASTP results against UniProt (http://www.uniprot.org/) and AraCyc (http://www.arabidopsis.org/biocyc/) enzyme protein sequences and their EC annotations as the basis. The second approach followed from mapping of Au5 gene models to annotated JGI v3.1 gene models, for which EC terms and Gene Ontology annotation were assigned using a combination of BLASTP, AutoFACT, InterProScan, and PRIAM. The comprehensive annotation is presented in Supplementary Table S3. See Supplementary information for full details.

Growth simulations

Simulation procedures consisted of FBA (Orth *et al*, 2010) and flux variability analysis (FVA) (Mahadevan and Schilling, 2003) as implemented in the COBRA toolbox (Becker *et al*, 2007), testing model capabilities while optimizing biomass functions to simulate growth (Supplementary Table S10) or subsistence on starch by optimizing ATP maintenance. FBA is a widely used simulation approach for large-scale, constraint-based metabolic models and has become a standard method in the systems biology field with a long history of success (Gianchandani *et al*, 2010). Different environmental conditions were modeled by appropriately setting reaction flux constraints in *i*RC1080 (Supplementary Table S6) including environmental exchanges, non-growth associated ATP maintenance, O_2 photoevolution, starch degradation, and light- or dark-regulated enzymatic reactions (Supplementary Table S5).

C. reinhardtii strains and growth conditions

For transcript verification experiments, *C. reinhardtii* strain CC-503 was grown in tris-acetate-phosphate medium containing 100 mg/l carbamicillin without agitation, at room temperature (22–25°C) and under continuous illumination with cool white light at a photosynthetic photon flux of 60 $\mu E/m^2/s$.

For growth experiments under 660 nm peak LED light (Supplementary Figure S5), *C. reinhardtii* strain UTEX2243 was grown in a bubble column photobioreactor at 23–27°C with P49 medium. The total volume of algal culture was 300 ml, and the gas supply was 180 ml/min air with 2.5% CO_2. The 660-nm peak LED light supply was set at 10 kHz frequency and different duty cycles to get varied average photon fluxes.

Transcript verification by sequencing

ORF amplicons were generated from *C. reinhardtii* cells by RT–PCR from RNA or PCR from Gateway clones. The Roche 454FLX Titanium sequencing system was used for sequencing of the generated ORF amplicons according to the manufacturer's instructions. The generated data were processed using the GS FLX data analysis software v2.3. Minimum overlap length of 40 nucleotides and minimum overlap identity of 90% were used to align the sequencing reads against the Au5 reference sequences. ORFs encoding transporter proteins were verified by capillary Sanger sequencing.

Prism reaction derivation

Spectral bandwidths that effectively drive each photon-utilizing reaction in *i*RC1080 were determined from published experimental activity spectral data or absorbance data. Effective spectral bandwidths were defined as the full width half maximum of activity, denoted by color-paired dashed lines in Figure 3A. The effective spectral bandwidths were used to derive stoichiometric coefficients of the prism reactions used to quantitatively represent different light sources from the composition of their published irradiance spectra, converted to photon flux units according to Supplementary Equations

3 and 4. Coefficients for each of the effective spectral bandwidths were computed based on Equation 1.

$$C_a^b = \frac{\int_a^b L(\lambda)\,\mathrm{d}\lambda}{\int_{380\,\mathrm{nm}}^{750\,\mathrm{nm}} L(\lambda)\,\mathrm{d}\lambda} \tag{1}$$

$$\begin{aligned} C_a^b &= \text{effective bandwidth coefficient} \\ L(\lambda) &= \text{photon flux as a function of wavelength} \\ a &= \text{effective bandwidth lower limit} \\ b &= \text{effective bandwidth upper limit} \end{aligned}$$

Each coefficient represents the ratio of photon flux in the defined effective bandwidth to total visible photon flux. Definite integrals in Equation 1 were approximated using the trapezoidal rule. For each light source, all effective bandwidth coefficients were compiled into a single reaction in the form of Equation 2.

$$\mathrm{photonVis} \longrightarrow \begin{aligned} &(C_{281\mathrm{nm}}^{306\mathrm{nm}})\mathrm{photon298} + (C_{406\mathrm{nm}}^{454\mathrm{nm}})\mathrm{photon437} \\ &+ (C_{378\mathrm{nm}}^{482\mathrm{nm}})\mathrm{photon438} + (C_{417\mathrm{nm}}^{472\mathrm{nm}})\mathrm{photon450} \\ &+ (C_{451\mathrm{nm}}^{526\mathrm{nm}})\mathrm{photon490} + (C_{608\mathrm{nm}}^{666\mathrm{nm}})\mathrm{photon646} \\ &+ (C_{661\mathrm{nm}}^{685\mathrm{nm}})\mathrm{photon673} + (C_{662\mathrm{nm}}^{691\mathrm{nm}})\mathrm{photon680} \end{aligned} \tag{2}$$

Constraints on prism reaction fluxes (Supplementary Table S6) were derived from the total visible photon flux, the definite integral of the spectrum from 380 to 750 nm. The total experimentally measured emitted visible photon flux was converted to model units of incident photon flux using the values in Supplementary Table S11 and Supplementary Equations 5 and 6. Prism reactions for 11 different light sources (Supplementary Figure S3) were generated.

