nature genetics

LETTERS

https://doi.org/10.1038/s41588-019-0370-6

A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis

Xiaobo Li[1,2,3], Weronika Patena[1,2], Friedrich Fauser[1,2], Robert E. Jinkerson[2,7], Shai Saroussi[2], Moritz T. Meyer[1], Nina Ivanova[2], Jacob M. Robertson[1,2], Rebecca Yue[2], Ru Zhang[2,8], Josep Vilarrasa-Blasi[2], Tyler M. Wittkopp[2,4,9], Silvia Ramundo[5], Sean R. Blum[2], Audrey Goh[1], Matthew Laudon[6], Tharan Srikumar[1], Paul A. Lefebvre[6], Arthur R. Grossman[2] and Martin C. Jonikas[1,2*]

Photosynthetic organisms provide food and energy for nearly all life on Earth, yet half of their protein-coding genes remain uncharacterized[1,2]. Characterization of these genes could be greatly accelerated by new genetic resources for unicellular organisms. Here we generated a genome-wide, indexed library of mapped insertion mutants for the unicellular alga *Chlamydomonas reinhardtii*. The 62,389 mutants in the library, covering 83% of nuclear protein-coding genes, are available to the community. Each mutant contains unique DNA barcodes, allowing the collection to be screened as a pool. We performed a genome-wide survey of genes required for photosynthesis, which identified 303 candidate genes. Characterization of one of these genes, the conserved predicted phosphatase-encoding gene *CPL3*, showed that it is important for accumulation of multiple photosynthetic protein complexes. Notably, 21 of the 43 higher-confidence genes are novel, opening new opportunities for advances in understanding of this biogeochemically fundamental process. This library will accelerate the characterization of thousands of genes in algae, plants, and animals.

The green alga *Chlamydomonas* has long been used for genetic studies of eukaryotic photosynthesis because of its rare ability to grow in the absence of photosynthetic function[3]. In addition, it has made extensive contributions to basic understanding of light signaling, stress acclimation, and metabolism of carbohydrates, lipids, and pigments (Fig. 1a)[4–6]. Moreover, *Chlamydomonas* has retained many genes from the plant–animal common ancestor, which has contributed to understanding of fundamental aspects of the structure and function of cilia and basal bodies[7,8]. Like *Saccharomyces cerevisiae*, *Chlamydomonas* can grow as a haploid, facilitating genetic studies. However, until now, the value of *Chlamydomonas* had been limited by the lack of mutants for most of its nuclear genes.

In the present study, we sought to generate a genome-wide collection of *Chlamydomonas* mutants with known gene disruptions to provide mutants in genes of interest for the scientific community and then to leverage this collection to identify genes with roles in photosynthesis. To reach the necessary scale, we chose to use random insertional mutagenesis and built on advances in insertion mapping and mutant propagation from our pilot study[9]. To enable mapping of insertion sites and screening of pools of mutants on a much larger scale, we developed new tools leveraging unique DNA barcodes in each transforming cassette.

We generated mutants by transforming haploid cells with DNA cassettes that randomly insert into the genome and inactivate the genes into which they insert. We maintained the mutants as indexed colony arrays on agar medium containing acetate as a carbon and energy source to allow recovery of mutants with defects in photosynthesis. Each DNA cassette contained two unique barcodes, one on each side of the cassette (Supplementary Fig. 1a–d). For each mutant, the barcode and genomic flanking sequence on each side of the cassette were initially unknown (Supplementary Fig. 1e). We determined the sequence of the barcodes in each mutant colony by combinatorial pooling and deep sequencing (Supplementary Figs. 1f and 2). We then mapped each insertion by pooling all mutants and amplifying all flanking sequences together with their corresponding barcodes, followed by deep sequencing (Supplementary Fig. 1g). The combination of these datasets identified the insertion site(s) in each mutant. This procedure yielded 62,389 mutants on 245 plates, with a total of 74,923 insertions that were largely randomly distributed over the chromosomes (Fig. 1b,c, Supplementary Figs. 3 and 4, and Supplementary Table 5).

This library provides mutants for ~83% of all nuclear genes (Fig. 2a–d). Approximately 69% of genes are represented by an insertion in a 5′ UTR, an exon, or an intron—the regions in which disruption is most likely to cause an altered phenotype. Many gene sets of interest to the research community are well represented, including genes encoding proteins phylogenetically associated with the plant lineage (GreenCut2)[1], proteins that localize to the chloroplast[10], and proteins associated with the structure and function of flagella or basal bodies[11,12] (Fig. 2b). Mutants in this collection are available through the CLiP website (see URLs). Over 1,800 mutants have already been distributed to over 200 laboratories worldwide in the first 18 months of prepublication distribution (Fig. 2e). These mutants are facilitating genetic investigation of a broad range of processes, ranging from photosynthesis and metabolism to cilia structure and function (Fig. 2f).

[1]Department of Molecular Biology, Princeton University, Princeton, NJ, USA. [2]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA, USA. [3]School of Life Sciences, Westlake Institute for Advanced Study, Westlake University, Hangzhou, China. [4]Department of Biology, Stanford University, Stanford, CA, USA. [5]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA. [6]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN, USA. [7]Present address: Department of Chemical and Environmental Engineering, University of California, Riverside, Riverside, CA, USA. [8]Present address: Donald Danforth Plant Science Center, St. Louis, MO, USA. [9]Present address: Salk Institute for Biological Studies, La Jolla, CA, USA. *e-mail: mjonikas@princeton.edu

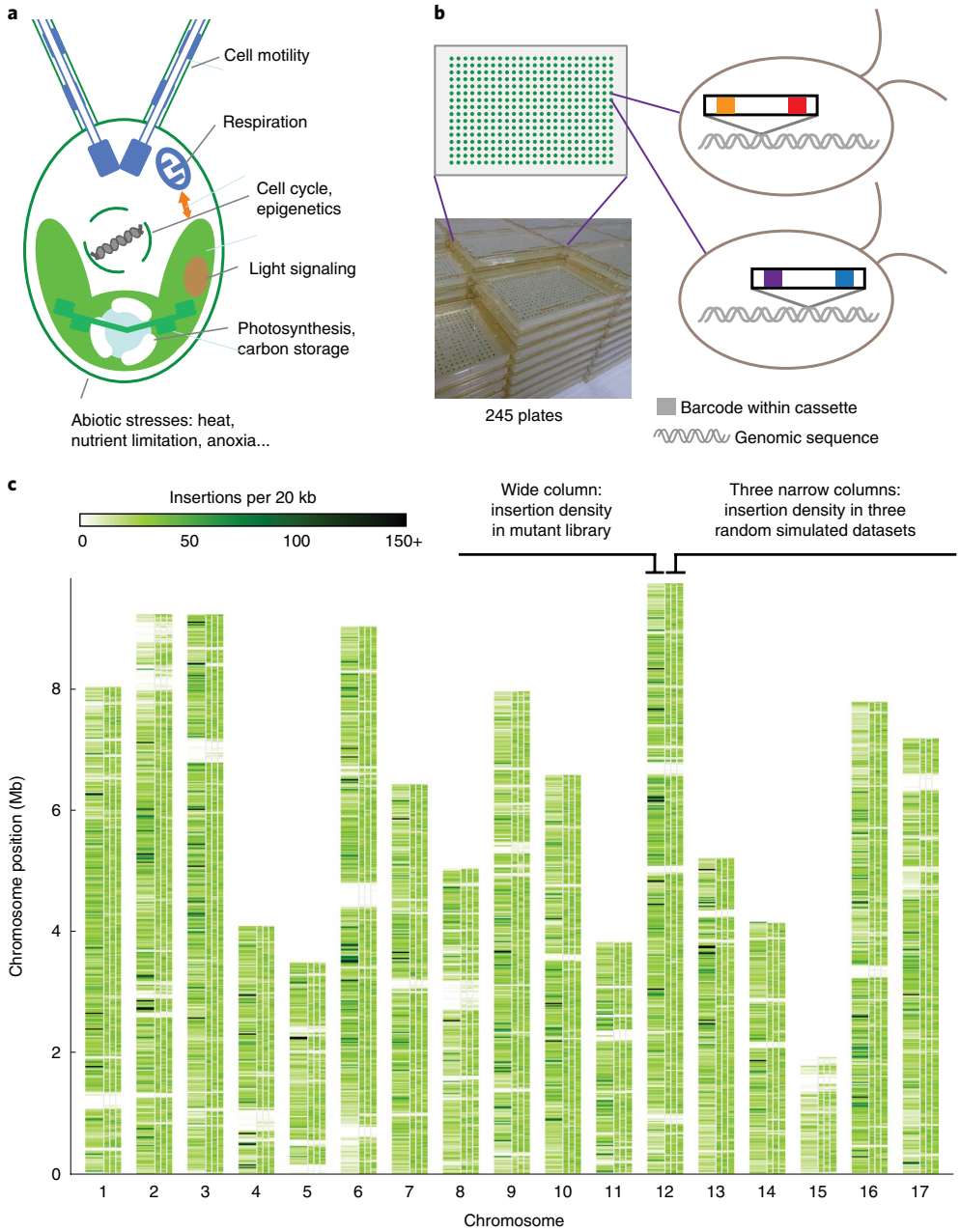


Fig. 1 | A genome-wide library of *Chlamydomonas* mutants was generated by random insertion of barcoded cassettes and mapping of insertion sites. **a**, *Chlamydomonas* is used for studies of various cellular processes and organism–environment interactions. **b**, Our library contains 62,389 insertional mutants maintained as 245 plates of 384-colony arrays. Each mutant contains at least one insertion cassette at a random site in its genome; each insertion cassette contains one unique barcode at each end (Supplementary Fig. 1a–c). **c**, The insertion density is largely random over the majority of the genome. This panel compares the observed insertion density over the genome (left column above each chromosome number) to the density in three simulations with insertions randomly distributed over all mappable positions in the genome (three narrow columns to the right for each chromosome). Areas that are white for all columns represent regions where insertions cannot be mapped to a unique genomic position owing to highly repetitive sequence. See also Supplementary Fig. 4.

