nature | **methods**

Metabolic network analysis integrated with transcript verification for sequenced genomes

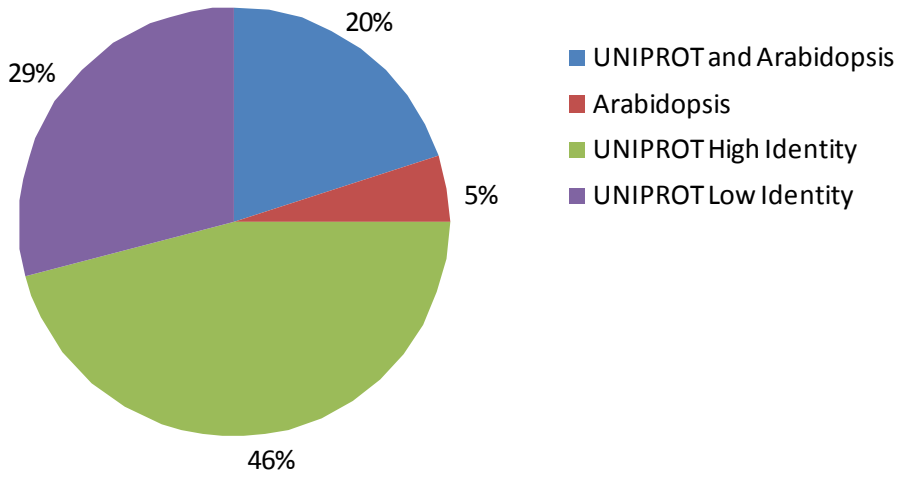
Ani Manichaikul, Lila Ghamsari, Erik F Y Hom, Chenwei Lin, Ryan R Murray, Roger L Chang, S Balaji, Tong Hao, Yun Shen, Arvind K Chavali, Ines Thiele, Xinping Yang, Changyu Fan, Elizabeth Mello, David E Hill, Marc Vidal, Kourosh Salehi-Ashtiani & Jason A Papin

Supplementary figures and text:

Supplementary Figure 1	Distribution of evidence (UNIPROT high confidence, UNIPROT low confidence, and *Arabidopsis*) for *Chlamydomonas* JGI v3.1 EC annotation.
Supplementary Figure 2	Distribution of the JGI v3.1 gene models having evidence from each annotation category (UNIPROT high confidence, UNIPROT low confidence, and *Arabidopsis* proteome).
Supplementary Figure 3	Comparison of simulated photosynthetic evolution (E), uptake (U) and net exchange of oxygen with experimental data, across a range of photon flux values.
Supplementary Figure 4	A well validated network model towards metabolic engineering.
Supplementary Figure 5	High resolution detailed map of the final metabolic network reconstruction.
Supplementary Table 2	Comparison of presence/absence of triacylglycerol synthesis pathway EC terms in existing JGI v3.0 annotation and our new JGI v3.1 annotation.
Supplementary Table 3	List of unique EC terms used to choose primers.
Supplementary Table 6	Summaries of *in silico* versus literature validation of physiological parameters.
Supplementary Table 7	Summaries of *in silico* knockout validations.
Supplementary Table 8	Summary of the central metabolic network reconstruction of *C. reinhardtii* after incorporating results of transcript verification experiments.
Supplementary Note	

Note: Supplementary Tables 1, 4–5, 9–10 and Supplementary Data 1 are available on the Nature Methods website.

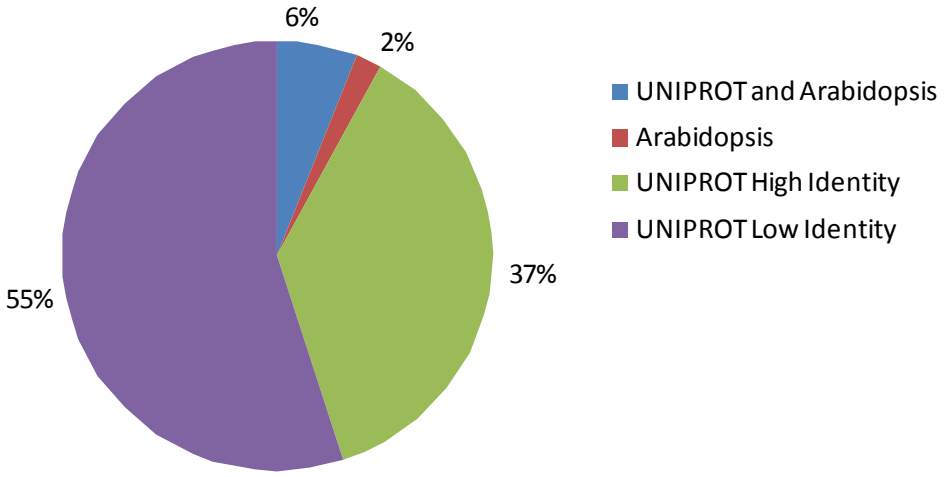
Supplementary Figure 1



Supplementary Figure 1: Distribution of evidence (UNIPROT high confidence, UNIPROT low confidence, and *Arabidopsis*) for *Chlamydomonas* JGI v3.1 EC annotation.