Random sampling of prism reaction space and significance test

For a given prism reaction, first the sum of the stoichiometric coefficients was calculated, representing the total quantity of metabolically active photons per incident photon from the specified light source. Next, to sample the space of prism reactions, 10 000 random prism reactions with the same sum of stoichiometric coefficients were generated and used in growth simulations. In these simulations, input photon flux was constrained to the reported experimental values, generating a set of simulated results (biomass or photosynthetically evolved O_2 flux, depending on the experimental parameter) with one value corresponding to each experimental data point. The Euclidean distance between the sampled and experimental results was calculated for each of the 10 000 randomized prism reactions (Figure 5). The significance of the experimental agreement with simulations reported for a given prism reaction derived directly from analysis of irradiance spectra was established by comparison between the corresponding Euclidean distance and the distribution of distances from the randomly sampled prism reactions. Probability of achieving equal or closer results to experiments by chance was computed as the proportion of smaller values in the randomly sampled distribution of 10 000 distances.

Procedure for efficient LED design

Multiple iterations of FVA were used to maximize growth while minimizing the energy of the sum of individual wavelengths of model photon flux. The ratios of these individual wavelength photon fluxes to total photon flux were set as stoichiometric coefficients for a theoretical maximum-efficiency prism reaction. The Euclidean vector distance was computed (Supplementary Figure S6) between this set of coefficients and prism reaction coefficients calculated for an LED spectrum of the same shape as the experimentally measured 674 nm peak LED but centered at varying wavelengths across the visible spectrum, with a total photon flux equal to the total theoretical maximum-efficiency photon flux. The spectrum corresponding to the minimum distance was taken as the solution and subsequently tested through growth simulation.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Harish Nagarajan for his critical assessment of this work. This research was supported by the Office of Science (Biological and Environmental Research), US Department of Energy, Grant DE-FG02-07ER64496 (to JP and KS-A), New York University Abu Dhabi Research funds (to KS-A), and Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative (to CCSB). RLC was supported by the National Science Foundation IGERT training Grant DGE0504645, and EFYH was supported by the Jane Coffin Childs Memorial Fund for Medical Research.

Author contributions: RLC wrote the manuscript and developed the light-modeling approach. RLC and AM reconstructed the metabolic network and performed model simulations and analysis. LG, YS and TH performed transcript verification experiments and associated computational analyses. EFYH and SB performed genome functional annotation. WF performed bioreactor growth experiments. BØP, JAP, and KS-A designed the study.

References

Akkerman I, Janssen M, Rocha J, Wijffels RH (2002) Photobiological hydrogen production: photochemical efficiency and bioreactor design. *Int J Hydrogen Energy* **27:** 1195–1208

Baroli I, Do AD, Yamane T, Niyogi KK (2003) Zeaxanthin accumulation in the absence of a functional xanthophyll cycle protects *Chlamydomonas reinhardtii* from photooxidative stress. *Plant Cell* **15:** 992–1008

Barta DJ, Tibbitts TW, Bula RJ, Morrow RC (1992) Evaluation of light emitting diode characteristics for a space-based plant irradiation source. *Adv Space Res* **12:** 141–149

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* **2:** 727–738

Berg J, Tymoczko J, Stryer L (2007) *Biochemistry*. New York, USA: W.H. Freeman

Bjorn L (2007) *Photobiology: The Science of Life and Light*. Dordrecht, Netherlands: Springer

Bohne F, Linden H (2002) Regulation of carotenoid biosynthesis genes in response to light in *Chlamydomonas reinhardtii*. *Biochim Biophys Acta* **1579:** 26–34

Boyle NR, Morgan JA (2009) Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Syst Biol* **3:** 4

Bro C, Regenberg B, Forster J, Nielsen J (2006) In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. *Metab Eng* **8:** 102–111

Cahoon AB, Timko MP (2000) yellow-in-the-dark mutants of *Chlamydomonas* lack the CHLL subunit of light-independent protochlorophyllide reductase. *Plant Cell* **12:** 559–568

Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213:** 73–88

Dent RM, Haglund CM, Chin BL, Kobayashi MC, Niyogi KK (2005) Functional genomics of eukaryotic photosynthesis using insertional mutagenesis of *Chlamydomonas reinhardtii*. *Plant Physiol* **137:** 545–556