To identify genes required for photosynthesis, we screened our library for mutants deficient in photosynthetic growth. Rather than phenotyping each strain individually, we pooled the entire library into one culture and leveraged the unique barcodes present in each strain to track the abundance of individual strains after growth under different conditions. This feature enables genome-wide screening with speed and depth unprecedented in photosynthetic eukaryotes. We grew the pool of mutants photosynthetically in the light in minimal Tris-phosphate (TP) medium with carbon dioxide (CO_2) as the sole source of carbon and heterotrophically in the dark in Tris-acetate-phosphate (TAP) medium, where acetate provides fixed carbon and energy[3] (Fig. 3a). To quantify mutant growth under each condition, we amplified and performed deep sequencing of the barcodes from the final cell populations. We then compared the ability of each mutant to grow under the photosynthetic and heterotrophic conditions by comparing the read counts for each barcode in the two conditions (Supplementary Table 10 and Supplementary Note). Mutant phenotypes were highly reproducible (Fig. 3b and

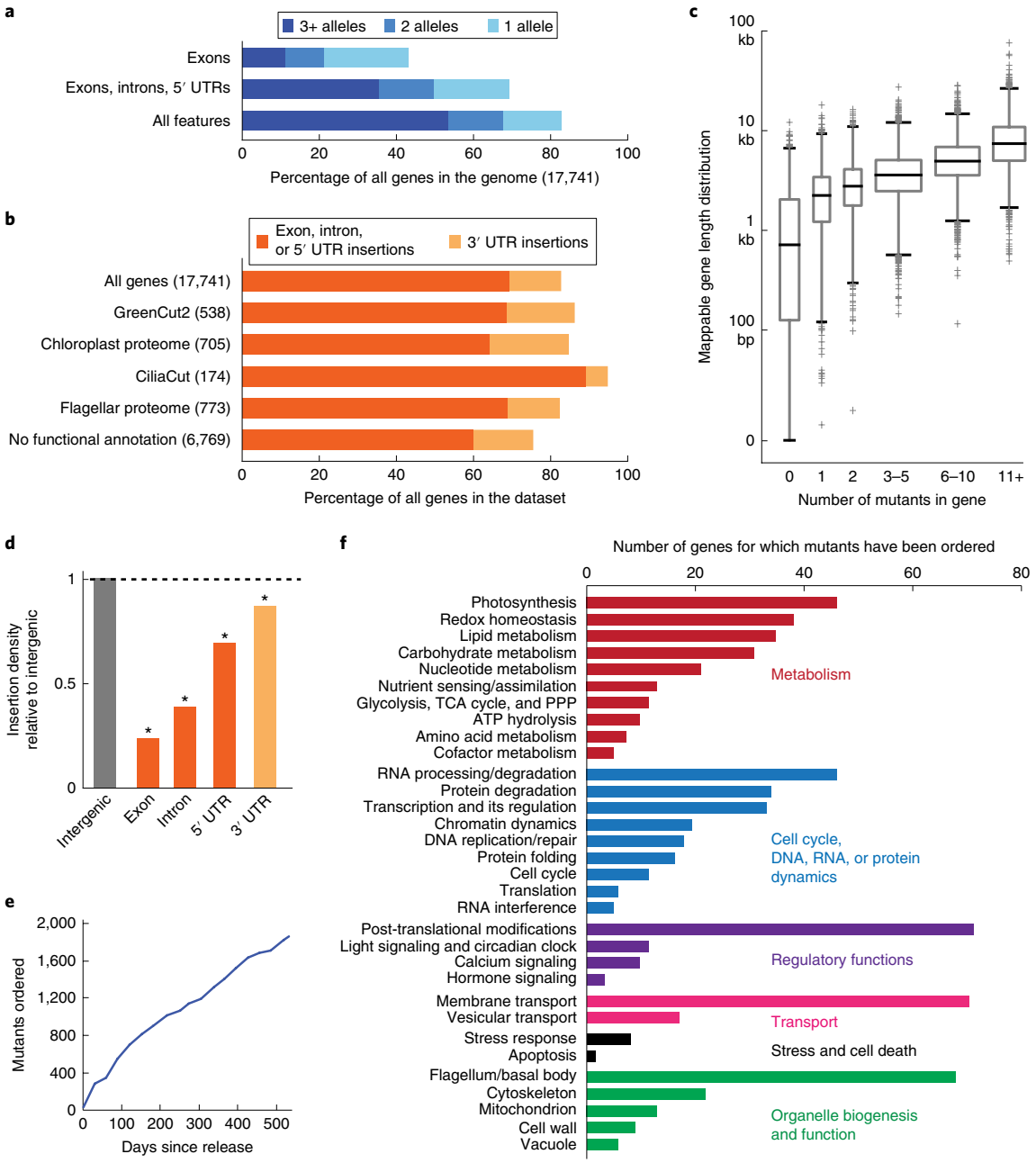
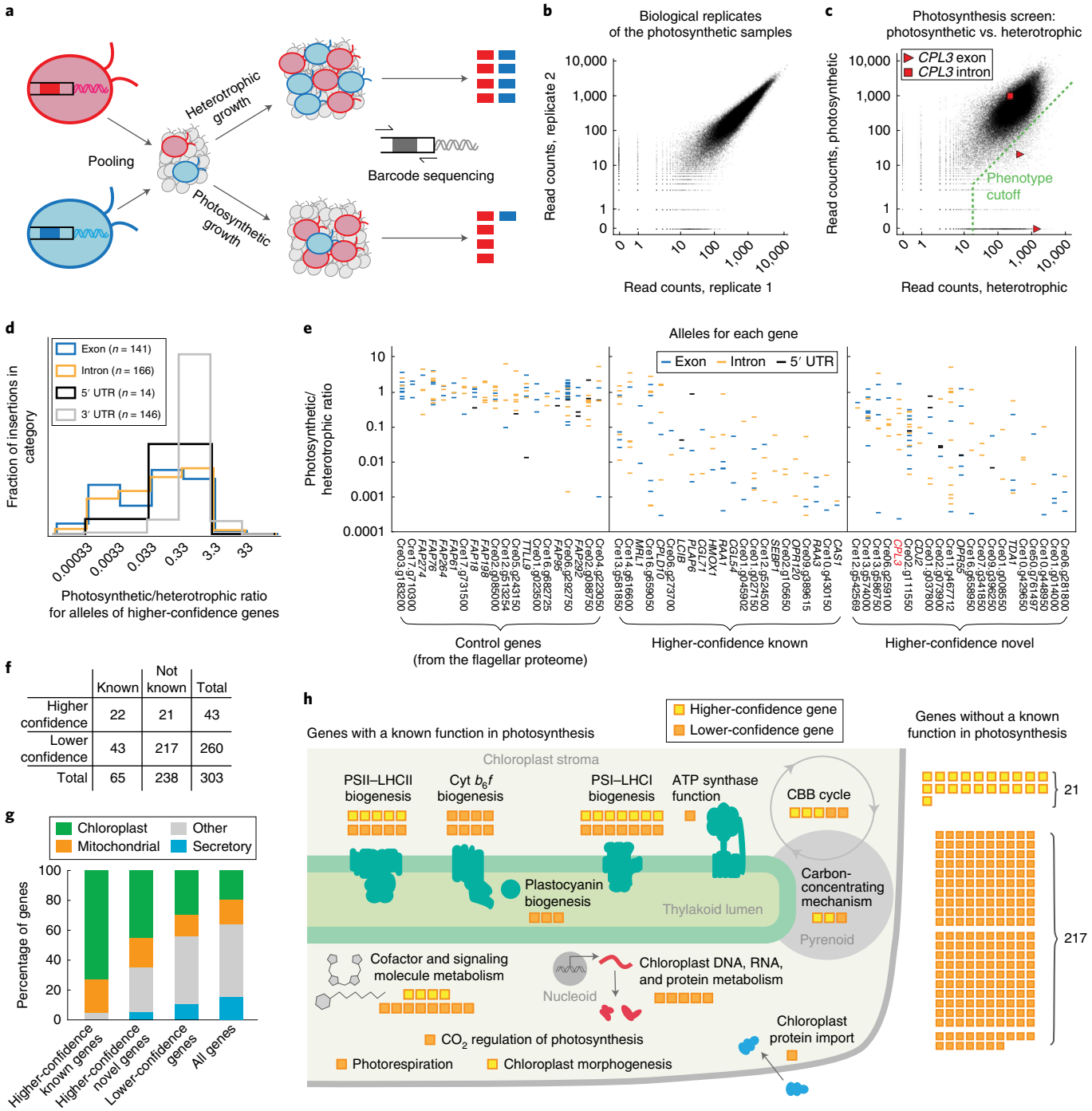


Fig. 2 | The library covers 83% of *Chlamydomonas* genes. a, 83% of all *Chlamydomonas* genes have one or more insertions in the library. **b**, In various functional groups, more than 75% of genes are represented by insertions in the library. **c**, The number of insertions per gene is roughly correlated with gene length. The middle bar of each box represents the median, box heights represent quartiles, the whiskers represent the first and ninety-ninth percentiles, and outliers are plotted as crosses. Box widths are proportional to the number of genes in each bin. **d**, Insertion density varies among different gene features, with the lowest density in exons. Asterisks denote a difference compared with intergenic insertions with $P < 10^{-78}$, with the chi-square test of independence. **e**, More than 1,800 mutants were distributed to approximately 200 laboratories around the world during the first 18 months of the library's availability. **f**, Distributed mutants are being used to study a variety of biological processes. Only genes with some functional annotation are shown.

Supplementary Fig. 5a,b). In total, we identified 3,109 mutants deficient in photosynthetic growth (Fig. 3c and Supplementary Note).

To identify genes with roles in photosynthesis, we developed a statistical analysis framework that leverages the presence of multiple alleles for many genes. This framework allows us to overcome several sources of false positives that have been difficult to account for with previous methods, including cases where the phenotype is not caused by the mapped disruption. For each gene, we counted the number of mutant alleles with and without a phenotype and evaluated the likelihood of obtaining these numbers by chance given the total number of mutants in the library that exhibited the phenotype (Supplementary Table 11 and Supplementary Note).