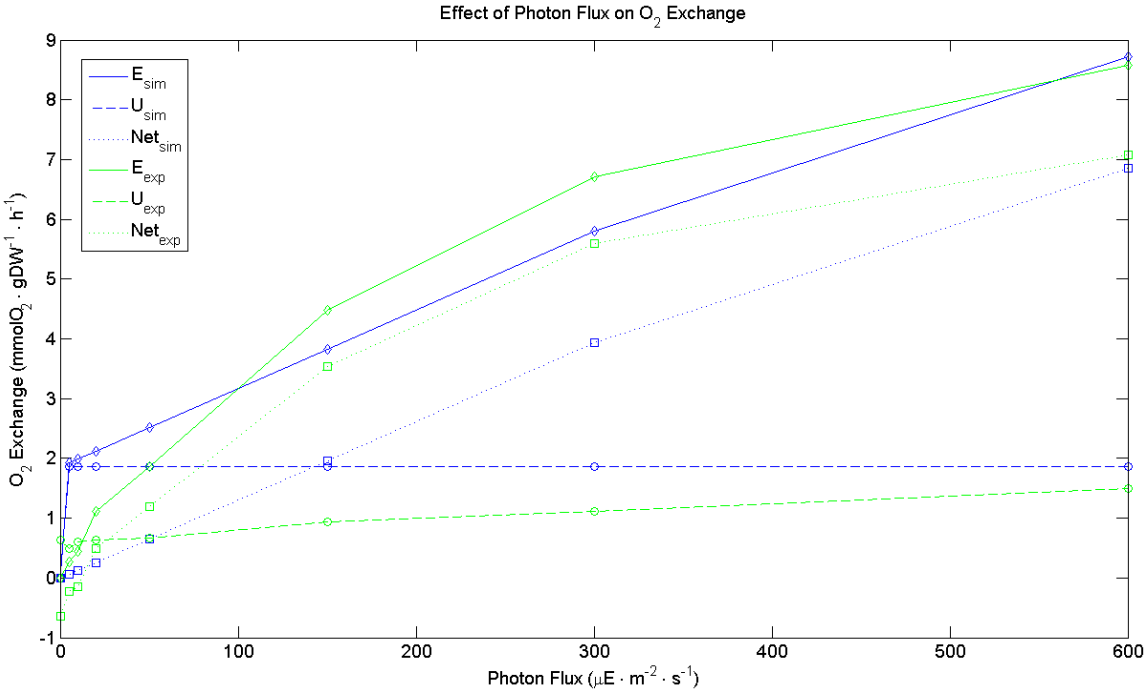
Nature Methods: doi: 10.1038/nmeth.1348

Supplementary Figure 2



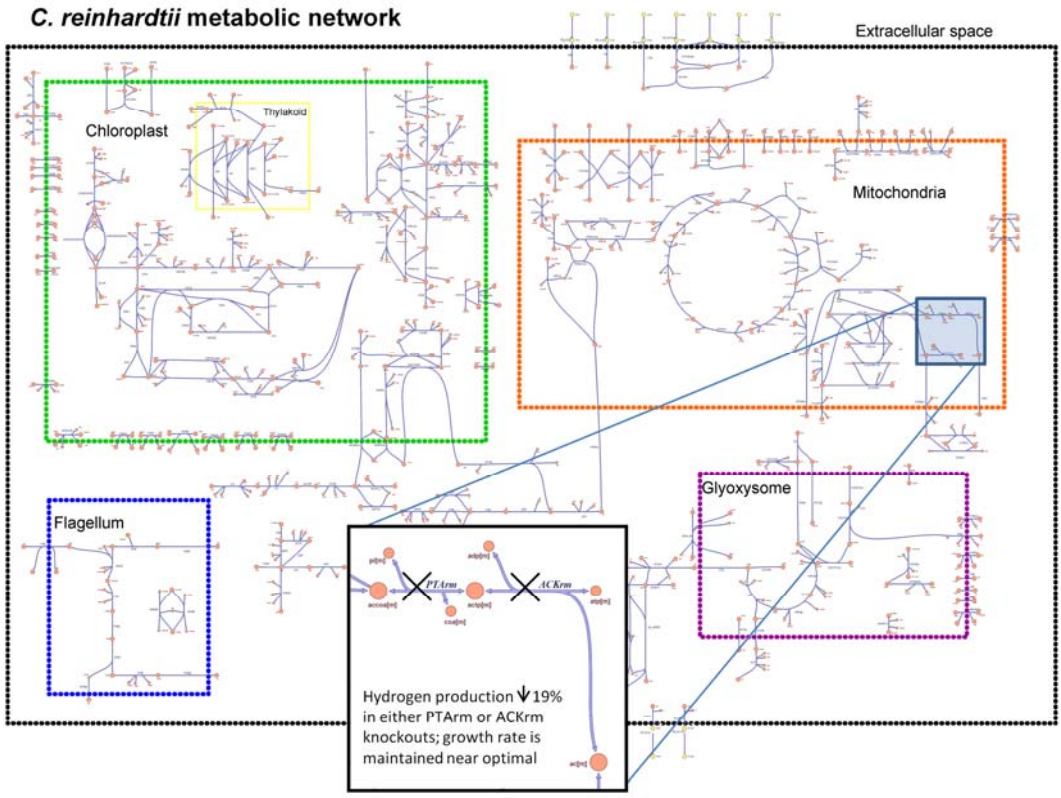
Supplementary Figure 2: Distribution of the JGI v3.1 gene models having evidence from each annotation category (UNIPROT high confidence, UNIPROT low confidence, and *Arabidopsis* proteome).

Supplementary Figure 3



Supplementary Figure 3: Comparison of simulated photosynthetic evolution (E), uptake (U) and net exchange of oxygen with experimental data, across a range of photon flux values.