Desplats C, Mus F, Cuine S, Billon E, Cournac L, Peltier G (2009) Characterization of Nda2, a plastoquinone-reducing type II NAD(P)H dehydrogenase in *chlamydomonas* chloroplasts. *J Biol Chem* **284:** 4148–4157

Drapier D, Rimbault B, Vallon O, Wollman FA, Choquet Y (2007) Intertwined translational regulations set uneven stoichiometry of chloroplast ATP synthase subunits. *EMBO J* **26:** 3581–3591

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* **104:** 1777–1782

Dubertret G, Gerard-Hirne C, Trémolières A (2002) Importance of trans-Δ3-hexadecenoic acid containing phosphatidylglycerol in the formation of the trimeric light-harvesting complex in *Chlamydomonas*. *Plant Physiol Biochem* **40:** 829–836

Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* **33:** 164–190

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:** 121

Fernandes BD, Dragone GM, Teixeira JA, Vicente AA (2010) Light regime characterization in an airlift photobioreactor for production of microalgae with high starch content. *Appl Biochem Biotechnol* **161:** 218–226

Funke RP, Kovar JL, Weeks DP (1997) Intracellular carbonic anhydrase is essential to photosynthesis in *Chlamydomonas reinhardtii* at atmospheric levels of CO_2. Demonstration via genomic complementation of the high-CO2-requiring mutant ca-1. *Plant Physiol* **114:** 237–244

Gianchandani EP, Chavali AK, Papin JA (2010) The application of flux balance analysis in systems biology. *Wiley Interdiscip Rev Syst Biol Med* **2:** 372–382

Giroud C, Eichenberger W (1989) Lipids of *Chlamydomonas reinhardtii*. Incorporation of [14C]acetate, [14C]palmitate and [14C]oleate into different lipids and evidence for lipid-linked desaturation of fatty acids. *Plant Cell Physiol* **30:** 121–128

Giroud C, Gerber A, Eichenberger W (1988) Lipids of *Chlamydomonas reinhardtii*. Analysis of molecular species and intracellular site(s) of biosynthesis. *Plant Cell Physiol* **29:** 587–595

Greenbaum E (1988) Energetic efficiency of hydrogen photoevolution by algal water splitting. *Biophys J* **54:** 365–368

Griffiths G, Leverentz M, Silkowski H, Gill N, Sanchez-Serrano JJ (2000) Lipid hydroperoxide levels in plant tissues. *J Exp Bot* **51:** 1363–1370

Griffiths WT, McHugh T, Blankenship RE (1996) The light intensity dependence of protochlorophyllide photoconversion and its significance to the catalytic mechanism of protochlorophyllide reductase. *FEBS Lett* **398:** 235–238

Grossman AR, Croft M, Gladyshev VN, Merchant SS, Posewitz MC, Prochnik S, Spalding MH (2007) Novel metabolism in *Chlamydomonas* through the lens of genomics. *Curr Opin Plant Biol* **10:** 190–198

Guschina IA, Harwood JL (2006) Lipids and lipid metabolism in eukaryotic algae. *Prog Lipid Res* **45:** 160–186

Harris E, Stern D, Witman G (2008) *The Chlamydomonas Sourcebook: Introduction to Chlamydomonas and Its Laboratory Use*. Amsterdam, Netherlands: Academic Press

Harris EH (2001) *Chlamydomonas* as a model organism. *Annu Rev Plant Physiol Plant Mol Biol* **52:** 363–406

Hegemann P, Marwan W (1988) Single photons are sufficient to trigger movement responses in *Chlamydomonas reinhardtii*. *Photochem Photobiol* **48:** 99–106

Hemschemeier A, Melis A, Happe T (2009) Analytical approaches to photobiological hydrogen production in unicellular green algae. *Photosynth Res* **102:** 523–540

Hills MJ, Roscoe TJ (2006) *Synthesis of Structural and Storage Lipids by the ER*. Berlin/Heidelberg: Springer

Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. *Plant J* **54:** 621–639

Janssen M, de Winter M, Tramper J, Mur LR, Snel J, Wijffels RH (2000) Efficiency of light utilization of *Chlamydomonas reinhardtii* under medium-duration light/dark cycles. *J Biotechnol* **78:** 123–137

Kajikawa M, Yamato KT, Kohzu Y, Shoji S, Matsui K, Tanaka Y, Sakai Y, Fukuzawa H (2006) A front-end desaturase from *Chlamydomonas reinhardtii* produces pinolenic and coniferonic acids by omega13 desaturation in methylotrophic yeast and tobacco. *Plant Cell Physiol* **47:** 64–73

Lang ID (2007) *New Fatty Acids, Oxylipins and Volatiles in Microalgae*, PhD Thesis, Göttingen: Mathematisch-naturwissenschaftliche Fakultäten, Georg-August-Universität Göttingen