We identified 303 candidate photosynthesis genes on the basis of our statistical analysis. These genes are enriched for membership in a diurnally regulated photosynthesis-related transcriptional cluster[13] ($P < 1 \times 10^{-11}$), are enriched for upregulation upon dark-to-light transitions[14] ($P < 0.003$), and encode proteins enriched for predicted chloroplast localization ($P < 1 \times 10^{-8}$). As expected[15], the candidate



	Known	Not known	Total
Higher confidence	22	21	43
Lower confidence	43	217	260
Total	65	238	303

Fig. 3 | A high-throughput screen using the library identifies many genes with known roles in photosynthesis and many novel components. a, Unique barcodes allow screening of mutants in a pool. Mutants deficient in photosynthesis can be identified because their barcodes will be less abundant after photosynthetic growth than they are after heterotrophic growth. **b**, Biological replicates were highly reproducible, with a Spearman's correlation coefficient of 0.982. Each dot represents one barcode. See also Supplementary Fig. 5. **c**, The phenotype of each insertion was determined by comparing its read count under the photosynthetic and heterotrophic conditions. Insertions that fell below the phenotype cutoff were considered to result in a defect in photosynthesis. *CPL3* alleles are highlighted. **d**, Exon and intron insertions are most likely to show strong phenotypes, whereas 3′ UTR insertions rarely do. The plot is based on all insertions for the 43 higher-confidence genes. **e**, The photosynthetic/heterotrophic ratios for all alleles are shown for higher-confidence photosynthesis-screen hit genes and control genes. Each column is a gene; each horizontal bar is an allele. **f**, The 303 candidate genes were categorized on the basis of statistical confidence in this screen and whether they had a previously known function in photosynthesis (Supplementary Note). **g**, Known higher-confidence genes, novel higher-confidence genes, and lower-confidence genes are all enriched in predicted chloroplast-targeted proteins ($P < 0.011$). **h**, A schematic summary illustrating the numbers of candidate genes in each category (as in **f**) and the specific functions of the genes with a known role in processes related to photosynthesis.

genes also encode a disproportionate number of GreenCut2 proteins ($P < 1 \times 10^{-8}$), which are conserved among photosynthetic organisms but absent from non-photosynthetic organisms[1]: 32 GreenCut2 proteins are encoded by the 303 candidate genes (11%), as compared to ~3% of genes in the entire genome.

Photosynthesis occurs in two stages: the light reactions and carbon fixation. The light reactions convert solar energy into chemical energy and require the coordinated action of photosystem II (PSII), cytochrome b_6f, photosystem I (PSI), ATP synthase complexes, and a plastocyanin or cytochrome c_6 metalloprotein, as well as small-molecule cofactors[16]. PSII and PSI are each assisted by peripheral light-harvesting complexes (LHCs) known as LHCII and LHCI, respectively. Carbon fixation is performed by enzymes in the Calvin–Benson–Bassham (CBB) cycle, including the CO_2-fixing enzyme Rubisco. In addition, most eukaryotic algae have a mechanism to concentrate CO_2 around Rubisco to enhance its activity[17].

Sixty-five of the genes we identified encode proteins that were previously shown to have a role in photosynthesis or chloroplast function in *Chlamydomonas* or vascular plants (Fig. 3f). These include 3 PSII–LHCII subunits (PSBP1, PSBP2, and PSB27) and 7 PSII–LHCII biogenesis factors (CGL54, CPLD10, HCF136, LPA1, MBB1, TBC2, and Cre02.g105650), 2 cytochrome b_6f complex subunits (PETC and PETM) and 6 cytochrome b_6 biogenesis factors (CCB2, CCS5, CPLD43, CPLD49, MCD1, and MCG1), 5 PSI–LHCI subunits (LHCA3, LHCA7, PSAD, PSAE, and PSAL) and 9 PSI–LHCI biogenesis factors (CGL71, CPLD46, OPR120, RAA1, RAA2, RAA3, RAT2, Cre01.g045902, and Cre09.g389615), a protein required for ATP synthase function (PHT3), plastocyanin (PCY1) and 2 plastocyanin biogenesis factors (CTP2 and PCC1), 12 proteins involved in the metabolism of photosynthesis cofactors or signaling molecules (CHLD, CTH1, CYP745A1, DVR1, HMOX1, HPD2, MTF1, PLAP6, UROD3, Cre08.g358538, Cre13.g581850, and Cre16.g659050), 3 CBB cycle enzymes (FBP1, PRK1, and SEBP1), 2 Rubisco biogenesis factors (MRL1 and RMT2), and 3 proteins involved in the algal carbon-concentrating mechanism (CAH3, CAS1, and LCIB), as well as proteins that have a role in photorespiration (GSF1), CO_2 regulation of photosynthesis (Cre02.g146851), chloroplast morphogenesis (Cre14.g616600), chloroplast protein import (SDR17), and chloroplast DNA, RNA, and protein metabolism (DEG9, MSH1, MSRA1, TSM2, and Cre01.g010864) (Fig. 3h and Supplementary Table 12). We caution that not all genes previously demonstrated to be required for photosynthetic growth were detectable by this approach, especially the ones with paralogous copies in the genome, such as *RBCS1* and *RBCS2*, which encode the small subunit of Rubisco[18]. Nonetheless, the large number of known factors recovered in our screen is a testament to the power of this approach.

In addition to recovering these 65 genes with known roles in photosynthesis, our analysis identified 238 candidate genes with no previously reported role in photosynthesis. These 238 genes represent a rich set of targets to better understand photosynthesis. Because our screen likely yielded some false positives, we divided all genes into 'higher-confidence' ($P < 0.0011$; false-discovery rate (FDR) < 0.27) and 'lower-confidence' genes on the basis of the number of alleles that supported each gene's involvement in photosynthesis (Fig. 3d–f, Tables 1 and 2, and Supplementary Note). The 21 higher-confidence genes with no previously reported role in photosynthesis are enriched in chloroplast localization (9/21, $P < 0.011$; Fig. 3g) and transcriptional upregulation during dark-to-light transition (5/21, $P < 0.005$), similarly to the known photosynthesis genes. Thus, these 21 higher-confidence genes are particularly high-priority targets for the field to pursue.

Functional annotations for 15 of the 21 higher-confidence genes suggest that these genes could have roles in regulation of photosynthesis, photosynthetic metabolism, and biosynthesis of the photosynthetic machinery. Seven of the genes likely have roles in regulation of photosynthesis: *GEF1* encodes a voltage-gated channel, Cre01.g008550 and Cre02.g111550 encode putative protein kinases, *CPL3* encodes a predicted protein phosphatase, the protein encoded by *TRX21* contains a thioredoxin domain, Cre12.g542569 encodes a putative glutamate receptor, and the protein encoded by Cre13.g586750 contains a predicted nuclear importin domain. Six of the genes are likely involved in photosynthetic metabolism: the *Arabidopsis thaliana* homolog of Cre10.g448950 modulates sucrose and starch accumulation[19], the protein encoded by Cre11.g467712 contains a starch-binding domain, Cre02.g073900 encodes a putative carotenoid dioxygenase, *VTE5* encodes a putative phosphatidate cytidylyltransferase, Cre10.g429650 encodes a putative alpha/beta hydrolase, and the protein encoded by Cre50.g761497 contains a magnesium transporter domain. Finally, two of the genes are likely to have roles in the biogenesis and function of photosynthesis machinery: the protein encoded by *EIF2* has a translation initiation factor domain and *CDJ2* encodes a protein with a chloroplast DnaJ domain. Future characterization of these genes by the community is likely to yield fundamental insights into photosynthesis.

As an illustration of the value of the genes identified in this screen, we sought to explore the specific function of one of the higher-confidence candidate genes, *CPL3* (conserved in plant lineage 3; Cre03.g185200, also known as *MPA6*), which encodes a putative protein phosphatase (Fig. 4a and Supplementary Fig. 6). Many proteins in the photosynthetic apparatus are phosphorylated, but the role and regulation of these phosphorylation events are poorly understood[20]. An insertion junction that mapped to the 3′ UTR of *CPL3* was previously found in a collection of acetate-requiring mutants, although it was not determined whether this mutation caused the phenotype[15]. In our screen, three mutants with insertion junctions in *CPL3* exons or introns exhibited a deficiency in photosynthetic growth (Fig. 3c and Supplementary Table 13). We chose to examine one allele (LMJ.RY0402.153647, referred to hereafter as *cpl3*; Fig. 4a and Supplementary Fig. 6a) for phenotypic confirmation, genetic complementation, and further studies.

Consistent with the pooled growth data, the *cpl3* mutant showed a severe defect in photosynthetic growth on agar, which was rescued under heterotrophic conditions (Fig. 4b). We confirmed that the *CPL3* gene was disrupted in the *cpl3* mutant and found that complementation with a wild-type copy of the *CPL3* gene rescued the phenotype, demonstrating that the mutation in *CPL3* was the cause of the growth defect of the mutant (Supplementary Note and Supplementary Fig. 6a–d).

We then examined photosynthetic performance, morphology of the chloroplast, and composition of photosynthetic pigments and proteins in *cpl3*. The photosynthetic electron transport rate was decreased under all light intensities, suggesting a defect in the photosynthetic machinery (Fig. 4c). The chloroplast morphology of *cpl3* appeared similar to that of the wild type on the basis of chlorophyll fluorescence microscopy (Supplementary Fig. 7a). However, we observed a lower chlorophyll *a*/chlorophyll *b* ratio in *cpl3* than in the wild type (Supplementary Fig. 7b), which suggests a defect in the accumulation or composition of the protein–pigment complexes involved in the light reactions[21]. By using whole-cell proteomics, we found that *cpl3* was deficient in accumulation of all detectable subunits of the chloroplast ATP synthase (ATPC, ATPD, ATPG, AtpA, AtpB, AtpE, and AtpF), some subunits of PSII (D1, D2, CP43, CP47, PsbE, and PsbH), and some subunits of PSI (PsaA and PsaB) (FDR < 0.31 for each subunit; Fig. 4d,f and Supplementary Table 14). We confirmed these findings with western blots for CP43, PsaA, and ATPC (Fig. 4e and Supplementary Fig. 7c). Our results indicate that CPL3 is required for normal accumulation of thylakoid protein complexes (PSII, PSI, and ATP synthase) involved in the light reactions of photosynthesis.

