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

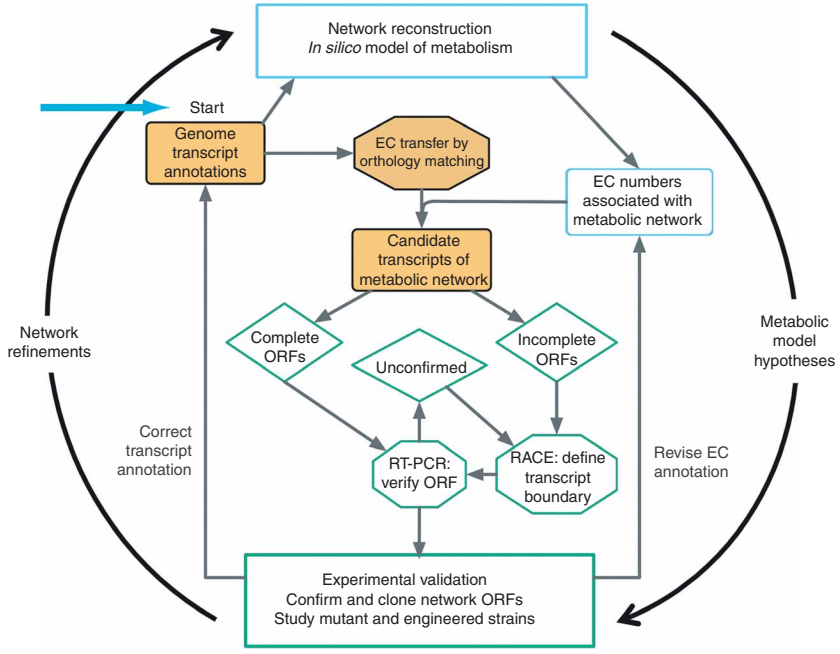

Figure 1 | Assessing and improving gene annotation for *C. reinhardtii*: iterative process integrating gene annotation experiments with metabolic network reconstruction and analysis. Starting with a draft network reconstruction, EC terms associated with model reactions are mapped to corresponding transcripts. 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

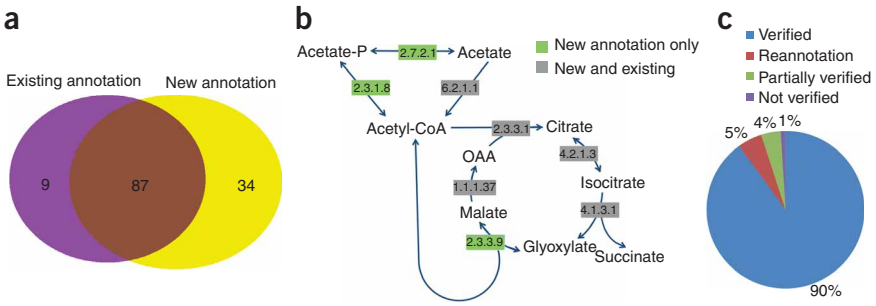

Figure 2 | Integrating the network model with transcript verification experiments. (**a**) Comparison of central metabolic EC terms annotated in existing JGI v3.0 and our annotation of JGI v3.1 (**Supplementary Note**). (**b**) Applying these two versions of EC annotation to inform the network reconstruction 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

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 'touchdown' 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