Supplementary Figure 4

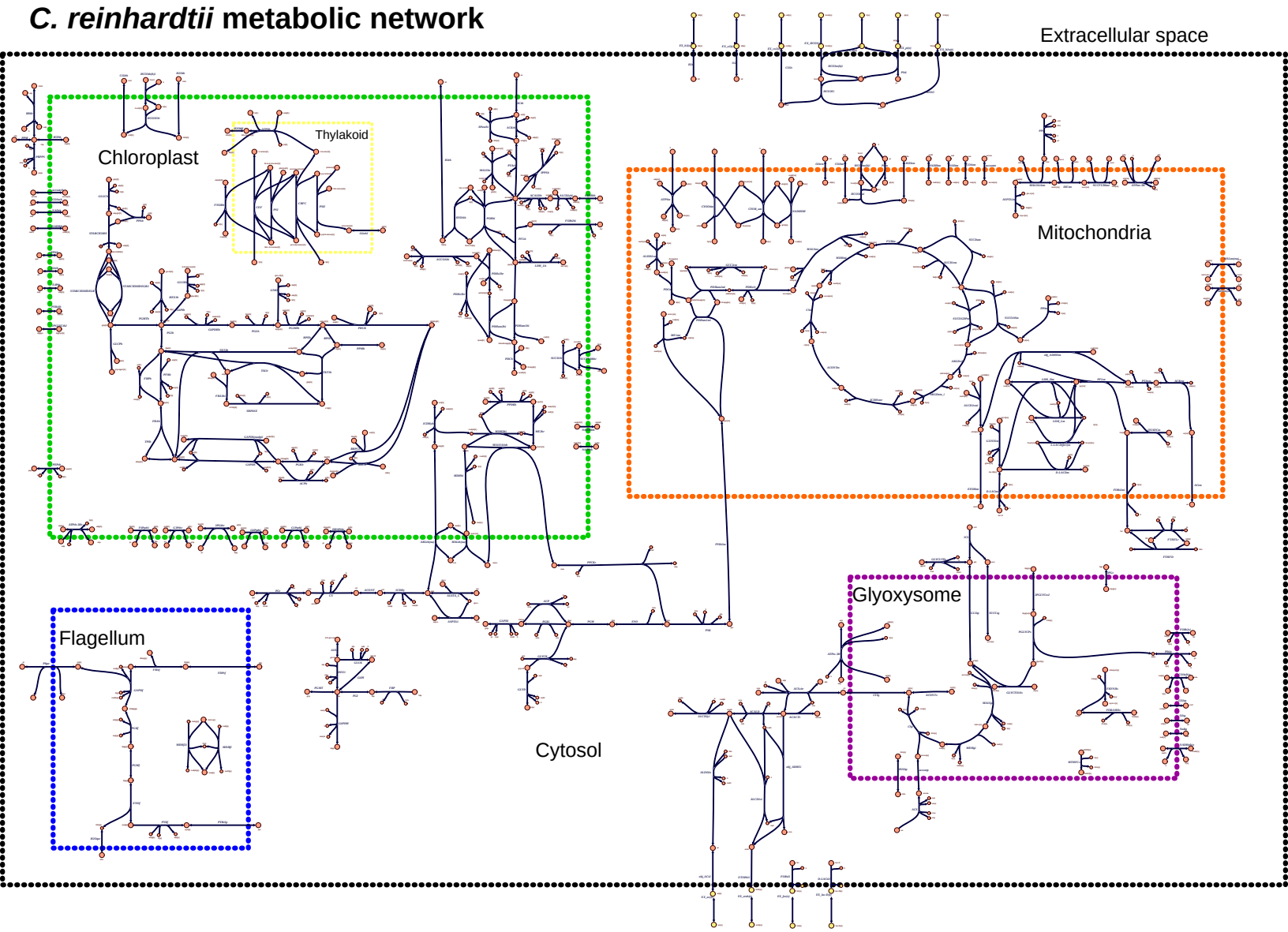


Supplementary Figure 4: A well validated network model towards metabolic engineering. The metabolic network reconstruction provides a critical tool for probing metabolic gene perturbations. We generated a complete set of *in silico* reaction knockouts under photosynthetic growth conditions, and examined variations in hydrogen production coupled with optimal biomass production (**Supplementary Note** and **Supplementary Table 10**). We predict 19% reduction in capacity for hydrogen production in knockouts of two pyruvate metabolism reactions located in the mitochondria, PTArm and ACKrm, compared to wild type. The hypothesized relationship is reasonable because the (Fe-Fe)-hydrogenase reaction catalyzing hydrogen production in the chloroplast is also part of the pyruvate metabolism pathway in *C. reinhardtii*, and suggests up-regulation of phosphotransacetylase and acetate kinase to ensure the limits on the corresponding reactions do not inhibit hydrogen

production *in vivo*. That perturbation of mitochondrial reactions can modulate hydrogen production, a process which occurs primarily in the chloroplast[1], demonstrates the utility of taking a whole-cell, systems-level approach to metabolic engineering.

Nature Methods: doi: 10.1038/nmeth.1348

Supplementary Figure 5



Supplementary Figure 5: High-resolution detailed map of the final metabolic network reconstruction.

Supplementary Table 2

	Existing v3.0 annotation	Our v3.1 annotation
2.7.1.31	No	Yes
1.2.1.3	Yes	Yes
1.1.1.72	No	No
1.1.1.21	No	No
1.1.1.2	Yes	Yes
2.7.1.30	Yes	Yes
3.1.3.21	No	No
2.3.1.15	Yes	Yes
2.3.1.51	No	Yes
3.1.3.4	No	Yes
2.7.1.107	No	Yes
3.1.1.34	No	No
2.3.1.158	No	Yes
2.3.1.20	No	Yes

Supplementary Table 2: Comparison of presence/absence of triacylglycerol synthesis pathway EC terms in existing JGI v3.0 annotation and our new JGI v3.1 annotation.