Our finding that 21 of the 43 higher-confidence photosynthesis genes identified were uncharacterized suggests that nearly half of

Table 1 | Higher-confidence genes from the photosynthesis screen with a previously known role in photosynthesis

Category	Gene	Definition or description in Phytozome[12]	PredAlgo[a]	Alleles in two replicates			*Arabidopsis* homolog[e]	Reference and corresponding organism(s)
				+[b]	−[c]	FDR[d]		
Calvin–Benson–Bassham cycle	Cre03.g185550 (*SEBP1*, *SBP1*)	Sedoheptulose-1,7-bisphosphatase	C	3	0	0.021	AT3G55800.1 (*SBPASE*)	*Arabidopsis*[29]
				3	0	0.018		
	Cre12.g524500 (*RMT2*)	Rubisco small subunit *N*-methyltransferase	O	3	0	0.021	AT3G07670.1	*Pisum*[30]
				3	0	0.018		
	Cre06.g298300 (*MRL1*, *PPR2*)	Pentatricopeptide-repeat protein, stabilizes *rbcL* mRNA	C	1	1	1.000	AT4G34830.1 (*MRL1*)	*Chlamydomonas* and *Arabidopsis*[31]
				2	0	0.239		
Carbon-concentrating mechanism	Cre12.g497300 (*CAS1*, *TEF2*)	Rhodanese-like calcium-sensing receptor	C	2	0	0.260	AT5G23060.1 (*CaS*)	*Chlamydomonas*[32]
				2	0	0.239		
	Cre10.g452800 (*LCIB*)	Low-CO_2-inducible protein	C	2	0	0.260	–	*Chlamydomonas*[33]
				1	1	1.000		
Chloroplast and thylakoid morphogenesis	Cre14.g616600	–	M	4	3	0.021	AT1G03160.1 (*FZL*)	*Arabidopsis*[34]
				4	3	0.018		
Cofactor and signaling molecule metabolism	Cre13.g581850	–	M	5	5	0.010	AT4G31390.1	*Arabidopsis*[35]
				2	8	1.000		
	Cre10.g423500 (*HMOX1*, *HMO1*)	Heme oxygenase	C	3	0	0.021	AT1G69720.1 (*HO3*)	*Chlamydomonas*[14]
				3	0	0.018		
	Cre03.g188700 (*PLAP6*, *PLP6*)	Plastid lipid-associated protein, fibrillin	C	3	1	0.070	AT5G09820.2	*Arabidopsis*[36]
				3	1	0.056		
	Cre16.g659050	–	C	4	6	0.098	AT1G68890.1	*Chlamydomonas*[37]
				4	6	0.075		
PSI protein synthesis and assembly	Cre12.g524300 (*CGL71*)	Predicted protein	C	2	0	0.260	AT1G22700.1	*Synechocystis*[38], *Arabidopsis*[39], and *Chlamydomonas*[40]
				2	0	0.239		
	Cre01.g045902	–	C	1	1	1.000	AT3G24430.1 (*HCF101*)	*Arabidopsis*[41,42]
				2	0	0.239		
PSI RNA splicing and stabilization	Cre09.g389615	–	M	5	0	0.0002	AT3G17040.1 (*HCF107*)	*Chlamydomonas*[43] and *Arabidopsis*[42,44 f]
				5	0	0.0002		
	Cre01.g027150 (*CPLD46*, *HEL5*)	DEAD/DEAH-box helicase	M	5	1	0.0004	AT1G70070.1 (*EMB25*, *ISE2*,	*Arabidopsis*[45]
				5	1	0.0003		
	Cre09.g394150 (*RAA1*)	–	M	5	1	0.0004	–	*Chlamydomonas*[46]
				5	1	0.0003		
	Cre12.g531050 (*RAA3*)	*psaA* mRNA maturation factor 3	C	3	0	0.021	–	*Chlamydomonas*[47]
				3	0	0.018		
	Cre10.g440000 (*OPR120*)	–	C	2	0	0.260	–	*Chlamydomonas*[48,49]
				2	0	0.239		
PSII protein synthesis and assembly	Cre13.g578650 (*CPLD10*, *NUOAF5*)	Similar to complex I intermediate-associated	C	3	3	0.260	AT1G16720.1 (*HCF173*)	*Arabidopsis*[42,50,51]
				3	3	0.208		
	Cre02.g073850 (*CGL54*)	Predicted protein	C	2	0	0.260	AT1G05385.1 (*LPA19*, *Psb27-H1*)	*Arabidopsis*[52]
				2	0	0.239		
	Cre02.g105650	–	C	2	0	0.260	AT5G51545.1 (*LPA2*)	*Arabidopsis*[53]
				2	0	0.239		
	Cre06.g273700 (*HCF136*)	–	C	2	0	0.260	AT5G23120.1 (*HCF136*)	*Arabidopsis*[42] and *Synechocystis*[54]
				1	1	1.000		
	Cre10.g430150 (*LPA1*, *REP27*)	–	C	2	0	0.260	AT1G02910.1 (*LPA1*)	*Arabidopsis*[55]
				1	1	1.000		

[a]Prediction of protein localization by PredAlgo[56]: C, chloroplast; M, mitochondrion; SP, secretory pathway; O, other. [b]The number of exon, intron, or 5′ UTR mutant alleles for the gene that satisfied our requirement of a minimum of 50 reads and showed at least ten times fewer normalized reads in the sample grown in TP in the light than in the sample grown in TAP in the dark. [c]The number of exon, intron, or 5′ UTR mutant alleles for the gene that satisfied our minimum read count requirement but not the requirement for at least tenfold depletion in the TP−light condition. [d]The FDR for the gene in comparison to all alleles for all genes (Supplementary Note). [e]*Arabidopsis* homolog, obtained from the 'best_arabidopsis_TAIR10_hit_name' field in Phytozome[12]. [f]AT3G17040.1 is required for functional PSII in *Arabidopsis*, whereas Cre09.g389615 was shown to be involved in PSI accumulation in *Chlamydomonas*.

Table 2 | Higher-confidence genes from the photosynthesis screen with no previously known role in photosynthesis

Gene	Definition or description in Phytozome	PredAlgo	Alleles in two replicates			*Arabidopsis* homolog
			+	-	FDR	
Cre01.g008550	Serine/threonine kinase related	O	2	0	0.260	AT1G73450.1
			1	1	1.000	
Cre01.g014000	-	C	3	0	0.021	-
			3	0	0.018	
Cre01.g037800 (*TRX21*)	ATP-binding protein; thioredoxin domain	O	3	3	0.260	AT2G18990.1 (*TXND9*)
			1	5	1.000	
Cre02.g073900	All-*trans*-10′-*apo*-β-carotenal 13,14–cleaving dioxygenase	C	3	1	0.070	AT4G32810.1 (*ATCCD8, CCD8, MAX4*)
			3	1	0.056	
Cre02.g111550	Serine/threonine kinase related	SP	10	8	$<10^{-6}$	AT4G24480.1
			6	12	0.015	
Cre03.g185200 (*CPL3, MPA6*)	Metallophosphoesterase/metallo-dependent phosphatase	C	3	4	0.260	AT1G07010.1
			3	4	0.239	
Cre06.g259100	-	C	1	4	1.000	-
			3	2	0.117	
Cre06.g281800	Domain of unknown function (DUF1995)	C	3	0	0.021	-
			3	0	0.018	
Cre07.g316050 (*CDJ2*)	Chloroplast DnaJ-like protein	M	2	0	0.260	AT5G59610.1
			1	1	1.000	
Cre07.g341850 (*EIF2, INFB*)	Translation initiation factor IF-2, chloroplastic	C	2	0	0.260	AT1G17220.1 (*FUG1*)
			2	0	0.239	
Cre08.g358350 (*TDA1, OPR34*)	Fast leucine-rich domain containing[a]	C	3	2	0.152	-
			3	2	0.117	
Cre09.g396250 (*VTE5*)	Phosphatidate cytidylyltransferase	SP	2	0	0.260	AT5G04490.1 (*VTE5*)
			1	1	1.000	
Cre10.g429650	Alpha/beta hydrolase family (Abhydrolase_5)	O	2	0	0.260	-
			1	1	1.000	
Cre10.g448950	Nocturnin	C	1	1	1.000	AT3G58560.1
			2	0	0.239	
Cre11.g467712	Structural maintenance of chromosomes smc family member; starch-binding domain	M	7	7	0.0003	AT5G05180.1
			7	7	0.0003	
Cre12.g542569	Ionotropic glutamate receptor	O	0	2	1.000	AT1G05200.1 (*ATGLR3.4, GLR3.4, GLUR3*)
			2	0	0.239	
Cre13.g566400 (*OPR55*)	Fast leucine-rich domain containing[a]	M	4	2	0.018	-
			4	2	0.015	
Cre13.g574000 (*GEF1, CLV1*)	Voltage-gated chloride channel	O	1	11	1.000	AT5G26240.1 (*ATCLC-D, CLC-D*)
			4	8	0.144	
Cre13.g586750	Transportin 3 and importin	O	3	4	0.260	AT5G62600.1
			2	5	1.000	
Cre16.g658950	-	C	2	2	0.909	-
			3	1	0.056	
Cre50.g761497	Magnesium transporter *mrs2* homolog, mitochondrial	M	2	0	0.260	AT5G22830.1 (*ATMGT10, GMN10, MGT10, MRS2-11*)
			2	0	0.239	

[a]The annotation of 'fast leucine-rich domain containing' cannot be confirmed by BLASTP analysis at NCBI[57].

the genes required for photosynthesis remain to be characterized. This finding is notable considering that genetic studies on photosynthesis extend back to the 1950s[22]. Our validation of the role of *CPL3* in photosynthesis illustrates the value of the uncharacterized genes identified in this study as a rich set of candidates for the community to pursue.