Lemaire SD, Guillon B, Le Marechal P, Keryer E, Miginiac-Maslow M, Decottignies P (2004) New thioredoxin targets in the unicellular photosynthetic eukaryote *Chlamydomonas reinhardtii*. *Proc Natl Acad Sci USA* **101:** 7475–7480

Levskaya A, Chevalier AA, Tabor JJ, Simpson ZB, Lavery LA, Levy M, Davidson EA, Scouras A, Ellington AD, Marcotte EM, Voigt CA (2005) Synthetic biology: engineering *Escherichia coli* to see light. *Nature* **438:** 441–442

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5:** 264–276

Manichaikul A, Ghamsari L, Hom EF, Lin C, Murray RR, Chang RL, Balaji S, Hao T, Shen Y, Chavali AK, Thiele I, Yang X, Fan C, Mello E, Hill DE, Vidal M, Salehi-Ashtiani K, Papin JA (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat Methods* **6:** 589–592

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H *et al* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318:** 245–250

Mo ML, Palsson BO, Herrgard MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* **3:** 37

Montagud A, Navarro E, Fernandez de Cordoba P, Urchueguia JF, Patil KR (2010) Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst Biol* **4:** 156

Nakamura N, Tanaka S, Teko Y, Mitsui K, Kanazawa H (2005) Four Na + /H + exchanger isoforms are distributed to Golgi and post-Golgi compartments and are involved in organelle pH regulation. *J Biol Chem* **280:** 1561–1572

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* **5:** 320

Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* **28:** 245–248

Park JH, Kim TY, Lee KH, Lee SY (2011) Fed-batch culture of Escherichia coli for L-valine production based on in silico flux response analysis. *Biotechnol Bioeng* **108:** 934–946

Polle JE, Kanakagiri SD, Melis A (2003) tla1, a DNA insertional transformant of the green alga *Chlamydomonas reinhardtii* with a truncated light-harvesting chlorophyll antenna size. *Planta* **217:** 49–59

Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4:** R54

Remacle C, Baurain D, Cardol P, Matagne RF (2001a) Mutants of *Chlamydomonas reinhardtii* deficient in mitochondrial complex I: characterization of two mutations affecting the nd1 coding sequence. *Genetics* **158:** 1051–1060

Remacle C, Duby F, Cardol P, Matagne RF (2001b) Mutations inactivating mitochondrial genes in *Chlamydomonas reinhardtii*. *Biochem Soc Trans* **29:** 442–446

Riekhof WR, Sears BB, Benning C (2005) Annotation of genes involved in glycerolipid biosynthesis in *Chlamydomonas reinhardtii*: discovery of the betaine lipid synthase BTA1Cr. *Eukaryot Cell* **4:** 242–252

Salvador ML, Klein U, Bogorad L (1993) Light-regulated and endogenous fluctuations of chloroplast transcript levels in *Chlamydomonas*. Regulation by transcription and RNA degradation. *Plant J* **3:** 213–219

Shimizu-Sato S, Huq E, Tepperman JM, Quail PH (2002) A light-switchable gene promoter system. *Nat Biotechnol* **20:** 1041–1044

Smart EJ, Selman BR (1991) Isolation and characterization of a *Chlamydomonas reinhardtii* mutant lacking the gamma-subunit of chloroplast coupling factor 1 (CF1). *Mol Cell Biol* **11:** 5053–5058

Spalding MH, Spreitzer RJ, Ogren WL (1983) Carbonic anhydrase-deficient mutant of *Chlamydomonas reinhardii* requires elevated carbon dioxide concentration for photoautotrophic growth. *Plant Physiol* **73:** 268–272

Spolaore P, Joannis-Cassan C, Duran E, Isambert A (2006) Commercial applications of microalgae. *J Biosci Bioeng* **101:** 87–96

Stern D, Harris E, Witman G (2008) *The Chlamydomonas Sourcebook: Organellar and Metabolic Processes*. Amsterdam, Netherlands: Academic Press

Suzuki K, Marek LF, Spalding MH (1990) A photorespiratory mutant of *Chlamydomonas reinhardtii*. *Plant Physiol* **93:** 231–237

Tatsuzawa H, Takizawa E, Wada M, Yamamoto Y (1996) Fatty acid and lipid composition of the acidophilic green alga *Chlamydomonas* sp. *J Phycol* **32:** 598–601

Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5:** 93–121

Weers PMM, Gulati RD (1997) Growth and reproduction of *Daphnia galeata* in response to changes in fatty acids, phosphorus, and nitrogen in *Chlamydomonas reinhardtii*. *Limnol Oceanogr* **42:** 1584–1589

Yashodhara BM, Umakanth S, Pappachan JM, Bhat SK, Kamath R, Choo BH (2009) Omega-3 fatty acids: a comprehensive review of their role in health and disease. *Postgrad Med J* **85:** 84–90