Supplementary Table 3

EC	EC (continued)	EC (continued)
1.1.1.2	2.2.1.1	4.1.3.1
1.1.1.28	2.2.1.2	4.2.1.11
1.1.1.284	2.3.1.12	4.2.1.2
1.1.1.37	2.3.1.54	4.2.1.3
1.1.1.39	2.3.1.61	5.1.3.1
1.1.1.40	2.3.1.8	5.3.1.1
1.1.1.41	2.3.3.1	5.3.1.6
1.1.1.42	2.3.3.8	5.3.1.9
1.1.1.44	2.3.3.9	5.4.2.1
1.1.1.49	2.4.1.1	5.4.2.2
1.1.1.82	2.6.1.1	6.2.1.1
1.10.2.2	2.6.1.2	6.2.1.4
1.12.7.2	2.7.1.1	6.2.1.5
1.18.1.2	2.7.1.11	
1.2.1.10	2.7.1.19	
1.2.1.12	2.7.1.40	
1.2.1.13	2.7.2.1	
1.2.1.3	2.7.2.3	
1.2.1.9	2.7.9.1	
1.2.4.1	3.1.1.31	
1.2.4.2	3.1.3.11	
1.2.7.1	3.1.3.37	
1.3.5.1	3.6.3.14	
1.3.99.1	4.1.1.1	
1.6.5.3	4.1.1.31	
1.6.99.3	4.1.1.39	
1.8.1.4	4.1.1.49	
1.9.3.1	4.1.2.13	

Supplementary Table 3: List of unique EC terms used to choose primers.

Supplementary Table 6

Growth condition	Objective function	Target quantity	Reference value (source)	*In silico* prediction
Dark aerobic	Precursor biomass	Acetate yield	0.009 g DW / mmol acetate (Sager and Granick, 1953)	0.012 g DW / mmol acetate
Dark anaerobic (starch fermentation)	ATP demand	Formate:Ethanol:Acetate production	2 : 1 : 1 (Gfeller and Gibbs, 1984)	2 : 1 : 1
Light (photosynthesis)	Precursor biomass	Grows, Produces H_2	Yes, Yes (Kruse *et al.*, 2005)	Yes, Yes

Supplementary Table 6: Summaries of *in silico* vs. literature validation of physiological parameters.[2 4]

Supplementary Table 7

EC number	Associated protein	*In vivo* knockout characterization (source)	*In silico* knockout
1.6.5.3	NADH:ubiquinone oxidoreductase Complex I	Reduced growth on acetate in the dark (Remacle *et al.*, 2001)	Confirmed
4.2.1.1	carbonic anhydrase	Increased CO_2 required for photoautotrophic growth (Spalding *et al.*, 1983; Funke *et al.*, 1997)	Confirmed
4.1.1.39	RuBisCO	Photosynthesis deficient (Khrebtukova and Spreitzer, 1996)	Confirmed
1.9.3.1	cytochrome *c* oxidase Complex IV	Obligate photoautotroph (Remacle *et al.*, 2001a)	Use of acetate severely inhibited
1.10.2.2	ubiquinol-cytochrome *c* oxidoreductase Complex III	Obligate photoautotroph (Remacle *et al.*, 2001a; Harris, 2001)	Use of acetate severely inhibited
3.2.1.68†	Isoamylase	Lowered O_2 and dramatically lowered H_2 photoevolution (Posewitz *et al.*, 2004)	Somewhat lowered O_2 and H_2 photoevolution
1.8.1.4	Dihydrolipoyl dehydrogenase	Acetate requiring* (Krishna Niyogi and Rachel Dent, personal communication, 2008)	Grows photoautotrophically

Supplementary Table 7: Summaries of *in silico* knockout validations.[5 10]

† EC 3.2.1.68 (isoamylase) is not currently represented in the metabolic network; however, an alternative starch debranching enzyme (EC 3.2.1.142) is present in the network and fulfills an overlapping function. Therefore, the *in silico* knockout used to simulate this mutant was for the reaction corresponding to EC 3.2.1.142.

*The *in vivo* characterization reported was for an insertion mutant rather than for a gene knockout. Genes perturbed by insertion may retain function similar to wild type, which may explain the discrepancy between the *in vivo* and *in silico* characterizations of this enzyme perturbation.

Nature Methods: doi: 10.1038/nmeth.1348

Supplementary Table 8

Metabolic network characteristics				
Pathways	**Reactions**	**Compartments**	**Reactions**	**Metabolites**
Glycolysis / gluconeogensis	42	Cytosol	43	68
Pyruvate metabolism	31	Chloroplast	59	71
TCA cycle	23	Lumen	3	10
Glyoxylate metabolism	19	Flagellum	8	19
Carbon Fixation	15	Glyoxysome	11	36
Pentose phosphate pathway	16	Mitochondria	35	52
Photosynthesis	9	Exchange (extracellular)	11	11
Oxidative phosphorylation	6	Membrane spanning	89	
Starch metabolism	6	**Total**	**259**	**267**
Intracellular transport	81	Unique metabolites		113
Exchange	11	Literature references		83

Supplementary Table 8: Summary of the central metabolic network reconstruction of *C. reinhardtii* after incorporating results of transcript verification experiments. Our reported counts include reactions and metabolites duplicated across multiple compartments. For example, the citrate synthase reaction (EC 2.3.3.1) is functional in the mitochondria, the cytosol, and the glyoxysome. Accordingly, the associated metabolites acetyl-CoA, oxaloacetate, citrate and CoA appear in each of these compartments. Our central metabolic network reconstruction represents a computationally functional collection of reactions based on those present in our annotation of JGI v3.1, and manually curated to incorporate existing knowledge of biochemical pathways, and literature evidence specific to *C. reinhardtii* and related species.

Nature Methods: doi: 10.1038/nmeth.1348

SUPPLEMENTARY NOTE

EC annotation of JGI version 3.1 proteins

The JGI version 3.1 (v3.1) transcripts do not carry EC annotations. To annotate these putative transcripts, we carried out BLAST sequence comparison using two reference data sets and respective strategies. We extracted EC-annotated proteins in UNIPROT-Swiss-Prot database, a manually annotated and reviewed protein database. The UniProt-SwissProt database contained a set of ~120,000 proteins from over 5,000 species carrying 2,321 EC terms. Despite its wide coverage, this data set does not contain any completely EC-annotated proteins from higher organisms. Another reference data set we used was a manually annotated and complete proteome of *Arabidopsis* (http://proteomics.arabidopsis.info/). It is a smaller data set containing ~32,000 proteins in total. The *A. thaliana* proteome data set was used to catalogue 1,800 enzymatic proteins which were assigned to 498 unique EC numbers.