More broadly, it is our hope that the mutant resource presented here will serve as a powerful complement to newly developed

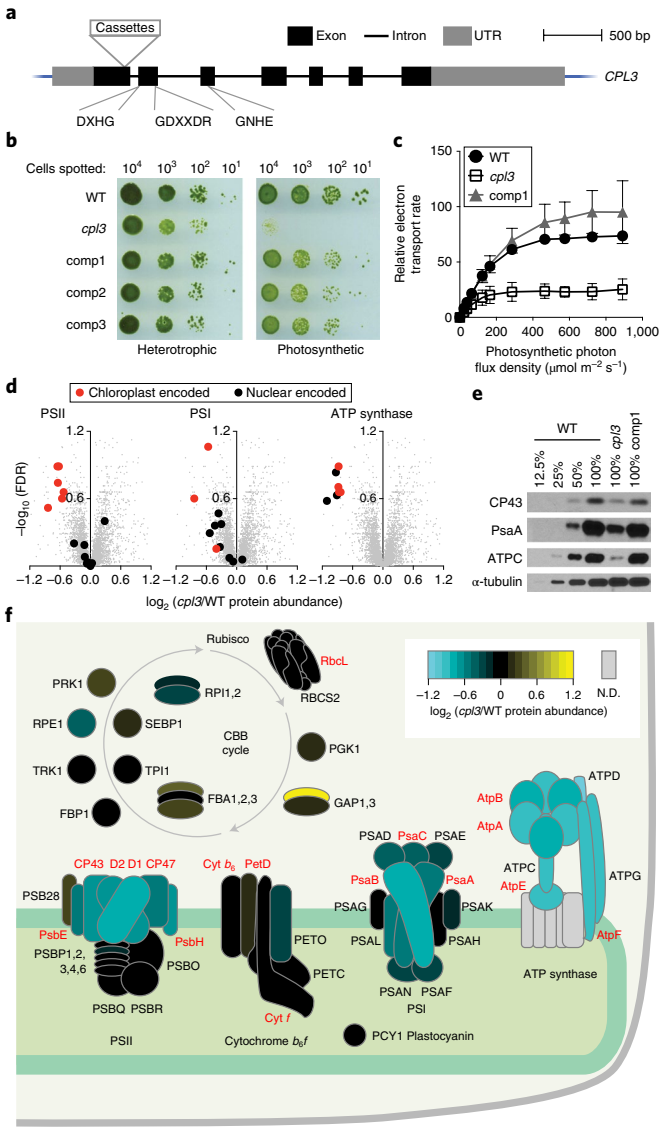


Fig. 4 | CPL3 is required for photosynthetic growth and accumulation of photosynthetic protein complexes in the thylakoid membranes. **a**, The *cpl3* mutant contains cassettes inserted in the first exon of *CPL3*. The locations of conserved protein phosphatase motifs are indicated (Supplementary Fig. 6e). **b**, *cpl3* is deficient in growth under photosynthetic conditions and can be rescued upon complementation with the wild-type *CPL3* gene (comp1–comp3 represent three independent complemented lines). WT, wild type. **c**, *cpl3* has a lower relative photosynthetic electron transport rate than the wild-type strain and comp1. Error bars, s.d. ($n = 3$ for WT and comp1; $n = 7$ for *cpl3*). **d**, Whole-cell proteomics (Supplementary Table 14) indicates that *cpl3* is deficient in accumulation of PSII, PSI, and the chloroplast ATP synthase. Each dot represents one *Chlamydomonas* protein; PSII, PSI, and ATP synthase subunits are highlighted as black and red symbols. **e**, Western blots showing that CPL3 is required for normal accumulation of the PSII subunit CP43, the PSI subunit PsaA, and the chloroplast ATP synthase subunit ATPC. α-tubulin was used as a loading control. To facilitate estimation of protein abundance in the *cpl3* and comp1 samples, 50%, 25%, and 12.5% dilutions of the wild-type sample were loaded. See also Supplementary Fig. 7c. **f**, A heat map showing the protein abundance of subunits in the light reaction protein complexes and enzymes in the CBB cycle in *cpl3* relative to the wild type based on proteomics data. Depicted subunits that were not detected by proteomics are filled in gray (N.D.). Nuclear- and chloroplast-encoded proteins are labeled in black and red, respectively. A stack of horizontal ovals indicates different isoforms for the same enzyme, such as FBA1, FBA2, and FBA3.

gene-editing techniques[23–28] and that, together, these tools will help the research community generate fundamental insights in a wide range of fields, from organelle biogenesis and function to organism–environment interactions.

URLs. CLiP website for mutant distribution, https://www.chlamylibrary.org/; Jonikas Lab GitHub repositories of scripts, https://github.com/Jonikas-Lab/Li-Patena-2019/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0370-6.

Received: 20 May 2018; Accepted: 8 February 2019;
Published online: 18 March 2019

References

1. Karpowicz, S. J., Prochnik, S. E., Grossman, A. R. & Merchant, S. S. The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. *J. Biol. Chem.* **286**, 21427–21439 (2011).
2. Krishnakumar, V. et al. Araport: the *Arabidopsis* information portal. *Nucleic Acids Res.* **43**, D1003–D1009 (2015).
3. Levine, R. P. Genetic control of photosynthesis in *Chlamydomonas reinhardi*. *Proc. Natl Acad. Sci. USA* **46**, 972–978 (1960).
4. Gutman, B. L. & Niyogi, K. K. *Chlamydomonas* and *Arabidopsis*. A dynamic duo. *Plant Physiol.* **135**, 607–610 (2004).
5. Harris, E. H., Stern, D. B. & Witman, G. B. *The Chlamydomonas Sourcebook* (Academic Press, 2009).
6. Rochaix, J. D. *Chlamydomonas reinhardtii* as the photosynthetic yeast. *Annu. Rev. Genet.* **29**, 209–230 (1995).
7. Li, J. B. et al. Comparative genomics identifies a flagellar and basal body proteome that includes the *BBS5* human disease gene. *Cell* **117**, 541–552 (2004).
8. Silflow, C. D. & Lefebvre, P. A. Assembly and motility of eukaryotic cilia and flagella: lessons from *Chlamydomonas reinhardtii*. *Plant Physiol.* **127**, 1500–1507 (2001).
9. Li, X. et al. An indexed, mapped mutant library enables reverse genetics studies of biological processes in *Chlamydomonas reinhardtii*. *Plant Cell* **28**, 367–387 (2016).
10. Terashima, M., Specht, M. & Hippler, M. The chloroplast proteome: a survey from the *Chlamydomonas reinhardtii* perspective with a focus on distinctive features. *Curr. Genet.* **57**, 151–168 (2011).
11. Pazour, G. J., Agrin, N., Leszyk, J. & Witman, G. B. Proteomic analysis of a eukaryotic cilium. *J. Cell Biol.* **170**, 103–113 (2005).
12. Merchant, S. S. et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–251 (2007).
13. Zones, J. M., Blaby, I. K., Merchant, S. S. & Umen, J. G. High-resolution profiling of a synchronized diurnal transcriptome from *Chlamydomonas reinhardtii* reveals continuous cell and metabolic differentiation. *Plant Cell* **27**, 2743–2769 (2015).
14. Duanmu, D. et al. Retrograde bilin signaling enables *Chlamydomonas* greening and phototrophic survival. *Proc. Natl Acad. Sci. USA* **110**, 3621–3626 (2013).
15. Dent, R. M. et al. Large-scale insertional mutagenesis of *Chlamydomonas* supports phylogenomic functional prediction of photosynthetic genes and analysis of classical acetate-requiring mutants. *Plant J.* **82**, 337–351 (2015).
16. Allen, J. F., de Paula, W. B., Puthiyaveetil, S. & Nield, J. A structural phylogenetic map for chloroplast photosynthesis. *Trends Plant Sci.* **16**, 645–655 (2011).
17. Giordano, M., Beardall, J. & Raven, J. A. CO_2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu. Rev. Plant Biol.* **56**, 99–131 (2005).
18. Goldschmidt-Clermont, M. & Rahire, M. Sequence, evolution and differential expression of the two genes encoding variant small subunits of ribulose bisphosphate carboxylase/oxygenase in *Chlamydomonas reinhardtii*. *J. Mol. Biol.* **191**, 421–432 (1986).
19. Suzuki, Y., Arae, T., Green, P. J., Yamaguchi, J. & Chiba, Y. AtCCR4a and AtCCR4b are involved in determining the poly(A) length of granule-bound starch synthase 1 transcript and modulating sucrose and starch metabolism in *Arabidopsis thaliana*. *Plant Cell Physiol.* **56**, 863–874 (2015).
20. Wang, H. et al. The global phosphoproteome of *Chlamydomonas reinhardtii* reveals complex organellar phosphorylation in the flagella and thylakoid membrane. *Mol. Cell. Proteomics* **13**, 2337–2353 (2014).
21. Bassi, R., Soen, S. Y., Frank, G., Zuber, H. & Rochaix, J. D. Characterization of chlorophyll *a*/*b* proteins of photosystem I from *Chlamydomonas reinhardtii*. *J. Biol. Chem.* **267**, 25714–25721 (1992).

22. Sager, R. & Zalokar, M. Pigments and photosynthesis in a carotenoid-deficient mutant of *Chlamydomonas*. *Nature* **182**, 98–100 (1958).
23. Baek, K. et al. DNA-free two-gene knockout in *Chlamydomonas reinhardtii* via CRISPR–Cas9 ribonucleoproteins. *Sci. Rep.* **6**, 30620 (2016).
24. Jiang, W., Brueggeman, A. J., Horken, K. M., Plucinak, T. M. & Weeks, D. P. Successful transient expression of Cas9 and single guide RNA genes in *Chlamydomonas reinhardtii*. *Eukaryot. Cell* **13**, 1465–1469 (2014).
25. Shin, S. E. et al. CRISPR/Cas9-induced knockout and knock-in mutations in *Chlamydomonas reinhardtii*. *Sci. Rep.* **6**, 27810 (2016).
26. Slaninová, M., Hroššová, D., Vlček, D. & Wolfgang, W. Is it possible to improve homologous recombination in *Chlamydomonas reinhardtii*? *Biologia* **63**, 941–946 (2008).
27. Greiner, A. et al. Targeting of photoreceptor genes in *Chlamydomonas reinhardtii* via zinc-finger nucleases and CRISPR/Cas9. *Plant Cell* **29**, 2498–2518 (2017).
28. Ferenczi, A., Pyott, D. E., Xipnitou, A. & Molnar, A. Efficient targeted DNA editing and replacement in *Chlamydomonas reinhardtii* using Cpf1 ribonucleoproteins and single-stranded DNA. *Proc. Natl Acad. Sci. USA* **114**, 13567–13572 (2017).
29. Liu, X. L., Yu, H. D., Guan, Y., Li, J. K. & Guo, F. Q. Carbonylation and loss-of-function analyses of SBPase reveal its metabolic interface role in oxidative stress, carbon assimilation, and multiple aspects of growth and development in *Arabidopsis*. *Mol. Plant* **5**, 1082–1099 (2012).
30. Klein, R. R. & Houtz, R. L. Cloning and developmental expression of pea ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit *N*-methyltransferase. *Plant Mol. Biol.* **27**, 249–261 (1995).
31. Johnson, X. et al. MRL1, a conserved pentatricopeptide repeat protein, is required for stabilization of *rbcL* mRNA in *Chlamydomonas* and *Arabidopsis*. *Plant Cell* **22**, 234–248 (2010).
32. Wang, L. et al. Chloroplast-mediated regulation of CO_2-concentrating mechanism by Ca^{2+}-binding protein CAS in the green alga *Chlamydomonas reinhardtii*. *Proc. Natl Acad. Sci. USA* **113**, 12586–12591 (2016).
33. Wang, Y. & Spalding, M. H. An inorganic carbon transport system responsible for acclimation specific to air levels of CO_2 in *Chlamydomonas reinhardtii*. *Proc. Natl Acad. Sci. USA* **103**, 10110–10115 (2006).
34. Gao, H., Sage, T. L. & Osteryoung, K. W. FZL, an FZO-like protein in plants, is a determinant of thylakoid and chloroplast morphology. *Proc. Natl Acad. Sci. USA* **103**, 6759–6764 (2006).
35. Martinis, J. et al. ABC1K1/PGR6 kinase: a regulatory link between photosynthetic activity and chloroplast metabolism. *Plant J.* **77**, 269–283 (2014).
36. Kim, E. H., Lee, Y. & Kim, H. U. Fibrillin 5 is essential for plastoquinone-9 biosynthesis by binding to solanesyl diphosphate synthases in *Arabidopsis*. *Plant Cell* **27**, 2956–2971 (2015).
37. Lefebvre-Legendre, L. et al. Loss of phylloquinone in *Chlamydomonas* affects plastoquinone pool size and photosystem II synthesis. *J. Biol. Chem.* **282**, 13250–13263 (2007).
38. Wilde, A., Lunser, K., Ossenbuhl, F., Nickelsen, J. & Borner, T. Characterization of the cyanobacterial ycf37: mutation decreases the photosystem I content. *Biochem. J.* **357**, 211–216 (2001).
39. Stockel, J., Bennewitz, S., Hein, P. & Oelmuller, R. The evolutionarily conserved tetratrico peptide repeat protein pale yellow green7 is required for photosystem I accumulation in *Arabidopsis* and copurifies with the complex. *Plant Physiol.* **141**, 870–878 (2006).
40. Heinnickel, M. et al. Tetratricopeptide repeat protein protects photosystem I from oxidative disruption during assembly. *Proc. Natl Acad. Sci. USA* **113**, 2774–2779 (2016).
41. Lezhneva, L., Amann, K. & Meurer, J. The universally conserved HCF101 protein is involved in assembly of [4Fe-4S]-cluster-containing complexes in *Arabidopsis thaliana* chloroplasts. *Plant J.* **37**, 174–185 (2004).
42. Meurer, J., Meierhoff, K. & Westhoff, P. Isolation of high-chlorophyll-fluorescence mutants of *Arabidopsis thaliana* and their characterisation by spectroscopy, immunoblotting and northern hybridisation. *Planta* **198**, 385–396 (1996).
43. Douchi, D. et al. A nucleus-encoded chloroplast phosphoprotein governs expression of the photosystem I subunit PsaC in *Chlamydomonas reinhardtii*. *Plant Cell* **28**, 1182–1199 (2016).
44. Felder, S. et al. The nucleus-encoded *HCF107* gene of *Arabidopsis* provides a link between intercistronic RNA processing and the accumulation of translation-competent *psbH* transcripts in chloroplasts. *Plant Cell* **13**, 2127–2141 (2001).
45. Carlotto, N. et al. The chloroplastic DEVH-box RNA helicase INCREASED SIZE EXCLUSION LIMIT 2 involved in plasmodesmata regulation is required for group II intron splicing. *Plant Cell Environ.* **39**, 165–173 (2016).
46. Perron, K., Goldschmidt-Clermont, M. & Rochaix, J. D. A factor related to pseudouridine synthases is required for chloroplast group II intron trans-splicing in *Chlamydomonas reinhardtii*. *EMBO J.* **18**, 6481–6490 (1999).
47. Rivier, C., Goldschmidt-Clermont, M. & Rochaix, J. D. Identification of an RNA–protein complex involved in chloroplast group II intron trans-splicing in *Chlamydomonas reinhardtii*. *EMBO J.* **20**, 1765–1773 (2001).
48. Jacobs, J. et al. Identification of a chloroplast ribonucleoprotein complex containing trans-splicing factors, intron RNA, and novel components. *Mol. Cell. Proteomics* **12**, 1912–1925 (2013).
49. Marx, C., Wunsch, C. & Kuck, U. The octatricopeptide repeat protein Raa8 is required for chloroplast trans splicing. *Eukaryot. Cell* **14**, 998–1005 (2015).
50. Link, S., Engelmann, K., Meierhoff, K. & Westhoff, P. The atypical short-chain dehydrogenases HCF173 and HCF244 are jointly involved in translational initiation of the *psbA* mRNA of *Arabidopsis*. *Plant Physiol.* **160**, 2202–2218 (2012).
51. Schult, K. et al. The nuclear-encoded factor HCF173 is involved in the initiation of translation of the *psbA* mRNA in *Arabidopsis thaliana*. *Plant Cell* **19**, 1329–1346 (2007).
52. Wei, L. et al. LPA19, a Psb27 homolog in *Arabidopsis thaliana*, facilitates D1 protein precursor processing during PSII biogenesis. *J. Biol. Chem.* **285**, 21391–21398 (2010).
53. Ma, J. et al. LPA2 is required for efficient assembly of photosystem II in *Arabidopsis thaliana*. *Plant Cell* **19**, 1980–1993 (2007).
54. Komenda, J. et al. The cyanobacterial homologue of HCF136/YCF48 is a component of an early photosystem II assembly complex and is important for both the efficient assembly and repair of photosystem II in *Synechocystis* sp. PCC 6803. *J. Biol. Chem.* **283**, 22390–22399 (2008).
55. Peng, L. et al. LOW PSII ACCUMULATION1 is involved in efficient assembly of photosystem II in *Arabidopsis thaliana*. *Plant Cell* **18**, 955–969 (2006).
56. Tardif, M. et al. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.* **29**, 3625–3639 (2012).
57. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Acknowledgements

We thank O. Vallon for helpful discussions; M. Cahn and G. Huntress for developing and improving the CLiP website; X. Ji at the Stanford Functional Genomics Facility and Z. Weng at the Stanford Center for Genomics and Personalized Medicine for deep sequencing services; A. Itakura for help in library pooling; S. Ghosh, K. Mendoza, M. LaVoie, L. Galhardo, X. Li, Y. Wang, and Q. Chen for technical assistance; K. Barton, W. Briggs, and Z.-Y. Wang for providing lab space; J. Ecker, L. Freeman Rosenzweig, and M. Kafri for constructive suggestions on the manuscript; and the Princeton Mass Spectrometry Facility for proteomics services. This project was supported by a grant from the National Science Foundation (MCB-1146621) awarded to M.C.J. and A.R.G., grants from the National Institutes of Health (DP2-GM-119137) and the Simons Foundation and Howard Hughes Medical Institute (55108535) awarded to M.C.J., a German Academic Exchange Service (DAAD) research fellowship to F.F., Simons Foundation fellowships of the Life Sciences Research Foundation to R.E.J. and J.V.-B., an EMBO long-term fellowship (ALTF 1450-2014 and ALTF 563-2013) to J.V.-B and S.R., a Swiss National Science Foundation Advanced PostDoc Mobility Fellowship (P2GEP3_148531) to S.R., and a Westlake University startup fund to X.L.

Author contributions

X.L. developed the method for generating barcoded cassettes. R.Y. and S.R.B. optimized the mutant generation protocol. R.Y., N.I., and X.L. generated the library. J.M.R., N.I., A.G., and R.Y. maintained, consolidated, and cryopreserved the library. X.L. developed the barcode sequencing method. N.I., X.L., R.Y., and W.P. performed combinatorial pooling and super-pool barcode sequencing. X.L. performed LEAP-Seq. W.P. developed the mutant mapping data analysis pipeline and performed data analyses for barcode sequencing and LEAP-Seq. W.P. analyzed insertion coverage and hot- and coldspots. R.Z. and J.M.R. performed insertion verification PCRs and Southern blots. F.F., R.E.J., and J.V.-B. developed the library screening protocol. F.F., J.V.-B., and X.L. performed the photosynthesis mutant screen and barcode sequencing. R.E.J. and W.P. developed data analysis methods and implemented them for the photosynthesis screen. X.L. and T.M.W. annotated the hits from the photosynthesis screen. X.L., J.M.R., and S.R. performed growth analysis, molecular characterizations, and complementation of *cpl3*. S.S. and T.M.W. performed physiological characterizations of *cpl3*. M.T.M. and S.S. performed western blots on the photosynthetic protein complexes. M.T.M. performed microscopy on *cpl3*. X.L., W.P., and T.S. performed proteomic analyses. M.L. and P.A.L. maintained, cryopreserved, and distributed mutants at the Chlamydomonas Resource Center. X.L., W.P., A.R.G., and M.C.J. wrote the manuscript with input from all authors. M.C.J. and A.R.G. conceived and guided the research and obtained funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0370-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Generation of the indexed and barcoded mutant library. A three-step pipeline was developed for generation of an indexed, barcoded library of insertional mutants in *Chlamydomonas* (Fig. 1b and Supplementary Fig. 1).