To search the UNIPROT set, we first defined the longest open reading frame of each of the *Chlamydomonas* JGI v3.1 transcripts, and then ran unidirectional BLAST of the translated transcripts against the UNIPROT reference set. The best hit of each translated JGI v3.1 transcript together with its hit identity and score was recorded and the EC annotation was transferred. Previous studies suggest that sequence similarity of at least 40% is necessary to accurately perform functional prediction[11, 12]. Therefore, we imposed a threshold of 40% identity, together with a score of 50 to ensure sufficient length of match, and we defined those sequences that meet both these criteria as "high confidence". We also defined "low confidence" hits as those sequences that have *either* 40% identity *or* BLAST score of 50.

Nature Methods: doi: 10.1038/nmeth.1348

To search the *Arabidopsis* reference set, we ran two directional BLAST of translated *Chlamydomonas* transcripts against the *Arabidopsis* proteome. We processed the two directional BLAST results using Inparanoid[13] to establish orthologous groups between the two species. EC annotation was then transferred from *Arabidopsis* to orthologous transcripts in *Chlamydomonas*. The Inparanoid program uses the best reciprocal match algorithm to identify orthologous groups between complete genomes. Although orthologs and paralogs are all homologous, orthologous genes are presumed to carry out the same function in different species while paralogous genes presumably carry out new functions in the same species[14]. Transfer of functional annotation to orthologous genes has higher likelihood of accuracy than single directional homology search.

We searched the UNIPROT and found 898 EC terms in *Chlamydomonas* JGI v3.1. When we searched the *Arabidopsis* database we obtained 236 terms of them 16 were not unambiguously defined. 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. **Supplementary Fig. 1** describes the distribution of the EC annotation in *Chlamydomonas* v3.1. For most of the EC terms, there are multiple corresponding v3.1 gene models. In total, the EC terms are associated with 3,368 *Chlamydomonas* v3.1 gene models. **Supplementary Fig. 2** illustrates the distribution of the gene models in each annotation category.

Nature Methods: doi: 10.1038/nmeth.1348

Functional assignments of the transcripts corresponding to the central metabolic pathways made using BLASTP program was further validated by assigning enzymatic domain families of Pfam database[15] to the protein products of these transcripts. Using HMMER, we assigned enzymatic domain families of Pfam database to 174 transcripts (**Supplementary Table 4**), which were further confirmed through PSI-BLAST and HHpred searches against non-redundant database. The library of profiles for various domains was prepared by extracting all alignments from the Pfam database and updating them by adding new members from the NR database. These updated alignments were then used to make HMMs with the HMMER package[16] or PSSMs with PSI-BLAST[17, 18]. Profile searches using the PSI-BLAST program were conducted either with a single sequence or a sequence with a PSSM used as the query, with a profile inclusion expectation (E) value threshold of 0.01 and were iterated until convergence. The E-value for cut-off of 10^{-3} was used as threshold confidence for assigning enzymatic domains to proteins products of the transcripts for HMMER searches. Further, these assignments were confirmed using PSI-BLAST[17, 18] and HHpred[19], HMM-HMM[20] comparison encoding programs.

Comparison existing JGI v3.0 and new v3.1 annotation

Because the set of EC terms included in our network reconstruction was generated using our JGI v3.1 annotation as the primary form of genomic evidence, we generated an unbiased set of ECs to compare coverage of JGI v3.0 and v3.1 annotation. To do so, we pooled all EC terms in KEGG from the following central metabolic pathways: glycolysis, citric acid cycle, pentose phosphate pathway, oxidative phosphorylation, photosynthesis, carbon fixation, starch metabolism, pyruvate metabolism and glyoxylate metabolism. Our annotation of JGI v3.1 had broader coverage than existing annotation of JGI v3.0 published online, which was generated primarily using Eukaryotic Orthologous Groups (KOGs)[21]. The discrepancy points to the possibility of false

Nature Methods: doi: 10.1038/nmeth.1348

negatives in the online annotation, with better coverage in our newly generated annotation (**Fig. 2a**).

Experimental verification of central metabolic transcripts

Verification by RT-PCR

In order to validate the central metabolic transcripts we performed RT-PCR experiments on 174 EC annotated gene models (**Supplementary Table 4**) as well as a reference set of 48 transcripts. The transcripts in the reference set were picked from individually studied well-annotated protein coding transcripts that were reported in literature and were assigned a GenBank accession number. The generated amplicons were cloned into our Gateway vector and were sequenced from both ends, either as minipools or as single colonies. The sequenced clones were then aligned against the corresponding JGI v3.1 predicted sequences. A transcript was defined "OST (ORF Sequence Tag) verified" if the sequenced ORF was identical, either in entire length or at both 5' and 3' ends, to the predicted gene model. Using this criterion, we could verify 136 (about 78%) of the hypothesized metabolic ORFs and 40 (83%) of the ORFs in the reference set (all other transcripts in the reference set were confirmed at least at one end). The metabolic ORF transcripts that were not captured using RT-PCR termini primers or were only verified at one end were subjected to RACE experiments for further evaluation.