To generate mutants, CC-4533[58] cells ('wild type' in the text and figures) were transformed with DNA cassettes that randomly insert into the genome, confer paromomycin resistance for selection, and inactivate the genes into which they insert. Each cassette contained two unique 22-nucleotide barcodes, one at each end of the cassette (Supplementary Fig. 1a–d and Supplementary Note). Transformants were arrayed on agar plates, and each insertion in a transformant would contain two barcodes. The barcode sequences as well as the insertion site were initially unknown (Supplementary Fig. 1e).

To determine the sequences of the barcodes in each colony, we generated combinatorial pools of the individual mutants; DNA was then extracted from each pool, and barcodes were amplified and deep-sequenced. The combinatorial pooling patterns were designed so that each colony was included in a different combination of pools, allowing us to determine the barcode sequences associated with individual colonies on the basis of which pools the sequences were found in (Supplementary Figs. 1f and 2a–e, and Supplementary Note). This procedure was similar in concept to the approach we used in our pilot study[9], but it consumed considerably less time because we used a simple PCR amplifying only the barcodes instead of a multistep flanking sequence extraction protocol (ChlaMmeSeq[58]) on each combinatorial pool.

To determine the insertion site associated with each barcode, the library was pooled into a single sample or divided into six separate samples. Barcodes and their flanking genomic DNA were PCR amplified using LEAP-Seq[9] (Supplementary Figs. 1g and 2f–j, and Supplementary Note). The flanking sequences associated with each barcode were obtained by paired-end deep sequencing[59,60].

The final product is an indexed library in which each colony has known flanking sequences that identify the genomic insertion site and barcode sequences that facilitate pooled screens in which individual mutants can be tracked by deep sequencing (Fig. 3a).

Insertion verification PCR. PCRs were performed in two steps to verify the insertion site[9] (Supplementary Table 6): (i) genome locus amplification and (ii) genome–cassette junction amplification. In the first step, genomic primers that were ~1 kb away from the flanking genomic sequence reported by LEAP-Seq were used to amplify the genomic locus around the flanking sequence. If the wild type produced the expected PCR band but the mutant did not or yielded a much larger product, this indicated that the genomic locus reported by LEAP-Seq might be disrupted by the insertional cassette and we proceeded to the second step. In this step, a primer binding to the cassette (oMJ913 for the 5′ side and oMJ944 for the 3′ side; Supplementary Table 6) and a second primer binding to flanking *Chlamydomonas* genomic DNA (one of the genomic primers from the first step) were used to amplify the genome–cassette junction. If the mutant produced a PCR band of the expected size that was confirmed by sequencing but the wild type did not, we categorized the insertion as 'confirmed'. For some mutants, genomic primers surrounding the site of insertion did not yield any PCR products in the wild type or mutant even after several trials, possibly owing to incorrect reference genome sequence or local PCR amplification difficulties. These cases were grouped as 'failed PCR' and were not further analyzed.

72 mutants (24 insertions each for confidence levels 1 and 2, confidence level 3, and confidence level 4) were chosen randomly from the library and tested. The genomic DNA template was prepared from a single colony of each mutant by using the DNeasy Plant Mini kit (69106, Qiagen). PCRs were performed using the Taq PCR Core kit (201225, Qiagen) as described previously[58]. PCR products of the expected size were verified by Sanger sequencing.

Southern blotting. Southern blotting was performed as previously described in detail[9]. Genomic DNA was digested with StuI enzyme (R0187L, New England Biolabs) and separated on a 0.7% Tris-borate-EDTA (TBE) agarose gel. The DNA in the gel was depurinated in 0.25 M HCl, denatured in a bath of 0.5 M NaOH and 1 M NaCl followed by neutralization in a bath of 1.5 M Tris-HCl (pH 7.4) and 1.5 M NaCl, and finally transferred onto a Zeta-Probe membrane (1620159, Bio-Rad) overnight, by using the alkaline transfer protocol given in the manual accompanying the membrane. On the next day, the membrane was gently washed with saline–sodium citrate (SSC) buffer (2× SSC: 0.3 M NaCl and 0.03 M sodium citrate), dried by paper towel, and UV cross-linked twice with a Stratalinker 1800 (Stratagene). For probe generation, the *AphVIII* gene on CIB1 was amplified by using primers oMJ588 and oMJ589 (Supplementary Table 1). The PCR product was purified and labeled according to the protocol of the Amersham Gene Images AlkPhos Direct Labeling and Detection System (RPN3690, GE Healthcare). The membrane was hybridized at 60 °C overnight with 10 ng ml^{-1} probe in hybridization buffer. On the next day, the membrane was washed with primary and secondary wash buffers and signal was visualized with CL-XPosure film (34093, Thermo Fisher).

Analyses of insertion distribution and identification of hotspots and coldspots. A mappability metric was defined to quantify the fraction of all possible flanking sequences from any genomic region that could be uniquely mapped to that region[58]. Calculation of mappability, hotspot and coldspot analysis, and simulations of random insertions were performed as described previously[58], except that a 30-bp flanking sequence length instead of a mixture of 20-bp and 21-bp lengths was used (because we were now using 30-bp flanking sequence data derived from LEAP-Seq rather than 20-bp and 21-bp ChlaMmeSeq sequences) and the v5.5 *Chlamydomonas* genome was used instead of the v5.3 genome[12]. This analysis was done on the original full set of mapped insertions, to avoid introducing bias from the choice of mutants for the consolidated set. The hotspot and coldspot analysis was performed on confidence level 1 insertions only, to avoid introducing bias caused by junk fragments and their imperfect correction. The full list of statistically significant hotspots and coldspots is provided in Supplementary Table 7.

Identification of under-represented gene ontology terms. For each Gene Ontology (GO) category, we calculated the total number of insertions in all genes annotated with the GO term and the total mappable length of all such genes, and we compared these values to the total number of insertions in and total mappable length of the set of flagellar proteome genes[11]. Comparison was performed by using Fisher's exact test with correction for multiple comparisons[61] to obtain the FDR. This analysis was done on the original full set of mapped insertions, to avoid introducing bias from the choice of mutants for the consolidated set. We decided to use the flagellar proteome as the comparison set because flagellar genes are unlikely to be essential; we did not use intergenic insertions or the entire genome because we knew that the overall insertion density differed between genes and intergenic regions. The statistically significant results are listed in Supplementary Table 8.

Prediction of essential genes. To predict essential genes in *Chlamydomonas*, we sought to generate a list of genes that had fewer insertions than would be expected randomly. Among them, those with no insertions were considered candidate essential genes.

For each gene, we calculated the total number of insertions in the gene and the total mappable length of the gene, and we compared these values to the total number of insertions in and total mappable length of the set of flagellar proteome genes[11], as was done for each GO category. The resulting list of genes with significantly fewer insertions than expected is discussed in the Supplementary Note and provided in Supplementary Table 9; the list includes 203 genes with no insertions and 558 genes with at least one insertion. However, only genes 5 kb or longer yielded an FDR of 0.05 or less when they had no insertion; our overall density of insertions was not high enough to detect smaller essential genes.

Pooled screens. Library plates that were replicated once every 4 weeks onto fresh medium were switched to a 2-week replication interval to support uniform colony growth before pooling. Cells were pooled from 5-d-old library plates. First, for each set of eight agar plates, cells were scraped using the blunt side of a razor blade (55411-050, VWR) and resuspended in 40 ml of liquid TAP medium in a 50-ml conical tube. Second, cell clumps were broken up by pipetting, by using a P200 pipette tip attached to a 10-ml serological pipette. In addition, cells were pipetted through a 100-µm cell strainer (431752, Corning). Third, subpools were combined into a master pool representing the full library.

The master pool was washed and resuspended in TP. Multiple aliquots of 2×10^8 cells were pelleted by centrifugation (1,000*g*, 5 min, room temperature), and the supernatant was removed by decanting. Some aliquots were used for inoculation of pooled cultures, whereas other aliquots were frozen at −80 °C as initial pool samples for later barcode extraction to enable analysis of reproducibility between technical replicates. For pooled growth, 20 liters of TAP or TP in a transparent Carboy container (2251-0050, Nalgene) was inoculated with the initial pool to a final concentration of 2×10^4 cells ml^{-1}. Cultures were grown at 22 °C, mixed by using a conventional magnetic stirbar, and aerated with air filtered by using a 1-µm bacterial air venting filter (4308, Pall Laboratory). The TAP culture was grown in the dark. For the two replicate TP cultures, the light intensity measured at the surface of the growth container was initially 100 µmol m^{-2} s^{-1} photons and was then increased to 500 µmol m^{-2} s^{-1} photons after the culture reached ~2×10^5 cells ml^{-1}. When the culture reached a final cell density of 2×10^6 cells ml^{-1} after seven doublings, 2×10^8 cells were pelleted by centrifugation (1,000*g*, 5 min, room temperature) for DNA extraction and barcode sequencing.

Molecular characterization of the *cpl3* mutant. Mutant genotyping PCRs were performed as previously described[9]. To complement the *cpl3* mutant, the wild-type *CPL3* gene was PCR amplified and cloned into the pRAM118 vector containing the *aph7″* gene[62], which confers resistance to hygromycin B. In this construct, expression of *CPL3* is under the control of the *PSAD* promoter. The construct was linearized before being transformed into the *cpl3* mutant. Transformants were robotically arrayed and assayed for colony size in the presence and absence of acetate (Supplementary Fig. 6c,d). Three representative lines that showed rescued photosynthetic growth were used in further phenotype analyses (Fig. 4).

Analyses of growth, chlorophyll, and photosynthetic electron transport. For all physiological and biochemical characterizations of *cpl3* described below,

we grew cells heterotrophically in the dark to minimize secondary phenotypes due to defects in photosynthesis.

For spot assays, cells were grown in TAP medium in the dark to log phase (~10^6 cells ml^{-1}). Cells were washed in TP and spotted onto solid TAP or TP medium. The TAP plates were incubated in the dark for 12 d before being imaged. The TP plates were incubated under 30 µmol m^{-2} s^{-1} photons for 1 d, under 100 µmol m^{-2} s^{-1} photons for 1 d, and then under 500 µmol m^{-2} s^{-1} photons for 4 d.

Chlorophyll *a* and chlorophyll *b* concentrations were measured as previously described[63] by using TAP-plated cells grown in the dark. We used cells grown on TAP in the dark instead of those grown on TP in the light for chlorophyll analyses, photosynthetic performance analyses, microscopy, proteomics, and western blot analysis to avoid observing secondary effects due to the photosynthetic defects of the *cpl3* mutant.

To measure the photosynthetic electron transport rate, cells grown in TAP in the dark were collected, resuspended in fresh TAP medium, and acclimated to the dark for 20 min. Cells were then measured for chlorophyll fluorescence under a series of increasing light intensities with the 'light curve' function on a DUAL-PAM-100 fluorometer (Walz). PSII quantum yield (ΦPSII) was quantified as previously described[64]. Relative electron transport rate (rETR) was calculated according to the following equation: $rETR = \Phi PSII \times I$, where I represents the emitted irradiance.

Proteomics. Cells grown in TAP in the dark were collected by centrifugation and flash frozen. Proteins were extracted from the frozen pellets by resuspension in lysis buffer (6 M guandium hydrochloride, 10 mM Tris(2-carboxyethyl) phosphine, 40 mM chloroacetamide, 100 mM Tris (pH 8.5), 1× MS-Safe protease inhibitor, and 1× phosphatase inhibitor cocktail II) and grinding in liquid nitrogen, followed by sonication. Protein lysates were then digested with trypsin (Promega) into peptides. Three biological replicates were processed for each strain.

Samples were labeled with tandem mass tags (TMTs), multiplexed, and then fractionated before tandem mass spectrometry (MS/MS) analyses. Briefly, each sample was labeled by using TMT labeling reagent (Thermo Fisher) according to the manufacturer's instructions. Samples were then mixed in equimolar amounts and desalted with C18 stage tips[65]. The dried peptide mixture was separated with strong cation exchange (SCX) stage tips[66] into four fractions. Each of the four fractions was diluted with 1% trifluoroacetic acid (TFA) and separated into three fractions with SDB-RPS stage tips. This procedure initially resulted in a total of 12 fractions. Fractions 1–3 (derived from the first SCX fraction) were pooled together, yielding ten final fractions. Each final fraction was diluted and injected into an Easy-nLC 1200 UPLC system (Thermo Fisher). Samples were loaded onto a nano capillary column packed with 1.9-µm C18-AQ (Dr. Maisch) mated to a metal emitter in line with a Fusion Lumos (Thermo Fisher). Samples were eluted using a split gradient of 10–20% solution B (80% acetonitrile and 0.1% formic acid) in 32 min and 20–40% solution B in 92 min, followed by column washing with 100% solution B for 10 min. The mass spectrometer was operated in data-dependent mode with the MS1 scan at 60,000 resolution (mass range of 380–1,500 m/z), an automatic gain control (AGC) target of 4×10^5, and a maximum injection time of 50 ms. Peptides above the threshold of 5×10^3 with charges of 2–7 were selected for fragmentation with dynamic exclusion after one run for 60 s with tolerance of 10 p.p.m. MS1 isolation windows of 1.6 m/z, MS2 isolation windows of 2 m/z, and higher-energy collisional dissociation (HCD) normalized collision energy (NCE) of 55% were selected. MS3 fragments were detected in the Orbitrap at 50,000 resolution in the mass range of 120–500 m/z with AGC at 5×10^4 and a maximum injection time of 86 ms. The total duty cycle was set to 3.0 s.

Raw files were searched with MaxQuant[67] while using default settings for MS3 reporter TMT 10-plex data. Files were searched against sequences of nuclear-, mitochondrial-, and chloroplast-encoded *Chlamydomonas* proteins supplemented with common contaminants[12,68,69]. Raw files were also analyzed within Proteome Discoverer (Thermo Fisher) by using the Byonic[70] search node (Protein Metrics). Data from MaxQuant and Proteome Discoverer were combined in Scaffold Q+ (Proteome Software), which was used to validate MS/MS-based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 80.0% probability by the scaffold local FDR algorithm. Protein identifications were accepted if they could be established at greater than 96.0% probability and contained at least two identified peptides. Scaffold Q+ non-normalized data were exported in the format of the $\log_2$ values for the reporter ion intensities, which reflect the relative abundance of the same protein among different samples multiplexed. Each sample was then normalized to a median of 0 (by subtracting the original median from the raw values, as the values are $\log_2$ transformed). For each gene and for each pair of samples, the normalized $\log_2$ intensity values from the three replicates for one sample were compared against those for the other sample using a standard t test. The resulting P values were adjusted for multiple testing[61], yielding an FDR for each gene in each pair of samples. We note that our calculation of FDR does not take into account the spectral count for each protein (provided in Supplementary Table 14), which is related to the absolute abundance of the protein and impacts the accuracy of proteomic measurements. Specifically, proteins with a low spectral count are likely of low abundance in cells and often exhibit large variation in the intensity value between biological replicates.

Western blotting. Cells grown in TAP in the dark were pelleted by centrifugation, resuspended in extraction buffer containing 5 mM HEPES-KOH (pH 7.5), 100 mM dithiothreitol, 100 mM Na_2CO_3, 2% SDS, and 12% sucrose, and lysed by boiling for 1 min. Extracted proteins were separated by SDS–PAGE (12% precast polyacrylamide gels, Bio-Rad) and α-tubulin was used as a loading and normalization control. Polypeptides were transferred onto PVDF membranes with a semidry blotting apparatus (Bio-Rad) at 15 V for 30 min. For western blot analyses, membranes were blocked for 1 h at room temperature in Tris-buffered saline with 0.1% Tween (TBST) containing 5% powdered milk followed by incubation for 1 h at room temperature with primary antibodies in TBST containing 3% powdered milk. Primary antibodies were diluted according to the manufacturer's recommendations. All antibodies were from Agrisera; the catalog numbers for the antibodies against CP43, PsaA, ATPC, and α-tubulin were AS11-1787, AS06-172-100, AS08-312, and AS10-680, respectively. Proteins were detected by enhanced chemiluminescence (K-12045-D20, Advansta) and imaged on a medical film processor (Konica) as previously described[9].

Additional methods. Additional method details are provided in the Supplementary Note.

Statistical analyses. The statistical methods and tests used are indicated throughout the manuscript. Fisher's exact test with Benjamini–Hochberg correction[61] for multiple comparisons was used to identify under-represented GO terms, essential genes, and hit genes in the photosynthesis screen and for the analysis of candidate gene enrichment. The binomial test with Benjamini–Hochberg correction for multiple comparisons was used for the hotspot and coldspot analysis. A chi-square test of independence was used for insertion density comparisons between features. A t test with Benjamini–Hochberg correction for multiple comparisons was used for analysis of the proteomics data. Please see the corresponding Methods or Supplementary Note section for details on each analysis.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

All programs written for this work have been deposited at GitHub (see URLs).

Data availability

Insertion details and distribution information for mutants are available through the CLiP website at https://www.chlamylibrary.org/. The mass spectrometry proteomics data on the *cpl3* mutant have been deposited to the ProteomeXchange Consortium via the PRIDE[71] partner repository with dataset identifier PXD012560. Other data that support the findings of this study are available from the corresponding author upon reasonable request.

References

58. Zhang, R. et al. High-throughput genotyping of green algal mutants reveals random distribution of mutagenic insertion sites and endonucleolytic cleavage of transforming DNA. *Plant Cell* **26**, 1398–1409 (2014).
59. Rubin, B. E. et al. The essential gene set of a photosynthetic organism. *Proc. Natl Acad. Sci. USA* **112**, E6634–E6643 (2015).
60. Wetmore, K. M. et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, e00306–e00315 (2015).
61. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
62. Berthold, P., Schmitt, R. & Mages, W. An engineered *Streptomyces hygroscopicus aph 7″* gene mediates dominant resistance against hygromycin B in *Chlamydomonas reinhardtii*. *Protist* **153**, 401–412 (2002).
63. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls *a* and *b* extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *BBA Bioenergetics* **975**, 384–394 (1989).
64. Saroussi, S. I., Wittkopp, T. M. & Grossman, A. R. The type II NADPH dehydrogenase facilitates cyclic electron flow, energy-dependent quenching, and chlororespiratory metabolism during acclimation of *Chlamydomonas reinhardtii* to nitrogen deprivation. *Plant Physiol.* **170**, 1975–1988 (2016).
65. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).

66. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
67. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
68. Maul, J. E. et al. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell* **14**, 2659–2679 (2002).
69. 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.* **223**, 211–216 (1990).
70. Bern, M., Kil, Y. J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics* **40**, 13.20 (2012).
71. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

nature research

Corresponding author(s): Martin C. Jonikas

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection in this study.
Data analysis	All commercial and open source software used in this manuscript is listed in the appropriate Supplemental Methods sections. All custom code used in this manuscript is deposited at: https://github.com/Jonikas-Lab/Li-Patena-2019

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The deep sequencing read count data for library super-pools and photosynthesis screen samples are provided in Supplementary Tables 4 and 10.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We empirically determined the sample sizes based on published research.
Data exclusions	We generated a library containing ~210,000 mutants and cherry-picked 62,389 mutants for long-term maintenance. We used these 62,389 mutants for analyses of library coverage.
Replication	We performed replicates and used orthogonal approaches where appropriate.
Randomization	We randomly picked mutants for validation of insertion site mapping.
Blinding	Blinding and randomization were not used for this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☐	☐ Palaeontology
☐	☐ Animals and other organisms
☐	☐ Human research participants

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	The mutants described in this manuscript are available from the Chlamydomonas Resource Center: https://www.chlamylibrary.org/

Antibodies

Antibodies used	Antibodies used in Fig. 4 are commercially available from Agrisera. The catalog numbers are provides in Methods.
Validation	Antibodies used in Fig. 4 have been tested in photosynthetic organisms (references are on the Agrisera website).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	We have been using the CC-4533 Chlamydomonas strain generated in our lab.
Authentication	When we generated the CC-4533 strain in 2012, we froze dozens of copies and now revive one stock each time. The identity of our line is validated by sequencing.
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Human research participants

Policy information about studies involving human research participants

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software | *Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.