Verification by RACE

Using the predicted gene models we designed gene-specific primers and carried out RACE in order to correct the predicted 5' and 3' boundaries. The sensitivity and specificity of our RACE experiments were increased by designing nested primers and performing touchdown PCRs. The nested primers were Gateway-tailed to permit the

Nature Methods: doi: 10.1038/nmeth.1348

cloning of RACE amplicons. The 5' RACE amplicons were subsequently sequenced from both ends, while the 3' RACE amplicons were sequenced unidirectionally from the 5' end to generate 5' and 3' RACE Sequence Tags (RST), respectively. We defined a gene model as "RST Verified" if a contig of the entire ORF could be assembled from the 5' and 3' RSTs, or if both the 5' and 3' RSTs completely matched the 5' and 3' ends of the predicted gene model, covering start and stop codons, respectively. Of the 38 RT-PCR-failed metabolic ORFs that were tested by RACE, 20 (53%) were RST verified. Together, RT-PCR and RACE experiments could verify 156 gene models corresponding to all 65 EC numbers in our central metabolic map.

RACE-defined ORFs

The sequence information provided by the RACE experiments was used to reannotate the unverified transcripts. The transcripts that were not fully verified using RT-PCR or RACE experiments were considered for reannotation only if they had a matched sequence of 25 nucleotides or more with the predicted sequence. Of the 18 unverified central metabolic transcripts, 16 met this criterion and were examined for reannotation as outlined below:

1) After alignment of 5' and 3' RSTs to the predicted gene models we determined the continuous sequences of RSTs that were completely matched to the predicted sequence and had a length of 25 nucleotides or more.
2) The 3' end of 5' RST matched sequence was bridged to the 5' end of 3' RST matched sequence using the predicted gene model sequence occurring in between.

Nature Methods: doi: 10.1038/nmeth.1348

3) To define the 5' boundaries, we first searched the 5' RST sequence upstream of the matched region for the occurrence of any start codon. We also determined all possible stop codons downstream of the matched region in 3' RST sequence.
4) Candidate ORFs were re-assembled considering any possible combination of start and stop codons which could produce an in frame and non-truncated protein coding ORF.
5) Candidate ORFs were BLATed against *Chlamydomonas* genomic sequence.
6) The longest sequence that could produce a protein coding ORF and had matched genomic sequence was picked as final reannotated sequence.

By following these steps, we could provide experimentally-based annotations for 9 of the 16 transcripts considered for reannotation (**Fig. 2c**). The reannotated sequences for these transcripts are shown in **Supplementary Table 4**. Although the other 7 transcripts were matched in part to their respective predicted sequences and we successfully defined protein coding ORFs for them, they were not completely matched to the *Chlamydomonas* genomic sequence and therefore were defined as "partially verified ORFs" (**Fig. 2c**).

Objective functions for *in silico* experiments

For the majority of *in silico* simulations characterizing different environmental conditions, we optimized a biomass reaction calculated such that one unit of flux through the reaction utilizes an amount (in mmol) of precursor metabolites corresponding to 1 g DW of biomass for the organism. The reaction stoichiometry, based on yeast precursor biomass, is as follows (I. Famili and S. Wiback, Genomatica Inc., personal communication):

Nature Methods: doi: 10.1038/nmeth.1348

[c] : (0.906) 3pg + (1.1128) accoa + (0.465) akg + (31.114) atp + (0.265) e4p + (0.809) f6p + (0.145) g3p + (1.713) g6p + (1.95) nad + (10.809) nadph + (1.237) oaa + (0.271) pep + (2.06) pyr + (0.313) r5p → (31.114) adp + (1.1128) coa + (7) h + (1.95) nadh + (10.809) nadp + (31.143) pi

For dark anaerobic conditions, *in silico* simulations were performed using the ATP demand reaction:

[c] : (31.114) atp → (31.114) adp + (31.114) pi

This reduced objective function was used to simulate the dark anaerobic physiology of *C. reinhardtii*, which calls for subsistence on starch reserves in the chloroplast rather than production of biomass. Although the objective functions were derived for yeast, many phenotypic predictions are robust to small variations in the biomass reaction[22].

Details about the current metabolic network reconstruction

Our final central metabolic reconstruction of *C. reinhardtii* accounts for key pathways of carbon flow common to eukaryotes, including glycolysis / gluconeogenesis, pyruvate metabolism, citric acid cycle, glyoxylate metabolism, pentose phosphate pathway, and oxidative phosphorylation (**Supplementary Fig. 5**). We have also included photosynthesis, carbon fixation and starch metabolism, capturing some of the photoautotrophic features of algae. The reconstruction accounts 259 reactions, the largest portion of which takes place in the chloroplast (24%, including reactions in the lumen subcompartment of the chloroplast), cytosol (17%) and mitochondria (14%), while the remaining reactions (45%) are included in the flagellum, the glyoxysome, extracellular exchange (to account for uptake of nutrients in the cell growth medium), or are membrane spanning (including intracellular transport and

Nature Methods: doi: 10.1038/nmeth.1348

oxidative phosphorylation) (see detailed summaries in **Supplementary Table 8**). The full reconstruction is detailed in **Supplementary Table 9**, and an SBML version is also included as **Supplementary Data 1**.

The count of 259 metabolic reactions includes some reactions duplicated across multiple intracellular compartments (for example malate synthase, localized in the glyoxysome, the chloroplast and the mitochondria), and other reactions which can be catalyzed by multiple EC numbers (like the hexokinase reaction, which can be catalyzed by EC 2.7.1.1 and 2.7.1.2). Further, some reactions such as transporters and photosystem I and II reactions are not mapped to any EC term. In all, the metabolic network reconstruction accounts for 106 unique EC terms. The majority of these EC terms could be mapped to one or more annotated transcripts (**Supplementary Table 1**); in total, these 106 EC terms were mapped to 303 transcripts.

Simulated growth optimizing biomass production under dark aerobic conditions resulted in an acetate yield of 0.012 g DW / mmol acetate, similar to the experimentally derived value of 0.009 (see derivation below). Simulated fermentation of starch reserves in the chloroplast under dark anaerobic conditions[23, 24] optimized ATP production and predicted a formate:acetate:ethanol output ratio of 2:1:1, corresponding exactly to the value reported in the literature[3]. Our simulations also supported hydrogen production associated with photosynthetic growth in the light (**Supplementary Table 6**). Simulated photosynthetic oxygen uptake, evolution, and net exchange also showed considerable agreement with experimental data[25] across a broad range of photon fluxes (**Supplementary Fig. 3**). The relationships among the simulated uptake, evolution, and net production rates closely paralleled the experimentally determined rates both qualitatively and quantitatively, reaching very nearly equivalent flux values at both extremes of the photon flux range and remaining close in the middle range as well.

Nature Methods: doi: 10.1038/nmeth.1348

In silico reaction deletion experiments for seven mutants documented in literature sources and elsewhere also showed good agreement with *in vivo* data (**Supplementary Table 7**). For example, a knockout of the mitochondrial gene *nd1* was reported to produce slow aerobic growth in the dark[5], and this result was duplicated by deletion of the corresponding reaction from our model. Similarly, *in vivo* knockouts of the mitochondrial *cox1* and *cob* genes were both reported to be obligate photoautotrophs[7, 6], and this phenotype was also observed through *in silico* experiments with our network reconstruction.

Derivation of literature-based acetate growth yield

The dark aerobic generation time for *C. reinhardtii* is approximately 18 hours[2]. This doubling time corresponds to a growth rate of $e^{\frac{\log(2)}{18}} - 1 = 0.039\,\mathrm{hr}^{-1}$. To obtain an acetate growth yield value, a value for the acetate uptake rate as a function of time is needed. We approximate this value using measurements reported in Sager and Granick (1953).

First, we express cell concentration as an exponential function of time:

$$C(t) = K_0 \cdot 2^{t/18}$$

After about 120 hours, at the end of log phase growth, Figure 1 from Sager and Granick (1953) shows a cell concentration of about $2^{20.8}$ cells/mL. From this measurement, we can conclude:

$$\begin{aligned} C(120) &= K_0 \cdot 2^{120/18} = 2^{20.8}\ \text{cells/mL} \\ \Rightarrow K_0 &= 2^{20.8-130/18} = 2^{14.1}\ \text{cells/mL} \end{aligned}$$

Nature Methods: doi: 10.1038/nmeth.1348

We can now write instantaneous acetate consumption as a function of time. Letting *A* denote the acetate uptake rate per cell · hr, we have:

$$\begin{aligned} A(t) &= A \cdot C(t) \\ &= A \cdot K_0 \cdot 2^{t/18} \end{aligned}$$

Sager and Granick reported the initial concentration of acetate in the medium used was 0.015 M and that growth on acetate in the dark continued until about 85% consumption of the initial acetate was achieved. From these values, we can infer the acetate uptake rate, *A*:

$$\begin{aligned} \int_0^{120} A(t)dt &= 0.015 \text{ mmol} \cdot 85\% \\ &= A \cdot K_0 2^{t/18} \cdot \frac{1}{(\log 2)/18}]_0^{120} \\ &= A \cdot K_0 \cdot \frac{1}{(\log 2)/18}]_0^{120}(2^{\frac{120}{18}} - 1) \\ &= A \cdot (4.69 \times 10^7) \\ \Rightarrow A &= 2.716 \times 10^{-10} \text{ mmol/cell} \cdot \text{hr} \end{aligned}$$

We can convert this acetate uptake rate to terms of cell mass rather than cell count using the approximate measurement of 63 pg dry weight per cell[26]:

$$A \cdot \frac{1}{63 \text{ pg/cell} \cdot 1 \text{ g}/10^{12} \text{ pg}} = 4.31 \text{ mmol acetate / g DW} \cdot \text{ hr}$$

Finally, we combine this acetate uptake rate with the growth rate of 0.039 g DW/hr to find the acetate growth yield of:

$$\frac{0.039 \text{ hr}^{-1}}{4.31 \text{mmol acetate/g DW} \cdot \text{hr}} = 0.009 \text{ mmol acetate / g DW.}$$

Nature Methods: doi: 10.1038/nmeth.1348

Validation of simulated photosynthetic oxygen exchange

Metabolic reaction flux bounds were set as in all other photosynthetic growth simulations performed in this study with the exception that maximum oxygen uptake was bounded at the experimental maximum[25]. This step was taken not only to model the simulations after the experimental conditions but also because oxygen is a growth-limiting factor under aerobic conditions. Since flux balance analysis performs a linear optimization maximizing an objective function, biomass in this validation, the simulation will always exhaust any limiting metabolic resources up to the maximum allowed extent. Therefore, the maximum oxygen uptake bound is reached in all simulations performed here.

Fluxes from experimental data and simulations were not initially in the same dimensional unit. The experimental units for oxygen exchange flux and photon flux were μmol·mg Chl^{-1}·hr^{-1} and μmol·m^{-2}·s^{-1}, respectively; the standard flux unit in our simulations is mmol·g DW^{-1}·hr^{-1}. Dimensional analysis was used to derive a conversion factor to relate experimental and simulated oxygen exchange flux units:

$$\frac{1\ \text{mmol}}{1000\ \mu\text{mol}} \cdot \frac{1.79 \times 10^{9}\ \text{mg chlorophyll}}{\text{cell}} \cdot \frac{\text{cell}}{4.8 \times 10^{11}\ \text{g dry weight}}$$

The measurement of chlorophyll per cell[3] and dry weight[27] of a log phase *C. reinhardtii* cell were taken from experimental data.

Conversion between experimental and simulated photon flux units is considerably more challenging. The experimental measure of photon flux consists only of the light radiating from the light source, not the actual rate of photon absorption by light harvesting complexes of the photosystems. Photon flux in the simulations represents exactly the rate of photons being absorbed and used to drive photosynthesis.

Accurately relating the experimental and simulated units therefore requires accounting for not only dimensional analysis but also all factors responsible for the difference between the light emitted from the light source and the actual photons absorbed by the light harvesting complexes. The dimensional analysis would simply require the measure of the total surface area of thylakoid membrane in a single cell and the dry weight of the cell; however, accounting for the discrepancy between emitted and absorbed light would require many additional measures such as the percent of photosynthetically active radiation from the given light source, the measure of both extra- and intracellular light scattering, the distribution density of light harvesting complexes in the thylakoid membrane, and the maximum rate of photon absorption of a single average harvesting complex.

Since a proper photon flux conversion factor has yet to be developed to account for all necessary considerations, a simplified approach was taken to derive a conversion factor for use in validating against this specific dataset. To derive this simplified conversion factor, photosynthetic growth simulations were run over a broad range of photon flux until the saturation level of net oxygen exchange was identified (82 $mmol \cdot g\ DW^{-1} \cdot hr^{-1}$). The experimental photon flux saturation point[25] was 500 $\mu mol \cdot m^{-2} \cdot s^{-1}$. Therefore, to relate the experimental and simulated photon flux units, the ratio of simulated to experimental photon flux saturation points was used. This simplification is reasonable in the scope of the current study and simply causes the simulations to saturate at the same photon flux level as in the experimental data.

Nature Methods: doi: 10.1038/nmeth.1348

References

1. Maione, T.E. & Gibbs, M. Hydrogenase-mediated activities in isolated chloroplasts of *Chlamydomonas reinhardii*. *Plant Physiol* **80**, 360–363 (1986).

2. Sager, R. & Granick, S. Nutritional studies with *Chlamydomonas reinhardi*. *Ann. N. Y. Acad. Sci.* **56**, 831–838 (1953).

3. Gfeller, R.P. & Gibbs, M. Fermentative metabolism of *Chlamydomonas reinhardtii*: I. Analysis of fermentative products from starch in dark and light. *Plant Physiol.* **75**, 212–218 (1984).

4. Kruse, O., Rupprecht, J., Bader, K.P., Thomas-Hall, S., Schenk, P.M. *et al.* Improved photobiological H_2 production in engineered green algal cells. *J. Biol. Chem.* **280**, 34170–34177 (2005).

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

6. Remacle, C., Duby, F., Cardol, P. & Matagne, R.F. Mutations inactivating mitochondrial genes in *Chlamydomonas reinhardtii*. *Biochem. Soc. Trans.* **29**, 442–446 (2001a).

7. Harris, E.H. Chlamydomonas as a model organism. *Annu. Rev. Plant. Physiol. Plant. Mol. Biol.* **52**, 363–406 (2001).

8. Spalding, M.H., Spreitzer, R.J. & Ogren, W.L. Carbonic anhydrase-deficient mutant of *Chlamydomonas reinhardii* requires elevated carbon dioxide concentration for photoautotrophic growth. *Plant Physiol.* **73**, 268–272 (1983).

9. Funke, R.P., Kovar, J.L. & Weeks, D.P. Intracellular carbonic anhydrase is essential to photosynthesis in *Chlamydomonas reinhardtii* at atmospheric levels of CO_2.

demonstration via genomic complementation of the high-CO_2-requiring mutant *ca-1*. *Plant Physiol.* **114**, 237–244 (1997).

10. Khrebtukova, I. & Spreitzer, R.J. Elimination of the Chlamydomonas gene family that encodes the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. *Proc. Natl. Acad. Sci. U S A* **93**, 13689–13693 (1996).

11. Rost, B. Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595–608 (2002).

12. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **333**, 863–882 (2003).

13. O'Brien, K.P., Remm, M. & Sonnhammer, E.L.L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**, D476–D480 (2005).

14. Studer, R.A. & Robinson-Rechavi, M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* **25**, 210–216 (2009).

15. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V. *et al.* The Pfam protein families database. *Nucleic Acids Res* **32**, D138–D141 (2004).

16. Zhang, Z. & Wood, W.I. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19**, 307–308 (2003).

17. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases. *Trends Biochem Sci* **23**, 444–447 (1998).

18. Altschul, S.F., Gertz, E.M., Agarwala, R., Schäffer, A.A. & Yu, Y.K. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* **37**, 815–824 (2009).

19. Söding, J., Biegert, A. & Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244–W248 (2005).

20. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).

21. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B. ***et al.*** The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

22. Varma, A. & Palsson, B. Metabolic Capabilities of *Escherichia coli* II. Optimal Growth Patterns. *Journal of Theoretical Biology* **165**, 503–522 (1993).

23. Levi, C. & Gibbs, M. Starch degradation in synchronously grown *Chlamydomonas reinhardtii* and characterization of the amylase. *Plant Physiol* **74**, 459–463 (1984).

24. Grossman, A.R., Croft, M., Gladyshev, V.N., Merchant, S.S., Posewitz, M.C. ***et al.*** Novel metabolism in Chlamydomonas through the lens of genomics. *Curr Opin Plant Biol* **10**, 190–198 (2007).

25. Sueltemeyer, D.F., Klug, K. & Fock, H.P. Effect of photon fluence rate on oxygen evolution and uptake by *Chlamydomonas reinhardtii* suspensions grown in ambient and CO_2-enriched air. *Plant Physiol* **81**, 372–375 (1986).

26. Nagel, K. & Voigt, J. In vitro evolution and preliminary characterization of a cadmium-resistant population of *Chlamydomonas reinhardtii*. ***Appl. Environ. Microbiol.*** **55**, 526–528 (1989).

27. Mitchell, S., , Trainor, F., Rich, P. & Goulden, C. Growth of *Daphnia magna* in the laboratory in relation to the nutritional state of its food species, *Chlamydomonas reinhardii*. *Journal of Plankton Research* **14**, 379–391 (1992).